# Proceedings of LFG09 

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/

## Contents

1 Editors' Note ..... 4
2 Ahmad Alsharif and Louisa Sadler: Negation in Modern Standard Arabic: An LFG approach ..... 5
3 Budour Al Sharifi and Louisa Sadler: The adjectival construct in Ara- bic ..... 26
4 Alex Alsina: The prepositional passive as structure-sharing ..... 44
5 Raúl Aranovich: Animacy effects and locative marking in Shona ap- plicatives ..... 65
6 I Wayan Arka, Avery Andrews, Mary Dalrymple, Meladel Mistica and Jane Simpson: A linguistic and computational morphosyntactic anal- ysis for the applicative - $i$ in Indonesian ..... 85
7 Ash Asudeh: Adjacency and locality: A constraint-based analysis of complementizer-adjacent extraction ..... 106
8 Eleftherios Avramidis and Jonas Kuhn: Exploiting XLE's finite state interface in LFG-based statistical machine translation ..... 127
9 Tina Bögel, Miriam Butt, Ron Kaplan, Tracy Holloway King and John Maxwell III: Prosodic phonology in LFG: A new proposal ..... 146
10 Gerlof Bouma and Jonas Kuhn: On the split nature of the Dutch laten- causative ..... 167
11 Miriam Bouzouita and Stergios Chatzikyriakidis: Clitics as calcified processing strategies: The case study of Spanish clitic placement and the PCC as a tree-logic restriction ..... 188
12 Aoife Cahill, Uli Heid, Christian Rohrer and Marion Weller: Using tri-lexical dependencies in LFG parse disambiguation ..... 208
13 Amy Dahlstrom: OBJ $_{\theta}$ without OBJ: A typology of Meskwaki objects ..... s222
14 Lachlan Duncan: A tale of two taqs: An OT-LFG account of plurals and distributives in K'ichee' Mayan ..... 240
15 Yehuda Falk: Islands: A mixed analysis ..... 261
16 Maria Flouraki and Despina Kazana: Constraining disjunctive con- structions in Modern Greek ..... 282
17 Martin Forst and Christian Rohrer: Problems of German VP coordi- nation ..... 297
18 Yvette Graham, Anton Bryl and Josef van Genabith: F-structure trans- fer-based statistical machine translation ..... 317
19 Dag Haug: med - the syntax and semantics of concomitance in Nor- wegian ..... 338
20 Annette Hautli and Tracy Holloway King: Adapting stochastic LFG input for semantics ..... 357
21 Anna Kibort: Intermediary agents and unexpressed pronouns ..... 378
22 Tibor Laczkó: Relational nouns and argument structure - evidence from Hungarian ..... 399
23 Helge Lødrup: Looking possessor raising in the mouth: Norwegian possessor raising with unergatives ..... 420
24 Maja Miličević: On the status of clitic reflexives and reciprocals in Italian and Serbian ..... 441
25 Györgi Rákosi: Beyond identity: The case of a complex Hungarian reflexive ..... 459
26 Ines Rehbein and Josef van Genabith: Automatic acquisition of LFG resources for German - as good as it gets ..... 480
27 Melanie Seiss: On the difference between auxiliaries, serial verbs and light verbs ..... 501
28 Ilona Spector: Hebrew floating quantifiers ..... 520
29 Tania Strahan: Outside-in binding of reflexives in Insular Scandina- vian ..... 541
30 Sebastian Sulger: Irish clefting and information-structure ..... 562
31 Lamia Tounsi, Mohammed Attia and Josef van Genabith: Parsing Arabic using treebank-based LFG resources ..... 583
32 Nigel Vincent: LFG and Dynamic Syntax: Two non-derivational theo- ries ..... 587
33 Michael Wescoat: Udi person markers and lexical integrity ..... 604
34 Beatriz Willgohs and Patrick Farrell: Case and grammatical functions in Imbabura Quechua: An LFG approach ..... 623
35 Annie Zaenen and Dick Crouch: OBLs hobble computations ..... 644
36 Sina Zarrieß and Jonas Kuhn: Paraphrases in LFG-based broad-cove- rage semantics ..... 655

## 1 Editors' Note

The program committee for LFG'09 were Kersti Börjars and Martin Forst. We would like to thank them again for putting together the program that gave rise to this collection of papers. Thanks also go to the executive committee and the reviewers, without whom the conference would not have been possible. The local organizing committee consisted of Anna Kibort who worked to put on yet another successful conference. We would like to thank Ash Asudeh for helping with the light editing and Dikran Karagueuzian for his and CSLI's unfailing support.

The table of contents lists all the papers presented at the conference. Some papers were not submitted to the proceedings. For these papers, we suggest contacting the authors directly.

# NEGATION IN MODERN STANDARD ARABIC: AN LFG APPROACH 

Ahmad Alsharif and Louisa Sadler<br>University of Essex University of Essex

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009
CSLI Publications
http://www-csli.stanford.edu/


#### Abstract

Modern Standard Arabic (MSA) uses five different particles to express sentential negation: the invariant particle maa, the particle laa and its tensed counterparts lam (PAST) and lan (FUT), and laysa which is marked only for SUBJ agreement. Partial analyses of these elements are offered in other frameworks, notably Minimalism (Shlonsky, 1997; Benmamoun, 2000), but have not to date received an analysis within LFG. We propose an approach to four of these particles: the fifth one, namely maa, raises a number of additional issues and we leave it to one side for reasons of space. laa, lam, lan show distinctions of TENSE, occur only with imperfective forms of the verb (excluding the perfective) and must immediately precede the verb itself. They are limited to occurrence in verbal sentences. We propose that the adjacency requirement follows from the fact that these negative particles are non-projecting words adjoined to the (imperfective) V. On the other hand, laysa is a fully verbal element, and is thus a negative verb, occurring only with present tense interpretation.


## 1 Data

### 1.1 Negative Particles

In Modern Standard Arabic (henceforth MSA) five different particles are used to express sentential negation: the (invariant) particle maa, the item laa and its (temporally) inflected counterparts lam and lan and (variously inflected) forms of laysa. Amongst these elements, laysa is unique in inflecting for SUBJ agreement. In the present paper, we will have nothing to say here about maa and concentrate uniquely on the forms of laa and laysa.

### 1.2 Laa, Lan, Lam

There are good grounds for distinguishing between laysa on the one hand, and laa, lam and lan on the other. For laa, lam and lan the basic facts are as follows. ${ }^{1}$ Firstly, all these negative forms occur in sentences which have a verbal element as the main predicate. There is a basic morphological opposition in Arabic between imperfective and perfective verbforms, and laa, lam, lan all co-occur only with imperfective forms of the verb: substituting perfective verbforms in all of the following examples would lead to ungrammaticality. The pairs in (1) - (3) exemplify the particle laa negating an imperfective indicative (with a present tense reading); (1) and (2) additionally illustrate SV(O) order and (3) shows vso word order. Note that irrespective of word order, the negative particle laa immediately precedes the imperfective verb in all of these examples.
(1) a. t-țullaab-u ya-drus-uu-n
the-students-NOM 3M-study.IPFV-3MP-IND
The students study/are studying.

[^0]b. t-ṭullaab-u laa ya-drus-uu-n the-students NEG 3M-study.IPFV-3MP-IND
The students do not study/are not studying.
(Benmamoun, 2000, 95)
(2) a. Zayd-un y-aktub-u al-yawm-a al-risalat-a

Zayd-NOM 3M-write.IPFV-3MS.IND the-day-ACC the-letter-ACC
Zayd is writing the letter today.
b. Zayd-un laa y-aktub-u al-yawm-a al-risalat-a

Zayd-NOM NEG 3M-write.IPFV-3MS.IND the-day-ACC the-letter-ACC
Zayd is not writing the letter today.
(3) a. Y-aktub-u Zayd-un al-yawm-a al-risalat-a

3M-write.IPFV-3MS.IND Zayd-NOM the-day-ACC the-letter-ACC
Zayd is writing the letter today.
b. Laa y-aktub-u Zayd-un al-yawm-a al-risalat-a NEG 3M-write.IPFV-3MS.IND Zayd-NOM the-day-ACC the-letter-ACC
Zayd is not writing the letter today.

The following set of data illustrate the basic facts with respect to the tensed forms of laa, namely lam and lan. (4) and (5) show that the future may be expressed by means of an imperfective (indicative) verb with the prefix $s a-$, and additionally that the future form verb is negated by using the particle lan in combination with a subjunctive mood imperfective (without the prefix $s a$-): again, adjacency is required between the particle and the main verb irrespective of sentential word order.
(4) a. $t$-tullaab-u sa-ya-dhab-uu-n the-students-NOM FUT-3M-go.IPFV-MP-IND
The students will go.
b. t-ṭullaab-u lan ya-dhab-u the-students-NOM NEG.FUT 3M-go.IPFV-MP.SBJV
The students will not go.
(Benmamoun, 2000, 95)
(5)
a. sa-ya-d.dhab-u t-ṭullaab-u

FUT-3M-go.IPFV-MSG-IND the-students-NOM
The students will go.
b. lan ya-dhab-a t-tullaab-u

NEG.FUT 3M-go.IPFV-MSG.SBJV the-students-NOM
The students will not go.

Finally (6) shows that the combination of the particle lam with an imperfective verb in jussive mood corresponds to an (affirmative) perfective verb. It should be noted that in the Arabic vernaculars, the basic constrast is between the marked form (IPFV.IND) in the affirmative and the unmarked form in the context of the tensed negative particle (that is, the JUSS/SBJV distinction in neutralised in the vernaculars).
(6)
a. t-ṭullaab-u $\quad \underline{d} a h a b-u u$
the-students-NOM go.PFV-3MP
The students left.
b. t-ṭullaab-u lam ya-dhab-uи
the-students-NOM NEG.PAST 3M-go.IPFV-MP.JUSS
The students did not go.
(Benmamoun, 2000, 95)
c. *lam t-ṭullaab-u ya-dhab-uu

NEG.PAST the-students-NOM 3M-go.IPFV-MP.JUSS
The students did not go.

To summarise, laa, lam and lan occur with verbal forms in the imperfective but not with perfective forms of the verb. In all cases, the negative particle must be adjacent to this form, see (6c). laa occurs with the indicative imperfective and cannot be used for sentences in the future or past. lam occurs with the jussive imperfective expressing negation in the past, and lan with the subjunctive imperfective, expressing negation in the future: thus lam and lan appear to be negative particles which carry temporal information.
(7)

| TENSE | AFFIRM FORM | NEG FORM |
| :--- | :--- | :--- |
| PRES | IPFV.IND | laa + IPFV.IND |
| PAST | PFV | lam + IPFV.JUSS |
| FUT | $s a$-IPFV.IND | lan + IPFV.SBJV |

### 1.3 Future Negation: A Further Data Point

It is generally claimed that laa canot co-occur with tensed verbs (Benmamoun, 2000; Bahloul, 1994). In fact, however, things are slightly more complicated. It is certainly true that 'double' expression of FUT is impossible (shown by (9) and (8)), but it is not completely accurate to state that laa cannot combine with a future marker. This is because there is an alternative analytic realization of future, namely the use of the particle sawfa with an (unprefixed) imperfective indicative form. As the data shows, laa can combine with safwa but not with prefixal future forms in sa- (hence the contrast between (11) and (12)).
(8) *sawfa lan $y$-aḥdur-a.

FUT NEG-FUT 3SM-come-SBJV
He will not come.
(9) *t-ṭullaab-u lan sa-ya-dhab-uun/-uu
the-students-NOM NEG.FUT FUT-3M-go.IPFV-MP.IND/-MP.SBJV
The students will not go.
(10) lan $y$-ahdur-a

NEG-FUT 3M-come.IPFV-Sm.SBJV
He will not come.
(11) *t-ṭullab-u laa sa-ya-d $h a b-u u-n$
the-students-NOM NEG FUT-3M-go.IPFV-3M-IND
The students will not go.(Benmamoun, 2000, 101)
(12) Sawfa laa y-ahdur-u

FUT NEG 3M-present.IPFV-3MS.IND
He will not come. (Fassi-Fehri, 1993, 82)

### 1.4 Laysa

laysa differs in several respects from the invariant forms laa, lan, lam. It realizes (SUBJ) agreement and is not required to be adjacent to the verb.
(13)

|  | SG | DU | PL |
| :--- | :--- | :--- | :--- |
| 1 | lastu |  | lasnaa |
| 2 M | lasta | lastumaa | lastum |
| 2 F | lasti | lastumaa | lastunna |
| 3 M | laysa | laysaa | laysuu |
| 3 F | laysat | laysataa | lasna |

(14) a. laysa ȟālid-un ya-ktub-u $\check{s}$-šifr-a NEG.3MS Khalid-NOM 3M-write.IPFV-3MS the-poetry-ACC

Khalid does not write/is not writing poetry.
b. laa ya-ktubu ȟālid-un š-šifr-a NEG 3M-write.IPFV-3SM Khalid-NOM the-poetry-ACC

Khalid does not write/is not writing poetry.
(Benmamoun, 2000, 103)

A third difference is that it occurs in both verbal and verbless sentences (unlike laa, lan, lam), that is, sentences with nominal and adjectival predicates.
(15) a. laysa Pahii muSalliman-an. NEG.3MS brother.my teacher-ACC
My brother is not a teacher.
b. laysa muYalliman-an.

NEG.3MS teacher-ACC
He is not a teacher.
laysa shows the typical behaviour of a verb in that number agreement is defective when it precedes the SUBJ:
(16) a. al-awlad-u lays-uu ya-ktub-uun.
the-boys-NOM NEG-3MP 3M-write.IPFV-3MP-IND
The boys do not write.
b. lays-a al-awlad-u ya-ktub-uun.

NEG-3MS the-boys-NOM 3M-write.IPFV-3MP-IND
The boys do not write.
laysa is compatible only with IPFV.IND verbs and receives a present interpretation.
(17) a. *laysa r-rağul-u ? $a k a l a$

NEG.3SM the-man-NOM eat.PERF.3SM
The man did not eat (Benmamoun, 2000, 105)
b. *laysa r-rağul-u sa-ya- ${ }^{\text {²kulu }}$ gadan

NEG.3SM the-man-NOM FUT-eat.IPFV.3SM tomorrow
The man will not eat tomorrow (Benmamoun, 2000, 105)

### 1.5 Compound Tenses

We use purely morphosyntactic glossing throughout. Verbs show a morphological distinction between PFV and IPFV forms: such forms are used to express both temporal and aspectual distinctions: the opposition between them in sentences containing a single analytic form broadly encodes a PAST/NONPAST temporal distinction. (See Fassi-Fehri (2004) for some discussion.) The INDIC imperfective further inflects for FUT (or combines with the particle sawfa). The imperfective stem also shows what are traditionally called distinctions of MOOD: INDIC, JUSS, SBJV. Compound tenses involve the combination of a finite auxiliary with the perfective and imperfective indicative (finite) forms. They are not required to be adjacent. The table below illustrates various compound tenses.
(18)

\begin{tabular}{|c|c|c|}
\hline FORM \& REALIZATION \& MEANING <br>
\hline \multirow[t]{2}{*}{PFV} \& katab-tu t-taqrīr-a \& \multirow[t]{2}{*}{PAST} <br>
\hline \& $I$ wrote the report. \& <br>
\hline IPFV \& ? aktub-u t-taqrīr-a \& PRES <br>
\hline FUT-IPFV \& I write/am writing the report. sa- ${ }^{\text {² aktub-u }}$ t-taqrīr-a I will write the report. \& FUT <br>
\hline PFV + PFV \& kun-tu qad katab-tu t-taqrīr-a I had written the report. \& PAST PRFT <br>

\hline PFV + IPFV \& | kun-tu ${ }^{`}$ aktub-u t-taqrīr-a |
| :--- |
| I was writing/used to write the report. | \& PAST PROG <br>

\hline PFV + FUT-IPFV \& kun-tu sa- ${ }^{ }$aktub-u t-taqrīr-a I was going to write the report. \& PAST FUT <br>
\hline IPFV.IND + PFV \& ${ }^{?}$ akūnu qad katab-tu t-taqrīr I (always) have written the report. \& PRES PRFT <br>
\hline FUT-IPFV + PFV \& sa- ${ }^{ }$akūnu qad katab-tu t-taqrīr I will have written the report. \& FUT PRFT <br>

\hline FUT-IPFV + IPFV \& | sa- ${ }^{\text {ªkūnu }}{ }^{\text {ªk aktub-u t-taqrīr-a }}$ |
| :--- |
| $I$ will be writing the report. | \& FUT PROG <br>

\hline
\end{tabular}

## 2 Minimalist Approaches

Negation in MSA (and in the Arabic vernaculars) has received a reasonable amount of theoretical attention within Minimalism (and its precursors), the major references being Benmamoun (2000); Ouhalla (2002) and Shlonsky (1997). Of these, the most extensive discussion is Benmamoun (2000), and for this reason we briefly present his approach here. The basic structural assumptions made in this account (which discusses negation in the vernaculars (concentrating on Moroccan Arabic (MA)) and MSA, involves a NegP projection situationed between TP and VP, as in (19). ${ }^{2}$


The crucial points of this analysis concern the assumptions about what features are inherent to each node. First, sentential negation (the Neg node), is taken to be specified for the categorial feature [+D] (Benmamoun, 2000, 69). The elements laa, lam and lan are generated in Neg. Second, Tenses are associated with different bundles of features generated on the T node, as follows (Benmamoun, 2000, 50):

[^1](20) $\mathrm{T} \rightarrow[+\mathrm{D}]$ (Present)
$\mathrm{T} \rightarrow[+\mathrm{D},+\mathrm{V}]$ (Past, Future)
$\mathrm{T} \rightarrow[+\mathrm{V}]$ (Imperative)

Suppose the node T is generated with the feaure bundle $[+$ Past, $+\mathrm{D},+\mathrm{V}]$ or $[+\mathrm{Fut},+\mathrm{D},+\mathrm{V}]$ ("the V feature must be checked by verbal heads, while the D feature can be checked by nominal heads or by verbs that carries (sic) agreement" (Benmamoun, 2000, 99)). By assumption, the Neg node is also specified for $[+\mathrm{D}]$. In order for both the +V and the +D features of the T node, to be appropriately "checked", it is necessary that both the V and the Neg move to the T node. A derivation such as the following will ensue, in which V raises to Neg and then Neg and V together raise to T. The spell out of the resultant T node is the combination of lam + verb, likewise if + Fut is generated on the T node, then the spell out will be lan $+v e r b$. As for Neg and V "they are both in tense supporting the tense feature and checking the categorial $[+\mathrm{V}]$ feature" (Benmamoun, 2000, 100).


The alternative might be to try to move the verb directly to deal with the +V feature (and spell out the tense): presumably such a verb could also check the D feature of the T node, as it carries subject agreement, but this violates Minimality, or take the Neg also but spell out the features on the verb, not the negation. This is ruled out by the assumption that tesne must be spelled out on the head of the complex, which is Neg (Benmamoun, 2000, 102).

Suppose now that the T node is generated with the feature bundle $[+$ Pres, +D ]. The +D feature can be checked by a nominal. Because there is no +V feature on T , neither the verb (nor the Neg ) is required to raise to T . However given that laa and the V are required to be adjacent, something must require this: "merger between laa and verb must be due to some property of laa itself. The property in question is the categorial feature [ +D ] feature of laa. The merger betwen laa and the verb, carrying subject agreement, allows the latter to check the categorial [+D] feature on the negative" (Benmamoun, 2000, 100).
(22)


In contrast to traditional accounts, which view laysa as a verbal element, Benmanoun takes it also to be a Neg particle (specified for [+D]). The idea is that since laysa itself inflects for SUBJ agreement, then this feature is checked by the SUBJ and so Neg (i.e. laysa) does not raise to T for purposes of feature checking. This means that in principle, it is free to be non-adjacent to the inflected verb (unlike laa).
(23)


Although it would take us too far afield to attempt here any substantial critique of this (or other Minimalist) proposals, we will make a number of brief observations about the account. The first is that it is far from complete in its present form. It does not explain how (by which mechanism) different negatives select different forms (moods) of the verb, and given that that there are no lexical differences postulated between laa, lam, lan (they result from the spell out of different sets of features in different tree locations, as far as we understand it), it is not obvious how this will be treated. Second, the account is radically incomplete in that there is no attempt to extend it to the more complicated facts of negation with compound tenses. Third, the assumption that Neg is categorially specified as +D plays a crucial role in terms of ensuring that forms of laa and the verb are strictly adjacent: the subject agreement features of V are required to check the +D specification of Neg heads. While this diacritic approach does indeed appear to produce the desired result, it is unclear what it actually represents (other than a diacritic). Moreover there is perhaps some unwelcome asymmetry in the treatment of the laa+V adjacency (which involves only this +D checking requirement) and that of the lam/lan +V adjacency, which additionally involves the verb checking the +V feature of T (and thus raising alongside Neg to $T)$. Fourth, it is unclear what checks the +D feature of the $\mathrm{T}[+$ Pres, +D$]$ node, in the case where laa + V occurs in Neg and in the case where laysa occurs in Neg. ${ }^{3}$ Fifth, there is no discussion or analysis

[^2]of the multiple agreements on the negative laysa and the following verb, while most of the previous approaches within this framework have postulated multiple functional (Agr) projections to account for this data.

## 3 Analysis of Laa, Lam and Lan

### 3.1 Adjacency and Selection

In short, we argue that adjacency follows because the negative particle and the verb form a small construction, that is, the particle is a non-projecting word in the sense of Toivonen (2003). Neg and V do not constitute a single morphological word. Unlike laysa, laa, lam and lan are non-projecting elements which occur as sister to I, and therefore occur with verbal elements. The behaviour of the negative particles laa, lam and lan is strongly reminiscent of the particles discussed in Toivonen (2003).
(24)

| Property | laa, lam, lan | Swedish Verbal Particles |
| :--- | :--- | :--- |
| Take complements | No | No |
| Can be modified | No | No |
| Bear stress | Yes | Yes |
| Adjoined to verb | Yes, left | Yes, right |
| Separable | No | Yes, but not by object |

(25)

$\longrightarrow \quad$| $\hat{\mathrm{I}}$ | I |
| :---: | :---: |
|  | $\uparrow=\downarrow$ |$\quad \uparrow=$

(26)


Each particle places certain co-occurrence requirements on its sister, and thus a question arises as to whether these are c - or f -structure constraints. We turn to this in the following subsection.

### 3.2 Selection

In order to discuss the matter of selection we will need to say much more about the tense and aspect system. There is some literature on this question, but accounts often appear to be driven more by theory-internal requirements than by the empirical facts. For the moment, we simply make the
following analytic assumptions. Firstly, although some researchers argue that MSA is a tenseless language (largely based on very theory-internal reasoning rather than data), we take it that MSA has tense as well as aspect and that TENSE involves distinctions of PAST/NON-PAST and FUT/NON-FUT. Additionally, as we have seen, the Arabic verb makes a morphosyntactic distinction between three moods, JUSS, SBJV and INDIC. Only the last of these, the INDIC, encodes distinctions of TENSE. JUSS and SBJV forms only occur when selected for. In principle, selection might be in terms of a MOOD feature or directly on c-structure form, and we return to this question.

With this in place we can formulate the lexical entries to capture the basic facts. The basic agreement information for 3MPL forms is provided in the template (27). Illustrative lexical entries for indicative verb forms (perfective, imperfective and future-imperfective) are in (28)-(30), and for the other moods in (32)-(33). ${ }^{4}$

(27) | $3 \mathrm{MPL} \equiv$ | $(\uparrow$ SUBJ NUM $)=$ PL |
| ---: | :--- |
|  | $(\uparrow$ SUBJ PERS $)=3$ |
|  | $(\uparrow$ SUBJ GEND $)=$ MASC |

(28) dahab-uu I ( $\uparrow$ PRED) $=$ go $<$ SUBJ $\rangle$

Perfective Form $(\uparrow$ TENSE PAST $)=+$ @3MPL
(29) ya-drus-uu-n I ( $\uparrow$ PRED) $=$ study $<$ SUBJ $>$ $(\uparrow$ TENSE PAST $)=-$ @ 3MPL
(30) sa-ya-dhab-uun I $(\uparrow$ PRED $)=$ go $<$ SUBJ $>$ $(\uparrow$ TENSE PAST $)=-$ $(\uparrow$ TENSE FUT $)=+$ $(\uparrow$ POL $)=$ POS @ 3MPL
(31) sawfa $\hat{I} \quad(\uparrow$ TENSE FUT $)=+$
(32)

$(\uparrow$ MOOD $)=$ JUSS
@ 3MPL

## Imperfective Form

Imperfective Form

|  | $(\uparrow$ TENSE PAST $)=-$ |
| :--- | :--- |
|  | $(\uparrow$ TENSE FUT $)=+$ |
|  | $(\uparrow$ POL $)=$ POS |
|  | $@ 3 \mathrm{MPL}$ |

To recap, the behaviour we need to capture is summarised in (34).
(34) laa coccurs with an imperfective indicative verbform
lam expresses $\mathrm{PAST}=+$ and selects the jussive verbform
lan expresses FUT=+ and selects the subjunctive verbform $s a$ - (and sawfa) express POL $=+$

Consider first the treatment of $s a$ - and sawfa. (30) limits the $s a$ - form to occurrence in a positive clause, whereas sawfa does not place this restriction. This will be used in accounting for (11) and (9) permitting laa to co-occur with sawfa (12). ${ }^{5}$
The entries for the particles are as follows. The TENSE specification in the entry for laa means it cannot combine with Perfectives, the POL specification prevents it combining with the $s a$-Imperfective. If it were to combine with JUSS or SBJV then there would overall be no TENSE which would be a problem. So the f-structure for (1b) is shown in (36)
(35) laa $\hat{I} \quad(\uparrow$ TENSE PAST $) \neq+$ $(\uparrow$ POL $)=$ NEG
$\left[\begin{array}{ll}\text { PRED } & \text { STUDY < SUBJ }> \\ \text { POL } & \text { NEG } \\ \text { TENSE } & {\left[\begin{array}{ll}\text { PAST } & -\end{array}\right]} \\ \text { SUBJ } & {\left[\begin{array}{ll}\text { PERS } & 3 \\ \text { PRED } & \text { STUDENT } \\ \text { SPEC } & \text { DEF } \\ \text { NUM } & \text { PL }\end{array}\right]}\end{array}\right]$
lam selects a JUSS and defines TENSE PAST,$+{ }^{6}$ whereas lan selects a SBJV. Note that these verbal forms are themselves tenseless, but TENSE information is expressed by the negative particle. We give the f-structure for (5b) by way of illustration.

$$
\begin{array}{rll}
\text { lam } \quad \hat{I} & (\uparrow \text { TENSE PAST })=+  \tag{37}\\
& (\uparrow \text { POL })=\text { NEG } \\
& (\uparrow \text { MOOD })={ }_{c} \text { JUSS }
\end{array}
$$

(38) lan $\hat{I} \quad(\uparrow$ TENSE FUT $)=+$ $(\uparrow$ POL $)=$ NEG $(\uparrow$ MOOD $)={ }_{c}$ SBJV

[^3](39)

$\left[\begin{array}{ll}\text { PRED } & \text { GO }<\text { SUBJ }> \\ \text { POL } & \text { NEG } \\ \text { MOOD } & \text { SBJV } \\ \text { TENSE } & {[\text { FUT }+]} \\ \text { SUBJ } & {\left[\begin{array}{ll}\text { PERS } & 3 \\ \text { PRED } & \text { STUDENT } \\ \text { SPEC } & \text { DEF } \\ \text { NUM } & \text { PL }\end{array}\right]}\end{array}\right]$

This accounts for all the simple tenses and their combinations with negative particles but there is rather a lot more data to account for, most of which the competing accounts seem to take account of.

### 3.3 Compound Tenses in MSA

We recall the table above which shows how compound tenses are formed in MAS. All three indicative verb forms can also occur in combination with a tensed auxiliary (e.g. forms of kāna 'be'): in this environment they express not TENSE but ASP. Aspectually, the verbforms express a three way distinction between PRF (completed), PROG (continuative) and PROSP (prospective). Aspectual qad is a non-projecting particle in V . Unlike the tensed (finite) forms, the aspectual version occurs in V . Therefore we have additional lexical entries as shown below.
(40) $1 \mathrm{SG} \equiv(\uparrow$ SUBJ NUM $)=\mathrm{SG}$
$(\uparrow$ SUbJ PERS $)=1$
(41) katab-tu
$V(\uparrow$ PRED $)=$ write $<$ SUBJ OBJ $>$
Perfective Form
$(\uparrow$ ASP $)=\mathrm{PRF}$
@ 1SG
(42) ${ }^{ }$aktub-u $\quad V$ ( $\uparrow$ PRED) $=$ write $<$ SUBJ OBJ $>\quad$ Imperfective Form
$(\uparrow$ ASP $)=$ PROG
@ 1SG
(43) $s a-{ }^{\text {² }}$ aktub-u $\quad V$ ( $\uparrow$ PRED) $=$ write $<$ SUBJ OBJ $>\quad$ Imperfective Form
$(\uparrow$ ASP $)=$ PROSP
@ 1SG

Unlike lexical verbs (which occur in I and V), (indicative) forms of auxiliary be occur only in I and hence are always tensed.
(44) kun-tu $I(\uparrow$ TENSE PAST $)=+\quad$ Perfective Form @ 1SG
(45) ${ }^{\top} a k u \bar{n}-u \quad I \quad(\uparrow$ TENSE PAST $)=-\quad$ Imperfective Form @ 1SG

```
(46) \(s a-{ }^{?} a k u \bar{n}-u \quad V \quad(\uparrow\) TENSE PAST \()=-\quad\) Imperfective Form \((\uparrow\) TENSE FUT \()=+\) @ 1SG
```


### 3.4 Exemplification

The following examples show how the basic data is accounted for by the analysis developed so far. In the following section we go on to look at the combination of negation and compound verbal forms.
(47) kun-tu qad katab-tu $t$-taqrīr-a
be.PFV-1SG PT write.PFV-1SG the-report-ACC
I had written the report.
PAST PRF
$\left[\begin{array}{ll}\text { PRED } & \text { WRITE < SUBJ, OBJ }> \\ \text { ASP } & \text { PERF } \\ \text { TENSE } & {\left[\begin{array}{ll}\text { PAST } & +\end{array}\right]} \\ \text { SUBJ } & {\left[\begin{array}{ll}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]}\end{array}\right]$
(48) kun-tu ${ }^{2} a k t u b-u \quad t$-taqrīr-a
be.PFV-1SG write-IPFV. 1 SG the-report-ACC
I was writing the report.
PAST PROG
$\left[\begin{array}{lc}\text { PRED } & \text { WRITE }<\text { SUBJ, OBJ }> \\ \text { ASP } & \text { PROG } \\ \text { TENSE } & {\left[\begin{array}{ll}\text { PAST } & +\end{array}\right]} \\ \text { SUBJ } & {\left[\begin{array}{cc}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]}\end{array}\right]$
(49) kun-tu $\quad$ sa- ${ }^{?}$ aktub-u $\quad$-taqrīr-a
be.PFV-1SG FUT-write-IPFV.1SG the-report-ACC
I was going to write the report.
PAST PROSP
$\left[\begin{array}{lc}\text { PRED } & \text { WRITE < SUBJ, OBJ }> \\ \text { ASP } & \text { PROSP } \\ \text { TENSE } & {\left[\begin{array}{ll}\text { PAST } & +\end{array}\right]} \\ \text { SUBJ } & {\left[\begin{array}{ll}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]}\end{array}\right]$
(50) ${ }^{?} a k u \bar{u} n u \quad q a d$ katab-tu $t$-taqrīr-a
be.IPFV.1SG PT write.PFV-1SG the-report-ACC
(When I see you on Tuesdays), I have (always) written the report.
PRES PRF

$$
\left[\begin{array}{lc}
\text { PRED } & \text { WRITE < SUBJ, OBJ }> \\
\text { ASP } & \text { PERF } \\
\text { TENSE } & {\left[\begin{array}{ll}
\text { PAST } & -
\end{array}\right]} \\
\text { SUBJ } & {\left[\begin{array}{ll}
\text { PERS } & 1 \\
\text { NUM } & \text { SG }
\end{array}\right]}
\end{array}\right]
$$

(51) $s a-{ }^{?} a k \bar{u} n u \quad q a d$ katab-tu $t$-taqrīr-a FUT-be.IPFV.1SG PT write.PFV-1SG the-report-ACC

I will have written the report.

$$
\left[\begin{array}{lc}
\text { PRED } & \text { WRITE < SUBJ, OBJ }> \\
\text { ASP } & \text { PERF } \\
\text { TENSE } & {\left[\begin{array}{ll}
\text { PAST } & - \\
\text { FUT } & +
\end{array}\right]} \\
\text { SUBJ } & {\left[\begin{array}{ll}
\text { PERS } & 1 \text { SG } \\
\text { NUM } & \text { SG }
\end{array}\right]}
\end{array}\right]
$$

### 3.5 Negation and Compound Tenses

We consider first the compound forms with lam in (52), (53) and (54), forming the negative past perfect, negative past progressive and negative past prospective (54b) respectively (we return to (54c) shortly).
(52) a. kun-tu qad katab-tu t-taqrīr-a
be.PFV-1SG PT write.PFV-1SG the-report-ACC
I had written the report.
PAST PRF
b. lam ${ }^{2}$ akun qad katabtu $t$-taqrīr-a

NEG.PAST be.JUSS.1SG PT write.PFV-1SG the-report-ACC
I had not written the report.
(53)
a. kun-tu $\quad ? a k t u b-u \quad t$-taqrīr- $a$
be.PFV-1SG write-IPFV.1SG the-report-ACC
I was writing the report.
PAST PROG
b. lam $\quad$ ªkun aktub-u t-taqrir-a

NEG.PAST be.JUSS.1SG write-IPFV.1SG the-report-ACC
I was not writing the report.
a. kun-tu sa- ${ }^{\mathbf{Z}} \mathbf{a k t u b - u} \quad t$-taqrīr- $a$
be.PFV-1SG FUT-write-IPFV. 1 SG the-report-ACC
I was going to write the report.
PAST PROSP
b. lam ${ }^{\text {ªkun }} \quad$ sa- ${ }^{3} a k t u b-u \quad t$-taqrir- $a$

NEG.PAST be.JUSS.1SG FUT-write-IPFV.1SG the-report-ACC
I was not going to write the report.
c. kun-tu lan ${ }^{\mathbf{T}} \mathbf{a k t u b - a} \quad t$-taqrrīr- $a$
be.PFV-1SG NEG.FUT write-SBJV.1SG the-report-ACC
I was not going to write the report.

The relevant lexical entries previously given are (37) (41), (42) and (43), that is the entries for lam (as $\hat{\mathrm{I}}$ ), and for $k a t a b-t u(\mathrm{~V}),{ }^{?} a k t u b-u(\mathrm{~V})$ and $s a-{ }^{?} a k t u b-u(\mathrm{~V})$. The new lexical entry is for the be auxiliary in the jussive form in (55).

```
?akun \(\quad\) I \(\quad\) १ MOOD \()=\mathrm{JUSS}\)
    @ 1SG
```

Notice that compound verbs may involve the combination of perfective form and imperfective form verbs. No feature clash results because the perfective/imperfective distinction is one of morphological form rather than f-structure feature content: as we have seen, a perfective form verb conveys distinctions of tense when it occurs initial in the verbal sequence, and conveys distinctions of aspect when it is non-initial. Similarly, where NEG markers which govern the SBJV or JUSS moods (of the imperfective verb) combine with indicative verbforms (whether in perfective or imperfective form) no clash in the MOOD feature arises, on the assumption that indicative verbs are not marked for this feature. ${ }^{7}$ (56) is the resultant f -structure for (52b).
$\left[\begin{array}{ll}\text { PRED } & \text { WRITE < SUBJ, OBJ > } \\ \text { POL } & \text { NEG } \\ \text { ASP } & \text { PRF } \\ \text { MOOD } & \text { JUSS } \\ \text { TENSE } & {[\text { PAST }} \\ \text { SUBJ } & {\left[\begin{array}{ll}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]}\end{array}\right]$

Turning now to the compound forms with lan for the (negative) future perfect, shown in (57), and also in principle for the (negative) future progressive. ${ }^{8}$ What is required is a lexical description for the subjunctive of auxiliary be, shown in (58).

[^4]a. sa- ${ }^{\mathbf{T}} \mathbf{a k} \mathbf{k} \mathbf{n u} \quad q a d$ katab-tu $t$-taqrīr-a

FUT-be.IPFV.1SG PT write.PFV-1SG the-report-ACC
I will have written the report
FUT PRF
b. lan $\quad$ ? a-kun-a qad katab-tu $t$-taqrīr-a

NEG.FUT be.SBJV.1SG PT write.PFV-1SG the-report-ACC
I will not have writen the report
(58) ? ${ }^{?} a k u \bar{u}-a \quad I \quad$ ( $\uparrow$ MOOD) $=$ SBJV
@ 1SG


Finally, we consider compound forms with laa: recall that laa negates the imperfective, and does not itself express TENSE. It is used in the negative present perfect shown in (59). The lexical entry for the imperfective indicative of auxiliary be was already given in (45) and repeated here as (60) for convenience.
(59)
a. ${ }^{ }$akūnu qad katab-tu $t$-taqrīr-a
be.IPFV.1SG PT write.PFV-1SG the-report-ACC
(When I see you on Tuesdays), I have (always) written the report
PRES PRF
b. laa ${ }^{\mathbf{2}}$ akūn-u qad katab-tu t-taqrīr-a NEG be.IPFV.1SG PT write.PFV-1SG the-report-ACC
.. I have not (already) written the report
(60) ${ }^{2} a k u \bar{u}-u \quad I \quad(\uparrow$ TENSE PAST $)=-$ @ 1SG
(61) $[$ PRED WRITE $<$ SUBJ, OBJ $>$ POL NEG
ASP PRF
TENSE [PAST - ]
SUBJ $\left[\begin{array}{cc}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]$

Before leaving laa and its tensed counterparts lam and lan, there is one further and intriguing data point, namely the example (54c), which appears to be an alternative to the (expected) (54b). It seems
that LAN + SBJV may occur in V position for semantic reasons which are not entirely clear to us. We incorporate this datum into our description by hypothesizing that only lan can adjoin to V (as well as I): tensed lan (but not lam) has an aspectual counterpart as shown in (62). This combines with a verb in subjunctive mood. We give the lexical entry for this verb in (63).
(62) lan $\hat{V}(\uparrow$ ASP $)=$ PROSP
$(\uparrow$ POL $)=$ NEG
$(\uparrow$ MOOD $)={ }_{c}$ SBJV
(63) ${ }^{2} a k t u b-a \quad V \quad(\uparrow$ PRED $)=$ write $<$ SUBJ OBJ $>$ $(\uparrow$ MOOD $)=$ SBJV @ 1SG
(64) kun-tu $I \quad(\uparrow$ TENSE PAST $)=+$ @ 1SG
(65)
$\left[\begin{array}{ll}\text { PRED } & \text { WRITE < SUBJ, OBJ }> \\ \text { POL } & \text { NEG } \\ \text { ASP } & \text { PROSP } \\ \text { MOOD } & \text { SBJV } \\ \text { TENSE } & {[\text { PAST }} \\ \text { SUBJ } & {\left[\begin{array}{cc}\text { PERS } & 1 \\ \text { NUM } & \text { SG }\end{array}\right]}\end{array}\right]$

The following summarises the data concerning negation with laa, lam, lan and compound tenses.
(66)

|  | FORM | REALIZATION | MEANING |
| :--- | :--- | :--- | :--- |
| 1. | LAM JUSS | lam ${ }^{~}$ aktub-a t-aqrīr-a <br> I did not write the report. <br> laa 'aktub-u t-aqrir-a | PAST |
| 2. | LAA IPFV |  | I am not writing the report. <br> lan ${ }^{?}$ aktub-a t-taqrīr-a <br> I will not write the report. |

## 4 Analysis of Laysa

After working through the details (of compound tense formation) necessary to give a reasonably full description of the negative particles laa, lam, lan, the analysis of laysa is quite straightforward. Rather than being a non-projecting word, laysa is a fully projecting I taking a range of complements. If its c-structure complement is verbal, then that complement must be in the indicative Imperfective form. As a fully projecting element, laysa is not subject to any adjacency restriction with respect to the verbal element. Note that as expected for a tensed verb, agreement with the SUBJ in VSO structures is partial (contrast (67a) and (67b)).
(67) a. al-awlad-u lays-uи ya-ktub-uun.
the-boys-NOM NEG-3MP 3M-write.IPFV-3MP-IND
The boys do not write/are not writing.
b. lays-a al-awlad-u ya-ktub-uun.

NEG-3MP the-boys-NOM 3M-write.IPFV-3MP-IND
The boys do not write/are not writing.
(68)

(69)


The lexical description for the negative tensed auxiliary laysa is given in (70):

```
(70) laysa I ( }\uparrow\mathrm{ TENSE PAST) = -
    (\uparrow TENSE FUT) = -
    (\uparrow POL) = NEG
    (\uparrow SUBJ PERS) = 3
    (\uparrow SUBJ GEND) = MASC
    V \in CAT }(\uparrow)=>(\uparrow\textrm{ASP})=\mp@subsup{}{c}{}\textrm{PROG
```

ya-ktub-uu-n $\quad V \quad(\uparrow$ PRED $)=$ study $<$ SUBJ $>$
$(\uparrow$ ASP $)=$ PROG
@ 3MPL

This accounts for the key aspects of the distribution of laysa which were noted above, namely, that it can occur in verbless and verbal sentences, it can be separated from the verb, and if it occurs with a verb, that verb is indicative imperfective in form.

## 5 Future Work and Open Questions

The approach outlined here is preliminary in very many ways, and there are a number of open questions which we intend to explore in future work. In particular, the approach to Tense and Aspect which we outline here is very preliminary. Further work is also needed on other possible non-projecting verbal particles (such as qad). In terms of the negative particles and the observed depedencies between particles and verbforms, the question remains as to whether selection between negative particles and verb forms should be dealt with in terms of c-structure (sub)categories: this seems to us to be quite an attractive alternative to the f-structure selection account (using the MOOD feature, which we outlined here. On such an alternative approach, one might encode the form selections as follows:

$$
\begin{align*}
\text { a. lan } \hat{I} & (\uparrow \text { TENSE FUT })=+  \tag{71}\\
& (\uparrow \text { POL })=\mathrm{NEG} \\
& \lambda(\hat{*} \mathrm{compl})={ }_{c} \mathrm{~V}_{[s b j v]}
\end{align*}
$$

b. lam $\hat{I} \quad(\uparrow$ TENSE PAST $)=+$
$(\uparrow$ POL $)=$ NEG
$\lambda\left({ }^{*}\right.$ compl $)={ }_{c} \mathrm{~V}_{[j u s s]}$
c. laa $\hat{I} \quad(\uparrow$ TENSE PAST $) \neq+$ $(\uparrow$ POL $)=$ NEG $\lambda(\hat{*} \mathrm{compl})={ }_{c} \mathrm{~V}_{[\text {indic }]}$

## References

Bahloul, Maher. 1994. Extending the NegP Hypothesis: Evidence from Standard Arabic. In M. Eid, ed., Perspectives on Arabic Linguistics: Papers from the annual symposium on Arabic Linguistics, vol. VIII, pages 31-47. Amsterdam: John Benjamins.

Benmamoun, Elabbas. 2000. The feature structure of functional categories: A comparative study of Arabic dialects. Oxford: Oxford University Press.

Falk, Yehuda N. 2008. Functional relations in the English auxiliary system. Linguistics 46/4:861-889.
Fassi-Fehri, Abdulkader. 1993. Issues in the Structure of Arabic Clauses and Words. Dordrecht, Holland: Kluwer Academic Publishers.

Ouhalla, Jamal. 2002. Negative sentences. In J. Ouhalla and U. Shlonsky, eds., Themes in Arabic and Hebrew Syntax, pages 299-320. Dordrecht, Holland: Kluwer.

Shlonsky, Ur. 1997. Clause Structure and Word Order in Hebrew and Arabic: An Essay in Comparative Semitic Syntax. Oxford: Oxford University Press.

Toivonen, Ida. 2003. Non-Projecting Words: A Case Study of Swedish Particles. Dordrecht: Kluwer Academic Publishers.

# THE ADJECTIVAL CONSTRUCT IN ARABIC 

Budour Al Sharifi and Louisa Sadler<br>University of Essex University of Essex

## Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://www-csli.stanford.edu/


#### Abstract

We propose an analysis of the adjectival construct in Arabic in LFG, building on previous work in LFG on a Welsh construction which shows several similarities to the Arabic (Mittendorf and Sadler, 2008) and work on the MSA and cognate Hebrew constructions by Hazout (2000); Kremers (2005); Siloni (2002); Heller (2002) and Kim (2002). The construction involves an adjective occurring with an immediately following definite nominal, which denotes a property, part or quality of the noun that the adjective modifies (in attributive use) or is predicated of (in predicative constructions). The major characteristics of this construction are that the post-adjectival nominal serves to delimit the respect in which the adjective denotes a property of the external nominal, and the adjectival head and the GEN complement are subject to a very strict adjacency requirement. We present a syntactic analysis, showing that the construction is formed in the syntax rather than the morphology, and sketch out a proposal as to how the semantics of the construction might be captured.


## 1 Introduction

We propose an analysis of the adjectival construct in Arabic in LFG, building on previous work in LFG on a Welsh construction which shows several similarities to the Arabic (Mittendorf and Sadler, 2008) and work on the Modern Standard Arabic MSA and cognate Hebrew constructions by Hazout (2000); Kremers (2005); Siloni (2002); Heller (2002) and Kim (2002). ${ }^{1}$

## 2 The Adjectival Construct

The adjectival construct consists minimally of an adjective or participle in construct with a noun which specifies the degree or manner of the property expressed by the adjective: an example appears in boldface in (1). It is traditionally described as ${ }^{ }$daafa gayr haqīqiyya or the false construct phrase. It is found in both Modern Standard Arabic (MSA) and in the Arabic vernaculars, although all the work that we are aware of concerning this construction discusses its MSA instantiation (basically Kremers (2005) and passing references in the literature on Hebrew.).
(1) imra ${ }^{2}$-at-un ğamil-at-u -l-wağh-i
woman-F-NOM beautiful-F-NOM the-face-GEN
a woman with a beautiful face
(Kremers, 2005)

It has the external distribution of an AP, occuring both attributively as in (1) and (2) and predicatively, shown in (3) and (4).
(2) bayt-un katīr-u -l-'abwāb-i
house-NOM.INDEF many-NOM the-doors-GEN
a house with many doors
(Kremers, 2005)

[^5](3) 'anta ${ }^{\text {c azūm-u }}$-l-hazz-i
you great-NOM the-fortune-GEN
You are very lucky.
(Kremers, 2005)
(4) $k a \bar{a} a$ huwlandiy-a l-?asl-i
was dutch-ACC the-origin-GEN
He was of Dutch origin.
(Ryding, 2005, 254)

The 'inner' N is obligatorily definite in MSA (and this is one aspect in which the construction differs from both its Hebrew and Welsh counterparts). In addition, the 'inner' N appears in GEN case in MSA (one area of difference between MSA and the vernaculars is that case marking is absent in the latter). A striking characteristic of the construction is that nothing can intervene between the Adj head and the GEN complement. As shown in (5), adjectival modifiers such as intensifiers normally come directly after the adjective they modify. However, if that adjective is itself in construction with a genitive complement, that complement obligatorily separates the adjectival head from its modifier, as illustrated by the grammaticality constrast between (6) and (7).
(5) $\breve{g} a m \bar{l} l-u n \quad \breve{g} i d d a n$
beautiful-NOM very
very beautiful
(6) $i^{2 m r a}{ }^{2}$-at-un ğamīl-at-u -l-wajh-i ğiddan
woman-F-NOM beautiful-F-NOM the-face-GEN very
a woman with a very beautiful face
(7) *imra ${ }^{3}$-at-un ğamil-at-u ğiddan -l-wajh-i
woman-F-NOM beautiful-F-NOM very the-face-GEN
a woman with a very beautiful face

The adjective in attributive use agrees with the head noun in CASE, NUM and GEN, which is the expected behaviour for adjectives in MSA. So for example, (8) is a noun phrase headed by a definite FSG noun which is itself modified by an adjective $l$-barlamāniyyat- $u$ and the adjectival construct $l$ $w \bar{a} s i^{c} a t-u l-n a f \bar{u} d \underline{d}-i$. In each modifier the adjective agrees with the head noun in CASE, NUM and GEN (9) is headed by a FSG indefinite noun, which is modified by an adjectival construct mutawassit-u $l$-hağm- $i$ in which the adjective agrees with the head noun in CASE, NUM and GEN

## (8) al-lağnat-u l-barlamāniyyat-u l-wāsi ${ }^{c}$ at-u l-nafūde-i

the-committee.FSG-NOM the-parliamentary.FSG-NOM the-wide.FSG-NOM the-influence-GEN
the widely influential parliamentary committee
(Ryding, 2005, 254)
(9) qidr-un mutawassit-u l-hağm-i
pot.FSG-NOM.INDEF medium.FSG-NOM the-size-GEN
a medium-sized pot
(Ryding, 2005, 254)

The examples above also show the adjectival construct showing definiteness agreement (alongside other attributive adjectives). However it differs in one respect in its agreement behaviour, in that it cannot show indefinite agreement (nunation) with an indefinite nominal head.

## (10) *imra ${ }^{2}$-at-u-n ğmil-at-u-n -l-wajh-i

woman-F-NOM-INDEF beautiful-F-NOM-INDEF the-face-GEN a woman with a beautiful face
(Kremers, 2005, 3)

Unlike compounds, the construction is productive and gives rise to compositionally predictable meanings. Furthermore, the sort of evidence that may be interepreted to support a morphological analysis for the adjectival construct in Hebrew (see Hazout (2000)), or in favour of a special prosodic status (see Siloni (2002)) is absent in MSA.

The construction expresses a particular relation between the two nouns. It typically occurs modifying a head noun within an NP "with the genitive noun specifying in which respect the adjective applies to the noun that it is predicated of" (Kremers, 2005). A descriptive grammar observes: "This kind of phrase is used to describe a distinctive quality of an item, equivalent to hyphenated expressions in English such as fair-haired, long-legged, many-sided" (Ryding, 2005, 254).

The adjectival construction can occur predicated or (or modifier to) the full range of NPs:
(11) Kul-u rağul-in tawīl-u l-s̆a ${ }^{\mathbf{c}} \mathbf{r} \mathbf{r} \mathbf{i}$
every-NOM man-GEN tall-NOM the-hair-GEN
every man long the-hair (every long-haired man)
(12) al-ră̆ul-u tawīl-u l-ša ${ }^{\mathbf{c}} \mathbf{r} \mathbf{r - i}$
the-man-NOM tall-NOM the-hair-GEN
the man long the hair
(13) la rağul-a tawīl-u l-ša ${ }^{\mathbf{c}} \mathbf{r - i}$
no man-ACC tall-NOM the-hair-GEN
no man long the hair
(14) talāt-u riğăl-in tawāl-u l-šá ${ }^{\mathbf{c}} \mathbf{r}$-i
three-NOM men-GEN tall-NOM the-hair-GEN
three man long the hair
(15) al-katir-u min al-riğăl-i tawāl-u l-ša ${ }^{\mathbf{c}} \mathbf{r}-\mathbf{i}$
the-many-NOM from the-man-GEN tall-NOM the-hair-GEN
many men long the hair

Traditional and contemporary descriptions and analyses observe a number of key similarities with the construct state construction which, unlike the adjectival construct, has received a good deal of theoretical attention (Ritter, 1988, 1991; Benmamoun, 2000; Kremers, 2003; Falk, 2007). In a nominal
construct state a nominal head is in close construction with a following NP. These similarities between the adjectival construct and the nominal construct state have motivated a number of analyses extending the approach from the latter to the former. In the MSA nominal construct state construction, the head noun is modified by an NP showing GEN case in MSA (the vernaculars have lost case marking), and has no formal definiteness marking (neither -n nor al-) unlike other nouns (a special construct form of both adjectives and nouns exists in Hebrew). The head and the construct argument (GEN complement) are inseparable: modifiers of the head appear after the construct NP (as with (6)).
(16) sayyār-at-u -l-rağul-i -l-ḥamrā'-u car-F-NOM the-man-GEN the-red-NOM
the man's red car

## 3 Previous Approaches and Related Work

In this section we describe what we take to be related constructions in Semitic and Celtic and briefly outline the analyses provided in the literature of these constructions. As noted above, there is not much work done on the Arabic adjectival construct itself (but see Kremers (2005)), but there has been significantly more work on the cognate Hebrew construction including Hazout (2000); Siloni (2002) and Kim (2002). The Hebrew construction, illustrated in (17) uses a special construct form of the adjective, yefat in the following examples.
(17) yalda yefat mar'e nixnesa la-xeder. girl.FSG beautiful.FSG look.MSG entered to.the-room
A good looking girl entered the room.
(Siloni, 2002, Hebrew)

Adjectival constructs do not show definiteness agreement with the external noun, rather the inner noun reflects the definiteness feature of the external noun:
(18) ha-yalda (*ha-)yefat *(ha-)mar'e nixnesa la-xeder.
girl.FSG the-beautiful.FSG the-look.MSG entered to.the-room
The good looking girl entered the room.
(Siloni, 2002, Hebrew)

### 3.1 Characteristics of Siloni's Approach

A key proposal in Siloni (2002) is that the construction is limited to inalienable nouns, and is found most typically (but not exclusively) with body parts. The claim is that alienable nouns are not found in this construction. However since parts of wholes can appear in the construction Siloni argues that they function as "extended inalienables", giving the following examples.
(19) xadarim gvohey tikra
rooms high ceiling
high-ceiling rooms
(20) sira gvohat toren
boat high mast
a high-mast boat

The basic idea is that adjectives have an open slot, which is saturated by the noun that they are predicated of or modify. In the construct construction, the empty slot is filled by the internal genitive. This would mean that an adjective in a construct construction would not be able to modify an external head noun. There has to be a slot for this to fill in the argument structure (thematic grid) of the adjective. The proposal (an assumption also shared by Hazout (2000)) is that this slot is in fact the possessor argument of the internal genitive argument. That is, "the particularity of inalienable nouns which qualifies them (and them only) to form adjectival constructs" is that they have a lexical possessor. In defence of this view, Siloni notes that with these (inalienable) nouns the possessor can only be missing in generic contexts, as in the following example:
(21) be-mitkan ze ha-roš zakuk le-hagana meyuxedet.
in-installation this the-head requires to-protection special
In this installation the head requires a special protection
(Siloni, 2002, Hebrew)
The idea (and the formal details of the treatment are not made clear) is that in the syntax "the inalienable noun and the adjectival head form a complex predicate ... which is saturated by the external noun". Hence, if alienable nouns do not have a possessor slot, then it will follow that they do not occur in the construction.

Since kinship terms are excluded from the construction (and these are arguably inalienable), Siloni must also point out ways in which the behaviour of kinship nouns is different from that of body part nouns, in order to maintain the assumption about lexical possessors. She notes that unlike body parts, you cannot have an external possessor (SUBJ or DAT) with a kinship noun, only an internal possessor, in Hebrew, and the possessorless kinship noun is also not permitted in a generic context (contrast with (21) above).

The characteristics taken to be typical of inalienable constructions are as follows.

## The distributivity effect

(22) ha-rofe badak lahem 'et ha-roš
the-doctor examined to.them ACC the-head
The doctor examined their heads.
(Siloni, 2002)

The singular constraint: parts which you have only one of are obligatorily singular, irrespective of the number of the external possessor in such constructions. Compare (22) to the examples below.
a. ha-yeladim herimum'et ha-yadayim
the-children raised ACC the-hands
The children raised their hands.
b. *ha-rofe badak lahem 'et ha-rašim the-doctor examined to.them ACC the-heads The doctor examined their heads.

## Limitation to restrictive modifiers only

(24) a. *ha-rofe badak lo 'et ha-roš ha-pacu'a the-doctor examined to.him ACC the-head the-wounded The doctor examined his wounded head.
b. *ha-rofe badak la 'et ha-yad ha-švura the-doctor examined to.her ACC the-head the-broken The doctor examined her broken hand.
(Siloni, 2002)

These constraints do not show up when the possessor is DP internal, since it is not the case that only lexical possessors are permitted in this construction, hence compare (25) to (22).
(25) hi badka 'et ha-rašim šel ha-yeladim.
she examined ACC the-heads of the-children
She examined the children's heads.
(Siloni, 2002)

These constraints also operate in the adjectival construct:
(26) ne'arim 'arukey xotem/*xotamim hištatfu ba-taxrut
guys long nose/*noses participated in.the-context
Long-nosed guys participated in the contest.
(Siloni, 2002)

Beyond this, there are further restrictions on the form of the Hebrew adjectival construct (some, but not all of which are shared by the MSA construction). No form of modification, either restrictive or appositive, is permitted in the Hebrew adjectival construct. Moreover the inalienable noun (internal complement) cannot take an of-complement and cannot itself take the form of a construct state construction. Siloni proposes that these restrictions follow from the fact that the genitive complement does not project a full referential DP. ${ }^{2}$ The article which appears on the complement nominal is argued to be the concordial article which would normally have appeared on the adjective, but cannot because heads of constructs never occur with an article.

### 3.2 Kim

Kim (2002) also proposes a complex predicate analysis, a proposal which largely accepts the syntactic assumptions of the analysis presented in Siloni (2002) and supplements it by providing further specification of the semantics associated with the construction, in order to account for the restriction to inalienable nouns.

The semantic translation that she gives for the construct adjective is as a function which maps a twoplace predicate into a one-place predicate.

$$
\text { (27) pretty.CS } \lambda \mathrm{R}_{<e e t>} \lambda \mathrm{x}[\operatorname{pretty}(\iota \mathrm{y}[\mathrm{R}(\mathrm{x})(\mathrm{y})])]
$$

[^6]This combines with the two place noun eyes (the assumption being that there is an internal argument denoting the possessor, and an external argument denoting the referent).

```
eyes \lambdau \lambdav [eyes(u)(v)]
```

The result of applying the construct state adjective to the nominal is as given in (29).
(29) pretty.CS eyes $\lambda \mathrm{x}[\operatorname{pretty}(\iota \mathrm{y}[\operatorname{eyes}(\mathrm{x})(\mathrm{y})])]$

This predicts also that any modification which does not impinge on the argument structure of the inner noun will in fact be grammatical.

On the remaining properties (the uninterruptibility of the ADJ N combination, the placement of the definiteness affix on the noun rather than the adjective), Kim follows Siloni (2002) in ascribing these to the prosody of Hebrew. Modifiers of the adjective appear "postposed" because the adjective is prosodically defective and gets stress via the complement, so nothing can intervene.

On the question of how and why the construction is limited to cases of inalienable possession, Kim is wary of following Siloni (2002) in attributing this to the need for a lexically specified possessor. In her account, in principle anything which can be appropriately typeshifted could in principle occur in the construction, but then essentially the idea is that they are filtered out by the syntax (the assumption being that the possessor slot of an inalienable is anaphoric while that of an "unrestricted" possessor is a pronominal (Koenig, 1999)). So you can produce the semantics but it is impossible to make, for example, "girl" the antecedent of the pronominal and thus the binding constraints would fail. ${ }^{3}$

### 3.3 Kremers

Kremers (2005) points out several empirical problems with the the analysis offered in Siloni (2002) and Kim (2002) (and by extension, the rather similar analysis of Hazout (2000)). They all take the construction as some form of complex adjective which is syntactically formed, with the inner nominal some type of inalienable noun which is the subject of the adjectival complex predicate. The external or head noun is taken to satisfy the possessor argument slot of the inner noun, which must be assumed to become an argument of the adjective by some sort of merger during the (syntactic) process of complex predicate formation. In fact, no details of the operation of complex predicate formation are given in Siloni (2002) and Kremers's observation that the process whereby "an argument of the genitive complement becomes the external argument of the adjectival construct" (Kremers, 2005) is opaque is certainly one we can agree with. ${ }^{4}$ He points out a number of further problems with the assumption that the inner nominal is the external argument of the entire construction. First, if this were the case, then the adjective should agree with this argument, whereas in fact, as we have seen, in attributive use the adjective shows concord with the external noun which it modifies. Second, the fact that the construction occurs predicatively is also incompatible with taking the genitive complement as its external argument, because the nominal it is predicated of fills this role. Finally, Kremers (2005)

[^7]notes that there are several problems with the assumption that the construction should be explained in terms of the occurrence of lexical possessors: kinship nouns also have lexical possessors, yet cannot occur in the construction (a restriction which Kim does take account of), while on the other hand, in MSA the use of alienably possessed complements is widespread. These considerations would suggest that the existence of a lexically specified possessor is not the main issue.

Kremer's alternative proposal is that the genitive (or inner) argument is not an external argement of the adjective but names some property or inherent part of the head noun and fulfills an internal thematic role of the adjectival head. He takes this role to be the attribute role of Higginbotham (1985). The genitive case which occurs in this construction (and a number of others in MSA) is a structural case that is assigned to an internal argument (and that internal argument can bear a wide range of thematic roles).

### 3.4 Welsh Genitive of Respect

In previous LFG work, Mittendorf and Sadler (2008) discuss a construction occurring in Welsh (and the other Celtic languages) which shows several resemblances to the Arabic construct adjectives. This is the construction illustrated in (30) - (33).
(30) dyn uchel ei gloch
man.MSG high.MSG his bell.FSG
a loud-mouthed man
(31) merch fyr ei thymer
girl.FSG short.MSG POSS. 3 SG temper.FSG
a short-tempered girl
(32) Mawr eu dawn yw'r gwŷr big.MSG their talent.FSG is the men.MSG
Hugely talented are the men.
(33) Mae'rferch yn fyr ei thymer.
is-the girl PRED short.MSG her temper.FSG
The girl is short-tempered.

The Welsh adjectival in-respect-of construct is a construction that is headed by the A and contains a (definite) NP: the AP can appear in syntactic environments that exclude definite NPs, showing that the construction is headed by the adjective not the NP. As the examples above show, the syntax of the inner NP differs in Welsh in that it contains an obligatory (possessor) clitic pronoun, which is obligatorily anaphoric to the head noun (in attributive use) and the SUBJ function in predicative use. In attributive position, the adjectival construct shows initial consonant mutation properties typical of APs, but it has slightly unusual agreement properties in that the A itself does not agree with either the head N or the following N . In terms of constituent structure, the NP appears (almost immediately) post-head in direct argument position, and in fact can be separated from the head adjective only by one of a small class of intensifying modifiers. Finally, the core of the relationship between the post-A

NP and the external N is one of inalienable possession: "The thing or quality denoted by the [post-A NP ] pertains to or is a part of the person or object denoted by [the SUBJ or head N ], the latter being represented by the poss[essive] pronoun" (Mac Cana, 1966, p. 91). Mittendorf and Sadler (2008) deal primarily with the morphosyntax of this construction and do not discuss in detail the semantic relations which must hold between the inner and outer nominals in the construction, but here too there are significant crosslinguistic similarities.

Mittendorf and Sadler (2008) propose that the attributive construction is analysed along the lines of (34) (for (31) and the predicative construction as shown in (35) for (33). ${ }^{5}$

$$
\begin{align*}
& {\left[\begin{array}{ll}
\text { PRED } & \text { GIRL }_{i} \\
\text { ADJ } & \left\{\left[\begin{array}{lll}
\text { PRED } & \text { SHORT }<\text { OBJ }> \\
\text { OBJ } & {\left[\begin{array}{ll}
\text { PRED } & \text { TEMPER }<\operatorname{POSS}> \\
\text { POSS } & {[\operatorname{PRED}} \\
\mathrm{PRO}_{i}
\end{array}\right]}
\end{array}\right]\right\}
\end{array}\right]}  \tag{34}\\
& {\left[\begin{array}{cl}
\text { PRED } & \text { SHORT }<\text { SUBJ OBJ }> \\
\text { SUBJ } & {[\text { PRED }} \\
\text { GIRL } \left._{i}\right]
\end{array}\right]} \tag{35}
\end{align*}
$$

In Welsh, only the default form of the adjective MSG permits this construction, by lexically selecting an OBJ. ${ }^{6}$

```
a. byr
```

```
\(\{(\uparrow\) PRED \()=\) SHORT
```

$\{(\uparrow$ PRED $)=$ SHORT
$\mid(\uparrow$ PRED $)=$ SHORT $<$ OBJ $>$
$\mid(\uparrow$ PRED $)=$ SHORT $<$ OBJ $>$
$\mid(\uparrow$ PRED $)=$ SHORT $<$ SUBJ $>$
$\mid(\uparrow$ PRED $)=$ SHORT $<$ SUBJ $>$
$\mid(\uparrow$ PRED $)=$ SHORT $<$ SUBJ OBJ $>\}$
$\mid(\uparrow$ PRED $)=$ SHORT $<$ SUBJ OBJ $>\}$
no GEND/NUM constraints

```
no GEND/NUM constraints
```

b. ber

$$
\begin{aligned}
& (\uparrow \text { PRED })=\text { SHORT } \\
& ((\text { ADJ } \in \uparrow) \text { GEND })={ }_{c} \mathrm{~F} \\
& ((\mathrm{ADJ} \in \uparrow) \mathrm{NUM})={ }_{c} \mathrm{SG}
\end{aligned}
$$

They assume that the 'special' occurrence of the grammatical function OBJ in lexical entries such as (36 a) would be associated with a particular semantics introducing the respect/quality property, but they do not formulate this. The linkage between the NP-internal bound pronoun and the modified head N/SUBJ can be established in the c-structure as shown in (37). ${ }^{7}$


[^8]
## 4 Analysis

Returning now to the Arabic data, there is good evidence in MSA that the construction is formed in the productive syntax (not in the lexicon/morphology). The inner NP (denoting the dimension in which the quality in question holds) is accessible to regular syntax, shown by the fact that this argument can be coordinated, as in (38), and that it can be syntactically modified, as in (39).
(38) Dahַla ră̆gl-un aswād-u l-šacr-i wa l-cāynayin-i
come.3SGM.PAST man-NOM black-NOM the-hair-GEN and the-eye.DUAL-GEN
A man black of hair and eyes came (A man with black hair and eyes came).
(39) bayt-un katir-u -l-ābwāb-i -lhdar-i
house-NOM many-NOM the-door.PL-GEN the-green-GEN
a house with many green doors.

This means that the inner nominal is neither a non-projecting word nor part of a morphological construction. Neither is it the case that the construct adjective is prosodically defective: the adjective which occurs in this construction is not a special form, but a regular adjective in all respect. Therefore, our conclusion is that the construction is simply and straightforwardly a product of the general phrase structure of Arabic. Arabic phrase structure must permit an (optional) GEN complement immediately adjacent to an adjectival head. We take it that there is good evidence that GEN is indeed a structural case in MSA (it is the case found on prepositional objects, for example).

### 4.1 External Behaviour of Adjectival Construct

As shown in (3), an adjectival construct may have the normal distribution of predicate adjectives. Attributively, it has the normal distribution of an AP modifier: it co-occurs with other NP nominal modifiers (as a modifier, it is unusual only in showing defective definiteness concord NP internally).
(40a) is stylistically preferred over (40b). Similarly (41a) is better than (41b) for the same reason. Both are, however, acceptable and thus we assume both are to be permitted alongside other orderings of nominal modifiers by the c-structure constraints.
(40) a. al-rağul-u l-muthaqqaf-u l-ṭawīl-u l-qāmat-i
the-man-NOM the-cultured-NOM the-tall-NOM the-height-GEN the cultured, tall man
(Ryding, 2005, 222)
b. al-rağul-u l-ṭaw̄̄l-u l-qāmat-i l-muthaqqaf-u
the-man-NOM the-tall-NOM the-height-GEN the-cultured-NOM
the cultured, tall man
(41)

[^9]b. imra²-at-un $\breve{g} a m i \bar{l}-a t-u \quad-l-w a \breve{g} h-i \quad t a w \bar{l} l-a t-u n$
woman-F-NOM.INDEF beautiful-F-NOM the-face-GEN tall-F-NOM.INDEF
a tall woman with a beautiful face (a tall beautiful-faced woman)
Note that examples such as (40a) show that the head N and the adjectival construct are not required to be linearly adjacent (or form a small construction, for example). The following rule then generates the adjectival construct alongside any other AP modifiers of the NP.
\[

$$
\begin{array}{lccc}
\mathrm{NP} & \rightarrow & \mathrm{~N} & \mathrm{AP}^{*}  \tag{42}\\
& \uparrow=\downarrow & \downarrow \in(\uparrow \text { ADJ })
\end{array}
$$
\]

### 4.2 Internal Structure of AP

The main question that this construction raises is that of the syntactic status of the post-adjectival NP. The "inner" NP immediately follows the adjectival head, suggesting it is a subcategorised (direct) argument of the adjective. The existence of the predicative use of the construction suggests that the GEN argument is not SUBJ (see Kremers (2005) and Mittendorf and Sadler (2008) on Welsh on this point). We propose that adjectives assign GEN case to their direct internal argument (adjectives may also take other types of complements in Arabic, including prepositional and verbal complements). Adjectives would not be alone in assigning a structural GEN case, as we find GEN marking the direct complements of prepositions, numbers, and (some) quantifiers, and also in the nominal CS construction.

These observations motivate the following rule. The construct NP is immediately posthead and maps to a direct argument. Here we call this OBJ but note that it could well be $\mathrm{OBJ}_{\theta}$. Since it is restricted essentially to attribute or quality, then this may well be more appropriate but we do not pursue that here. The requirement that the construct nominal must be syntactically definite is captured by the constraining equation.

```
(43) A' }->\mathrm{ A NP PP*
    \uparrow=\downarrow\quad(\uparrow OBJ)=\downarrow (\uparrow OBL)=\downarrow
    (\downarrow\textrm{DEF})=\mp@subsup{}{c}{}+
```


### 4.3 $\quad$ F-structure

Consider now the f-structure that follows from this proposal, shown for both attributive and predicative examples below. Adjectives (or more specifically, a subclass of adjectives) may (optionally) subcategorise for an object, which will be subject to a particular interpretation in the semantics.
(44) imra ${ }^{2}$-at-un $\quad$ gamīl-at-u -l-wağh-i
woman-F-NOM beautiful-F-NOM the-face-GEN
a woman with a beautiful face
(Kremers, 2005)
(45)

$$
\left[\begin{array}{ll}
\text { PRED } & \text { WOMAN } \\
\text { ADJ } & \left\{\left[\begin{array}{ll}
\text { PRED } & \text { BEAUTIFUL }<\text { OBJ }> \\
\text { OBJ } & {\left[\begin{array}{ll}
\text { PRED } & \text { FACE } \\
\text { DEF } & +
\end{array}\right]}
\end{array}\right]\right\}
\end{array}\right.
$$

(46) kāna huwlandiy-u l- ${ }^{3}$ aṣl-i was.3SGM dutch-NOM the-origin-GEN
He was of Dutch origin.
(47)
$\left[\begin{array}{ll}\text { PRED } & \text { BE } \\ \text { TENSE } & \text { PAST } \\ \text { SUBJ } & 1: \\ \text { XCOMP } & {\left[\begin{array}{ll}\text { PRED } & \text { PRO }]\end{array}\right.} \\ {\left[\begin{array}{ll}\text { PRED } & \text { DUTCH }<\text { SUBJ, OBJ }> \\ \text { OUBJ } & 1: \\ \text { OBJ } & {\left[\begin{array}{ll}\text { PRED } \\ \text { DEF }\end{array}+\right.}\end{array}\right]}\end{array}\right]$

In terms of its f-structure, a construct adjective is also special in one further aspect which is that in attributive use it does not show agreement in indefiniteness (nunation). Otherwise, it exhibits normal NP internal concord in CASE, DEF, GEND and NUM. So as this following example shows, when it occurs with other AP modifiers of an indefinite nominal head, it will be the only one to fail to show full indefiniteness agreement. In (48), the head of the nominal within which the adjectival construct agrees is itself INDEF and GEN (it is GEN because it is the structural complement of the numeral ? $a w w a l-u$ ). The adjective rafic ${ }^{c}-i$ is defective in showing GEN but not nunation, the marker of indefiniteness. The standard agreement pattern is as shown in (49).

he first-NOM official-GEN.INDEF american-GEN.INDEF high-GEN the-level-?
ya-zuur-u l-bahrayn-a.
3s?-visit-NOM the-bahrain-ACC
He is the first high-level American official to visit Bahrain.
(Ryding, 2005, 222)
(49)
a. bayt-u-n ǧamı̄l-u-n
house-NOM-INDEF beautiful-NOM-INDEF
a beautiful house
(Kremers, 2003, 167)
b. al-bayt-u -l-ǧamīl-u
the-house-NOM the-beautiful-NOM
the beautiful house
(Kremers, 2003, 167)
c. al-riğal-u -l-ṭiwāl-u
the-men-NOM the-tall.PL-NOM
the tall men
(Kremers, 2003, 58)
d. al-nisā-u -l-ṭaw $\bar{l} l-\bar{a} t-u$
the-women-NOM the-tall-F.PL-NOM
the tall women
(Kremers, 2003, 58)

If all adjectives and nouns in MSA are fully specified for GNCD, then definiteness agreement is handled along the lines of CASE and PNG agreement. The templates in (50) and (51) hence show the agreement information for MPL.NOM.DEF and MPL.NOM.INDEF adjectives respectively. ${ }^{8}$
(50) MPL-DEF-NOM-ADJ $\equiv((A D J \in \uparrow) D E T)=$ DEF
$(($ ADJ $\in \uparrow) \mathrm{CASE})=\mathrm{NOM}$
$((\operatorname{ADJ} \in \uparrow)$ CONC GEND $)=$ MASC
$((A D J \in \uparrow) \operatorname{CONC} N U M)=$ PLUR
(51) MPL-INDEF-NOM-ADJ $\equiv((A D J \in \uparrow)$ DET $)=$ INDEF
$((\operatorname{ADJ} \in \uparrow) \mathrm{CASE})=\mathrm{NOM}$
$((\operatorname{ADJ} \in \uparrow) \operatorname{CONC}$ GEND $)=$ MASC
$((\operatorname{ADJ} \in \uparrow) \operatorname{CONC} N U M)=$ PLUR

The lack of nunation on indefinite adjectives within this particular construction may then be handled lexically by specifying that only the definite and the bare adjective (without tanwiin) permit the CS construction. ${ }^{9}$

```
-l-ṭiwāl-u {(\uparrow PRED) = 'TALL' | (\uparrow PRED ) = 'TALL (< SUBJ >)'}
    @ MPL-DEF-NOM-ADJ
```

(53) tiwāl-un ( $\uparrow$ PRED) $=$ 'TALL'
@MPL-INDEF-NOM-ADJ
(54) ti tiwāl-u $\{(\uparrow$ PRED $)=$ 'TALL' $\mid(\uparrow$ PRED $)=$ 'TALL $(\langle$ SUBJ $\rangle)$ ' $\}$
@ MPL-INDEF-NOM-ADJ

## 5 Semantics of the Construction: Some Initial Thoughts

As pointed out in connection with the Welsh examples (55) and (56) in Mittendorf and Sadler (2008), the Arabic examples (57) and (58) also indicate that the adjective does not apply directly to the head noun, but is restricted in its interpretation to the dimension indicated by the inner nominal.
(55) merch dal byr ei thymer
girl tall short her temper a tall short-tempered girl Welsh
(56) menyw lân frwnt ei thafod woman clean dirty her tongue a clean foul-mouthed woman

[^10]imra²t-un nazuifat-un washat-u l-sān-i
woman-NOM.INDEF clean-NOM.INDEF dirty-NOM the-tongue-GEN
a clean foul-mouthed woman
(58) imra${ }^{2}$ t-un twīlat-un qasīrat-u l-nazar-i
woman-NOM.INDEF tall.FSG-NOM.INDEF short.FSG-NOM the-sight-GEN
a tall woman short of sight

### 5.1 Treating the Inner NP as a Property

Given that the inner or complement nominal cannot be referred to in the following text, a possibility is that it corresponds semantically to a property rather than a full generalized quantifier: the idea is that the property denoted by this nominal serves to restrict the interpretation of the adjective to the appropriate dimension. The semantics for an attributive adjective would be as in (59), repeated in more convenient shorthand form in (60) (see Dalrymple (2001) for an accessible introduction to glue semantics in LFG). ${ }^{10}$

$$
\begin{array}{ll}
\text { ğamil-at-u } & (\uparrow \text { PRED })=\text { 'BEAUTIFUL }(<>) ’  \tag{59}\\
& \lambda \mathrm{P} . \lambda \mathrm{x} . \text { beautiful }(\mathrm{x}) \wedge \mathrm{P}(\mathrm{x}): \\
& {\left[\left((\mathrm{ADJ} \in \uparrow)_{\sigma} \mathrm{VAR}\right) \multimap\left((\mathrm{ADJ} \in \uparrow)_{\sigma} \operatorname{RESTR}\right)\right] \multimap} \\
& {\left[\left((\mathrm{ADJ} \in \uparrow)_{\sigma} \mathrm{VAR}\right) \multimap\left((\mathrm{ADJ} \in \uparrow)_{\sigma} \operatorname{RESTR}\right)\right]}
\end{array}
$$

(60) ğamil-at-u $\quad \lambda \mathrm{P} . \lambda$ x.beautiful $(\mathrm{x}) \wedge \mathrm{P}(\mathrm{x})$ :

$$
[\mathrm{v} \multimap \mathrm{r}] \multimap[\mathrm{v} \multimap \mathrm{r}]
$$

(61) imra²-at-un $\lambda \mathrm{X}$. $\operatorname{woman}(\mathrm{X}):[\mathrm{v} \multimap \mathrm{r}]$

The idea is that the phrase $\breve{g} a m i l-a t-u-l-w a \breve{g} h-i$ 'beautiful the-hair' would also be a function from N meanings to N meanings: $\breve{g} a m i l a t u$ consumes the meaning of $-l w a \breve{g} h i$ to produce this meaning. (62) gives the meaning constructor associated with the GEN form wağhi 'face' and (63) a preliminary formulation of the meaning constructor for a construct state adjective as a function from an OBJ meaning to a function from properties to properties ( N meanings to N meanings). The meaning constructor in (63) would consume that in (62) to produce the meaning constructor shown in (64). Glue constructors are abbreviated as above.
(62) $w a \breve{g} h-i \quad \lambda \mathrm{x} . \operatorname{face}(\mathrm{x}):[\mathrm{v} \multimap \mathrm{r}]$
(63) ğamil-at-u(CS) $\quad \lambda \mathrm{Q} \lambda \mathrm{P} \lambda \mathrm{x} . \exists \mathrm{y}[p \operatorname{cart-of}(\mathrm{y}, \mathrm{x}) \wedge \mathrm{P}(\mathrm{x}) \wedge \mathrm{Q}(\mathrm{y}) \wedge$ beautiful(y)]:

$$
(\uparrow \mathrm{OBJ})_{\sigma} \multimap[[\mathrm{v} \multimap \mathrm{r}] \multimap[\mathrm{v} \multimap \mathrm{r}]]
$$

(64) $\breve{g}$ amil-at-u $-l-w a \breve{g} h-i \quad \lambda \mathrm{P} \lambda \mathrm{x} . \exists \mathrm{y}[\mathrm{part}-\mathrm{of}(\mathrm{y}, \mathrm{x}) \wedge \mathrm{P}(\mathrm{x}) \wedge$ face $(\mathrm{y}) \wedge$ beautiful $(\mathrm{y})]$ :

$$
[\mathrm{v} \multimap \mathrm{r}] \multimap[\mathrm{v} \multimap \mathrm{r}]
$$

[^11]The derivation of $i m r a$ '-at-un $\breve{g} a m i l-a t-u-l-w a \breve{g} h-i$ 'woman beautiful of face' is shown in the proof below. The meaning constructor for the construct state adjective consumes that of its direct complement, producing a function from properties to properties (that is, an adjectival meaning). This can then combine with the nominal meaning associated with the head noun, and then finally with the determiner. ${ }^{11}$

There are many details here which need further consideration. We have here specified that the 'partof' restriction comes from the adjective in construction, and that the adjective is looking to combine with something which has a nominal, rather than an NP, meaning. So here we are assuming that we associate a simple property meaning with the definite form of a noun, in addition to any other meanings, possibly restricted to this construction. An alternative would be to associate the definite/indefinite marked noun with a pair of meaning constructors, one for the nominal core of its meaning and one corresponding to the determiner, and then have the construction (or the construct adjective itself) consume (i.e. dispose of) the determiner meaning.

```
\(\lambda Q \lambda P \lambda x . \exists y[P(x) \wedge Q(y) \wedge \operatorname{beautiful}(y) \wedge \operatorname{part}-o f(y, x)]:\)
\((\uparrow O B J)_{\sigma} \multimap((v \multimap r) \multimap(v \multimap r)) \quad \lambda x . f a c e(x): v \multimap r\)
    \(\lambda P \lambda x . \exists y[P(x) \wedge \lambda x . f a c e(x)(y) \wedge\) beautiful \((y) \wedge \operatorname{part}-o f(y, x)]:\)
    \(((v \multimap r) \multimap(v \multimap r))\)
    \(\lambda P \lambda x . \exists y[P(x) \wedge\) face \((y) \wedge\) beautiful \((y) \wedge\) part-of \((y, x)]::\)
    \(\frac{((v \multimap r) \multimap(v \multimap r))}{\lambda x . \exists y[\lambda x . w o m a n}(x)(x) \wedge\) face \((y) \wedge\) beautiful \(\left.(y) \wedge \operatorname{part-of}(y, x)\right]: \quad \multimap_{\mathcal{E}}\)
    \(\lambda x . \exists y[\lambda x . \operatorname{woman}(x)(x) \wedge\) face \((y) \wedge\) beautiful \((y) \wedge \operatorname{part-of}(y, x)]:\)
    \((v \multimap r)\)
    \(\lambda x . \exists y[\operatorname{woman}(x) \wedge\) face \((y) \wedge \operatorname{beautiful}(y) \wedge \operatorname{part}-o f(y, x)]:(v \multimap r)\)
```


### 5.2 An alternative semantics

In the above approach, the entire 'constructional burden' was essentially located in the lexical entry for the adjective itself which occurs in the construct state. Such a view might be at least partly motivated by the fact that in cognate languages, such as Hebrew, as we have seen, a special form of the adjective is required in this construction, and equally, it is natural that the special subcategorisation properties of adjectives in this construction are associated with a special meaning constructor. It is at least plausible however that more of the specifications are actually associated directly with the construction itself, alongside the constraint that the direct complement is definite, or that the NP complement itself plays a more important role, introducing some sort of possession relation, with its meaning modelled on that of whose book (Dalrymple, 2001, p421) (recall that the direct complement NP must denote a property or quality associated with the head noun or NP that the construction is predicated of). On this alternative view, then, the meaning constructor associated with the complement nominal might be along the lines shown in (65). The construct adjective introduces two meaning constructors (similar to the approach taken to attributive adjectives to permit recursive modification in Dalrymple (2001)).

$$
\begin{equation*}
-l-w a \breve{g} h-i \quad \lambda \mathrm{Q} . \lambda \mathrm{x} . \text { the }(\mathrm{f}, \text { face }(\mathrm{f}) \wedge \operatorname{poss}(\mathrm{x}, \mathrm{f}) \wedge \mathrm{Q}(\mathrm{f})):\left[\mathrm{v}_{\sigma} \multimap \mathrm{r}_{\sigma}\right] \multimap\left[\mathrm{c}_{\sigma} \multimap \mathrm{d}_{\sigma}\right] \tag{65}
\end{equation*}
$$

[^12]\[

$$
\begin{array}{ll}
\text { ğamil-at-u } & \lambda \text { b.beautiful }(\mathrm{b}): \mathrm{v} \multimap \mathrm{r}  \tag{66}\\
& \lambda \mathrm{P} . \lambda \mathrm{Q} . \mathrm{P}(\mathrm{x}) \wedge \mathrm{Q}(\mathrm{x}):[\mathrm{c} \multimap \mathrm{~d}] \multimap[[\mathrm{v} \multimap \mathrm{r}] \multimap[\mathrm{v} \multimap \mathrm{r}]]
\end{array}
$$
\]

We now apply -l-wağh-i to the basic meaning constructor for ğamil-at-u and then apply the second constructor of ğamil-at-u to the result. This gives us a function from $[[\mathrm{v} \multimap \mathrm{r}] \multimap[\mathrm{v} \multimap \mathrm{r}]]$ (a set of properties).

```
\(\lambda Q \lambda x . \operatorname{the}[f, f a c e(f) \wedge \operatorname{poss}(x, f) \wedge Q(f)]:\)
\((v \multimap r) \multimap(c \multimap d) \quad \lambda b\). beautiful \((b): v \multimap r\)
    \(\underline{\lambda x . t h e[f a c e(f) \wedge \operatorname{poss}(x, f) \wedge(\lambda b . \text { beautiful }(b))(f)]:(c \multimap d)} \Rightarrow_{\beta}{ }^{\mathcal{E}}\)
        \(\lambda\) x.the \([\) face \((f) \wedge \operatorname{poss}(x, f) \wedge\) beautiful \((f)]:(c \multimap d) \quad \lambda P \lambda Q \cdot P(z) \wedge Q(z)\)
            \(\lambda Q[\lambda\) x.the \((\) face \((f) \wedge \operatorname{poss}(x, f) \wedge\) beautiful \((f))](z) \wedge Q(z)\)
                \(\overline{\lambda Q[\lambda x . t h e ~}(\) face \((f) \wedge \operatorname{poss}(z, f) \wedge\) beautiful \((f))] \wedge Q(z):(v \multimap r) \multimap(v \multimap r) \quad \Rightarrow_{\beta}\)
```


## 6 Conclusion and Further Work

In this paper we have presented the main characteristics of the adjectival construct construction as it occurs in MSA. The construction shows a number of important similarities both to the better known nominal construct state construction in Semitic and also to the Celtic adjectival construction (so-called genitive of respect), for which a syntactic analysis in LFG is presented in Mittendorf and Sadler (2008). That paper argues that the Celtic construction should be recognised as a case in which adjectives appear with their own direct complements, and here we adopt essentially that approach to the Arabic data. Among the syntactic differences between the Arabic and the Welsh construction, however, is the fact that the complement NP is obligatorily morphologically definite in MSA while it occurs with an obligatory possessive pronominal clitic/inflection in Welsh. As a first step toward providing an account of the semantics of this family of constructions in LFG we present some preliminary thoughts as to how it might be formalized using glue. Building on this preliminary sketch will be one focus of our future work on the construction. We also do not yet have a sufficient understanding of how this construction relates to a number of subtly different adjectival constructions in MSA, nor of what the facts are in the Arabic vernaculars, both of which are topics which require further research.

## References

Benmamoun, Elabbas. 2000. The feature structure of functional categories: A comparative study of Arabic dialects. Oxford: Oxford University Press.

Dalrymple, Mary. 2001. Lexical Functional Grammar. San Diego, CA: Academic Press.
Falk, Yehuda. 2007. Constituent Structure and GFs in the Hebrew Nominal Phrase. In A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Maling, and C. Manning, eds., Architectures, Rules and Preferences: A Festschrift for Joan Bresnan, pages 103-126. Stanford, CA: CSLI Publications.

Hazout, Ilan. 2000. Adjectival genitive constructions in Modern Hebrew. The Linguistic Review 17:29-52.

Heller, Daphna. 2002. Possession as a Lexical Relation: Evidence from the Hebrew Construct State. In L Mikkelsen and C. Potts, ed., Proceedings of the 21st WCCFL, pages 127-140.

Higginbotham, James. 1985. On semantics. Linguistic Inquiry 16(4):547-593.
Kim, Jiyung. 2002. Adjectives in construct. In G. Katz, S. Reinhard, and P. Reuter, eds., Sinn and Bedeutung VI: Proceedings of the Sixth Annual Meeting of the Gesellschaft für Semantik.

King, Tracy Holloway and Mary Dalrymple. 2004. Determiner agreement and noun conjunction. Journal of Linguistics 40(1):69-104.

Koenig, Jean-Pierre. 1999. French body-parts and the semantics of binding. Natural Language \& Linguistic Theory 17(3):219-265.

Kremers, Joost. 2003. The Arabic noun phrase: a Minimalist approach. Ph.D. thesis, University of Utrecht, Utrecht, The Netherlands.

Kremers, Joost. 2005. Adjectival Constructs in Arabic. Linguistische berichte 203:331-348.
Mac Cana, Proinsias. 1966. An old nominal relative sentence in Welsh. Celtica 7:91-115.
Mittendorf, Ingo and Louisa Sadler. 2008. NP would like to meet GF: A Welsh Adjectival Construction. In M. Butt and T. H. King, eds., Proceedings of the LFG08 Conference, pages 373-393. Stanford, CA: CSLI Publications.

Ritter, Elizabeth. 1988. A head movement approach to construct state nominals. Linguistics 26:909929.

Ritter, Elizabeth. 1991. Two Functional Categories in Noun Phrases: Evidence from Modern Hebrew. In S. Rothstein, ed., Perspectives on Phrase Structure, vol. Syntax and Semantics 25, pages 37-62. Academic Press.

Ryding, Karin. 2005. A Reference Grammar of Modern Standard Arabic. Cambridge, UK: Cambridge University Press.

Siloni, Tal. 2002. Adjectival constructs and inalienable constructions. In J. Ouhalla and U. Shlonsky, eds., Themes in Arabic and Hebrew Syntax. Dordrecht, Holland: Kluwer.

# THE PREPOSITIONAL PASSIVE AS STRUCTURESHARING 

Alex Alsina<br>Pompeu Fabra University<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

Evidence is reviewed that requires abandoning verb-preposition reanalysis in the derivation of prepositional passives ( P -passives). An alternative is proposed in which P-passives are a type of raising construction, where the object of the prepositional complement is structure-shared with the subject of the passive verb. This places the construction within the scope of the Theory of Structure-Sharing (Alsina 2008), which helps explain some of its properties. A proposal is made to account for the phrase-structure of P passives and for the crosslinguistic variation regarding the availability of P -passives and preposition stranding.


## 1 Introduction

Prepositional passive, also known as pseudopassive, referred to here as P passive, for short, has all the features of a canonical passive construction, except for one: in a P-passive, the subject does not correspond to an object of the verb in the active form, as it would in a canonical passive, but to the object of a prepositional complement, as the following examples illustrate. ${ }^{\dagger}$
(1) a. We all counted on Kim for this job.
b. Kim was counted on for this job (by all of us).
(2) a. The car bumped into the wall.
b. The wall was bumped into (by the car).

The standard LFG analysis of P-passives, since Bresnan (1982), has been to assume an optional process of lexical, or morphological, incorporation of the preposition into the verb, whereby a verb and the head of its prepositional complement form a lexical unit, a word, to which the lexical rule of passivization can apply. This approach has been criticized by Postal 1986 and Baltin and Postal (1996) on the basis of English and by Lødrup (1991) taking into account Norwegian data. The arguments from these works against verbpreposition reanalysis (V-P reanalysis)-the idea that the verb and the adjacent preposition in examples like (1)-(2) optionally form a lexical unit or word-are so strong that they require abandoning this idea and thinking of a different way to account for P-passives.

In addition, P-passive has an effect also found in long-distance dependency (LDD) constructions: the creation of structures with preposition stranding (P-stranding), illustrated in (3a) for P-passive and in (3b) for LDD:
(3) a. This article was talked about in the workshop.
b. Which article did you talk about in the workshop?

While English and other languages such as Norwegian have P-passive and LDD as sources for P-stranding, some languages, such as French and other

[^13]Romance languages, ${ }^{1}$ do not allow P-stranding of either kind. The following examples ((4a-b) from Postal (1986:204)) illustrate this for French. (4b) shows that, in French, an argument that corresponds to the object of a prepositional complement, as in (4a), cannot be the subject of a passive clause, and (4c) shows that it cannot appear in clause-initial position by LDD.
(4) a. On ne peut pas compter sur Marc pour faire ce boulot. One can not count on Marc for do this job
b. * Marc ne peut pas être compté sur/dessus pour faire ce boulot. Marc can not be counted on for do this job
c. * Qui est-ce qu'on ne peut pas compter sur/dessus pour faire ce boulot? who can one not count on for do this job?
The contrast between (3) and (4) has suggested to some linguists that languages split into exactly two types with respect to P-stranding: those that allow P-stranding from both sources, like English, and those that don't allow it from either source, like French. Hornstein and Weinberg (1981) and Kayne (1981) developed a unified account of P-stranding to capture this assumed tight correlation between the two potential sources of P-stranding. However, Maling and Zaenen (1985) argue that there is no correlation between the availability of P-passive and of P-stranding in a given language because there are languages, such as Icelandic, Swedish, and Danish, that allow P-stranding arising from LDD, but not as a result of passivization. Maling and Zaenen (1985:162) illustrate this point for Swedish, where P-passives are ungrammatical, as in (5a), although topicalization of the object of a PP is fine, as in (5b), where the presence of the dummy subject det shows that the sentenceinitial NP is not a subject, but a topicalized phrase.
a. * Hon skrattades at. she was-laughed at
b. Henne skrattades det at. her was-laughed it at 'People laughed at her.'
Maling and Zaenen (1985) claim that P-stranding is determined by two separate parameters: (1) whether prepositions can govern the tail-end of an LDD (or an empty category, in their terms) and (2) whether a verb and a preposition can be reanalyzed as a word. If we formulate the parameters that regulate the availability of P-stranding as statements of different subtheories of grammar (a statement on LDD and a statement on word formation), we predict that there should be four types of languages corresponding to the
${ }^{1}$ Prince Edward Island French is an exception to the claim that Romance languages lack P-stranding, since, according to Law (2006:672) (citing King and Roberge (1990)), this dialect of French has both types of P-stranding.
positive or negative setting of each of these two parameters. If we assume that languages may vary depending on whether they allow P-stranding by LDD and whether they allow P-stranding by passivization, we predict the existence of four possibilities, as shown in the table in (6):
(6) Types of languages according to availability of P-stranding:

$\left.$|  |  | P-stranding by LDD |  |
| :--- | :--- | :--- | :---: |
|  |  | yes | no |
| P-stranding by <br> passivization | yes | English <br> Norwegian | Icelandic <br> Swedish | | French, other Ro- |
| :--- |
| mance languages | \right\rvert\,

Table (6) shows that one of the four types of languages predicted by the claim that the two sources of P -stranding are independent of each other is not attested: no language has been argued to allow P -stranding by passivization, while disallowing it by LDD. ${ }^{2}$ In order to explain this gap, assuming that it is not accidental, P-passives and LDD cannot be as unrelated as they are standardly assumed to be in LFG.

This paper presents an analysis of P-passives that, on the one hand, does not assume a morphological process of V-P reanalysis and, on the other hand, assumes that both types of P-stranding referred to in (6) involve structure-sharing of the object of the preposition with a structurally more prominent grammatical function. This proposal brings P-passives within the scope of the Theory of Structure-Sharing of Alsina (2008), which will allow us to explain the covariation of the two phenomena, namely, the claim that only languages that allow P-stranding by LDD allow P-passives.

In what follows, first, I summarize the standard LFG approach to Ppassives involving V-P reanalysis and the arguments that have been presented for and against this assumption. I then present the Theory of StructureSharing showing that many of the properties of P-passives follow straightforwardly from assuming that P -passives are a kind of raising construction, an assumption also made in Lødrup (1991) and Tseng (2007). Next, I propose the phrase-structure of P-passives, which is independently needed for verbparticle constructions, and the constraints relating properties of the phrasestructure with structure-sharing configurations. Two parameters, stated as such constraints, derive the claim about crosslinguistic facts embodied in (6).

[^14]
## 2 V-P reanalysis

### 2.1 Arguments for V-P reanalysis

One of the motivations, if not the main one, for assuming V-P reanalysis as part of the account of P-passives is that, in the version of LFG of Bresnan (1982), passivization is a lexical rule that operates on the list of grammatical functions (GFs) selected by a lexical item. It is a rule that takes a lexical item (an active verb form) as input and produces a different lexical item (a passive verb form) by changing the GFs assigned to arguments. In that version of the framework, the assignment of GFs to arguments is assumed to be fully specified in the lexical entries of words. The passive lexical rule replaces the GF name OBJ in a semantic form by SUBJ and replaces the name SUBJ by $\varnothing$ or OBL by. This means that this rule can only apply to a lexical item whose argument list includes a SUBJ and an OBJ.

Thus, the semantic form of a transitive verb like read includes both a SUBJ and an OBJ: 'read $\langle(\uparrow \operatorname{SUBJ})(\uparrow \mathrm{OBJ})\rangle$ '. The passive lexical rule can apply to it yielding the semantic form 'read $\langle\varnothing(\uparrow$ SUBJ $)\rangle$ ' or 'read $\left\langle\left(\uparrow \mathrm{OBL}_{\mathrm{by}}\right)(\uparrow\right.$ SUBJ $\left.)\right\rangle$ '. However, if we take an intransitive verb, like those used in (1)-(3), its semantic form lacks an OBJ and, so, it should not be possible for the passive lexical rule to apply to it. The V-P reanalysis rule is a way to turn an intransitive verb into a transitive verb. This rule, according to Bresnan 1982: 51, involves a morphological change, which analyzes a verb-preposition sequence as a single word of category V , and an operation that replaces the GF P OBJ in a lexical form by OBJ. In this way, the output lexical item includes the GF OBJ in its semantic form and can be the input to the passive lexical rule. In other words, the object of the preposition becomes the object of the derived verb, morphologically composed of a verb and a preposition.

A consequence of this analysis is that V-P reanalysis is restricted to prepositions that head a complement of the verb, i.e., cannot involve prepositions that head adjuncts. This is because the grammatical function P OBJ required in the input to the V-P reanalysis rule is one of the GFs assigned to arguments. This restriction accounts for the ungrammaticality of passive forms such as (7), where the preposition heads an adjunct of the verb (from Bresnan (1982:51)):
(7) a. $\quad *$ No reason was left for. (They left for no reason.)
b. * The operation was died after. (The patient died after the operation.)
Another consequence of this analysis is that the output of the V-P reanalysis rule can undergo the lexical process of Adjective Conversion. The italicized sequences in (8) (from Bresnan (1982:53)) are adjectives. On the assumption that word formation rules cannot take multi-word structures as input, those sequences have to be single words, to be input to Adjective Conversion.
(8) a. Each unpaid for item will be returned.
b. His was not a well-looked on profession.

Yet another consequence of the V-P reanalysis rule is that it provides an immediate explanation for the fact that adverbs cannot appear between the verb and the preposition in P-passives, as in (9). This is because, under this view, the verb and the preposition constitute a word, which cannot be broken up in the syntax. Notice that an adverb can appear immediately before a stranded $P$ if it is stranded by LDD, as in (10) (exx. from Bresnan (1982:54)):
(9) a. * Everything was paid twice for.
b. * Your books were gone most thoroughly over.
(10) a. That is something that I would have paid twice for.
b. These are the books that we have gone most thoroughly over.

The existence of double passives, such as those illustrated in (11) (Bresnan 1982:60), could be taken to be a problem for the V-P reanalysis rule, because, if advantage is the object in (11a), the passive form in (11b) is to be expected, but not (11c), and, if her talents is the object in (11a), then we expect (11c) and not (11b):
(11) a. No one took advantage of her talents.
b. Not much advantage was taken of her talents.
c. Her talents weren't taken advantage of.

Bresnan's (1982) proposal to deal with these cases is to assume a dual analysis. There is an idiomatic object analysis, in which advantage in take advantage of is an object, although not an argument in the argument structure of the verb, and an incorporated object analysis, in which take advantage is treated as a morphologically complex verb. V-P reanalysis can apply to the incorporated object structure yielding the complex verb take advantage of, in which the apparent object of the preposition is the object of this complex verb. Passivization can apply either to this complex verb, making the apparent object of the preposition the passive subject, as in (11c), or to the idiomatic object structure, making the idiomatic noun the passive subject, as in (11b).

A positive consequence of this analysis is that it provides an explanation for the contrast in (12) (Bresnan 1982:61):
(12) a. How much advantage was taken of John?
b. * How much advantage was John taken of?

John can only be the subject in (12b) if advantage is incorporated with the verb, but then advantage cannot be separated from the verb, as it is in (12b).

However, given this explanation, the fact that the putative incorporated noun can be modified by adjectives, as in (13), is a bit of a problem:
(13) a. John was taken unfair advantage of.
b. Her talents were made good use of.

Since the postverbal noun is assumed to be incorporated, Bresnan (1982) is forced to assume that the adjective is too. This yields fairly complex morphological structures in which each of the component elements coincidentally occupies the same position it would occupy if it was positioned by syntactic rules in a [V NP P] structure. We might reasonably expect a different order of elements in the incorporated structure from the order in the idiomatic object structure.

### 2.2 Arguments against V-P reanalysis

Postal (1986) and Baltin and Postal (1996) present several arguments against assuming V-P reanalysis. V-P reanalysis makes the claim that the P-object (the apparent object of a preposition in a putatively reanalyzable V-P sequence) is really (at least optionally) the object of a verb morphologically composed of a verb and a preposition and, therefore, should behave as the object of a verb, while both the verb and the preposition should be syntactically inert, as they are components of a word. The arguments that follow, based on those works, show that this claim is not correct.

An NP object of a verb can follow a PP or adverb when the former is focused or consists of a large number of lexical items, in the construction known as heavy NP shift, as in (14a). A P-object should behave like an ordinary object of a verb, given V-P reanalysis, and potentially appear following a PP. But this is not what happens, as we see in (14b). (Examples (14)-(20) are from Baltin and Postal (1996:129-133).)
(14) a. I discussed $\qquad$ 1 with Lorenzo [the problems he was having with deliveries] ${ }_{1}$.
b. *I argued with $\qquad$ 2 about such problems [the drivers' union leader] ${ }_{2}$.

The preposition cannot be deleted without its object under gapping. If the verb and the preposition formed a word, there would be no reason not to expect deletion of the [V-P] verb:
(15) a. Frank called Sandra and Arthur $\qquad$ Louise.
b. Frank talked to Sandra and Arthur $\qquad$ *(to) Louise.

The V and P that supposedly make up a reanalyzed verb in passive structures like (16) should not be coordinated with another V or P respectively and should not be gapped independently of each other, as coordination and gapping, being syntactic processes, have no access to the internal structure of words, but (16) shows that this is possible:
(16) a. The bridge was flown (both) over and under.
b. Fascism was fought for by Goebbels and (then) against by De Gaulle.

The object of a preposition behaves differently from the object of a verb with respect to subdeletion, as (17) shows. Passivizable P-objects behave like objects of prepositions and not like objects of verbs, as illustrated in (18):
(17) a. Larry screamed more of those words than he did $\qquad$ of these words.
b. *Larry screamed about more of those words than he did $\qquad$ about of these words.

* Jane talked to more of these people than Sally talked to $\qquad$ of those people.
Passivizable P-objects behave like objects of prepositions, not like objects of verbs, with respect to floating quantifiers:
(19) a. Mike handed the photostats all to Louise.
b. The air force struck (*at) those targets both in the morning.
(20) a. Mike warned those employees all about speaking to reporters.
b. * Mike talked to those employees all about speaking to reporters.

In Lødrup's (1991) analysis of Norwegian P-passives, the main argument against V-P reanalysis as a step in the derivation of P-passives is the non-adjacency of the passive verb and the preposition in many structures. Not only are the V and the P not adjacent in Norwegian P-passives such as (21a-b), but it would be ungrammatical to have them next to each other in (21c) (from Lødrup (1991:118-119)). The V and the P are italicized in (21).
(21) a. De må bli passet bedre på.
'They must be looked better after'
b. Hvorfor passes de ikke bedre på?
'Why look-PASSIVE they not better after'
c. * Hvorfor passes på de ikke bedre?
'Why look-PASSIVE after they not better'
This evidence shows that the V and the P cannot be treated as a lexical unit in the Norwegian P-passive. In fact, according to Lødrup (1991), the V and the P never behave as a syntactic unit in the Norwegian P-passive and, so, the claim that the Norwegian P-passive cannot involve V-P reanalysis is a widely accepted conclusion (see references in Lødrup (1991)).

One of the striking pieces of evidence for V-P reanalysis is the claim that P-passives undergo Adjective Conversion. However, Postal (1986) and Baltin and Postal (1996:142) claim that the instances of adjectives composed of a participle and a preposition do not constitute evidence for V-P reanalysis: there are many P-passive participles without a corresponding adjectival form. Relevant examples are: *argued-about proposals, *sat-on tables, *spoken-to students. Thus, the mechanism that allows P-passives has to be much more general than the process that allows P-incorporation in adjectives
and the two processes cannot be assumed to be related. In addition, Lødrup (personal communication) notes that, in Norwegian, adjectives derived from P-passive forms are completely impossible.

The evidence presented by Postal (1986) and Baltin and Postal (1996) shows that, on the basis of the English facts alone, V-P reanalysis has to be rejected as a requirement for P-passive. Lødrup (1991) also shows that V-P reanalysis has to be rejected for P-passive in Norwegian. If we assume that P passive is essentially the same construction in English and Norwegian, ${ }^{3}$ it is clear that we need to find an alternative to V-P reanalysis for both languages.

## 3 P-passive as structure-sharing

If we assume there is no V-P reanalysis, it is reasonable to suppose that the preposition bears the same GF in the active structure as in the P-passive: an oblique complement. If so, the subject of the P-passive is not a thematic argument of the verb, but of the prepositional complement. This is, therefore, a kind of raising construction, in which the subject of the matrix predicate is a thematic argument of a complement of this predicate, only this complement, instead of being verbal, as in standard raising constructions, is prepositional (see Lødrup (1991) and Tseng (2007) for a similar claim). A raising construction involves structure-sharing (S-S) of two GFs: two GFs having the same value. Given this, the structure falls within the scope of the Theory of Structure-Sharing, proposed in Alsina (2008).

Since S-S is a feature common to both LDD and raising, the Theory of Structure-Sharing provides a unified treatment to LDD and raising by constraining the structures in which S-S arises. The theory consists of a set of well-formedness conditions on the f-structure: all instances of S-S must satisfy these conditions. A positive consequence of this theory is that it allows us to dispense with control equations, both the kind that are specified in lexical entries and the kind that are specified in c-structure rules. ${ }^{4}$ In what follows, I will briefly present this theory, following Alsina (2008), in order to show that it accounts for crucial facts of P-passives.

This theory consists of four principles or conditions. The first of these-the Nonthematic Condition on Structure-Sharing-is given in (22). The notion of f-prominence, which is used in the formulation of condition (22), is defined in (23).

[^15](22) Nonthematic Condition on Structure-Sharing: In every f-structure containing structure-sharing, a nonthematic GF is the most f prominent GF in the structure-sharing relation.
(23) F-prominence: GF $\alpha$ is more f-prominent than GF $\beta$ iff $\alpha$ f-commands $\beta$ and either $\beta$ does not f -command $\alpha$ or $\alpha$ is higher than $\beta$ in the GF Hierarchy DF > OBJ > OBL. ${ }^{5}$
LDD constructions always satisfy condition (22), as they involve an OP as the most f-prominent GF in the S-S relation. If we take a nonthematic GF to be a GF that is not licensed by its mapping to a semantic participant, whether argument or adjunct, we see that OP is always nonthematic, because semantic participants are always mapped onto other GFs (SUBJ, OBJ, OBL). The other structure-shared GF in LDD is either lower in the GF Hierarchy or does not fcommand the OP. (24) schematically represents the structure-sharing relation in two sentences involving LDD. It shows that the nonthematic OP is more fprominent than the other GF it is structure-shared with:

(24) a. What do you think $\left[\begin{array}{l}\text { is in that box } \\ \text { SUBJ }\end{array}\right]$
b. Who did you keep that secret from


In raising constructions, there is always a nonthematic SUBJ or OBJ in the matrix clause that asymmetrically f-commands the other GF in the structure-sharing relation, as schematized in (25):


In canonical passive constructions, the subject is not part of a S-S relation, as it is a thematic GF (a GF licensed by its mapping to an argument of the verb), but in P-passives the subject is not a an argument of the verb, but of the prepositional complement of the verb. In these constructions, the Nonthematic Condition is satisfied because the subject is nonthematic and more f-prominent than the GF it is structure-shared with, namely, the object of the preposition. Consider the active-passive pair in (26):
(26) a. John paid for the tickets.
b. The tickets were paid for.

[^16]The argument structure of pay has an external argument as its logical subject (the most prominent argument at argument structure) and an oblique-for argument. Passivization, as in all passive forms, suppresses the logical subject, which means it makes it unavailable for mapping to a direct GF. Thus, the external argument is unexpressed or expressed as an oblique phrase in the passive form of pay. The clause whose predicate is pay needs to satisfy the Subject Condition (see Bresnan and Kanerva (1989), among others), by which every verbal f-structure must include a subject, but neither the external argument nor the oblique argument can map onto the subject. The structure would violate this condition, were it not for the possibility of having a subject structure-shared with the object of the oblique complement. Thus, in the active (26a), the subject is the external argument of pay, whereas in the passive (26b), the subject the tickets is a nonthematic GF structure-shared with the object of the preposition, as shown schematically in (27):


The second condition that a structure-sharing relation needs to satisfy is the Locality Condition:
(28) Locality Condition on Structure-Sharing: Every f-structure that contains a GF G and is f-commanded by another GF structure-shared with G has a DF involved in this relation.
The representation in (27) does not satisfy Locality, because the embedded fstructure has a GF structure-shared with a GF in the matrix f-structure, but there is no DF (SUBJ or OP) in the embedded f-structure that is also part of this S-S relation. Notice that (24b) does not satisfy this condition either, whereas (24a) and (25) do because there is a DF in the embedded f-structure that is part of the S-S relation. In both (24a) and (25) the DF in question is a SUBJ. In other cases, Locality is satisfied because the embedded f-structure has an OP taking part in the S-S relation that has its tail-end in this f-structure. Whether Locality is satisfied by having a SUBJ or an OP be the DF referred to in (28) depends on the Non-Subject Binding Condition (30). The statement of this condition requires the definition of the F-Binding relation given in (29):
(29) F-Binding: $\alpha$ f-binds $\beta$ iff
a. $\quad \alpha$ and $\beta$ are different GFs with the same value, and
b. there's an argument in $\alpha$ 's f -structure that is equal to $\beta$ or has $\beta$ as a feature and is not higher than $\alpha$ in the GF hierarchy.
(30) Non-Subject Binding Condition (Non-SUBJ Bind): The closest fbinder $^{6}$ of a non-SUBJ is a DF and, in a different f-structure, an OP.

[^17]When LDD spans two (or more) f-structures, Locality and Non-SUBJ Bind are satisfied if the S-S relation includes in the embedded f-structure either a SUBJ, as in (24a), or an OP, as in (31). Notice that (24b) does not satisfy Locality because it lacks the OP included in (31).


As for P-passives, Locality requires there to be a DF in the prepositional fstructure. Since this DF has the SUBJ of the matrix f-structure as its closest fbinder, it cannot be an OP by Non-SUBJ Bind (30). It has to be a SUBJ; as it is nonthematic, the Nonthematic Condition (22) is satisfied. The representation in (27) is thus replaced by (32):


The evidence for the SUBJ in the prepositional f-structure is that it has to satisfy the fourth Condition of the theory: The SUBJ Binding Condition:
(33) SUBJ Binding Condition (SUBJ Bind): A SUBJ that is structureshared with a more f-prominent GF is
a. f-bound in a non-SUBJ of a clause and
b. in a tenseless f-structure if its closest f-binder is not OP.

By this condition, the embedded SUBJ in (32) must be f-bound and an fbound GF is an argument or the feature of an argument in the f-structure of the f-binder. It follows from this that P-passive is only possible if the prepositional f-structure is an argument of the passive verb. In other words, there is no P -passive if the P introduces an adjunct, as in (7). This prediction follows from assuming a nonthematic subject in the prepositional f-structure of P-passive, which is required by the Theory of Structure-Sharing.

Some verbs that do not require a prepositional argument allow Ppassive, as in the following examples:
(34) a. This chair has been sat on by Fred.
b. The room looks like it's been lived in.
c. This spoon has been eaten with.
(Davison 1980:45)
Davison (1980) observes that locative and instrumental phrases are very likely to participate in P-passives. We can just assume that certain verbs can augment their argument structures with a locative or instrumental argument. Notice that these verbs often allow locative inversion (e.g., On that rock sat a huge angry lion, In that space lived a whole family of capybaras).

We do need to say that nominal and adjectival ( $[+\mathrm{N}]$ ) f-structures do not allow nonthematic subjects, unlike verbal and prepositional ( $[-\mathrm{N}]$ ) f-
structures. This is reflected in the following condition:
(35) Nonthematic subject constraint: Nonthematic subjects are only possible in [-N] f-structures.
This explains the non-existence of P-passives out of NPs (Lødrup 1991:124):
(36) a. Reisen til Drøbak ble tenkt på.

The trip to Drøbak was thought of.
b. * Drøbak ble tenkt på reisen til.

* Drøbak was thought of the trip to.

To summarize, this section has argued that P-passive is just a type of raising construction: raising out of a P complement. As such, it involves a structure-sharing relation and has to satisfy the principles of the Theory of Structure-Sharing. Nothing specific to P-passives needs to be assumed. And it follows from the theory that there is no P-passive involving adjuncts.

## 4 The c-structure of P-passives

At this point we still need to account for the following facts:
a) Adjacency of V and P in P -passives (no intervening adverbs);
b) Unavailability of P-passives in the presence of a thematic object;
c) Fixed V $\qquad$ position of nonthematic idiomatic objects in P-passive.
In order to capture these facts, we need to make explicit some assumptions about c-structure and the correspondence between c-structure and f-structure, first of all, the following function-category mapping principle:
(37) No PP in P-passives: A prepositional f-structure whose subject is nonthematic cannot map onto a phrasal category.
As argued earlier, the preposition in P-passives maps onto an f-structure (a prepositional f-structure) that has a nonthematic subject (see (32)). By principle (37), it cannot map onto a PP (or a P'); it maps onto a P that projects no further. This P is, therefore, a non-projecting category, to use Toivonen's (2003) term, indistiguishable from a particle in verb-particle constructions. So, it occupies the P position that we independently need for particles in verb-particle constructions. The phrase-structure rules for the verb phrase assumed here are based on Kiparsky (1988): ${ }^{7}$
(38) a. $\quad \mathrm{VP} \rightarrow \mathrm{V}^{\prime} \mathrm{XP}^{*}$
b. $\quad \mathrm{V}^{\prime} \rightarrow \mathrm{V}(\mathrm{NP}) \mathrm{P}^{*}$
(where $X P^{*}=$ a sequence of 0 or more maximal projections of any category, and $\mathrm{P}^{*}=$ a sequence of 0 or more instances of P )
There is evidence for the phrase structures that these rules license. In the first

[^18]place, they predict a fixed V-NP-P-NP order in ditransitive verb-particle constructions, illustrated in the following examples (Farrell 2005:131):
(39) a. She likes to give her students away candy.
b. * She likes to give away her students candy.
(40) a. I want you to send my brother over a cake.
b. ${ }^{\%}$ I want you to send over my brother a cake.
( $\%=$ ungrammatical for some speakers)
If we assume that there can be no more than one unrestricted and one restricted object per clause in English and the function-category correspondence rules in (41), we explain the facts in (39)-(40).
(41) a. NP daughter of $\mathrm{V}^{\prime}$ is for unrestricted object
b. NP daughter of VP is for restricted object

The unrestricted object NP is licensed by rule (38b) and the other NP is licensed by (38a). As the P, the particle away or over, has to appear at the right edge of $\mathrm{V}^{\prime}$, it follows that it has to come after one NP and precede the other one in a c-structure such as this:
(42) [vp [v' give [np her students] away] [np candy]]

In the second place, the alternative order of particle and object NP in examples like (43) is also predicted by this theory.
(43) a. look up the word / look the word up
b. take out the garbage / take the garbage out
c. turn off the lights / turn the lights off

If we don't assume any principle forcing the object of a monotransitive construction to be either restricted or unrestricted, it follows that it can be either. As a restricted object, it will appear as sister of $\mathrm{V}^{\prime}$, therefore following the particle, and, as an unrestricted object, if will appear as sister of V , therefore preceding the particle. Thus, the c-structures for the two orders in (43a) would be as shown in (44), where the NP is a restricted object in (44a) and an unrestricted object in (44b):
(44) a. [vp [v' look up] [np the word] ]
b. [vp [v' look [ NP the word] up]]

In the third place, there is no heavy NP shift over an NP. An object NP can follow a PP or other phrase, provided the NP is focused or heavy, but it cannot follow another NP, however heavy the former NP may be, as shown in (45)-(46):
(45) a. I sent the message to a reporter.
b. I sent to a reporter the defamatory message that everyone is talking about.
(46) a. I sent that reporter a message.
b. * I sent a message those reporters who kept following me.

The alternative orders of NP and PP in (45) follow from assuming that these constituents are ordered by Linear Precedence (LP) constraints. As in Gazdar and Pullum (1981), Falk (1983), Gazdar, Klein, Pullum, and Sag (1985), and Sag (1987), among others, LP constraints are statements on the order of sister constituents. The relevant LP constraints here are: NP < PP (an NP precedes a PP) and XP[light] < XP[heavy] (a light phrase precedes a heavy phrase). In case of conflict between the two LP constraints, ordering by weight wins over ordering by category.

Whereas the NP the message in (45a) can be licensed by either rule (38a) or (38b), the PP can only be licensed by (38a). In a structure where both the NP and the PP complements of sent are sister constituents, they are subject to the LP constraints. In such a structure, if the NP complement is heavy, as in (45b), it follows the PP, as shown in (47a). On the other hand, in structures with two NP complements, like (46), the two NPs are not sister constituents, as shown in (47b), and so they are not subject to the LP constraints stated above and the weight of the two objects is irrelevant for their relative order. This explains the ungrammaticality of (46b).
(47) a. [ $\mathrm{vp}^{\mathrm{v}} \mathrm{v}$ sent] [pp to a reporter] [ NP the defamatory message...]]
b. [ $\mathrm{vP}[\mathrm{v}$ sent [ NP that reporter]] [ NP a message]]

Having established (38) as the rules for the VP in English, the P in Ppassives, which does not project a PP by mapping constraint (37), has to appear in the P position licensed by (38b). This provides an immediate explanation for the claim illustrated in (9) that adverbs cannot appear between the verb and the P in a P-passive. There is no position in rule (38b) for an adverb. Notice that a preposition stranded by LDD is unaffected by constraint (37) and thus projects a PP (even though it includes nothing but a P ); this PP is licensed by rule (38a) and therefore can appear following an adverb, as seen in (10).

In order to explain the claim that P-passives are only possible with verbs that don't have a thematic object in the active form (intransitive verbs and verbs with nonthematic objects), we need to assume an additional constraint: constraint (48), on the argument-GF mapping, by which the preferred choice for the subject function is an argument (as opposed to a nonargument) and an argument of its local predicate (as opposed to an argument of a nonlocal predicate). See Alsina (2001:380) for independent evidence for this constraint.
(48) Subject Selection Principle: If possible, the subject should be (a) an argument (b) of its local argument structure.
A principle such as this is implicitly assumed in general, as it explains why nonthematic subjects are a last resort option. For example, as a default, a
monadic predicate is intransitive; only under very specific conditions can an expletive be used as the subject, overriding this principle. In the case of P passives, it explains why there are no P-passives of transitive verbs in general, as shown in (49b) (see also Zwicky (1987:648)), the exception being when the object is nonthematic, as in (50): ${ }^{8}$
(49) a. The cookie has been paid for.
b. * The cookie has been paid 50 cents for.
(50) a. I caught sight of Mary in the crowd.
b. Mary was caught sight of in the crowd.

There is no argument of the verb that can be a subject in (49a) or (50b), but there is in (49b). Once the logical subject of pay is suppressed by the passive morphology and therefore made unavailable for mapping to the subject, there is no other argument of the verb in (49a) that can be the subject; in this case, an NP that is not an argument of the verb can be chosen as the subject. In (49b), on the other hand, the argument structure of pay is not the same as in (49a): in this case, there is another argument in addition to the logical subject (the amount argument) and that is the preferred choice for subject. This explains the ungrammaticality of (49b), where a nonargument of the verb is the subject, while an argument of the verb is available. In contrast, (50) has an NP in object position (sight) that is clearly not a semantic argument of the verb. The verb catch and the noun sight form a noncompositional idiom in catch sight (of) in which the noun does not correspond to an argument. Consequently, by principle (48), sight, as a nonargument, loses out against the argument of the preposition as the preferred choice for subject.

The assumption that only arguments (expressions bearing thematic roles) can be restricted objects (see Alsina (1996a,b, 2001), among others) explains why nonthematic objects must appear in immediately postverbal position in P-passives. This is illustrated by the contrast in (51):
(51) a. John was taken advantage of.
b. * John was taken of advantage.

On the assumption just argued for that P-passives are possible in the presence of an object NP only if this object is nonthematic, advantage has to be nonthematic in (51). Therefore, it can only be an unrestricted object and occupy the position of sister of V ( not sister of $\mathrm{V}^{\prime}$, reserved for restricted

[^19]objects). The preposition of in (51) is at the right edge of $\mathrm{V}^{\prime}$ and, so, must follow the object, accounting for the contrast in (51).

But if advantage in (51) is a nonargument, how come we have the alternative form in (52), where this N is the passive subject?
(52) Advantage was taken of John.

We have to assume a dual analysis of the idiomatic expression take advantage of: one in which advantage is an argument of take and one in which it is a nonthematic object of the idiom take advantage of. As an argument, it can passivize, as in (52), and undergo Heavy NP Shift:

They took of John such an amazingly unfair advantage that everyone complained.
In its nonargument analysis, advantage cannot be part of an LDD, as the bottom of an LDD must be a thematic argument. This explains the contrast in (12), repeated as (54):
(54) a. How much advantage was taken of John?
b. * How much advantage was John taken of?

The only way John can be the subject in (54b) is if advantage is a nonargument, but then advantage cannot be part of an LDD. In addition, as a nonargument, advantage cannot undergo Heavy NP Shift:

* John was taken of such an amazingly unfair advantage that everyone complained.
An object that follows a prepositional complement has to be restricted, but the object in (55) cannot be restricted because only arguments can be restricted and advantage would have to be a nonargument in order for an NP that is not an argument of the verb to be the subject, given principle (48).

The lexical entries for take, needed for examples like (52), (53), and (54a), and for the noncompositional take advantage, needed for (51a), are given in (56a) and (56b) respectively. (56a) has an argument structure with an external argument, an internal argument and an oblique case argument, whereas (56b) has an argument structure with only an external argument and an oblique of-argument. Subscripted numbers, or indices, in (56) and (57) signal the correspondence between elements of different structures (astructure, c-structure and f-structure). (56b) is an instance of a multi-word idiom, where each word is referred to by the other word.
(56) a. take: $\quad \mathrm{V}_{3}$ [PRED 'take $\langle$ Ext Int [case:OBL] $\rangle$ '] $]_{3}$
$\begin{array}{ll}\text { b. take: } & \mathrm{V}_{1} \\ \text { advantage: } & \mathrm{N}_{2}\end{array}\left[\begin{array}{l}\text { PRED 'take-advantage }\langle\text { Ext [case:of] }\rangle ' \\ \mathrm{GF}_{2}\end{array}\right]$
Entry (56a) is needed for passive forms like (52) and entry (56b) is needed for passives like (51a). C- and f-structures for these two examples are given
in (57); being passive forms, the external argument is suppressed, which is notated by subscripting this argument with the emptyset symbol.



## 5 Crosslinguistic variation

We need to account for the following range of crosslinguistic variation:
a) Norwegian allows P-passives with some thematic objects, unlike English;
b) Icelandic, Danish and Swedish allow P-stranding, but no P-passive;
c) French, Catalan, Spanish (and other Romance languages) do not allow P-stranding of either the LDD or the P-passive type.
A) Norwegian allows thematic objects in P-passives, provided they are nonspecific (Lødrup 1991: 126-127):
(58) a. Barna ble skiftet bleier/*bleiene på.
'The children were changed napkins/the-napkins on.'
b. Brevet ble klistret frimerker på.
'The letter was pasted stamps on.'
c. "? Brevet ble klistret noen grønne frimerker på.
'The letter was pasted some green stamps on.'
d. * Brevet ble klistret de grønne frimerkene på.
'The letter was pasted the green stamps on.'

In order to account for these facts, we need a different subject selection principle for Norwegian from the one proposed for English in (48):
(59) Norwegian Subject Selection Principle: If possible, the subject should be (a) a specific argument (b) of its local argument structure.
In (58b), (59) is satisfied because the object (frimerker) is a non-specific argument; so a nonargument can be the subject. In ( $58 \mathrm{c}, \mathrm{d}$ ), there is a specific argument of the verb that could be chosen as the subject and, so, making another NP (not an argument of the verb) the subject violates this principle.
B) In order to explain the fact that Icelandic, Danish, and Swedish (ID\&S) allow P-stranding arising from LDD, but not from passivization, as argued by Maling and Zaenen (1985) (see ex. (5)), we need to disallow nonthematic subjects in prepositional f-structures. This can be done by introducing a parameter of variation in constraint (35): while some languages (like English and Norwegian) allow nonthematic subjects in verbal and prepositional ( $[-\mathrm{N}]$ ) f-structures, other languages (ID\&S) only allow them in verbal ( $[-\mathrm{N}$, $+\mathrm{V}])$ f-structures, as indicated in (60):

## (60) Parametrized nonthematic subject constraint:

Nonthematic subjects are only possible in $\left\{\begin{array}{ll}\mathrm{a}) & {[-\mathrm{N}]} \\ \mathrm{b}) & {[-\mathrm{N},+\mathrm{V}]}\end{array}\right\}$-structures.
If the $f$-structure corresponding to a preposition cannot have a nonthematic subject, it will be impossible for the object of the preposition to raise up to the embedding f-structure without violating a principle of the Theory of Structure-Sharing such as Locality (28). Thus, ID\&S, with parameter setting (60b), disallow P-passives. This parameter setting does not exclude Pstranding by LDD, because the prepositional f-structure does not involve a nonthematic subject in this case; it involves an OP instead.
C) The Romance facts exemplified in (4) can be accounted for by a constraint that rules out structure-shared GFs in a prepositional f-structure:
(61) Structure-Sharing in prepositional f-structures: A prepositional fstructure cannot have two structure-shared GFs.

This constraint prevents P -stranding of both kinds: in either case, P -stranding is only possible if the object of the P is structure-shared with another GF of the same f-structure-a SUBJ, for P-passive, or an OP, for LDD. This is necessary for the structure to satisfy Locality (28).

The parameters of variation embodied in (60) and (61) predict the existence of three types of languages and the non-existence of the unattested type of language in table (6), where P-stranding is possible in P-passives but not in LDD. Active (61) gives no P-stranding; active (60a) gives both types of P-stranding; active (60b) gives P-stranding by LDD only.

## 6 Conclusions

This paper argues for the following claims. First, P-passives are an instance of the general passive operation, which suppresses the logical subject: There is no special rule for P-passives. Second, there is structure-sharing between the passive subject and the prepositional oblique's object, which allows the structure to satisfy the Subject Condition. Third, this S-S relation is possible because it satisfies the Theory of Structure-Sharing: No functional control equation or equivalent lexical device is needed to ensure this S-S. Fourth, GFs are not assigned to arguments in the argument-structure at the lexical level. This assignment arises in the syntax as part of the mapping between argument-structure and f-structure, which allows a passive clause to have a subject that is not an argument of the verb, but of a dependent of this verb.

Finally, since P-passive (a type of raising) and LDD both have S-S in common, we expect some covariation between the two constructions. This is what we see: No P-stranding is possible in French and other Romance languages in either construction, whereas it is possible in Germanic, either in both constructions, as in English and Norwegian, or only in LDD, as in ID\&S. If P-passive and LDD were unrelated phenomena, as they are usually assumed to be in LFG, we would expect a fourth type of language in which P -stranding is possible in passives but not in LDD. The present proposal predicts the non-existence of this type of language.

## References

Alsina, Alex. 1996a. The role of argument structure in grammar: Evidence from Romance. Stanford, CA: CSLI Publications.
Alsina, Alex. 1996b. Passive types and the theory of object asymmetries. Natural Language and Linguistic Theory 14:673-723.
Alsina, Alex. 2001. On the nonsemantic nature of argument structure. Language Sciences 23:355-389.
Alsina, Alex. 2008. A theory of structure-sharing: Focusing on long-distance dependencies and parasitic gaps. In Miriam Butt and Tracy H. King (eds.), Proceedings of the LFG2008 Conference, 5-25. Stanford, CA: CSLI Publications.
Baltin, Mark and Paul Postal. 1996. More on reanalysis hypothesis. Linguistic Inquiry 27:127-145.
Bolinger, Dwight. 1975. On the passive in English. In A. Makkai and V. B. Makkai (eds.), The first LACUS Forum 1974, 57-77. Columbia, SC: Hornbeam Press.
Bresnan, Joan. 1982. The passive in lexical theory. In Joan Bresnan (ed.), The mental representation of grammatical relations, 3-86. Cambridge, MA: MIT Press.
Bresnan, Joan. 2001. Lexical-functional syntax. Oxford: Blackwell.
Bresnan, Joan, and Jonni Kanerva. 1989. Locative inversion in Chichewa: A case study of factorisation in grammar. Linguistic Inquiry 20:1-50.
Davison, Alice. 1980. Peculiar passives. Language 56:42-66.
Gazdar, Gerald, and Geoffrey Pullum. 1981. Subcategorization, constituent order and
the notion 'Head'. In Michael Moortgat, Harry van der Hulst, and Teun Hoekstra (eds.), The scope of lexical rules, 107-127. Dordrecht: Foris.
Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan A. Sag. 1985. Generalized Phrase Structure Grammar. Oxford: Blackwell.
Falk, Yehuda. 1983. Constituency, word order, and phrase structure rules. Linguistic Analysis 11:331-360.
Falk, Yehuda N. 2001. Lexical-Functional Grammar. Stanford, CA: CSLI Publications.
Farrell, Patrick. 2005. English verb-preposition constructions: Constituency and order. Language 81:96-137.
Hornstein, Norbert, and Amy Weinberg. 1981. Case Theory and preposition stranding. Linguistic Inquiry 12:55-91.
Kayne, Richard. 1981. ECP extensions. Linguistic Inquiry 12:93-133.
King, Ruth, and Yves Roberge. 1990. Prepositional stranding in Prince Edward Island French. Probus 2:351-369.
Kiparksy, Paul. 1988. Agreement and linking theory. Unpublished ms., Stanford University.
Law, Paul. 2006. Preposition stranding. In Martin Everaert and Henk van Riemsdijk (eds.), The Blackwell companion to syntax, vol. III, 631-684. Malden, MA: Blackwell Publishing.
Lødrup, Helge. 1991. The Norwegian pseudopassive in lexical theory. Working Papers in Scandinavian Syntax 47:118-129.
Maling, Joan, and Annie Zaenen. 1985. Preposition stranding and passive. Nordic Journal of Linguistics 8:197-209. [Cited from Joan Maling and Annie Zaenen (eds.), Modern Icelandic Syntax (Syntax and Semantics 24), 153-164. San Diego, CA: Academic Press.]
Postal, Paul. 1986. Studies of passive clauses. Albany, NY: State University of New York Press.
Sag, Ivan A. 1987. Grammatical hierarchy and linear precedence. In Geoffrey Huck and Almerindo Ojeda (eds.), Discontinuous constituency (Syntax and Semantics 20), 303-340. New York, NY: Academic Press.

Toivonen, Ida. 2003. Non-projecting words. A case study of Swedish particles. Dordrecht: Kluwer.
Tseng, Jesse. 2007. English prepositional passive constructions. In Stefan Müller (ed.), Proceedings of the HPSG07 Conference, 271-286. Stanford, CA: CSLI Publications.
Ziv, Yael, and Gloria Sheintuch. 1980. Passives of obliques over direct objects. Lingua 54:1-17.
Zwicky, Arnold M. 1987. Slashes in the passive. Linguistics 25:639-669.

# ANIMACY EFFECTS AND LOCATIVE MARKING IN SHONA APPLICATIVES 

Raúl Aranovich<br>University of California Davis

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

As in other symmetrical languages, the relative order of theme and goal in Shona applicatives is free. However, when the two complements are human, word order freezes inside the VP, with the applied argument realized as the primary object. This paper develops an analysis of this phenomenon in Bidirectional Optimality Theory. It is argued that Object Freezing results from the absence of formal differences between alternative candidates. Additional data from locative marked applied objects provides evidence in favor of the weak version of B-OT, since more marked forms constitute superoptimal pairs with more marked meanings.


"The harder I try the more difficult I find it to say which nominal phrases are syntactic objects in Bantu."

Thilo C. Schadeberg (1995: 173)

## 1 Locative applied objects in Bantu

Locative-marked objects are well-documented among the members of the Bantu family (Bresnan 1999). A less noticed phenomenon is the occurrence of locative prefixes on the objects of applied verbs (Marten 2003, Rugemalira 2004), as shown in (1a). In Shona, a symmetrical language (Harford 1993), locative-marked objects can also occur as secondary objects (O2). This is shown in (1b). ${ }^{1}$
(1) a. [Ngoni, Rugemalira 2004:287]

Ag-il-a $\quad k u$-dasi.
get.lost-APPL-F LOC-wild
'get lost in the wild'

[^20]b. [Shona]

Murume a-kand-ir-a chimuti ku-imbwa.
man 3Sg-throw-APPL-F stick LOC-dog
'the man threw a stick at the dog.'

Locative-marked O2 matter to a theory of applicative constructions because of the way in which they interact with animacy. Hawkinson \& Hyman (1974) observe the following constraint on Shona ditransitives: When the two complements of a ditransitive verb are human, the applied argument can only be linked to a primary object (O). However, locative marked applied objects are free from this constraint. In this paper, I will argue that a structure with a locative-marked O 2 is a marked form that allows the listener to retrieve a marked argument structure. This analysis, in turn, offers evidence for a Bidirectional Optimality-Theoretical approach to argument linking (Blutner 2000, Blutner et al. 2006).

## 2 Shona as a symmetrical language.

In symmetrical languages, either complement of a ditransitive can be an O (Bresnan \& Moshi 1990). Evidence that Shona is symmetrical comes from pronominalization, word order alternations, and passivization.

In an applied construction, the applied argument (i.e. a beneficiary) can be placed immediately after the verb (2a), and it can also be cross-referenced by an object prefix (2b). These are two of the general characteristic properties of primary objects across the Bantu languages. But themes can also have these properties, as the examples in (3a-b) show. In Shona applicative constructions with two complements, then, either complement can display primary object properties at one time. This makes Shona a symmetrical language.
a. Nda-vig-ir-a mwana chipo.

1Sg.PAST-hide-APPL-F child gift
'I have hidden the gift for the child.
b. Mwana nda-mu-vig-ir-a chipo.
child 1Sg.PAST-3Sg-hide-APPL-F gift
'The boy, I have hidden the gift for him.'
(3)
a. Nda-vig-ir-a chipo mwana.

1Sg.PAST-hide-APPL-F gift child
'I have hidden the gift for the boy.'
b. Chipo nda-chi-vig-ir-a mwana.
gift 1Sg.PAST-3Sg-hide-APPL-F child
'The gift, I have hidden it for the child.'
According to Lexical Mapping Theory, symmetrical languages arise when the theme and the applied argument can be intrinsically marked $[-\mathrm{r}]$. The $[-\mathrm{r}]$ argument, whichever it is, is linked to a primary object, as in (4) and (5).


Passivization provides additional evidence that Shona is symmetrical. When a transitive applied verb is passivized, either one of its internal arguments can be realized as the subject, as shown in ( $6 \mathrm{a}-\mathrm{b}$ ). Moreover, when one of the internal arguments is linked to the subject of the passive, the other one still displays primary object properties, like being able to topicalize and be crossreferenced by an object prefix, as in (6c).
(6) a. Chipo cha-vig-ir-w-a mwana.
gift 3Sg-hide-APPL-PASS-F boy
'the gift has been hidden for the boy.'
b. Mwana a-ka-vig-ir-w-a chipo. boy 3Sg-PAST-hide-APPL-PASS-F gift 'the boy has been hidden the gift for.'
c. Mwana a-ka-chi-vig-ir-w-a chipo. boy 3 Sg-PAST-3Sg-hide-APPL-PASS-F gift 'the gift, it has been hidden for the boy.'

As in a truly symmetrical language, the two internal arguments of a Shona ditransitive predicate can be intrinsically marked [-r] concurrently. ${ }^{2}$ Because of the Biuniqueness Principle, this only occurs when the agent is suppressed or mapped onto an oblique (Bresnan \& Moshi 1990). Thus, when one of the internal arguments is linked to the subject, the other one is linked to the primary object. The linking of the two possible f-structures corresponding to the passive of a ditransitive are shown in (7).

(7) \begin{tabular}{lll}

Agent \& \begin{tabular}{l}
Ben <br>
-- <br>
{$[-r]$}

 \& 

Theme <br>
{$[-\mathrm{r}]$}
\end{tabular} <br>

--------------------
\end{tabular}

## 3 The status of locative applicatives

Locative-marked NPs can function as obliques (Obl), as in (8a). These locative arguments can also be realized as applied objects. When they are realized as primary objects, as in (8b), locative applied arguments lose their locative prefix. There is also an alternative construction in which the locative argument retains the locative prefix, and it follows the theme. Interestingly, the verb has the applicative suffix, as shown in (8c). In this sense, the construction in (8c) contrasts with the one on (8a).
(8) a. Murume a-kand-a chimuti ku-imbwa.
man 3Sg-throw-F stick LOC-dog
'the man threw a stick towards the dog.'
b. Murume a-kand-ir-a imbwa chimuti.
man 3Sg-throw-APPL-F dog stick
'the man threw a stick at the dog.'
c. Murume a-kand-ir-a chimuti ku-imbwa.
man 3Sg-throw-APPL-F stick LOC-dog
'the man threw a stick at the dog.'

2 Alsina (1996) makes a distinction between alternating and symmetrical languages. In alternating languages, more than one internal argument can display primary object properties, but only one argument at a time can do so. In symmetrical languages, on the other hand, more than one internal argument can display primary object properties at the same time. Shona is a language of the latter type

My claim is that the locative-marked NP in (8c) is an O2, not an oblique. As expressed in the quote from Thilo Schadeberg's (1995) paper on Bantu objects, the diagnostics for objecthood are not always decisive. The evidence to distinguish a locative-marked O 2 from an Obl is also elusive, but convincing evidence for analyzing the locative argument in (8c) as an object comes from the behavior of locative-marked NPs in causatives. When a monotransitive predicate is expanded with the causative suffix -is-, as in (9a), the causee is realized as a primary object. If a causative construction is built on a ditransitive predicate, on the other hand, the causee is realized as an oblique, marked with the instrumental preposition ne. This is shown in (9b). The reason for this contrast is quite clear: ditransitives already have two complements, so the causee cannot be mapped onto an objective function when the external instigator takes over the function of sentential subject. If the locative-marked argument of (8c) is a complement, then the clause should behave as a ditransitive when it is expanded with the causative suffix. This is indeed what can be observed. The causee cannot be linked to a primary object (10a), it must be realized as an oblique instead, to avoid 'overcrowding' (10b).

## varume.

Pl.man
'the chief made the men build a house for his wife.'
'I made the farmer throw a stick at the dog.'

## 4 Animacy constraints on Shona applicatives

Goals, sources, and other locative arguments, then, can be realized as primary objects or as secondary objects in applicative constructions, as expected in a
symmetrical language. Unlike beneficiaries, however, locative arguments are marked with a locative prefix when they occur as applied secondary objects. There is another difference between applied beneficiaries and applied locative arguments, with respect to animacy. Hawkinson \& Hyman (1974) show that when an applicative construction has two human complements, Shona ceases to behave like a symmetrical language: the primary object is always interpreted as the beneficiary, and the secondary object as the theme/patient. This is shown in (11). ${ }^{3}$
[Hawkinson \& Hyman 1974:151]
Murume a-ka-chek-er-a mukadzi mwana.
man 3 Sg-PAST-cut-APPL-F woman child
'the man cut the child for the woman.'
(not: 'the man cut the woman for the child.')

Human locative applied objects can also be realized as primary objects when the theme is human. This is shown in (12a) and (13a). But unlike beneficiaries, human applied goals and sources can be realized as O2s even if the theme is human, as in (12b) and (13b). In these cases, however, the locative marker on the applied object is obligatory.
a. Nda-si-ir-a murume mhandara.

1Sg-PAST-leave-APPL-F man maiden
'I have left the maiden to the man.'
b. Nda-si-ir-a mhandara ku-murume.

1Sg-PAST-leave-APPL-F maiden LOC-man
'I have left the maiden to the man.'
(13)
a. Nda-vig-ir-a murume mhandara.

1Sg-PAST-hide-APPL-F man maiden
'I have hidden the maiden from the man.'
3 MChombo \& Firmino (1999) report a similar pattern in Gitonga. In fact, animacy effects in ditransitives are not uncommon in the Bantu family. In Runyambo, for instance, human complements precede non-human ones, and animates precede inanimates (Rugemalira 1991). The same constraint is observed in Sesotho (Morolong and Hyman 1977). According to Wald (1994, 1998), postverbal animacy effects of this type are characteristic of the Southern Bantu languages. To the North, animacy effects are observed in the behavior of verbal object markers. Thus, in Swahili, only the complement that is higher in animacy can be cross-referenced by an object marker. Wald notices that Shona displays a transitional pattern of animacy, between North and South.
b. Nda-vig-ir-a mhandara ku-mигите. 1Sg-PAST-hide-APPL-F maiden LOC-man 'I have hidden the maiden from the man.'

Object Freezing occurs in Shona when a sentence is potentially ambiguous, then. Symmetrical mappings are possible when the internal arguments are formally differentiated by means of animacy or locative marking. When the two complements are human and unmarked, the only possible linking ensures that the first argument is interpreted as the beneficiary. In recent years, a bidirectional approach to optimization has been developed to account for such cases of disambiguation (Blutner 2000, Blutner et al. 2006). Bouma (2008), for instance, applies Bidirectional Optimality Theory (B-OT) to explain restrictions on the relative order of subject and object in Dutch. In the following sections, I will develop a B-OT analysis of Object Freezing in Shona. In my analysis, as in many other applications of B-OT, markedness will play a central role. The logic of bidirectional optimization is that marked candidates are not discarded as sub-optimal if they can be interpreted as the expression of a marked input. A B-OT account of disambiguation will then find a way to pair up a marked meaning with a marked form. The OT literature features a number of proposals about animacy effects in syntax based on a sub-hierarchy of markedness constraints (Aissen 1999, 2003, Bresnan....). I will argue that these proposals cannot account for Object Freezing in Shona in their original unidirectional, interpretive optimization framework, but that they provide a useful foundation for the bidiriectional analysis I develop here. ${ }^{4}$

## 5 Bidirectional OT: Basic concepts and refinements

An Optimality-Theoretical grammar (Prince and Smolensky 1993) consists of a ranked set of violable constraints, that evaluate a set of candidate forms for a given input structure. The candidate that incurs the least severe violations of the constraint set is selected as the output. Applications of OT to syntax often reflect a 'productive optimization' bias, assuming that the input is some

[^21]sort of semantic representation, or argument structure, and the output some sort of fully formed syntactic structure. Bidirectional OT complements this kind of productive optimization with an interpretive optimization, in which the input is a syntactic structure and the output a semantic representation. The result of bidirectional optimization is a super-optimal form-meaning pair $<f, m>$ that is optimal in either direction. In addition, B-OT (in its weak version) may select some pairs $\left\langle f^{\prime}, m^{\prime}\right\rangle$ as super-optimal even if they are not optimal in any direction. Marked candidates, which normally lose to less marked forms in a uni-directional competition, may then be rescued by the logic of bidirectional optimization as the output for a marked input. This aspect of B-OT makes it a natural framework to formalize analyses of differential argument marking.

In its weak version (Jäger 2002), the definition of a super-optimal pair in B-OT is formulated as follows:
(14) A form-meaning pair $\langle f, m\rangle$ is called super-optimal if and only if:
i. there is no distinct super-optimal pair $\left\langle f^{\prime}, m\right\rangle$ such that $\left.\left\langle f^{\prime}, m\right\rangle\right\rangle$ $<f, m>$, and
ii. there is no distinct super-optimal pair $\left\langle f, m^{\prime}\right\rangle$ such that $\left.\left\langle f, m^{\prime}\right\rangle\right\rangle$ $<f, m>$.

The selection of super-optimal candidates is represented in the diagram in (15), in which the arrows point to the more harmonic candidate in a productive optimization (horizontal) or in an interpretive one (vertical). A black arrow indicates that a candidate loses to (or is blocked by) a superoptimal one.


The pair $\langle f, m\rangle$ is super-optimal by virtue of having the most harmonic (i.e. least marked) productive output $f$ and interpretive output $m$. In a productive optimization that has $m$ as input, for instance, $f^{\prime}$ will lose to $f$, so the pair $<f^{\prime}, m>$ cannot be super-optimal. In an interpretive optimization that has $f^{\prime}$ as input, on the other hand, $m$ will outperform $m^{\prime}$. But since $\left\langle f^{\prime}, m>\right.$ is not superoptimal, this leaves $\left\langle f^{\prime}, m^{\prime}\right\rangle$ as the super-optimal pair, according to the definition in (14). The same reasoning applies to $\left\langle f, m^{\prime}\right\rangle$, the other corner of the diagram in (15). In other words, the pair $<f, m>$ blocks the pairs $\left\langle f^{\prime}, m>\right.$
and $\left\langle f, m^{\prime}\right\rangle$, clearing up the field for the more marked pair $\left\langle f^{\prime}, m^{\prime}\right\rangle$.
Because of its representational nature and its parallel architecture, LFG can be advantageously used to formalize the constraints of OT syntax (Bresnan 2000). Kuhn (2003) applies B-OT to LFG, considering correspondences between f-structure and c-structure. I propose to evaluate pairs $\langle f, a>$ of an f-structure $f$ and an a-structure $a$. I divide markedness constraints in two families: (a) productive constraints, which rule against combinations of f-structure features like grammatical functions, case, etc.; and (b) interpretive constraint, which rule against combinations of a-structure features like semantic roles. Since gender, animacy, and other inherent grammatical features of syntactic constituents are not specified as part of the argument structure of a predicate, constraints on animacy and gender marking (i.e. nominal class prefixes in Bantu) will only be active on the productive side of optimization.

## 6 A B-OT account of locative OBJ2

In informal terms, my analysis is based on two assumptions: (a) an astructure with a restricted goal/source is marked, ${ }^{5}$ and (b) an f-structure with a locative object is also marked. Following the OT treatment of Lexical Mapping Theory in Aranovich (2009), mapping constraints are stated in terms of marked associations of the functional features $[ \pm \mathrm{r}, \pm \mathrm{o}]$ with other linguistic features (animacy, thematic roles, etc.). The constraints I propose as part of my analysis of Shona applicatives are stated in (16).
(16) a. $*[+o] /$ LOC: objective functions are not marked for locative gender. ${ }^{6}$
b. *[+r]/goal: goals (and sources) are not restricted arguments.

5 This assumption is based on the limited distribution of restricted applied objects in East Bantu. According to Wald (1994), West Bantu languages like Umbundu are direct object languages (i.e. they allow for passivization of the theme instead of the beneficiary), while East Bantu languages are innovative in moving towards a primary object pattern, as exemplified by Swahili. In a language like Umbundu, the unmarked option is to assign the feature [-r] to the theme, not to the beneficiary, but in a language like Swahili the markedness values are reversed. Matters are complicated somewhat by the emergence of a symmetrical pattern in Interior Bantu and Southeast Coast Bantu. But even there the unmarked status of the primary object pattern is apparent: in Central Chewa, the theme can passivize with instrumental applied objects, but not with beneficiary applied objects.
6 LOC refers to locative gender as an f-structure feature, not to a locative role in argument structure.

Tableaux 1 and 2 show how one of the super-optimal pairs is selected. This is the pair linking the unmarked a-structure to the unmarked f-structure. In this way, applied goals are encoded as primary objects (without locative marking), and these in turn are interpreted as goals.
TABLEAU 1a: productive optimization

| a: V <ag, goal[-r], th[+r]> | $*[+\mathrm{o}] /$ LOC |
| :---: | :---: |
| $\mathrm{f}: \mathrm{S}, \mathrm{O}, \mathrm{O} 2$ |  |
|  | $*!$ |

TABLEAU 1b: interpretive optimization

| f: S, O, O2 | *[+r]/goal |
| :---: | :---: |
| \% a: V <ag, goal[-r], th[+r]> |  |
| $\mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], goal[+r]> | *! |

In addition, bidirectional optimization links the marked a-structure with the marked f-structure (i.e. it encodes a [+r] goal/source as a locative-marked O2). The following tableaux show the selection of the disharmonic candidates as super-optimal (the optimal candidates are indicated with ©, the dark pointing hand). These are not super-optimal candidates, however, being blocked by the winning ones in tableaux 1 and 2 . For instance, the pair $\left\langle f, a^{\prime}\right\rangle$, which is the most harmonic one in tableau 2a, is blocked by $\langle f, a\rangle$ in tableau 1 b . Since there is no competition in which $\left\langle f^{\prime}, a^{\prime}\right\rangle$ is less harmonic than another super-optimal pair, $\left\langle f^{\prime}, a^{\prime}\right\rangle$ is itself super-optimal.

TABLEAU 2a: productive optimization

| $\mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], goal[+r]> | $*[+\mathrm{o}] / \mathrm{LOC}$ |
| :--- | :---: |
| - $\mathrm{f}: \mathrm{S}, \mathrm{O}, \mathrm{O} 2$ |  |
| \& $\mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}$, Loc-O2 | $*!$ |

TABLEAU 2 b : interpretive optimization

| $\mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}$, Loc-O2 | $*[+\mathrm{r}] /$ goal |
| :--- | :---: |
| $\sim \mathrm{a}: \mathrm{V}<\mathrm{ag}$, goal[-r], th[+r]> |  |
| \& $\mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], goal[+r]> | $*!$ |

## 7 A B-OT account of beneficiary applied objects

The set of productive and interpretive competitions discussed in the previous section insure that when a locative argument (goal or source) is realized as an O it does not bear a locative marker, but that this locative marker is retained when the locative argument is realized as an O 2 . This accounts, for instance for the alternation between (12a) and (12b), and also between (13a) and (13b). Unlike locative arguments, beneficiaries are never marked with locative prefixes, not even as O2. ${ }^{7}$ Locative marking, then, cannot formally distinguish an applicative structure with a beneficiary O from a structure with a beneficiary O 2 . With two human complements, the interpretation neutralizes to the unmarked argument structure in which O is always interpreted as the beneficiary, resulting in Object Freezing. At this point, an attentive reader may ask how it is possible for a beneficiary to ever be realized as an O2, even when the theme is not human. What I argue in this section is that the inanimate feature of the theme helps to disambiguate the structure, forcing the interpretation of the human argument as a beneficiary, even when it is an O2.
(In)animacy, like locative marking, provides the marked formal features that allows the listener to retrieve the marked argument structure. In informal terms, f-structures with inanimate unrestricted arguments are marked, and can be linked to a marked a-structure with a [+r] beneficiary in a bidirectional optimization. To account for the markedness of these structures, I propose the additional constraints in (17).
(17) a. $*[-r] /$ Inan(imate): Unrestricted functions are not assigned to NPs with inanimate features. ${ }^{8}$
b. $\quad *[+r] /$ ben: Beneficiaries are not restricted. ${ }^{9}$

7 In fact, because they cannot have locative prefixes, beneficiaries can only be realized as applied objects, never as obliques.
8 The motivation for this constraint is discussed in more detail in section 9. Notice that $[-r]$ functions include $S$ and the primary object $O$, so this constraint is not qualifying an inanimate O as more marked than an S . Rather, it ensures that an inanimate O is paradigmatically more marked than an inanimate O 2 . Notice, too, that this constraint does not state that the feature [-r] is marked, only that its association with an NP with inanimate features is marked.
9 A reviewer remarks that beneficiaries, because of the nature of their semantics, are restricted. However, in primary object languages, beneficiaries are assigned to [-r] functions, not themes. This is because when there is more than one internal argument the one that is unrestricted is the one that ranks higher in the thematic hierarchy, the beneficiary in this case.

The following tableaux show the selection of a super-optimal pair that matches a marked a-structure with a restricted applied beneficiary ( $a^{\prime}$ ) and a marked f -structure in which the O is inanimate $\left(f^{\prime}\right)$.

TABLEAU 3a: productive optimization

| $\mathrm{a}: \mathrm{V}<\mathrm{ag}, \mathrm{th}[-\mathrm{r}]$, ben[+r]> | $*[-\mathrm{r}] /$ Inan |
| :--- | :---: |
| $* \mathrm{f}: \mathrm{S}, \mathrm{O}, \mathrm{O} 2_{[\text {inan }]}$ |  |
| \& $\mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}_{[\text {inan }]}, \mathrm{O} 2$ | $*!$ |

TABLEAU 3b: interpretive optimization

| $\mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}_{\text {[inan] }}, \mathrm{O} 2$ | $*[+\mathrm{r}] / \mathrm{ben}$ |
| :---: | :---: |
| - $\mathrm{a}: \mathrm{V}<\mathrm{ag}$, ben[-r], th[+r]> |  |
| अ $\mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], ben[+r]> | $*!$ |

The bidirectional optimization in tableaux (3a) and (3b), then, accounts for the contrast between examples like (2a) and (3a), in which an inanimate theme can be realized as either an O 2 or as an O . The beneficiary can be realized as an O 2 in (3a) without any additional marking, since animacy distinguishes it from the theme. When both the beneficiary and the theme are human, on the other hand, there are not two formally distinct f-structures that can enter the competition, but there are two distinct a-structures. The competition for super-optimal status is confined to two pairs, then: the unmarked $\langle f, a\rangle$, and the less harmonic $\left\langle f, a^{\prime}\right\rangle$. Since $\langle f, a\rangle$ is already superoptimal, it blocks $\left\langle f, a^{\prime}\right\rangle$. This is shown in tableau 4.

TABLEAU 4: interpretive optimization

| $\mathrm{f}: \mathrm{S}, \mathrm{O}_{\text {[hum] }}, \mathrm{O} 2_{\text {[hum] }}$ | $*[+\mathrm{r}] /$ ben |
| ---: | :---: |
| a: V <ag, ben[-r], th[+r]> |  |
| $\mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], ben[+r]> | $*!$ |

A clause with an applied human beneficiary and a human theme, then, has only one possible interpretation, one in which the beneficiary is the O and the theme is the O 2 . The contrast between an a-structure with a [-r] beneficiary and an a-structure with a [+r] beneficiary is neutralized in favor of the former, since there is no formal means to express the distinction. This is the
desired result, given the observations about sentences like (11).

## 8 Locative marking in applied constructions with inanimate themes

The last case to consider is probably the hardest. It involves sentences (8b) and (8c), repeated below as (18a-b), with the goal in boldface. Since the theme is inanimate, and therefore distinct from a human goal, the locative marking on the goal seems redundant when the goal is an O 2 .
a. Murume akand-ir-a imbwa chimuti.
man 3Sg-throw-APPL-F dog stick
'the man threw a stick at the dog.'
b. Murume a-kand-ir-a chimuti ku-imbwa
man 3Sg-throw-APPL-F stick LOC-dog
'the man threw a stick at the dog.'

The observation that explains the presence of the locative prefix in (18b) is that (18b) also competes against sentences in which the O 2 is a morphologically unmarked applied argument, i.e. a beneficiary. The logic of multiple-way competitions in B-OT is such that the more marked the form, the more marked the meaning it is going to be paired with. Once a superoptimal pair is identified, any less harmonic pair with which it competes is blocked. This is illustrated in the following diagram:


An f-structure with an inanimate O and a locative marked O 2 is the most marked. A bidirectional optimization will pair it up with the most marked astructure. For this to be the case, a restricted locative applied argument must be more marked than a restricted beneficiary. This is achieved by ordering the markedness constraints I have introduced in (16b) and (17b) in the following way:

$$
\begin{equation*}
*[+\mathrm{r}] / \text { goal } \gg *[+\mathrm{r}] / \mathrm{ben} \tag{20}
\end{equation*}
$$

As shown in Tableau 5a, the f -structure $f$ is the most harmonic one, but it
cannot make a super-optimal pair with the a-structure $a^{\prime \prime}$ because $f$ is interpreted as the expression of the unmarked argument structure $a$ (i.e. the astructure in which the goal is [-r]). The next f-structure in the harmonic ranking is $f^{\prime}$, but the pair $\left\langle f^{\prime}, a^{\prime \prime}\right\rangle$ is blocked by another super-optimal pair, $\left\langle f^{\prime}, a^{\prime}\right\rangle$ (where $a^{\prime}$ is the a-structure that has a beneficiary marked as [+r]). This kind of blocking is indicated here by the mark. This leaves $f^{\prime \prime}$ as the only viable candidate to form a super-optimal pair with $a^{\prime \prime}$. A similar reasoning applies to the interpretive optimization in tableau 5b. Even though $a$ is the unmarked interpretation, it cannot form a super-optimal pair with $f^{\prime \prime}$ because the pair $\langle f, a\rangle$ blocks it. The next most harmonic a-structure $a^{\prime}$ has an restricted beneficiary, but this is the super-optimal interpretation for an f structure $f^{\prime}$, with an O2 that has no locative prefix. The pair $\left\langle f^{\prime}, a^{\prime}\right\rangle$, then, blocks the pair $\left\langle f^{\prime \prime}, a^{\prime}\right\rangle$. The only expression for $f^{\prime \prime}$ that is not blocked by another super-optimal pair is $a^{\prime \prime}$.

TABLEAU 5a: productive optimization

| a": V <ag, th[-r], goal[+r]> | $*[+\mathrm{o}] /$ LOC | $*[-\mathrm{r}] /$ Inan |
| :--- | :---: | :---: |
| - $\mathrm{f}: \mathrm{S}, \mathrm{O}, \mathrm{O} 2_{\text {inan }}$ |  |  |
| $* \mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}_{\text {inan }}, \mathrm{O} 2$ |  | $*!$ |
| \& $\mathrm{f}^{\prime}: \mathrm{S}, \mathrm{O}_{\text {inan }}$, Loc-O2 | $*!$ | $*$ |

TABLEAU 5b: interpretive optimization

| $\mathrm{f}^{\prime \prime}: \mathrm{S}, \mathrm{O}_{\text {[inan }}$, Loc-O2 | $[+\mathrm{r}] /$ goal | $*[+\mathrm{r}] /$ ben |
| :--- | :--- | :--- |
| © $\mathrm{a}: \mathrm{V}<\mathrm{ag}$, goal[-r], th[+r]> |  |  |
| $\boldsymbol{*} \mathrm{a}^{\prime}: \mathrm{V}<\mathrm{ag}$, th[-r], ben[+r]> |  | $*!$ |
| \& a': V <ag, th[-r], goal[+r]> | $*!$ |  |

Even in symmetrical languages, then, one of the ditransitive structures is marked. In Shona, locative gender and animacy are clues that make it possible for the listener to assign the marked interpretation to a ditransitive applicative. When these surface clues are absent, only the unmarked interpretation is available to the listener.

## $9 \quad$ B-OT and the 'harmonic alignment' analysis compared

The bidirectional model of Object Freezing in Shona builds on earlier OT approaches to animacy. Aissen $(1999,2003)$ uses harmonic alignment to
model the implicational effects of person-animacy scales on voice alternations and differential object marking. In this approach, a personanimacy scale like the one in (21a) is aligned with the binary scale of grammatical functions in (21b), yielding the constraint sub-hierarchies in (22)
(21) a. 1 st, 2 nd $>3$ rd $>\operatorname{Hum}($ an $)>\operatorname{Anim}($ ate $)>\operatorname{Inan}($ imate $)$
b. $\quad$ S(ubject) $>\mathrm{O}($ bject $)$
a. $\quad *$ S/Inan $\gg$ *S/Anim $\gg$ SS/Hum $\gg *$ S/3rd $\gg *$ S/1st, $*$ S/2nd
b. $\quad * \mathrm{O} / 1 \mathrm{st}, 2 \mathrm{nd} \gg$ O/3rd $\gg$ O/Hum $\gg$ *O/Anim $\gg$ *O/Inan

The effect of these constraint sub-hierarchies is that pronominal objects are more marked than human objects, human objects are more marked than nonhuman objects, and so on. For subjects, the markedness relations are inverted. When a patient is human, then, the grammar of a particular language may prefer to link it to a subject (as in a passive construction, for instance), or else to realize the patient as a morphologically marked object, as it happens in Spanish and other Romance languages.

It is not difficult to conceive an extension to ditransitives of Aissen's work on animacy effects on transitives, assuming the binary scale $O>O 2$ of grammatical functions. This analysis predicts that a sentence in which the patient is mapped to a more prominent function than the goal (i.e. the O ) is marked. ${ }^{10}$ This is what happens, for instance, in the English prepositional dative. The marked structure is preferred, however, when the unmarked competitor (i.e. the English double object construction) would have a lowranked function (an O2) with more prominent animacy features. This is precisely what Bresnan et al. (2005) conclude in their study of the occurrence of ditransitives in the Switchboard corpus of spoken English. They found that, among other factors, pronominal complements tend to precede nominal complements, and animate complements tend to precede inanimate complements. Thus, a prepositional dative sentence like (23a) would be more likely to occur than a double object sentence like (23b). ${ }^{11}$

10 In addition, prepositional dative sentences are more marked than double object sentences because of the additional structural complexity contributed by the prepositional phrase (i.e. they violate *STRUC, the same constraint penalizing passives).
11 This study is a follow-up to Bresnan et al. (2001), who argue that the effects of the constraint sub-hierarchies proposed by Aissen are also observed in patterns of syntactic variation. They analyze the occurrence of passive in the Switchboard corpus to find that the ratio of active to passive is the smallest for clauses in which the patient outranks the agent in person features.
(23) a. give it to the child
b. $\quad$ give the child it

One way to account for this kind of animacy effects on ditransitives is to harmonically align the person-animacy scale in (21a) with the binary scale $O$ $>O 2$, yielding the constraint sub-hierarchies in (24). The constraint subhierarchy in (24b) has the effect of penalizing those candidates whose secondary objects rank higher in the scale of person-animacy.

$$
\begin{align*}
& \text { a. } * \mathrm{O} / \mathrm{Inan} \gg * \mathrm{O} / \text { Anim } \gg * \mathrm{O} / \mathrm{Hum} \gg * \mathrm{O} / 3 \mathrm{rd} \gg * \mathrm{O} / 1 \mathrm{st}, * \mathrm{O} / 2 \mathrm{nd}  \tag{24}\\
& \text { b. } * \mathrm{O} 2 / 1 \mathrm{st}, 2 \mathrm{nd} \gg * \mathrm{O} 2 / 3 \mathrm{rd} \gg * \mathrm{O} 2 / \mathrm{Hum} \gg * \mathrm{O} 2 / \text { Anim } \gg \\
& * \mathrm{O} 2 / \mathrm{Inan}
\end{align*}
$$

This extension of Aissen's harmonic alignment approach to ditransitives has to be ammended, however, because of two shortcomings. The first one is of a general nature, and it can be solved by harmonically aligning semantic roles and animacy features with the functional features $[ \pm 0]$ and $[ \pm r]$. The second one arises when the unidirectional, productive-oriented optimization implied in Aissen's model is applied to Object Freezing in Shona. This, too, can be solved, by adopting the bidirectional approach I have developed in the preceding sections.

The first shortoming is that the sub-hierarchy in (24a) seems to contradict the hierarchy previously discussed in (22b): it favors candidates whose primary objects rank higher on the animacy scale. One possible solution to this problem is to replace the grammatical functions in the constraints in (22) and (24) by their distinctive functional features, $[ \pm 0]$ and $[ \pm r]$ respectively. The sub-hierarchies in (24) are then replaced by the ones in (25).

```
a. \(\quad *[-r] /\) Inan \(\gg *[-r] /\) Anim \(\gg *[-r] / H u m \gg *[-r] / 3 r d \gg *[-r] / 1 s t\),
    *[-r]/2nd
    b. \(\quad\) [ \([\mathrm{r}] / 1 \mathrm{st}\), 2 nd \(\gg\) *[rr]/3rd \(\gg\) *[+r]/Hum >> *[+r]/Anim >>
    *[+r]/Inan
```

According to this model, animacy features high in the hierarchy are unmarked for $[-\mathrm{o}]$ and [-r] functions. Notice that the constraint *[-r]/[Inan], introduced in (17a), now finds its place among the constraints in the subhierarchy (25a).

The second shortcoming concerns the unidirectional application of
these constraints to Object Freezing in Shona. The animacy effects on Shona ditransitives discussed in Hawkinson and Hyman (1974), which are the focus of my paper, are often cited as the categorical counterpart to the English preference for clauses in which the first complement outranks the second in animacy. However, on close examination, the harmonic alignment model of animacy effects cannot be straightforwardly applied to Object Freezing in Shona. A markedness sub-hierarchy of alignment constraints determines the outcome of optimization only when the candidates have arguments that differ in animacy. Object Freezing represents the opposite situation. When the two complements are [+human], only one argument mapping is possible: the O can only be a goal, not a theme. An alignment constraint of the form *O2/Hum (or *[+r]/Hum) cannot decide between a candidate in which the O 2 is the goal (or beneficiary), and a competitor in which the O 2 is a theme. Conversely, if the candidates' complements have different animacy features, a markedness constraint should rule out one in favor of the other. But this is the case in which either candidate is a possible output: when the theme is [animate], it can be mapped onto an O or onto an O 2 .

For the harmonic alignment approach to Object Freezing in Shona to succeed both directions of optimization have to be taken into account. The central hypothesis in this analysis is that marked f-structures have an O 2 that outranks the O in a scale of animacy. ${ }^{12}$ Due to the logic of bidirectional optimization, the marked f-structures are not discarded. Instead, they are recycled as the expression of a marked a-structure. Object Freezing occurs when there is no competing, more marked f-structure that can be interpreted as the expression of a marked a-structure. This occurs when the overt resources of the grammar, in terms of animacy features or morphological marking, are too poor to mark two alternative structures as formally distinct. Object Freezing in Shona, then, offers a compelling argument in favor of bidirectional optimization. It also offers indirect but convincing evidence that the same soft constraints that are at work in English have a categorical effect in other languages.

## References

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. Natural Language \& Linguistic Theory 17: 673-711.

12 There are applicative constructions in some other Bantu languages, however, which seem to behave according to the unidirectional, productive model of optimization (cf. footnote 2). It is open to future research how to integrate these languages in the bidirectional model I am proposing for Shona.

Aissen, Judith. 2003. Differential object marking: iconicity vs. economy. Natural Language \& Linguistic Theory 21: 435-483.
Alsina, Alex. 1996. Passive types and the theory of object asymmetries. Natural Language \& Linguistic Theory 14: 673-723.
Aranovich, Raúl. 2009. Feature-based argument mapping and animacy optimization in impersonal passives. Linguistics 47: 619-652 .
Blutner, Reinhardt. 2000. Some aspects of optimality in natural language interpretation. Journal of Semantics 17: 189-216.
Blutner, Reinhard; Helen De Hoop, and Petra Hendriks. 2006. Optimal Communication. Stanford, CSLI Publications.
Bouma, Gerlof. 2008. Starting a Sentence in Dutch: A Corpus Study of Subject- and Object-Fronting. Ph.D. Dissertation, University of Groningen.
Bresnan, Joan. 1999. Locative case vs. locative gender. In Laurel A. Sutton and Christopher Johnson (eds.), Proceedings of the XVII Annual Meeting of the Berkeley Linguistic Society, pp. 53-58. Berkeley, Berkeley Linguistic Society.
Bresnan, Joan. 2000. Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), Optimality Theory: Phonology, Syntax and Acquisition, pp. 334-385. Oxford, Oxford University Press.
Bresnan, Joan; Anna Cueni, Tatiana Nikitina, and Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Krämer, and J. Zwarts (eds.). Cognitive Foundations of Interpretation, pp. 69-94. Amsterdam, Royal Netherlands Academy of Science.
Bresnan, Joan; Shipra Dingare, and Christopher Manning. 2001. Soft constraints mirror hard constraints: voice and person in English and Lummi. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG01 Conference. Stanford, CSLI Publications.
Bresnan, Joan, and Lioba Moshi. 1990. Object asymmetries in comparative Bantu syntax. Linguistic Inquiry 21: 147-185.
Fortune, G. 1955. An Analytical Grammar of Shona. London, Longmans, Green \& Co.
Fortune, G. 1980. Shona Grammatical Constructions II. Salisbury [Harare], University of Rhodesia [Zimbabwe].
Harford, Carolyn. 1993. The applicative in Chishona and Lexical Mapping Theory. In S. Mchombo (ed.), Theoretical Aspects of Bantu Grammar, pp. 93-111. Stanford. CSLI Publications.
Hawkinson, Annie, and Larry Hyman. 1974. Hierarchies of natural topic in Shona. Studies in African Linguistics 5: 147-170.

Jäger, Gerhard. 2002. Some notes on the formal properties of Bidirectional Optimality Theory. Journal of Logic, Language, and Information 11: 427-451.
Kuhn, Jonas. 2003. Optimality-Theoretic Syntax: A Declarative Approach. Stanford, CSLI Publications.
Marten, Lutz. 2003. The dynamics of Bantu applied verbs: an analysis at the syntax-pragmatics interface. In Kézie K. Lébikaza (ed.), Proceedings of the III World Congress of African Linguistics, pp. 207-221. Köln, Rüdiger Köppe.
MChombo, Sam, and Gregório Firmino. 1999. Double object constructions in Chichewa and Gitonga: a comparative analysis. Linguistic Analysis 29: 214-233.
Morolong, Malillo, and Larry Hyman. 1977. Animacy, objects and clitics in Sesotho. Studies in African Linguistics 8: 199-217.
Prince, Alan \& Smolensky, Paul. 1993. Optimality Theory: Constraint Interaction in Generative Grammar [Technical Report TR-2, Rutgers Center for Cognitive Science]. New Brunswick, Rutgers University.
Rugemalira, Josephat. 1991. What is a symmetrical language? Multiple object constructions in Bantu. In Kathleen Hubbard (ed.) Special Session on African Language Structures, Proceedings of BLS XVII, pp. 200-209. Berkeley, Berkeley Linguistic Society..
Rugemalira, Josephat. 2004. Locative arguments in Bantu. In A. Akinlabi and O. Adesola (eds.), Proceedings of the IV World Congress of African Linguistics, pp. 285-295. Köln, Rüdiger Köppe.
Schadeberg, Thilo C. 1995. Object diagnostics in Bantu. In E. 'Nolue Emenanjo and Ozo-mekuri Ndimele (eds.), Issues in African Languages and Linguistics: Essays in Honour of Kay Williamson, pp. 173-180. Aba, Nigeria, National Inst. for Nigerian Languages.
Wald, Benji. 1994. East Coast Bantu and the evolution of constraints on passivization. Sprache und Geschichte in Afrika 15: 211-315.
Wald, Benji. 1998. Issues in the North/South syntactic split of East Bantu. In Ian Maddieson and Thomas J. Hinnebushc (eds.), Language History and Linguistic Description in Africa [Trends in African Linguistics 2], pp. 95-106. Trenton/Asmara, Africa World Press.

# A LINGUISTIC AND COMPUTATIONAL MORPHOSYNTACTIC ANALYSIS FOR THE APPLICATIVE -I IN INDONESIAN 

 Avery Andrews ${ }^{\star}$, Jane Simpson $\downarrow$<br>- Australian National University, $\boldsymbol{*}$ Oxford University, $\boldsymbol{\bullet}$ University of Sydney<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009
CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

We present a precise LFG-based analysis of the suffix $-i$ in Indonesian, addressing the issues of applicative-causative polysemy of the suffix and its alternation with $-k a n$. We also show how the analysis can be integrated into an implementation of an existing computational grammar. Our computational implementation of applicativisation and related phenomena is the first of its kind for Indonesian, and also provides evidence for the robustness of LFG as a theory and XLE as a computational implementation of the theory in handling this linguistically complex phenomena.

Building on work on predicate composition (Alsina 1996; Butt 1995) and using the restriction operator (Kaplan and Wedekind 1993; Butt, King, and Maxwell III 2003; Butt and King 2006), we demonstrate that our novel unified a-structure based approach to verbal derivation in Indonesian can handle not only valence changing $-i$, giving rise to applicative-causative polysemy, but also valencepreserving $-i$. We argue that different types of $-i$ (and also $-k a n$ ) result from different possibilities of argument fusion: double or single fusion. Double fusion, which is typical for -i, results in applicativisation, whereas single fusion of $-i$ results in causativisation.


## 1 Introduction

Indonesian is one of the most extensively studied Austronesian languages (Chung 1976; Myhill 1988; Purwo 1989, 1995; Macdonald 2001; Musgrave 2001, among others), yet the precise linguistic analysis of the suffix $-i$ has not been investigated in detail. Linguistically, the analysis must address the issues of applicative-causative polysemy or homonymy of $-i,{ }^{1}$ and, from a computational point of view, we must address how the analysis can be integrated into an implementation of an existing computational grammar. The suffix $-i$, like the suffix -kan (Arka 1993), can appear as applicative or causative, as seen in (1). There are other uses of $-i$ for which the proper analysis as applicative or causative is not clear (as we discuss in section 2, examples (5)-(6)).
a. Applicative - $i$

| i. | $\mathrm{X}_{\text {SUBJ }}$ | datang | $[\mathrm{ke} \mathrm{Y}]_{\text {OBL }}$ | ' X come to $\mathrm{Y} '$ |
| :--- | :--- | :--- | :--- | :--- |$\quad$ (intransitive)

b. Causative $-i$ :

| i. | X | panas |  | ' X is hot' | (intransitive) |
| :--- | :--- | :--- | :--- | :--- | :--- |
| ii. | Y | panas- i | X | ' Y heat up X ' | (transitive) |

Traditional grammars of Indonesian (Moeliono and Dardjowidjojo 1988; Sneddon 1996, among others) typically simply list the uses of $-i$ without explicit argumentation as to whether there is one (polysemous) $-i$, or more than one (homonymous) - $i$. Our proposal addresses this issue; we claim that Indonesian $-i$ is polysemous, as further explicated in section 3.

The discussion of morphological applicativisation and causativisation in the literature has typically focused on clear cases with verbal/adjectival stems, i.e. stems like datang or panas as in (1). Such stems are argument-taking predicates, and causativisation/applicativisation can be clearly identified by checking argument alternations of the stems. However, the same causative/applicative affix may also take stems of other categories, and in those cases the analysis of the form as causative or applicative is not straightforward. The suffix $-i$ can be productively used with a noun stem, e.g. kantong- $i$ 'pocket- $i=$ put X in (own) pocket' and garam- $i$ 'salt- $i=$ put salt in X '. In these cases the noun stem is not understood as an argument-taking predicate, but rather as an argument (a location or theme) at some underlying semantic level. Again, this kind of derivation is often only mentioned in passing, and not given a precise analysis. Its significance is often overlooked as part of a wider family of transitivising processes that include applicativisation and causativisation. In section 3, we propose an argument-structure based analysis of $-i$ which can be easily extended to account for $-i$ verbs formed with noun stems.

[^22]From a computational point of view, we need to capture the syntactic, semantic and aspectual characteristics of $-i$ so that our computational grammar can produce correct parses of sentences headed by verbs with $-i$.

This paper proposes a novel unified a-structure based approach to verbal derivation, which can handle not only valence changing $-i$, giving rise to applicative-causative polysemy, but also valencepreserving $-i$ (discussed later in this paper). We analyse $-i$ as carrying its own PRED(ICATE) argument structure. Word-formation with $-i$ involves predicate composition of the PRED of the suffix with the PRED of its stem, similar to complex predicate formation as described in Alsina (1996) and Butt (1995). We adopt the LFG-based predicate composition approach of complex predicate formation (Alsina 1996, Butt 1995), and extend it to handle Indonesian data. The implementation makes use of the restriction operator (Kaplan and Wedekind 1993; Butt, King, and Maxwell 2003; Butt and King 2006). The implementation of the analysis is discussed in section 4.

Computational morphosyntactic treatments addressing the issues of applicativisation and related phenomena have not been previously proposed for Indonesian. Applicativisation (and its complex interaction with other kinds of word-formation such as voice selection and reduplication) has not been previously implemented in XLE either. The implementation of the LFG analysis of $-i$ in XLE therefore provides the first evidence for the robustness of XLE and LFG in handling this linguistically complex phenomena.

## 2 Indonesian verbal morphology and properties of -i

The Indonesian suffix- $i$, like - kan, has traditionally been described as a valence-increasing morpheme. The suffix $-i$ has been given less attention than its -kan counterpart, as it is often regarded as 'simpler' than -kan (Vamarasi 1999). However, as we shall show, its behaviour is equally complex, as we see in the following sections, which outline Indonesian verbal morphology and describe the basic properties of $-i$.

### 2.1 Verbal template in Indonesian

Indonesian verbs can be morphologically simple or complex. The verbal template in Indonesian consists of a root, possibly with one or more affixes: (Prefix*-)Root(-Suffix*). The outermost prefix is a voice-related prefix and the outermost suffix is typically a transitiviser suffix -kan/-i. Between the outermost affix and the root, there may be another affix, e.g., the causative prefix per- as in memperlihatkan 'show (<(lit.) 'cause $X$ to be seen'), or a loan suffix, e.g. -isasi as in memfungsionalisasikan 'functionalize'. Reduplication may also add complexity.

The verbal root can be free or bound. A free root such as datang 'come' and pergi 'go' can appear in its affixless form in syntax. A bound root, however, must be affixed to appear in syntax. The bound root often has a vague meaning and no clear grammatical category in isolation, which has led to the claim that it is 'precategorial' (Verhaar 1984). The root only gets a specific meaning and specific grammatical category when it is affixed; e.g., -alih- 'change position or course' $\rightarrow$ mengalih $(\mathrm{V})$ 'change to (a different position, topic, etc.), alihkan (V) 'distract, shift', peralihan ( N ) 'transfer, transition', and pengalihan $(\mathrm{N})$ 'diversion'.

### 2.2 Basic properties of - $\boldsymbol{i}$

The suffix $-i$ is a derivational suffix with the following properties. Firstly, $-i$ can be affixed to stems of different categories. The following table shows that $-i$ can appear with a noun, an adjective or a verb. The verb stem can be intransitive or transitive.

| Roots | Derived -i verbs | Roots | Derived -i verbs |
| :---: | :---: | :---: | :---: |
| air (N) 'water' | air-i 'water' | lompat 'jump' ( $\mathrm{V}_{\mathrm{ITR}}$ ) | lompat-i 'jump over' |
| kulit (N) 'skin’ | kulit-i 'peel' | tidur 'sleep' ( $\mathrm{V}_{\mathrm{ITR}}$ ) | tidur-i 'sleep on’ |
| gula (N) 'sugar' | gula-i 'put sugar in' | diam 'stay' ( $\mathrm{V}_{\mathrm{ITR}}$ ) | diam-i 'dwell in' |
| ketua '(N) chair (of an organization) | ketua-i 'chair or lead in a meeting/organization' | tulis 'write' ( $\mathrm{V}_{\text {TR }}$ ) | tulis-i 'write on something.' |


| panas (A) 'hot' | panas-I 'heat (water)' | kirim 'send' ( $\mathrm{V}_{\mathrm{TR}}$ ) | kirim-i 'send' (ditrns) |
| :---: | :---: | :---: | :---: |
| basah (A) 'wet' | basah-i 'dampen' | siram 'spray' $\left(\mathrm{V}_{\mathrm{TR}}\right)$ | siram-i 'spray with' |
| lengkap (A) <br> 'complete' | lengkap-i 'complete' | cium 'kiss' ( $\mathrm{V}_{\mathrm{TR}}$ ) | cium-i 'kiss repeatedly' |
| jauh 'far' (A) | jauh-i 'make oneself far from' | pegang 'hold' ( $\mathrm{V}_{\mathrm{TR}}$ ) | pegang-i 'hold tightly' |

Secondly, affixation with -i may or may not result in a change of valence of the stem. The examples in (2)-(4) show that $-i$ can increase the syntactic valence of the stem by promoting an oblique to Object:
(2)
$\begin{array}{lllll}\text { a. } \begin{array}{lll}\text { Ia } & \text { duduk } & \text { di }\end{array} \begin{array}{l}\text { kursi } \\ 3 \mathrm{~s} \\ \text { sit } \\ \text { LOC } \\ \text { chair }\end{array} & \text { itu } \\ \text { that } \\ \text { 'S/he sat on the chair' } & & \end{array}$
b. Ia menduduk-i kursi itu 3s AV.sit-i chair that 'S/he was sitting on the chair'
(3)
a. Ia melempar batu ke saya (go) (Vtr $\rightarrow \mathrm{Vtr})$
3 s AV-throw stone to 1 s
'S/he threw stones at me.'
a. Ia melempar-i saya dengan batu. 3s AV.throw-i 1s with stone 'S/he pelted me with stones.'
(4)
a. Ayah mengirim uang kepada $\{$ dia $\mid=n y a\}$ (go) (Vtr $\rightarrow$ Vdtr)
father AV.send money to $3 \mathrm{~s}=3 \mathrm{~s}$
'Father sent money to him/her.'
b. Ayah mengirim-i $\{$ dia $\mid=n y a\}$ uang
father AV.send-i $3 \mathrm{~s}=3 \mathrm{~s}$ money
'Father sent her/him money.'
The examples in (5) and (6), however, show no valence increase with $-i$. Rather -i merely adds some aspectual meaning (repetition, intensity).
(5)
a. Ia memukul saya
3 s AV.hit
'S/he hit me'
b. Ia memukul-i saya

3s AV.hit-i 1s
'S/he was hitting me'
(6) a. Ia memegang pencuri itu.

3s AV.hold thief that
'S/he held the thief.'
b. Ia memegang-i pencuri itu. 3s AV.held-i thief that 'S/he was holding the thief tightly.'

It is useful to compare the syntax and semantics of $-i$ to its $-k a n$ counterpart. As noted by Kaswanti (1995), Sneddon (1996) and Kroeger (2007), -i alternates with -kan to provide different possibilities for object linking, similar to the spray-load alternation in English (Levin 1993). Verbs affixed with $-i$ have a locative/goal object, whereas those with -kan have what Kroger (2007) calls a displaced theme/patient object:
(7)
a. Buruh itu memuat-kan beras ke kapal. worker that AV.load-kan rice to ship 'The workers loaded the rice onto the ship.'
b. Buruh itu memuat-i kapal dengan beras. worker that AV.load- $i$ ship with rice 'The workers loaded the ship with rice.'
a. Anak-anak menempel-kan poster ke tembok. child-REDUP AV.stick-kan poster to wall 'The children stuck a picture to the wall.'
b. Anak-anak menempeli tembok dengan poster child-REDUP AV.stick-i wall with poster 'The children stuck pictures all over the wall.'
a. Ia memuntah-kan darah segar.

3s AV.vomit-kan blood fresh 'S/he vomited fresh blood.'
b. Ia memuntah-i baju-nya.

3s AV.vomit-i shirt-3s
'S/he vomited on his/her shirt'
The following examples also show the different effects of $-i$ and $-k a n$ :
(10)
a. Polisi datang
police come
'The police arrived/came.'
b. Mereka men-datang-kan polisi

3p AV-come-kan police
'They arrived with the police' (comitative-applicative -kan)
'They called for/made the police come.' (causative -kan)
c. Mereka mendatang-i polisi (applicative-i)

3p AV.come-i police
'They came to/approached the police'
a. Dia men-jatuh-kan buku saya. (causative-kan)
3s AV-fall-kan book my
'S/he dropped my book.' (Lit: S/he made my book fall)
b. Dia men-jatuh-i buku saya. (applicative-i)
he AV-fall-i book my
'S/he fell on my book.'

In these examples, the (causative) locative $-i$ alternates with the causative displaced-theme $-k a n$.
(12)
a. Buku-nуa menumpuk / bertumpuk
book-3s AV.pile.up ber.pile.up
'His/her books piled up.'
b. Mereka menumpuk-kan buku di meja. (causative-kan)

3p AV.pile.up-kan book on table
'They piled up books on the table.'
c. Mereka menumpuk-i meja itu dengan buku. (causative-i)
$3 p$ AV.pile.up-i table that with book
'They were piling up books on the table.'
(13) (from (Kroeger 2007); adapted from Dardjowidjojo 1971)
a. Air itu sedang meng-alir ke sawah. water that PROG AV.flow to rice.field 'The water is flowing to the rice field.'
b. Dia meng-alir-kan air itu ke sawah=nya. 3s AV.flow-kan water that to rice.field=3sg 'S/he caused the water to flow to his/her rice field.'
c. Dia meng-alir-i sawah=nya dengan air itu. 3s AV.flow- $i$ rice.field $=3 \mathrm{sg}$ with water that 'S/he flooded his/her rice field with the water.'

The following shows that causative-permissive $-i$ alternates with benefactive $-k a n$. That is, $-i$ encodes the object as borrower with the $\mathrm{A} /$ subject understood as the source/owner of the thing borrowed. With -kan, the A is the borrower and $\mathrm{s} / \mathrm{he}$ borrows it from someone else for the benefit of the U saya.

| a. Ia meminjam- $i$ | saya | uang. |
| :--- | :--- | :--- | :--- |
| $3 \mathrm{~s} \quad$ AV.borrow-i | 1 s | money |
| 'S/he lent me money.' |  |  |

b. Ia meminjam-kan saya uang. (saya=benefactive; source is not the actor) 3s AV.borrow-kan 1s money
'S/he borrowed money for me.'
When the root is a noun, the root is often understood as a '(displaced) theme' associated with the location designated by the object (in a real or metaphorical sense). The location can be understood as part of a 'static' situation as in (15a), or the goal or source of action as in (15b-d). The stem itself can be also understood as the location as in (15e).
a. Sungai ini membatas-i Malang dan Lumajang (root = theme)
river this AV.border-i Malang and Lumajang
'This river becomes the border of Malang and Lumajang.'
b. Mereka mengair-i sawah-nya. 3p AV.water-i rice.field-3s
'They were watering their rice-field.'
c. Dia mengulit-i pisang itu.

3s AV.skin-i banana that
'S/he peeled the bananas (Lit. removed the skin from the bananas)
d. Dia mengulit-i buku itu.

3s AV.skin-i book that
'S/he added a cover to the book.' (adding cover to the book)
$\begin{array}{lllll}\text { e. Pihak } & \text { China } & \text { hendak } & \text { memenjara-i } & \text { Hồ Chi Minh... } \\ \text { side } & \text { Chinese } & \text { want } & \text { AV.prison-i } & \text { Hồ Chí Minh }\end{array}$
'The Chinese wanted to imprison Hồ Chí Minh.'
Though -i may show both applicative and causative functions, some roots do not allow both functions. The following patterns are observed. Firstly, there are roots which allow only the applicative function, as exemplified with datang 'come' $\rightarrow$ datangi (10c) and jatuh 'fall' $\rightarrow$ jatuhi (10b). The verb datang- $i$ can only mean 'come to $\mathrm{X}[\mathrm{loc}]$ ', not *'make X come'. To derive a causative meaning with these verbs, the suffix -kan is used, datang-kan 'make X come' (10b), jatuhkan (11a). Secondly, denominal $-i$ verbs with noun roots conceptualised as displaced themes typically have only an applicative function. The Object is understood as a location; e.g., kutu-i'louse-i= delouse X ', gula-i 'put sugar in X', etc. The -kan form is typically not attested; e.g., *kutu-kan, *gula-kan. Finally, there are roots that allow both applicative and causative functions, such as the root takut 'afraid', as in (16a)
for the applicative $-i$ reading and the causative $-i$ in (16b). Unlike the second pattern, $-k a n$ is also often commonly used for this third type of root.

a. | Sebut | 1 | hewan yang |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| mention | one | animal | REL | 2 |

'Name one animal that you are scared of'
'if there is no one/thing that makes them afraid'
To sum up, the suffix - $i$ can take roots of different categories with applicative and/or causative functions. The analysis of the precise nature of these functions is outlined in the next section.

## 3 Analysis

### 3.1 Patterns of alternations

The following properties must be captured by the analysis of $-i$ :
a. Syntax
i. valence increasing:
intransitive $\rightarrow$ transitive (examples (2), (10)), monotransitive $\rightarrow$ ditransitive (example (4))
ii. no valence changing effect (examples (5)-(6))
b. Semantics
i. Locative applicative and causative
ii. Iterative/intensifying/progressive
c. Related to (a) and (b), alternations with -kan.

The first question is whether we have one $-i$ or more than one $-i$. We argue that we have one $-i$, and that the different properties as summarized in (17) are predictable from the interaction of the core information carried by $-i$ and the information carried by the stem. We begin with a characterisation of the facts in theoretical terms, appealing to a Jackendoff-style semantic structure (Jackendoff 1990) to express our analysis. This forms the theoretical underpinning of our computational analysis, although our implementation is different and simpler than the analysis we present here in that it does not appeal to a separate level of argument structure.

Central to the semantics of $-i$ is its locative meaning component in the state of affairs (SOA) it encodes. This can be informally represented as in (18). ${ }^{3}$

## (18)

A AFFECT U ${ }_{i}(\{T O \mid F R O M\})$ BE.AT([LOC] $\left.]_{i}\right)$
AFFECT (henceforth, AFF; see Jackendoff 1990) is a general semantic primitive which is intended to cover different degrees of affectedness associated with causativisation and applicativisation. The exact interpretation of AFFECT is determined by the semantics of the stem. ${ }^{4}$

The locative meaning which is always added by $-i$ can in certain cases be thematically interpreted as part of PATH (i.e. TO or FROM in (18)). Thus it can sometimes be interpreted as 'goal' or 'source', depending on the meaning of the stem and world knowledge. For example, menguliti ( $<$ 'AV.skin.APPL') can mean 'remove the skin FROM' for an object understood to have skin to begin

[^23]with (e.g. a crocodile). Alternatively, it can mean 'to put skin ON': e.g. for 'books' because books are created with skin (i.e. cover) added in the production process; see 3.4.4.

The aspectual meaning - iterative, progressive, intensifying; see examples (5)-(6) - is arguably also related to the locative meaning of $-i$. That is, the locative $U$ is conceptually understood as having a spatial surface to which the action is applied. Affectedness applied to an unbounded space leads to a repetitive or progressive meaning, e.g. mengecat X 'paint X ' $\rightarrow$ mengecati X 'paint all over the surface of $X$ '. Other cases in which the same marking is used for locative alternation and aspectual distinctions such as telicity have been noted in the literature (Levin and Rappaport Hovav 1998).

The representation of the semantics of $-i$ as shown in (18) is sufficient for the purposes of this paper. The locative and aspectual meaning of the $-i$ verb, which is tied to OBJ, will be simply represented by its realisation as OBJ linked to a locative role.

## $3.2-i$ as a head PRED verbalizer

In line with work on causativisation and applicativisation (Alsina and Joshi 1991; Alsina 1992; Butt 1995; Austin 2005 [1996]), we propose an a-structure-based analysis with the following key points.

First, we claim that there is one $-i$ (i.e., a polysemous suffix). We analyse $-i$ as a three-place predicate with its own argument structure, as shown in (19). Affixation with $-i$ involves complex predicate composition, with argument fusion of the matrix and embedded arguments. Central to the analysis of $-i$ is that the matrix's second argument (ARG2) is thematically a locative (LOC)-related argument (i.e., possibly Goal or Source, in addition to Locative). ARG2 fuses with the LOC argument of the base wherever possible. ARG1 is thematically higher than ARG2, though not necessarily an agent. This representation, as we shall see, allow us to capture both causative and applicative uses of $-i$, as well as other uses.

## A-structure of $-i$ and the associated semantic roles

$$
\begin{equation*}
' \mathrm{PRED}_{1}<\mathrm{ARG}_{1}, \mathrm{ARG}_{2}, \quad \mathrm{PRED}_{2}<\ldots, \ldots \gg ' \tag{19}
\end{equation*}
$$

(A) (U: $\left.:_{\text {LOC }}\right) \quad$ where $\overline{\operatorname{argument}}(\mathrm{s})$ of $\mathrm{PRED}_{1}$ fuse(s) with arguments of $\mathrm{PRED}_{2}$
The second key point of our analysis is underspecified fusion. While the ARG2 of $-i$ is thematically specified as LOC-related, the overall fusion is underspecified, constrained by the semantics of the root, possibly with lexicalisation for certain verbs. For example, the verb baui (<bau 'odour'-i' (lit.) cause the odour of X to be in Y ') is lexicalised to mean ' Y smells X '.

We argue that this underspecified a-structure allows two different types of argument fusion, single or double, which then gives rise to different applicative and/or causative effects. We propose a general rule for $-i$ composition: arguments of thematically similar types tend to fuse. Thus, the Actorlike ARG1 of the matrix PRED tends to fuse with the actor-like ARG1 of the embedded PRED. Likewise, the Undergoer-like ARG2 of the matrix PRED fuses with the Undergoer-like ARG2 of the embedded PRED. However, since PRED2 may be a one-place predicate, ARG2 of the matrix PRED may fuse with the sole argument (i.e. ARG1) of the embedded PRED. Different possibilities are further discussed and illustrated in 3.4 below.

The third key point is that the derived a-structure is constrained by a set of a-structure wellformedness properties: core arguments outrank non-core arguments, and within these groups arguments are ordered thematically (Manning 1996; Arka 2003). This constraint determines the derived syntactic transitivity, and also possible $-i$ and -kan alternations, as we show in the next subsection.

Due to space limitations, we cannot outline the full details of the linking mechanism in this paper. We adopt a version of a-structure-based linking as described in Arka (2003:148-158), which is applicable to Indonesian (Arka and Manning 2008). An argument in the a-structure is represented as ARG or simply a "_" within angle brackets. If necessary, core status is indicated by nested bracketed groupings, with core arguments on the left group, and associated thematic roles placed within brackets underneath; see (21a) below. 5 Voice alternations may or may not alter argument structure. Actor Voice (AV) maps the most prominent core argument (ARG1 for $-i$ ) to SUBJ and the second core

[^24]argument (ARG2) to OBJ whereas the Undergoer Voice (UV) has a reverse mapping. Passive Voice alters the a-structure associated with the Actor. This is further discussed in 4.1.

### 3.3 Transitiviser: -i and -kan alternation

Our proposal that $-i$ introduces two core arguments (ARG1, ARG2) as shown in (19) is fully consistent with the traditional analysis of $-i$ as a transitiviser. Moreover, given our theory of a-structure (core $>$ non-core; matrix > embedded; thematically-based ordering within core/non-core), we also predict the valency of the resulting structure: this is determined by the thematic role of the argument of the embedded predicate that is not fused with a matrix argument (ARG1 or ARG2). The chart in (20) gives the subcategorisation frames resulting from combining $-i$ and $-k a n$ with predicates with different inventories of thematic roles, where one argument of the predicate remains unfused. The unfused argument is labelled ARG3 in the linking configuration of $-i$ and $-k a n$ in (20). It is the least prominent argument in the derived structure, and it can be either core (OBJ2) (20a) or non-core (OBL) (20b).
(20) Subcategorisation frames and associated thematic roles for verbs with $-i$ and $-k a n$
(a)

| Ditransitive |  |  |
| :--- | :--- | :--- |
| NP $_{\text {SUBJ }}$ | $\mathrm{NP}_{\text {OBJ }}$ | $\mathrm{NP}_{\text {OBJ2 }}$ |
| ARG1 | ARG2 | ARG3 |
| agt | go/src/loc | th |
| agt | ben/rec | th |

(b)

| Monotransitive |  |  |
| :--- | :--- | :--- |
| NP $_{\text {SUBJ }}$ | $\mathrm{NP}_{\text {OBJ }}$ | $\mathrm{PP}_{\text {OBL }}$ |
| ARG1 | ARG2 | ARG3 |
| ag | go/src/loc | instr |
| ag | $\mathrm{pt} / \mathrm{th}$ | go/src/loc |

The patterns in (20) are expected because the two matrix arguments, ARG1 and ARG2, are linked to the two arguments that are highest on the grammatical function hierarchy, namely SUBJ and OBJ respectively. The actor-like argument (e.g. Agent) is linked to ARG1/SUBJ by default. Each of the two suffixes $-i$ and $-k a n$ locks the linking of ARG2 to a certain role: the suffix $-i$ links ARG2 to a locative-related role (including goal/source), whereas -kan links it to beneficiary/recipient or theme.

The unfused argument (ARG3) is therefore the third in the list. When ARG3 is a displaced entity (i.e. a Theme) as in (20a), it is realised as OBJ2 in both $-i$ and $-k a n$ verbs. A theme is ranked low in the thematic hierarchy (Bresnan and Kanerva 1989; Bresnan and Moshi 1990; Simpson 1991), and fits well as the least prominent core argument among the three core arguments that make up a ditransitive structure.

However, when the unfused argument (ARG3) is not a theme, it is realised as an oblique and the structure is monotransitive, as in (20b). Non-theme arguments (instrument, goal, source, locative) cannot naturally appear as (second) objects (i.e. third least-prominent core arguments) in Indonesian. This constraint can be taken as reflecting the cross-linguistically common generalization that a locative role is intrinsically classified as a non-objective function (Bresnan and Kanerva 1989). The instrument role can be analysed as a causer-like argument which is also classified as a non-objective function.

The proposed analysis, capturing similarities and differences between $-i$ and $-k a n$ as depicted in (20), accounts for the fact that $-k a n$ and $-i$ can alternate with certain stems. A key aspect of the analysis is the idea that $-i$ and $-k a n$ are both matrix predicates with their own a-structure, including matrix ARG1 and ARG2 (cf. (19)), and that they lock the linking of ARG2. The suffix -i links ARG2 to a Goal/Loc-related role, whereas -kan links it to a Benefactive/Recipient or theme. The resulting (ditransitive or monotransitive) structures follow from independent principles. For instance, kirim 'send' is a monotransitive verb stem with the a-structure shown in (21a). This is precisely what -kan also specifies; i.e. ARG2 is a theme (see (20b)). Hence, it is not surprising that $-k a n$ is optionally present in this type of structure and does not change the basic argument linking, indicated by $-k a n$ in brackets in (21b).

[^25]b. Ayah mengirim(-kan) \begin{tabular}{l}
uang <br>
father AV.send(-kan) <br>
'Father sent money to me.'

 

kepada <br>
to
\end{tabular}

to

Since kirim 'send' also has a displaced theme and a Goal argument, it also allows the ditransitive structure with $-i$, as in (20a). This is shown in (22):

| a. Ayah | mengirim- $i$ | saya | uang |
| :---: | :--- | :---: | :--- |
| father | AV.send-i | 1s | money | 'Father sent me money.'

b. Ada dua pihak yang rutin mengirim-i saya SMS exist two side REL regular AV.send-i 1s short.message
'There are two groups of people who regularly send me short messages.'

Note that in this case the Goal is a human being, saya. It is therefore also interpreted as a Beneficiary/Recipient. For this reason, the ditransitive beneficiary structure (i.e., that of -kan, (20) a) is also expected to be available with this verb. This is correct, and in fact attested with high frequency on the Internet; a naturally-occurring example is shown in (23) ${ }^{6}$.

| Jika | anda | bermaksud | mengirim-kan | saya | satu | pesan, |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| if | 2 | intend | AV.send-kan | 1 s | one | message |
| Klik | "Contact" $d i$ | atas |  |  |  |  |
| click contact | at | above |  |  |  |  |
| 'If you want to send me a message, click "Contact" above' |  |  |  |  |  |  |

The alternation of $-i$ and $-k a n$ (mengirimi vs. mengirimkan) in some ditransitive structures exhibits aspectual differences. The suffix $-i$ signals progressive, iterative, or plural events. Thus, mengirimi in (22) implies that the act of sending took place on numerous occasions. In (22b), it is also signalled by the adverbial rutin 'regularly'. In contrast, -kan is typically associated with one-off events. In (23), the object/ARG2 is explicitly coded with satu 'one'. Replacing mengirimkan with mengirimi in this example would mean sending the same message over and over again to the same recipient.

### 3.4 Types of fusion

In our analysis, different results of combining predicates with -i (and also -kan) result from different possibilities of argument fusion: double or single fusion.

A derivation with $-i$ typically involves double fusion. That is, both matrix arguments (ARG1 and ARG2) are fused with arguments of the embedded predicate. Double fusion may result in applicativisation, in which an embedded non-core argument is promoted to core status due to its fusion with ARG2, or simply additional meaning without valence-changing effect.

In single fusion, only one of the matrix arguments is fused with an argument of the embedded predicate. Single fusion typically results in causativisation.

In the following, we discuss and exemplify different types of derived $-i$ structures.

### 3.4.1 Type 1

Type 1 involves derived monotransitive $-i$ verbs undergoing a valence-changing applicativisation effect. With a two-place intransitive base (with Goal/Locative second argument) such as jatuh 'fell (on)to X', datang 'come to X', and lewat 'pass at X', the result is a strictly monotransitive $-i$ verb. ${ }^{7}$ This is exemplified by (24a-b), which can be represented as (24c).

[^26]| a. Mangga lang besar jatuh | ke |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- |
| mango | REL | big | fall | to | house-3s |

'A big mango fell onto his house.'
b. Mangga yang besar men-jatuh-i rumah-nya (*menjatuhkan)
mango REL big AV-fall-i house-3s
'A big mango fell onto his house.'


The same promotion effect is also observed for certain two-place bound verbal roots, e.g. -kunjung'visit (to) X ' and -saing- 'compete (with) X '. ${ }^{8}$ These roots are 'bound' because they must be affixed in order to appear as verbs in a clause. The second argument (ARG2) is realised either as an OBL (noncore) or OBJ (core) depending on the affix. For example, affixed with the intransitive ber-, the verb berkunjung comes with its Goal argument as an OBL marked by ke (25a). Affixed with $-i$, the Goal is OBJ in (25b), exhibiting the applicativisation effect. The derived a-structure of kunjungi is shown in (25c).
(25)


### 3.4.2 Type 2

This type is associated with three-place predicates with a displaced theme such as kirim 'send'. There are two sub-types (Type 2a and Type 2b), depending on the core status of the displaced theme.

In Type 2a, the displaced theme is unfused and realised as OBJ2 (cf., (20a)). The example given earlier in (22) is mengirimi (AV.kirim-i), formed out of a free root, kirim 'send'. A bound root can also appear in this pattern; e.g., serah- 'transfer', sodor- 'offer', and suguh- 'serve'. For this type of root, the affix marking transitivity ( $-i$ or $-k a n$ ) is obligatory. Consider the AV verb menyuguhi in (26a). Note that for this root, there is no alternative monotransitive structure with the verb stem without a transitiviser, as seen by the unacceptability of menyuguh (26b) where the displaced theme is ARG2OBJ. If this structure is intended, then -kan must be used (26c). The a-structures for (26a) and (26c) are given in (26a') and (26c') respectively. The alternation of $-i(26 a)$ and $-k a n(26 c)$ is expected for

[^27]reasons discussed in 3.3. That is, given the subcategorisation frame of $-i$ shown in (20), the (unfused) theme can only be mapped onto OBJ2, because the other two arguments are fused with ARG1 and ARG2, realised as SUBJ and OBJ respectively.

| a. Engkau menyuguh-i | aku | minuman lezat |  |  |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 2s | AV.serve-i | 1 s | drink | tasty | FOR 'You served a very tasty drink to me.'

c. Engkau menyuguh-kan minuman lezat kepada aku 2s AV.serve-kan drink tasty to 1s 'You served a very tasty drink to me.'


Type 2 b is when the displaced theme of the underlying three-place predicate is non-core and hence surfaces as OBL. This is the type shown in (20b). The verb menyuguhi can be ditransitive (26a), or monotransitive ( 26 c ). Again, for reasons discussed in 3.3, these alternative structures are expected in our a-structure based analysis. The theme shows up with a prepositional instrumental marker dengan, as seen in example (27a) with the bound root -suguh- . More examples of this pattern, taken from the internet, are given in $(27 \mathrm{~b}-\mathrm{c})^{9}$.
$\begin{array}{llllll}\text { a. Engkau } & \text { menyuguh- } i & \text { aku } & \text { dengan } & \text { minuman } & \text { lezat } \\ 2 \mathrm{~s} & \text { AV.serve- } \mathrm{i} & 1 \mathrm{~s} & \text { with } & \text { drink } & \text { tasty }\end{array}$ 'You served a very tasty drink to me'
b. ... menyodor-i Juno dengan pertanyaan tentang nasib sang bayi.. AV.offer-i Juno with question about fate ART baby - ...asked Juno questions about the fate of the baby ...
$\begin{array}{lllllll}\text { c. Dia menyerah-i saya dengan tugas } & \text { yang } & \text { berat } \\ 3 \mathrm{~s} & \text { AV.give-i } & 1 \mathrm{~s} & \text { with } & \text { duty } & \text { REL } & \text { heavy }\end{array}$ 'He burdened me with heavy duties'

Our proposed analysis predicts that a sentence like (28) is unacceptable. This sentence is headed by $-i$ but has an a-structure with ARG2/OBJ linked to a non-LOC argument, in violation of the restrictions imposed by $-i$.

* Engkau menyuguhi minuman lezat aku (ARG2=theme)
2s $\quad$ AV.serve-i drink tasty 1s
(FOR: 'You served me with a very tasty drink')

[^28]
### 3.4.3 Type 3

Type 3 shows no valence change in $-i$ derivation, exemplified by pukul 'hit' $\rightarrow$ pukuli (29a-b). Verbs that are of the same type as 'hit' can be represented as having an a-structure shown in (29a'), involving a 'hitter' (agent) and a 'hittee' understood as the patient/target (pt/go). The hittee, saya in (29), is the affected participant (i.e., patient) that is also the target of (i.e., in contact with) an understood movable instrument. The second argument of the stem meets the specification for $-i$ in that it is a goal, and hence can fuse with the matrix ARG2. Double fusion of memukuli, as seen in (29b'), produces no valence change effect in the derivation.

```
\(\begin{array}{lll}\text { a. Ia memukul } & \text { saya } \\ 3 \mathrm{~s} & \text { AV.hit } & 1 \mathrm{~s}\end{array}\)
    'S/he hit me
b. Ia memukul-i saya
    3s AV.hit-i 1s
    'S/he was hitting me, \(\mathrm{s} /\) he hit me repeatedly'
a.'
pukul 'hit <_, _>'
        \((\mathrm{ag})(\mathrm{pt} / \mathrm{go})\)
b' '3s' '1s'
    \(\left.{ }_{-i}^{\text {SUBJ OBJ }} \quad\right|_{-} \mid\)
        (U:go) (ag) (pt/go)
```

While not involving a change in transitivity, - $i$ affixation does bring about a change in aspect; it indicates that the event is in progress or completed with iterative meaning as seen in the translation of (29b). This aspectual property of $-i$ is not surprising on the analysis that ARG2/OBJ is semantically linked to a locative-goal role. That is, the $-i$ verb leads to the interpretation that the action is applied to a surface of an object. For inherently punctual verbs like 'hit', actions affecting the surface of an object are given a repetitive interpretation. For other verbs where OBJ measures event completion (Tenny 1992, 1994), e.g., bunuh 'kill' (where the object being dead measures the event), $-i$ also gives rise to pluralisation or individuation of the object. For example, the object of membunuh in (30a) is by default singular, though plural is possible. However, the object of membunuhi in (30b) must be plural (i.e., reading i). For mass noun objects, $-i$ gives rise to individuation; e.g., tanah 'land' sold in (31) is understood as 'fractions' of the land.

| a. Dia membunuh binatang | itu |
| :--- | :--- |
| 3s AV.kill animal | that |
| 'S/he killed the animal(s)' |  |

b. Dia membunuhi binatang itu.

3s AV.kill-I animal that
i) 's/he killed the animals one by one'
ii) * 's/he killed the animal'

> Ia menjual-i tanah orang tua-nya
> 3s AV.sell-i land person old-3sPOSS
> 'He sold his parents' land, bit by bit.'

### 3.4.4 Type 4

This is a single fusion type where ARG2, like the other types discussed earlier, fuses with a LOC argument of the base wherever possible. However, ARG1 is newly introduced in Type 4. This $-i$ affixation then results in causativisation. Consider the bound root -alir- 'flow' in (32). It appears as an intransitive verb in (32a) with the locative/goal sawah as an oblique. The derived $-i$ verb can be monotransitive (32b) (with the displaced theme showing up as an oblique instrument marked by dengan) or ditransitive (32c) (with the displaced theme being OBJ2). These two $-i$ structures involve
single fusion as depicted in (32d), the only difference being the realisations of the unfused embedded displaced theme. ${ }^{10}$
a. Air itu sedang meng-alir ke sawah. SUBJ OBJ water that PROG AV-flow to rice.field 'flow $<$, , $>$ '
'The water is flowing to the rice field, 'The water is flowing to the rice field.'

$$
\begin{equation*}
\left.(\overline{\mathrm{th}}) \overline{(l o c}^{(\mathrm{loc}}\right) \tag{32}
\end{equation*}
$$

b. Dia meng-alir-i sawah=nya dengan air itu. 3s AV-flow-i rice.field=3sg with water that 'S/he flooded his/her rice field with the water.'
c. Dia meng-alir-i sawah=nya air itu. 3s AV-flow- $i$ rice.field=3sg water that 'S/he flooded his/her rice field with the water.'
d.


Denominal $-i$ verbs with noun roots such as sinar 'light', air 'water' and kutu 'louse' are common examples of Type $4-i$ verbs. The referent of the stem is understood as a displaced theme. The issue here that the embedded predicate of $-i$, i.e., the predicate that assigns 'displaced theme' to the stem, is not overtly realised. We posit an unexpressed predicate 'be.at' as part of the predicate astructure of the verb. For example, the verb mengatapi (33a) can be interpreted as depicting an event where atap 'roof' is the displaced theme (i.e., placed on the house).
a. Mereka mengatap-i rumahnya

3p AV.roof-i house-3s
'They roofed the house.'
b. '3p' 'house'


Adjective roots (e.g., sakit 'sick', panas 'hot', and kotor 'dirty') can also derive $-i$ verbs with causative meaning. This is exemplified in (34a). The fusion of the theme-locative argument shown in (34b) captures the meaning that jalan 'road' is understood as the surface of the road.
a. Jangan kotor-i jalan itu

NEG dirty-i road that
'Don't (you) make (the surface of) the road dirty.'
b.


[^29]
### 3.4.5 More than one type

A derived $-i$ verb may be of more than one fusion type. This is exemplified by the $-i$ verb derived from the root tumpuk 'pile up'. First consider the intransitive verb bertumpuk (35), in which the locative shows up as an oblique marked by $d i$ :

| Buku-nya ber-tumpuk di meja. |
| :--- |
| book-DEF ber-pile.up on |
| 'The books piled up on the table.' |

The $-i$ verb menumpuki can appear in three structures. It can be a three-place monotransitive verb, with the displaced theme appearing as an instrument. This is a derivation of Type 4 with single fusion, exemplified in (36a), whose fusion is represented in (36b).
a. Mereka menumpuk-i meja itu dengan buku. $3 p$ AV.pile.up-I table that with book 'They piled the table with books.'

Menumpuki can also involve double fusion, with two variants. The first variant is a two-place monotransitive transitive verb, exemplified in (37a). ARG1 and ARG2 fuse with the embedded displaced theme and locative arguments respectively.
a. Buku itu menumpuk-i meja itu book that AV.pile.up-i table that 'The books piled up on the table.'


The second variant is just like the first variant, the difference being that double fusion gives rise to a reflexive reading (38).That is, ARG1/SUBJ instantiated by a human-denoting argument mereka 'they' invokes a volitional/agent reading, producing the meaning 'cause themselves to pile up'.
a. Mereka menumpuk-i meja itu
3p AV.pile.up-i table book
'They piled themselves up on the table.'
d.'

(double fusion)

## 4 Implementation

We implement the analysis using XLE, ${ }^{11}$ an LFG-based grammar development environment for large-scale grammars. Applicativisation as predicate composition has not been implemented in XLE before. Our analysis and implementation serve as evidence that (classic) LFG can adequately handle all types of $-i$, including both applicative and causative properties. It can robustly deal with derivation with or without valence increase, as well as with noun stems with no clear a-structure.

The relevant components of the system include a tokenizer and morphological analyser, as well as phrase structure and sublexical rules, which include annotation with relevant constraints as defined in templates for reusability. We omit description of the morphological analyser here; see Pisceldo et al. (2008), and Mistica et al. (2009) for full description and discussion. For present purposes, the relevant function of the morphological analyser is to decompose the words of a sentence into a stem plus morphological tags, often representing morphemes, which are analysed by sublexical phrase structure rules which are annotated in the same way as standard LFG phrase structure rules. For example, the morphological analyser breaks the verb menduduki into the four component parts: AV+ (Actor Voice), the stem $d u d u k,+\mathrm{I}$ (the $-i$ suffix), and the part of speech tag + Verb. A sublexical phrase structure rule combines these components into a lexical V constituent.

### 4.1 Annotated phrase structure and sub-lexical rules

Our syntactic rules for Indonesian consist of phrase structure rules (c-str) and sublexical rules. The relevant (partial, somewhat simplified) c-str rules needed for $-i$ verbs are shown in (39)-(40). The c-structure rules regulate clausal structure, and the sublexical rules are used to analyse the upper-side output of the morphological analyser, and hence regulate word-internal hierarchical structures. They are annotated with grammatical functions and templates. The templates for $-i$ are those regulating VOICE and predicate composition (APPL and CAUS). Template calls are indicated by @ in the rules, as is standard.

## Clausal c-str rules:

a. $\quad \mathrm{S} \rightarrow \mathrm{NP}$ VP
b. $\mathrm{VP} \rightarrow \mathrm{V}^{\prime} \mathrm{PP}$
c. $V^{\prime} \rightarrow(N P) V(N P)$
(40) Sublexical rules:


It should be noted that the sublexical structure is not flat, as is often the case in XLE grammars. This is linguistically motivated by word formation patterns in Indonesian, and is also practically useful, e.g. the rule for V_STEM' (used in the analysis of the suffix - $i$ (with the tag +I ; see (44e))) is specified only once for all types of voice selections. Following the notation and structure in XLE, affix positions are represented in (40) as V_VOICE_BASE for the verbal prefix (e.g., $m e N-$ ) and V_I_BASE for the applicative/causative suffix - $i$. The V_SFX_BASE is also added to provide a slot for verbal category information. These sublexical positions are filled in by the tags of the relevant affixes; e.g., tag AV+ for the AV prefix meN- and +I for the causative/applicative $-i$ (see (44) sample entries with tags).

The template for Voice, given in (41), specifies three voice types in Indonesian: ACTORVOICE, UNDERGOER-VOICE, and PASSIVE-VOICE. Each comes with a template (not shown here) and imposes changes in the input subcategorisation frame, formulated here as classic lexical rules in LFG (Bresnan 1982). For example, PASSIVE-VOICE replaces OBJ by SUBJ, and specifies three

[^30]options for how the underlying SUBJ is realised ${ }^{13}$ (suppressed as NULL, OBL, or OBJ). Note that ACTOR-VOICE does not change the subcategorisation frame of the base, so if the verb is recognised as AV (prefixed with $m e N$-), the grammar just passes back the subcategorisation frame as is.
(41) Templates for Voice and Voice types in Indonesian
\[

$$
\begin{gathered}
\text { VOICE }(\text { SCHEMATA })= \\
\{\text { _SCHEMATA } \\
@ \text { ACTOR-VOICE } \\
\text { | _SCHEMATA } \\
@ \text { UNDERGOER-VOICE } \\
(\uparrow \text { OBJ }) \rightarrow(\uparrow \text { SUBJ }) \\
(\uparrow \text { SUBJ }) \rightarrow(\uparrow \text { OBJ }) \\
\text { _SCHEMATA } \\
@ \text { PASSIVE-VOICE } \\
(\uparrow \text { OBJ }) \rightarrow(\uparrow \text { SUBJ }) \\
\{\text { SUBJ }) \rightarrow \text { NULL } \\
\mid(\uparrow \text { SUBJ }) \rightarrow(\uparrow \text { OBL }) \\
\mid(\uparrow \text { SUBJ }) \rightarrow(\uparrow \text { OBJ })\}\} .
\end{gathered}
$$
\]

The template in (42) encodes the contribution of the applicative $-i$ affix:
(42) $\quad$ APPL_I $=$

$$
\begin{aligned}
& \left\{(\uparrow \text { PRED })=\text { 'V_Appl_i }<(\uparrow \text { SUBJ })\left(\uparrow \text { OBJ) \%PRED3> }{ }^{\prime}\right.\right. \\
& \uparrow \backslash \text { PRED } \backslash \text { GF }=\downarrow \text { PRED } \backslash \text { GF } \\
& \{(\downarrow \text { SUBJ })=(\uparrow \text { SUBJ }) \\
& \text { Type 1: IntrRoot } \rightarrow \text { Vtr } \\
& (\downarrow \text { OBL-LOC })=(\uparrow \text { OBJ }) \\
& (\downarrow \text { SUBJ })=(\uparrow \text { SUBJ }) \\
& (\downarrow \text { OBL-LOC })=(\uparrow \text { OBJ }) \\
& (\downarrow \mathrm{OBJ})=(\uparrow \text { OBL-INST }) \\
& (\uparrow \text { OBL-INST CASE })=\mathrm{c} \text { obl-inst } \mid \\
& (\downarrow \text { SUBJ })=(\uparrow \text { SUBJ }) \\
& (\downarrow \mathrm{OBJ})=(\uparrow \mathrm{OBJ}) \\
& \text { Type 3: TrRoot } \rightarrow \text { Vtr } \\
& \text { ( } \uparrow \text { TNS-ASP PROG) }=+ \\
& \sim(\uparrow \text { OBL-INST) "just for the iterative meaning of }-i \text { " }\} \\
& (\downarrow \text { PRED })=(\uparrow \text { PRED ARG3) } \\
& (\uparrow \text { PRED })=\text { 'V_Appl_i }<(\uparrow \text { SUBJ })(\uparrow \text { OBJ) }(\uparrow \text { OBJ2) \%PRED4 }>\text { ' } \\
& \uparrow \backslash \mathrm{PRED} \backslash \mathrm{GF}=\downarrow \backslash \mathrm{PRED} \backslash \mathrm{GF} \\
& (\downarrow \text { SUBJ })=(\uparrow \mathrm{OBJ}) \\
& (\downarrow \text { OBL-LOC })=(\uparrow \text { OBJ }) \\
& \text { Type 4: IntrRoot } \rightarrow \text { Vtr } \\
& (\downarrow \mathrm{OBJ})=(\uparrow \mathrm{OBJ} 2) \\
& (\downarrow \text { PRED })=(\uparrow \text { PRED ARG4) }\} \\
& (\uparrow \text { APPLICATIVE })=+ \text {. }
\end{aligned}
$$

This template implements the analysis discussed in section 3.2, using the restriction operator (Butt, King, and Maxwell 2003; Butt and King 2006) . The implementation is somewhat constrained by the current setup of XLE. Given that subcategorisation/a-structure frames are represented by GFs in XLE, rather than having a separate argument structure representation encoding thematic roles, ARG1 and ARG2 in our earlier representation in section 3 correspond to SUBJ and OBJ respectively in (42). Additionally, rather than using underspecification, each type of combination of $-i$ with a predicate is treated separately by a set of equations. \%PRED represents the base PRED, which is treated as the rightmost ARG of the V_Appl_I predicate; hence \%PRED3 or \%PRED4. When $-i$ appears as an affix

[^31]to a verb, this results in predicate composition. For example, simplifying the sublexical tree somewhat, we show the structure of the verb duduki 'sit.APPL' in (43). The effect of predicate composition is shown in the box. In addition to supplying the information that the $-i$ verb is applicative (i.e. ( $\uparrow$ APPLICATIVE $)=+$ ), the template specification $(\downarrow$ PRED $)=(\uparrow$ PRED ARG3) in the template results in the embedded PRED being the third (subordinate) argument of the applicative verb. The equation ( $\downarrow$ SUBJ $)=(\uparrow$ SUBJ $)$ of the template captures the fusion of the subordinate argument with the matrix argument, which is indicated by a connecting line in the box.


duduk
$(\uparrow$ PRED $)=$ 'duduk $<(\uparrow$ SUBJ $)>$ '

As shown in (43), the restrictions associated with $-i$ (the template shown in (42)) appear in the sublexical c-structure rule introducing $-i$ as seen in (43). That achieves the desired effect: when $-i$ is affixed, the composition affects the stem's a-structure specification. It should be noted that the template is actually specified as @(VOICE @(APPL_I VApp_I) in the sublexical rule shown in (40b). The internal bracketing of @(VOICE @(APPL_I VApp_I)) determines how applicativisation interacts with voice selection: Predicates are first composed by applicativisation, then a particular VOICE-
TYPE is selected.

### 4.2 Tags and lexical entries

Apart from the annotated rules just described, at the heart of the grammar is the lexical entries. Information specified in the lexical entries is of different kinds. This includes information about the grammatical category ( $\mathrm{N}, \mathrm{V}$, Affix, etc.) needed by c-structure or sublexical rules to place the relevant item in the correct position, semantic information (i.e. the presence of PRED) and other functionally related information (e.g. voice specification bundled in templates as seen in (41)). Sample lexical entries are given in (44). (The functional specifications associated with $-i$ (i.e., @APPL_i shown in (42)) are annotated in the sublexical rule of $-i$ (41b), and are therefore not shown in the lexical entry.)
(44) Sample entries: free forms

| a. | kursi | N | $@($ CN chair $)$. |
| :--- | :--- | :--- | :--- |
| b. | dia | PRON | $@($ PPRO 3 sg) |
| c. | duduk | V | $\{@$ (INTRANS sit) $\mid @($ INTRANS_SEM_PP sit loc) $\}$ |
| d. | di | P | $@$ (PREP in loc) $;$ |
|  |  | Pcase | $@($ PCASE to obl-dir dir) |

Sample entries: bound forms

| e. | +I | V_I |  |
| :--- | :--- | :--- | :--- |
| f. | AV + | V_VOICE | @(VOICE-TYPE AV). |
| g. | UV+ + | V_VOICE | $@(V O I C E-T Y P E ~ U V) . ~$ |
| h. | PASSdi+ | V_VOICE XLE | @(VOICE-TYPE PASSIVE). |

Note that affixes are listed as lexical entries for their abstract tags, rather than morphological forms. That is, in the two-level morphology adopted here where an affix form such as meN- is treated as a string realised as the prefix $m e N$ - (lower-side) and corresponding to the morphological tag AV+, it is the tag AV+ and not the morphological form $m e N$ - that is listed in the lexical entry. It is then associated with the AV restriction as specified by the template @(VOICE-TYPE AV).

Within the templates CN, PPRO, INTRANS, INTANS_SEM_PP, PREP and PCASE are defined features related to the lexical entry they annotate.

### 4.3 Parsing and sample parses

To illustrate how sentences with $-i$ verbs are parsed, we provide a few output parses with a brief description of their analyses.

For (45a), the input string is first broken into tokens by the tokenizer. The output is then fed into the morphological analyser so that morphologically complex words such as menduduki and kursi can be analysed and assigned morpheme and category tags as in (45b). Since the relevant tags and forms are listed in the lexical entries, e.g. AV+ (for $m e N-$ and $-i$ (see (44)), the XLE parser is able to recognise the tags, and use the information to assign the word a hierarchical structure on the basis of the sublexical rules as formulated in (40). In addition, given the functional constraints carried by the morphemes and the structures (cf. template calls signalled by @ in the entries and in the sublexical rules), the parser can also build functional structures involving predicate composition for the $-i$ verb based on the defined grammar rules. The output c - and f - structures are displayed in ( 45 c )..
(45) Tokenizing and Morpheme identification:
a. Input string: Ia menduduki kursi
b. Morphologically analysed string: Ia $\mathrm{AV}+d u d u k+\mathrm{I}+$ Verb kursi


Example (46) shows a sentence with the monotransitive base pukul 'hit'. The derivation does not change the transitivity, but marks progressive aspect (indicated by $[\mathrm{PROG}+]$ in the f -str).

'They were hitting us.'

## 5 Concluding remarks

We have presented an a-str based analysis for the applicative-causative polysemy of $-i$. We have shown that different properties of - i - valency-increasing or no valence changing effects as well as the related interative/progressive meaning - and the alternation of $-i$ with - kan, are predictable from the interaction of the core information carried by the suffix and the information carried by the stem.

Building on earlier work (Alsina 1996; Butt 1995), we present a predicate composition analysis involving underspecified argument structure, which allows two types of argument fusion: single and double. We claim that, at least for the Indonesian $-i$, the fusion follows a natural general rule in which arguments of thematically similar types tend to fuse.

We have demonstrated the implementation of the analysis in XLE. The computational grammar we are developing shows promise how predicate composition can handle verbal derivation with $-i$ as part of verb formation in Indonesian. The grammar can correctly identify a range of different aspects of $-i$, in particular its interaction with the VOICE system and the different possibilities of the syntax of $-i$.

## 6 References

Alsina, A. 1992. On the argument structure of causatives. Linguistic Inquiry 23:517-555.
-. 1996. The role of argument structure in grammar. Stanford: CSLI.
Alsina, Alex, and Smitha Joshi. 1991. Parameters in causative constructions. In CLS 27:2-16.
Arka, I Wayan. 1993. The -kan causative in Indonesian. MPhil Thesis, University of Sydney, Sydney.
——. 2003. Balinese morphosyntax: a lexical-functional approach. Canberra: Pacific Linguistics.
Arka, I Wayan, and Christopher Manning. 2008. Voice and grammatical relations in Indonesian: a new perspective. In Voice and grammatical relations in Austronesian Languages, edited by P. K. Austin and S. Musgrave, 45-69. Stanford: CSLI.
Austin, P. 2005 [1996]. Causatives and applicatives in Australian Aboriginal languages, edited by K. Matsumura. Tokyo: Hitsuji Shobo.
Beesley, Kenneth R., and Lauri Karttunen. 2003. Finite State Morphology. Stanford: CSLI.
Bresnan, Joan. 1982. The passive in lexical theory. In The Mental Representation of Grammatical Relations edited by J. Bresnan, 3-86. Cambridge, MA.: The MIT Press.
Bresnan, Joan, and Jonni M. Kanerva. 1989. Locative inversion in Chichewa: a case study of factorization in grammar. Linguistic Inquiry 20:1-50.
Bresnan, Joan, and L. Moshi. 1990. Object asymmetries in comparative Bantu Syntax. Linguistic Inquiry 21 (2):147-185.
Butt, Miriam. 1995. The Structure of Complex Predicates in Urdu. Stanford: CSLI Publications.
Butt, Miriam, and Tracy Holloway King. 2006 Restriction for morphological valency alternations: the Urdu causative. In Intelligent Linguistic Architectures: Variations on Themes., edited by R. M. Kaplan. Stanford: CSLI.
Butt, Miriam, Tracy Holloway King, and John T Maxwell III. 2003. Complex predicates via restrictions. In Proceedings of the LFG'03 Conference, edited by M. Butt and T. H. King. CSLI, http://csli- publications.stanford.edu/LFG/8/lfg03.html.
Chung, Sandra. 1976. On the subject of two passives in Indonesian. In Subject and topic, edited by C. N. Li, 57-99. NewYork: Academic Press.

Jackendoff, Ray. 1990. Semantic Structures. Cambridge, Massachusetts: MIT Press.
Kaplan, Ronald M., and Jürgen Wedekind. 1993. Restriction and correspondence-based translation. In Proceedings of the Sixth European Conference of the Association for Computational Linguistics.
Kroeger, Paul. 2007. Morphosyntactic vs. morphosemantic functions of Indonesian -kan. In In Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan, edited by A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Maling and C. Manning, 229-251. Stanford, CA: CSLI Publications.
Levin, B, and M Rappaport Hovav. 1998. Morphology and Lexical Semantics. In The Handbook of Morphology, edited by In, A. Spencer and A. M. Zwicky. Oxford: Blackwell.
Levin, Beth. 1993. English verb classes and alternations : a preliminary investigation. Chicago University of Chicago Press.
Macdonald, R. Ross and Dardjowidjojo, Soenjono. 2001. A student's reference grammar of modern formal Indonesian. Georgetown: Georgetown University Press.
Manning, Christopher D. 1996. Ergativity: argument structure and grammatical relations. Stanford: CSLI.

Mistica, Meladel, Wayan Arka, Timothy Baldwin and Avery Andrews. 2009. Double Double, Morphology and Trouble: Looking into Reduplication in Indonesian, Paper presented at the 2009 Australasian Language Technology Workshop (ALTW 2009), Sydney, Australia
Moeliono, Anton M, and Soenjono Dardjowidjojo, eds. 1988. Tata bahasa baku Bahasa Indonesia Jakarta: Balai Pustaka.
Musgrave, S. 2001. Non-subject arguments in Indonesian. PhD thesis, Melbourne University.
Myhill, J. 1988. Agent incorporation in Indonesian. Journal of Linguistics 24 (1):111-136.
Pisceldo, Femphy, Rahmad Mahendra, Ruli Manurung, and I Wayan Arka. 2008. A Two-Level Morphological Analyser for Indonesian. http://rspas.anu.edu.au/linguistics/projects/iwa/IndonParGram/papers/Alta.morpho.pdf
Purwo, Bambang Kaswanti. 1989. Voice in Indonesian : A Discourse Study. In Serpih -serpih telaah pasif bahasa Indonesia, edited by B. K. Purwo, 344-442. Jogyakarta: Kanisius.
——. 1995. The Two Proto-types of Ditransitive Verbs: The Indonesian Evidence. In Discourse grammar and typology, edited by W. Abraham, T. Givon and S. A. Thompson, 77-99. Amsterdam: John Benjamins Publishing Company.
Simpson, J. 1991. Warlpiri Morphosyntax. Dordrecht: Kluwer.
Sneddon, James. 2006. Colloquial Jakartan Indonesian. Canberra: Pacific Linguistics.
Sneddon, James N. 1996. Indonesian reference grammar. St Leonards: Allen and Unwin.
Tenny, C. 1992. The Aspectual Interface Hypothesis. In Lexical Matters, edited by I. Sag and A. Szabolcsi, 1-28. Stanford: CSLI.
1994. Aspectual Roles and the Syntax-Semantics Interface. Dordrecht: Kluwer.

Vamarasi, Marit Kana. 1999. Grammatical relations in Bahasa Indonesia. Canberra: Pacific Linguistics.
Verhaar, John W. M. 1984. The Categorial System in Contemporary Indonesian: Verbs. In Towards a Description of Contemporary Indonesian: Part I, edited by K. Purwo, 27-64. Jakarta: NUSA , Universitas Atma Jaya.
Wolff, John U., Dede Oetomo, Daniel Fietkiewicz. 1992. Beginning Indonesian Through SelfInstruction: Book 3, Lessons 16-25. New York: Cornell Southeast Asia Program Publications.

# ADJACENCY AND LOCALITY: A CONSTRAINT-BASED ANALYSIS OF COMPLEMENTIZER-ADJACENT EXTRACTION 

Ash Asudeh<br>Carleton University

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

This paper provides a new explanation of phenomena related to extraction following an overt complementizer ('that-t effects'), for which the theory-neutral term complement-adjacent extraction is adopted. The analysis stems from the Correspondence Architecture of Lexical-Functional Grammar, making formally explicit certain implicit, native relations of the architecture. No reference is made to traces. The key insight is that complement-adjacent extraction effects concern linear string adjacency, where the string is understood as part of the syntax-phonology interface. A new metavariable, $\succ$, is introduced and formally defined; $\succ$ identifies the next word's f-structure. A single constraint is proposed that accounts for a wide range of relevant phenomena.


## 1 Introduction

This paper provides a new explanation of phenomena related to subject extraction following an overt complementizer (often called 'that-t effects' or 'Comp-trace effects'). This normally leads to ungrammaticality (Perlmutter 1968), as demonstrated in (1), and has received numerous theoretical treatments (Bresnan 1972, 1977, Chomsky and Lasnik 1977, and many others since).
a. Who do you think saw Kim?
b. *Who do you think that saw Kim?

I adopt the theory-neutral descriptive term 'Complementizer-Adjacent Extraction' (CAE). ${ }^{1}$ The novel analysis stems from the Correspondence Architecture of Lexical-Functional Grammar (LFG), making formally explicit certain implicit, native relations of the architecture. The key insight is that CAE effects do not concern structural superiority, but instead concern linear string adjacency, where the string is understood as a representation of part of the syntax-phonology interface.

The paper is structured as follows. Section 2 briefly presents the grammatical architecture of LFG and highlights aspects that will be particularly relevant. Section 3 delves further into the CAE phenomenon and presents some relevant complications. In section 4, I present my proposal and discuss the inadequacy of an alternative based on f-precedence. Section 5 presents the formal analysis and applies it to some examples. Section 6 considers some previous proposals, especially in light of the data discussed in section 3. Section 7 concludes. There is also an appendix that considers and rejects f-precedence-based alternatives to the main proposal.

## 2 The Correspondence Architecture

LFG's Correspondence Architecture (Kaplan and Bresnan 1982, Kaplan 1987, 1989, Halvorsen and Kaplan 1988, Asudeh 2006) divides the form-meaning mapping into a series of simultaneously-present, discrete modules, each of which represents distinct linguistic information. The part of the architecture that is relevant here is shown in Figure 1. C(onstituent)-structure represents word order, dominance and constituency, as modelled by a standard (non-tangled) tree - i.e., a phrase-structural parse of the

[^32]

Figure 1: Relevant part of LFG's Correspondence Architecture
phonological string. F (unctional)-structure represents more abstract aspects of syntax, such as predication and grammatical functions, null pronominals, local and unbounded dependencies, etc. F-structure is modelled as a feature structure. The $\phi$ correspondence function maps elements of c-structure to elements of f-structure. The syntactically unparsed string and the $\pi$ correspondence function from the string to the c-structure are two components of the Correspondence Architecture that have received little attention since their proposal by Kaplan (1987). They form the heart of this analysis.

The theory of unbounded dependencies assumed here (Kaplan and Zaenen 1989) does not posit any null element (trace or copy) in the extraction site. This means that a key representational device that is standardly used in accounts of CAE is unavailable. An analysis of CAE that does not posit traces or copies is simpler, since it posits fewer entities. In addition to achieving a simplification and removing a theoretical motivation for traces/copies, this analysis captures a wide range of empirical data with a single constraint.

## 3 Phenomena and problems

The basic Complementizer-Adjacent Extraction phenomenon is as follows:
(2) a. Who do you think saw Kim?
b. *Who do you think that saw Kim?
(3) a. Who do you think Kim saw?
b. Who do you think that Kim saw?

The basic observation is that there cannot be a subject extraction from a complement clause if the clause is introduced by a complementizer (2b); I will call this the 'basic CAE effect'. In the absence of a complementizer, the extraction is grammatical (2a). Object extraction is grammatical whether the complementizer is present or not. There is considerable dialectal variation in these judgements (Pesetsky 1982, Sobin 1987, 2002), but there are speakers who robustly demonstrate this contrast and their grammars need to be accounted for. The variation must also be properly accounted for; I return to this issue in section 5.1.

CAE is more complex than these facts alone would indicate, due to a phenomenon often called the 'Adverb Effect' (Bresnan 1977, Culicover 1991, 1993):
a. Who did you say that, just a minute ago, sneezed?
b. Who does Kim think that, with Sandy out of the picture, might receive the nomination?

Insertion of an adverbial element immediately after the complementizer neutralizes the basic CAE effect for many speakers. This is unexpected, because the structural relation between the complementizer and the extraction site is not affected by the adverbial, which is adjoined to IP in c-structure and is an ADJUNCT at f-structure: ${ }^{2}$

[^33](5)



If the impossibility of subject extraction in CAE is to be attributed to the presence of an intervening complementizer, then the presence of other intervening structure that does not affect intervention by the complementizer should not improve things, but it does. However, the adverbial element does linearly separate the complementizer and the extraction site.

The Adverb Effect facts gained new life when Culicover (1993) brought them to centre stage in a criticism of then-standard transformational accounts of CAE, which were based on the Empty Category Principle (Chomsky 1981). Another previously well-known complication was, however, subsequently largely ignored in responses to Culicover's work, with some notable exceptions (e.g., Sobin 2002). This second complication is that subject extraction after the complementizer that in a relative clause is grammatical (Bresnan 1977):
(7) This is the person that sneezed.

I will call this the Relative Clause Paradox. Some analyses assume that the that in relative clauses is a different lexical item from the complementizer that (Gazdar 1981, Pollard and Sag 1994:220-222). On such accounts, there is arguably no paradox, if CAE is associated with the complementizer that, but not with the relativizer that. However, there are long-standing empirical arguments that there is only one that (Bresnan 1972, 1977, Emonds 1976) and an analysis that posits no ambiguity is also to be preferred on grounds of parsimony.

There are in fact further complications with relative clauses, because the CAE effect re-emerges if the that-relative is embedded:
(8) *This is the person who Kim thinks that sneezed.

An account that reconciles CAE and the Relative Clause Paradox is therefore only successful if it can also account for the contrast between (7) and (8).

CaE, the Adverb Effect and the Relative Clause Paradox are the key phenomena of interest and present a puzzling enough set of problems on their own. However, there is yet another relevant phenomenon, that of Embedded VP Topics:
(9) Mary knows that doubt her John never could.

There is a fronted element here, the VP topic, that appears between the complementizer and the subject of the complementizer's clause. Crucially, unlike the case with the Adverb Effect, this intervening material does not mitigate the basic CAE effect:
(10) * Who does Mary know that doubt her never could?

Embedded VP Topics thus show that a reconciliation of the Adverb Effect and CAE cannot rely on a naive notion of intervening structure between the complementizer and the extraction site. ${ }^{3}$

## 4 Proposal

In this section, I sketch the proposed explanation of CAE, including the related phenomena of the Adverb Effect, the Relative Clause Paradox and Embedded VP Topics. I also consider a potential alternative LFG account based on f-precedence (Bresnan 1984, Kaplan 1987) and show that the pattern of data observed in the previous section is fundamentally incompatible with an f-precedence account of CAE.

The analysis that I propose uses two parts of LFG's Correspondence Architecture that appeared in the original presentation of the (extended) architecture (Kaplan 1987), but which have subsequently been largely ignored:

1. The syntactically unparsed string that is the input to c-structure. I assume that this string is phonologically parsed, i.e. tokenized into words (Forst and Kaplan 2006). The string is therefore a representation of linear phonology.
2. The $\pi$ correspondence function that maps the string to c -structure.

The key observations are the following:

1. The Adverb Effect indicates that the relevant grammatical notion for CAE is linear adjacency, not structural superiority. Linear adjacency is a relation from the syntax-phonology interface, whereas structural superiority is a properly syntactic relation.
2. The Relative Clause Paradox and Embedded VP Topic data show that a notion of phonological realization of the head of the unbounded dependency (FOCUS or TOPIC) is also relevant.

These observations will be put into effect in a single constraint that can be summarized as follows:

## (11) CaE Constraint (informal)

It is not the case that the string element that immediately follows the complementizer maps to an f-structure that contains a subject that is both phonologically realized and is the head of an unbounded dependency.

The constraint will be part of the lexical entry for complementizers, which permits a standard lexicalist account of the observed variation in CAE dialects.

The notion of realization is already available through the inverse of the $\phi$ mapping from c-structure to f-structure (Halvorsen and Kaplan 1988). The notion of simultaneously being a subject and the head of an unbounded dependency is also already available, through inside-out functional application (Halvorsen and Kaplan 1988). The new part of the proposal concerns linear adjacency, for which I introduce a function on string elements and, based on this, a new metavariable, $\succ$, which denotes the next word's f-structure.

## 5 Analysis

The CAE Constraint was informally presented in section 4 as follows:

## (12) CAE Constraint (informal)

It is not the case that the string element that immediately follows the complementizer maps to an f-structure that contains a subject that is both phonologically realized and is the head of an unbounded dependency.

[^34]I will now formalize this constraint.
The string projection in the Correspondence Architecture must contain a native ordering relation for linear precedence. This can be represented as a function on string elements, which are characterized as words:
(13) $\mathrm{N}: \mathrm{W} \rightarrow \mathrm{W}$, where W is the set of words in the string

The string is assumed to be 'phonologically parsed' or tokenized into units. The function name, N , is meant to be mnemonic for 'next'.

Phonological realization can be defined through the inverse of the $\phi$ mapping from c-structure to f-structure:
(14) For any f-structure $f, \operatorname{REALIZED}(f)$ is true iff $\phi^{-1}(f) \neq \emptyset$.

In other words, the predicate REALIZED is just packaging, in a shorter and more intuitive form, the independently available equation $\phi^{-1}(f) \neq \emptyset$, which states that the set of c-structure nodes that map to $f$ (the c -structure correspondent of $f$ ) is not empty. The predicate REALIZED will thus return false if applied to a null pronominal.

The last piece of formalization is obtained by using the string function N , along with the correspondence functions $\pi$ and $\phi$, to identify the next string element's f-structure, notated as a metavariable $\succ$ :

$$
\begin{equation*}
\succ:=\phi\left(\mathcal{M}\left(\pi\left(N\left(\pi^{-1}(*)\right)\right)\right)\right) \text {, where } * \text { is a terminal node; undefined otherwise } \tag{15}
\end{equation*}
$$

The variable * picks out the current c-structure node (i.e., the node bearing an annotation involving $\succ)$. We take the inverse of the $\pi$ correspondence function from string to c-structure and apply it to *, returning a string element (a word). We then apply N to the obtained word to get the next word. Having obtained the word that follows the word that corresponds to the current c-structure node, we apply the $\pi$ function to get back into c-structure. We are then sitting at a terminal node, which may not be directly mapped to f -structure. We therefore apply the c-structure mother function $\mathcal{M}$ to get the pre-terminal node. Lastly, we apply the $\phi$ correspondence function from c-structure to f-structure. The metavariable $\succ$ can therefore be used in a lexical entry to refer to the f-structure of the word that immediately follows the relevant terminal node. ${ }^{4}$

The CAE constraint can now be defined as follows:
(16) CAE Constraint (formal)

```
\neg [ \operatorname { R E A L I Z E D } ( \succ \text { SUBJ ) ^(UDF } ( \succ \text { SUBJ ))]}
where UDF is an unbounded dependency function (FOCUS or TOPIC)
```

The inside-out existential constraint (UDF( $\succ$ SUBJ)) is true just in case the next word's subject is in an unbounded dependency, i.e. extracted. The constraint states that it cannot be the case that the next word's subject is both realized and extracted. I next turn to some examples, which show the constraint in effect.

### 5.1 Variation

The CAE constraint is associated with lexical entries of complementizers and is not a general structural constraint, e.g. associated with the category C. This lexicalist view has also been argued for independently by Falk (2006:130-131). For example, the lexical entry for that would look something like

[^35](17). I am not positing any ambiguity between that in relative clauses and the complementizer; both occurrences are in fact the complementizer and there is just the single lexical entry.

```
that C, ( }\uparrow\mathrm{ TENSE)
    (\uparrow MOOD) = DECLARATIVE
    \neg[REALIZED (\succ SUBJ) ^(UDF (\succ SUBJ))]
```

The lexical analysis explains variation between complementizers as to whether they block CAE or not.
For example, Sobin $(1987,2002)$ has shown that some English speakers allow CAE with that but not with whether. The impossibility of CAE with whether seems to require an explanation over and above an appeal to the status of whether-clauses as weak islands, although caution has to be exercised here, because Sobin does not report explicitly whether the differences in the relevant conditions are significant or not. In sum, it seems that some English speakers do not have a CAE effect with that but do have one with whether. This can be accounted for if the lexicons of such speakers do not contain the CAE constraint in the entry for that. Similarly, Shlonsky (1988) has observed that Hebrew $\int e$ ('that') allows CAE, but im ('if') blocks it. Again, this can be explained as lexical variation with respect to the CAE constraint. There are quite a few other cases of cross-linguistic and dialectal variation for CAE reported in the literature; see Kandybowicz (2009:329) for further references.

Lastly, there is also variation in whether the Adverb Effect ameliorates CAE effects (Sobin 2002). This variation can be explained with respect to differences between the $\uparrow$ and $\succ$ metavariables. If a speaker has the CAE constraint realized with the $\succ$ metavariable, then the Adverb Effect holds in the speaker's grammar, as outlined below. However, if a speaker has the CAE constraint with the two instances of $\succ$ replaced by $\uparrow$, then the Adverb Effect does not hold in the speaker's grammar, since interpolation of the adverbial does not affect the relationship when stated in terms of $\uparrow$, as sketched in the discussion of f-precedence in the appendix below.

### 5.2 Examples

In this section I show how the CAE constraint accounts for various cases that have been under discussion.
First, let us look at an example of how the constraint correctly blocks basic CAE:
(18) * Who do you think that sneezed?


The complementizer that is the fifth word $\left(w_{5}\right)$ in the string for (18). The next word is the head of the complementizer's clause, sneezed $\left(w_{6}\right)$. The $\succ$ metavariable in the lexical entry for that is therefore realized as $f_{5}$, which is the f -structure of the mother of the word immediately following that (i.e., sneezed). The SUBJ of $f_{5}$ is $f_{2}$, which is REALIZED, as who, and is also a UDF, since the SUBJ is also the FOCUS of the main clause's f-structure. The CAE constraint is therefore violated and the example is correctly blocked.

Second, let us look at a simple Adverb Effect example:
(20) Who do you think that probably left?


The $\succ$ metavariable in the lexical entry for that is here realized as $f_{6}$, which is the f -structure of the mother of the word immediately following that, the adverb probably. The adverb has no SUBJ, so the CAE constraint is trivially satisfied.

Third, let us consider a more complex adverbial in an Adverb Effect example:
(22) Who does Kim think that, with Sandy out of the picture, might receive the nomination?


The $\succ$ metavariable in the lexical entry for that is here realized as $f_{7}$, which is the f -structure of the mother of the word immediately following that, the preposition with. In contrast to the simpler Adverb Effect example, with arguably does have a SUBJ, if it is to be analyzed as having a small clause (i.e., predicative) complement (Pollard and Sag 1987, 1994:110). I am not interested in defending such an analysis here, but rather in showing that the CAE constraint would derive the correct result even if an adverbial does have a subject. The SUBJ of with's f -structure, $f_{7}$, is REALIZED as Sandy, but it is not a UDF, since Sandy is not the top of an unbounded dependency (i.e., the SUBJ is not extracted). The CAE constraint is therefore non-trivially satisfied by this kind of adverbial, again accounting for the Adverb Effect.

Fourth, let us consider a simple relative clause:
(24) the person that sneezed


The $\succ$ metavariable in the lexical entry for that is here realized as $f_{5}$, which is the f -structure of the mother of the word immediately following that, the verb sneezed. This verb does have a SUBJ and the SUBJ is a UDF, since it is the TOPIC in the relative clause. However, the SUBJ is the null relative pronoun and is not REALIZED - there is no c-structure correspondent of SUBJ. Therefore, the left conjunct in the CAE constraint is false and the constraint is satisfied as a result. The CAE constraint therefore accounts for simple cases of the Relative Clause Paradox.

Fifth, let us consider embedded relative clauses, in which the CAE effect re-emerges:
(26) * the person who Kim thinks that sneezed


The $\succ$ metavariable in the lexical entry for that is here realized as $f_{7}$, which is the f -structure of the mother of the word immediately following that, again the verb sneezed. This verb does have a SUBJ and the SUBJ is a UDF, since it is the TOPIC in the relative clause. This time, the SUBJ is in fact realized by the overt relative pronoun who. The CAE constraint is therefore violated in the more complex case, because the relative pronoun is REALIZED. The CAE constraint therefore also accounts for complex cases of the Relative Clause Paradox.

Lastly, the CAE constraint accounts for the Embedded VP Topic contrast, repeated here, although there is insufficient room to show the relevant structures:
(28) Mary knows that doubt her John never could.
(29) * Who does Mary know that doubt her never could?

This contrast shows that not just any intervening material blocks a CAE violation. In (28) the string element following that is the verb doubt, which has a REALIZED SUBJ, John. However, John is not a UDF, because there is no extraction of the subject; the CAE constraint is not violated. In contrast, the SUBJ of doubt is both REALIZED and a UDF in (29) and the CAE constraint is violated. As mentioned
briefly in footnote 3, the version of (29) without that is independently ungrammatical in standard English dialects. However, the account makes the specific prediction that (29) could be ungrammatical in a language even if the version of (29) without that is grammatical.

## 6 Previous proposals

In this section, I briefly review a number of previous proposals for capturing the CAE phenomena. There have been too many particular proposals to do them all justice. I will discuss the proposals as natural classes where possible, even though this risks obscuring differences. I will primarily focus on the empirical issues of whether the proposals capture the data (basic CAE effects, the Adverb Effect, the Relative Clause Paradox, and variation). ${ }^{5}$

### 6.1 An alternative proposal based on the syntax-phonology interface

Kandybowicz $(2006,2009)$ provides a theory of CAE based on PF, ${ }^{6}$ which constitutes the syntaxphonology interface in the Minimalist Program (MP; Chomsky 1995). There are other PF-based approaches to the phenomena, but I will only discuss Kandybowicz's proposals; see Kandybowicz (2009:329) for further citations. Kandybowicz (2009:328-329) also briefly reviews several non-PFbased Minimalist accounts. Based on my understanding of some of the non-PF MP accounts (Pesetsky and Torrego 2001, Ishii 2004) and on Kandybowicz's review of the others, they cannot account for the Relative Clause Paradox (without positing multiple thats) or cross-linguistic and dialectal variation in CAE effects (as stressed by Kandybowicz himself), including lexical variation.

Although Kandybowicz's proposal assumes the Minimalist framework, the underlying intuition of his account and the present account is shared: CAE effects ought to be captured at the syntax-phonology interface. Kandybowicz (2006) presents a theory of CAE in light of a careful consideration of prosodic data from English and Nupe. Kandybowicz (2009) further elaborates the account of Nupe. Unfortunately, the analysis of CAE in Nupe is insufficiently formalized in Kandybowicz (2006, 2009:334-339) to allow ready comparison with the CAE constraint. However, it seems that the proposal accounts for not only basic CAE effects, but also the Adverb Effect and the Relative Clause Paradox. It seems that the proposal would have trouble with lexical variation, as it offers a structural account based on properties of $\mathrm{C}^{0}$.

Kandybowicz (2006:223) proposes the following for English:

$$
\begin{equation*}
*\left\langle\mathrm{C}^{0}, \mathrm{t}\right\rangle \text { iff: } \quad \text { i. } \mathrm{C}^{0} \& \mathrm{t} \text { are adjacent within a prosodic phrase AND } \tag{30}
\end{equation*}
$$

ii. $\mathrm{C}^{0}$ is aligned with a prosodic phrase boundary

This raises theory-internal questions if PF in MP is to be understood as follows:
Consider a representation $\pi$ at PF. PF is a representation in universal phonetics, with no indication of syntactic elements or relations among them (X-bar structure, binding, government, etc.). To be interpreted by the performance systems A-P [Articulatory-Perceptual - AA], $\pi$ must be constituted entirely of legitimate PF objects, that is, elements that have a uniform, language-independent interpretation at the interface. (Chomsky 1995:194; emphasis in orginal)

It would seem that trace (or unpronounced parts of copy chains) should not constitute "legitimate PF objects", so it is unclear how (30) could even be stated as a PF constraint. However, PF is generally construed as a syntactic level, despite Chomsky's original conception (Jason Merchant, p.c.). But then this raises the question of why a syntactic level contains prosodic phrases. The tension remains.

[^36]
### 6.2 Alternative constraint-based proposals

There have been numerous previous constraint-based analyses of CAE. Some of these analyses such as Gazdar (1981), Pollard and Sag (1994) and Ginzburg and Sag (2000) - capture CAE, but do not capture the Adverb Effect and only capture the Relative Clause Paradox by postulating both a complementizer that and a relativizer that. I focus on two recent accounts that capture a broader range of data: the HPSG account of Levine and Hukari (2006) and the LFG account of Falk (2006).

### 6.2.1 The Intervention Constraint

Levine and Hukari (2006:99) propose the following constraint in their explanation of CAE:

## (31) Intervention Constraint

No complementizer may immediately precede the finite head of the clause marked by that complementizer.

Levine and Hukari point out that their Intervention Constraint is operational even where there is no subject extraction, unlike accounts of CAE that rely on somehow banning a Comp-trace sequence, where the trace in question is that of subject extraction. The Intervention Constraint is similar to the CAE constraint that I proposed above, in that both involve precedence. It should be clear that, in order to capture the Intervention Constraint formally, some precedence-based device like the precedence metavariable that I introduced is still necessary.

In support of their account, Levine and Hukari (2006:100) note the following contrast (the parenthetical remark after the second example appears in the original):
(32) a. *I wonder if could you move your car from in front of my driveway?
b. I wonder if at one point could you move your car from in front of my driveway? (with no comma intonation after point)

In (32a), there is no subject extraction, but the sentence is nonetheless ungrammatical. In (32b), we apparently see the ameliorating Adverb Effect, even in the absence of subject extraction. The Intervention Constraint accounts for this contrast.

However, the ungrammaticality of (32a) is also explained straightforwardly by the fact that verbs like wonder never embed a direct question:
(33) * Kim wondered if did Sandy snicker?
(34) *Robin pondered whether should Kim care?
(35) *I doubt if could you be quiet?

The issue then becomes explaining the grammaticality of (32b) for those speakers who perceive it as such.

Some light is cast on the issue by considering whether the verb in question supports a parenthetical usage with a direct question. Wonder is such a verb, whereas doubt is not:
(36) a. Could you be quiet, I wonder?
b. I wonder: could you be quiet?
(37) a. *Could you be quiet, I doubt?
b. *I doubt: could you be quiet?

The amelioration effect in (32b) is completely absent with doubt:
(38) a. *I doubt if could you be quiet.
b. *I doubt if, even with strong incentives, could you be quiet.

The parenthetical even with strong incentives is perfectly fine with doubt when its complement is not a direct question:
a. I doubt if you could be quiet.
b. I doubt if, even with strong incentives, you could be quiet.

If the Intervention Constraint in (31) is correct, the contrast between (38b) and (32b) is mysterious, particularly in light of the other data adduced in this section.

A further empirical inadequacy of the Intervention Constraint is that it does not resolve the Relative Clause Paradox, since the complementizer equally immediately precedes the finite head in an example like (7), repeated here:
(40) This is the person that sneezed.

The Intervention Constraint wrongly predicts these cases to be ungrammatical, unless the problematic assumption is made that the that in a relative clause is not the complementizer.

In sum, the Intervention Constraint analysis does not account for the full range of facts and the constraint itself arguably rests on a misanalysis of the facts in (32). The adverbial in (32b) is not ameliorating a complementizer-head adjacency, but rather supporting a parenthetical parse of I wonder if, which is otherwise impossible, since the normal parenthetical use of wonder does not take a complementizer. This effect still requires explanation, and such an explanation may shed further light on the Adverb Effect in CAE, but there is reason to doubt that the data in (32) should be conflated with the CAE data.

### 6.2.2 The Pivot Immediate Dominance Constraint

Falk $(2000,2001,2006)$ provides an account of CAE in light of his more general theory of pivots, which introduces a new grammatical function PIV, such that "The PIV is the element with the function of connecting its clause to other clauses in the sentence" (Falk 2006:74). Informally, his account of CAE is that the complementizers that show CAE effects contain a lexical constraint that states that "The clause [introduced by the complementizer - AA] has its own PIV" (Falk 2006:132). This constraint is formalized as follows (Falk 2006:133):

$$
\begin{equation*}
\phi^{-1}(\uparrow \mathrm{PIV}) \Rightarrow \uparrow \rightarrow_{f}(\uparrow \mathrm{PIV}) \tag{41}
\end{equation*}
$$

This constraint depends on a definition of the relation $\rightarrow_{f}$, which is functional immediate dominance (f-ID, on analogy to f-precedence):
(42) Functional immediate dominance (f-ID)
(Falk 2006:133, (49))
For any f-structures $f_{1}$ and $f_{2}, f_{1}$ f-IDs $f_{2}\left(f_{1} \rightarrow_{f} f_{2}\right)$ iff there exists a node $n_{1}$ in $\phi^{-1}\left(f_{1}\right)$ and a node $n_{2}$ in $\phi^{-1}\left(f_{2}\right)$ such that $n_{1}$ immediately dominates $n_{2}$.

Let us call constraint (41) the PIVOT Immediate Dominance (PID) constraint. The constraint is intended to have the consequence that "If $\phi^{-1}\left(\uparrow\right.$ PIV) exists, one of the nodes in $\phi^{-1}(\uparrow)$ must immediately dominate on the nodes in $\phi^{-1}$ ( $\uparrow$ PIV)" (Falk 2006:132, (48)). Falk (2006:133) shows that, in a basic CAE example, the constraint is not satisfied because there is no node in the c-structure correspondent of the complementizer's f-structure that immediately dominates the extracted subject (since he also assumes that there is no subject trace in c-structure).

The PID constraint is similar to the CAE constraint. This is more obvious if the left side of (41) is restated as $\phi^{-1}(\uparrow$ PIV $) \neq \emptyset$, which is just REALIZED $\left(\uparrow\right.$ PIV). ${ }^{7}$ The constraint thus not only accounts for basic CAE effects, but also accounts for the Relative Clause Paradox (and Embedded VP Topics), as discussed by Falk himself (Falk 2006:134). It is also a lexical constraint, so it can account for variation (Falk 2006:130-134).

[^37]However, the constraint cannot account for the Adverb Effect, because it is stated in terms of $\uparrow$ and, as we have seen, a constraint stated in terms of $\uparrow$ fails to capture the Adverb Effect, since interpolation of an adverbial does not affect the relation between the complementizer's f-structure and the f-structure of its other grammatical functions (whether SUBJ or PIV). In other words, the PID constraint ignores the evidence that CAE is a precedence-based phenomenon, not a dominance-based phenomenon. The PID constraint and the CAE constraint could easily be reconciled if SUBJ in the CAE constraint is replaced by PIV. Lastly, the relation of f-ID is potentially computationally exacting, like f-precedence, since f-ID requires comparison of two sets of c-structure nodes. ${ }^{8}$

### 6.3 Other proposals

### 6.3.1 The Fixed Subject Constraint and the Complementizer Constraint on Variables

Bresnan (1972) generalizes a previous proposal by Ross (1967) such that nothing can be extracted in the environment [COMP _ VP]. This is the Fixed Subject Constraint:
(43) Fixed Subject Constraint (FSC)

No NP can be crossed over an adjacent COMP


The FSC accounts for basic CAE effects and also accounts for the Adverb Effect (an interpolated adverbial disrupts adjacency), but does not predict the Relative Clause Paradox, since the banned configuration obtains in relative clauses. This latter problem was one of the motivations for the subsequent generalization of the FSC to the Complementizer Constraint on Variables (CCV) (Bresnan 1977:173), which accounts for the Relative Clause Paradox, without losing the FSC's account of CAE effects or the Adverb Effect. Despite the success of the CCV in accounting for much of the CAE phenomena, it relies on theoretical notions, such as conditions on transformations and structural descriptions, that are no longer part of even transformational theory and are obviously not part of constraint-based theories such as LFG.

### 6.3.2 The that-t filter

Chomsky and Lasnik (1977) propose the surface filter in (44) to capture CAE. Surface filters restrict the transformational component of a transformational grammar by marking as ungrammatical a subset of the set of outputs of the component.
(44) * $\left[{ }_{\bar{S}}\right.$ that $\left[{ }_{\mathrm{NP}} \mathrm{e}\right] \ldots$, unless $\overline{\mathrm{S}}$ or its trace is in the context $\left[_{\mathrm{NP}} \mathrm{NP} \ldots\right]$

The term that-t filter is still commonly used as a descriptive term, even though the filter itself is no longer adopted.

The filter does capture the Adverb Effect, because it is stated in terms of adjacency, not structural superiority. It also captures the Relative Clause Paradox, but only by directly stipulating relative clauses as an exception to the filter (the "unless" clause). The filter does not capture variation; even if it is generalized to the category C , it would still be a structural constraint that is incapable of capturing lexical variation.

Chomsky and Lasnik (1977) build on work by Perlmutter (1968), who first observed CAE effects. Perlmutter (1968) postulated a universal to the effect that the constraint that blocks CAE (e.g., a filter

[^38]like the one in (44)) is valid for all and only languages that lack Subject Pronoun Deletion (pro-drop). Chomsky and Lasnik (1977) design their filter to entail Perlmutter's universal. However, the universal is not true; for example, CAE effects do not hold in dialects of English and in certain Scandinavian dialects (Lohndal 2009), even though the dialects in question do not allow subject deletion in the intended sense of Romance pro-drop. Since the that-t filter entails a false claim, it cannot be correct.

### 6.3.3 The Empty Category Principle

There are many accounts of CAE that ultimately attempt to relate it to the Empty Category Principle (ECP). The ECP can be defined as follows, based on Chomsky (1981:274) and Chomsky (1986:88):
(45) Empty Category Principle (ECP): Traces must be properly governed

The essential insight common to ECP approaches is that the complementizer blocks proper government of a trace in CAE (Chomsky 1981, Kayne 1981, Pesetsky 1982, Rizzi 1990, among others). Culicover (1993) argues convincingly that the Adverb Effect data is fundamentally incompatible with ECP approaches. The reason is plain: if the complementizer blocks proper government, adjunction of an adverbial cannot undo this. Furthermore, if the complementizer blocks proper government in CAE, it must equally do so in relative clauses, unless the relativizer that is a distinct item. ECP approaches thus resolve the Relative Clause Paradox only at the expense of an otherwise unmotivated and empirically problematic ambiguity. Lastly, structural accounts such as these ECP accounts cannot explain the apparent lexical variation displayed in CAE effects.

### 6.3.4 CP expansion and CP contraction

Browning (1996) and Rizzi (1997) both propose analyses of CAE in which the explanatory mechanism involves an expansion of the CP structure, in some manner. Browning (1996:241ff.) proposes that the Adverb Effect obtains because the adverbial is in SpecCP, which forces 'CP Recursion', i.e. creation of another CP layer. Consider example (46) from Browning (1996:241). Due to the adverbial in SpecCP, the complementizer must move, targeting its own CP. The relative operator subsequently moves through the SpecCP created by movement of the complementizer, yielding (47):
(46) Robin met the man that Leslie said that for all intents and purposes was the mayor of the city.
(47) $\mathrm{OP}_{\mathrm{i}} \ldots\left[_{\mathrm{CP}} \mathrm{t}_{i}^{\prime}\left[{c^{\prime}}^{\text {that }}{ }_{c}\left[_{\mathrm{CP}}\right.\right.\right.$ for all intents and purposes $\left[{ }_{c^{\prime}} \mathrm{t}_{\mathrm{c} / \mathrm{i}}\left[{ }_{I P} \mathrm{t}_{\mathrm{i}}\right.\right.$ was the mayor $\left.\left.\left.\left.\left.\ldots\right]\right]\right]\right]\right]$

There are severe problems with this proposal. First, it is crucial that the adverb in question be in SpecCP, but this is problematic from a theory-internal perspective, because that position is an operator position and is not appropriate for adverbials. Browning states that she argues for this position (Browning 1996:241), but she seems to just assume it. Second, it is crucial that the complementizer not have an index (hence the subscripted c), but it is also crucial that the trace of the complementizer govern the subject trace. This is contradictory. Furthermore, in other cases it seems that the complementizer should have a (real) index according to the assumptions of the theory in question (Sobin 2002). Third, it is not clear why the complementizer must move rather than the structure just being ruled out. The theory provides no a priori baseline for this and therefore risks making no predictions regarding grammaticality. Lastly, the theory does not account for the Relative Clause Paradox, unless an additional relativizing that is assumed.

Rizzi (1997) presents a different sort of expanded CP analysis in which CP is split into two obligatory projections of Force and Finiteness, with intervening optional Topic and Focus projections: ForceP $>($ TopicP $)>($ FocusP $)>($ TopicP $)>$ FinP. The complementizer that occupies Force ${ }^{0}$ and a null counterpart occupies Fin ${ }^{0}$. Sobin (2002:534-535) raises a number of theory-internal problems for Rizzi's proposal, the most pernicious of which is how to ensure that the overt and covert complementizers interact properly. Rizzi himself acknowledges a variant of this problem and proposes that an economy constraint ("Avoid structure") is at play (Rizzi 1997:314). Nevertheless, empirical problems remain,
because the analysis seems not to capture the Relative Clause Paradox and cannot adequately account for variation (see also Sobin 2002:534-536).

Sobin (1987, 2002) builds on work by Pesetsky (1982) to instead argue for an analysis that contracts CP in relevant cases, rather than expanding it. Sobin (2002) proposes an operation called Fuse, following a proposal by Carnie (2000), which is an update of his previous notion of Fusion (Sobin 1987). The basic idea is that, under certain conditions, the specifier and head elements of CP can collapse into a single indexed head, i.e. SpecCP and C Fuse. The adverbial in an Adverb Effect example Fuses with the complementizer through adjunction: the adverbial first adjoins to C and then C and the phrase Fuse to create a new C .

There are a number of problems with Sobin's analysis. First, as already noted, it requires adjunction of a phrase to a head and subsequent treatment of the head-phrase adjunction structure as a head. This is poorly motivated and also risks undermining fundamental aspects of the theory of phrase structure. Second, in order to properly account for the Adverb Effect and to resolve the Relative Clause Paradox, Sobin (2002) must postulate two distinct variants of Fuse, one for chain heads and one for traces. Third, the two variants of Fuse entail two variants of that. Fourth, Sobin (2002:546) is compelled to postulate that the relative that is a kind of subject place holder bound by the modified nominal, but that cannot in general perform this function, even as a deictic pronoun. Contrast the putative binding of that by nobody in the grammatical relative clause example (48) with the ungrammatical examples in (49).
(48) There is nobody ${ }_{i}$ that $_{i}$ believes the claim.
(binding postulated in Sobin 2002)
a. Nobody ${ }_{i}$ said that he ${ }_{i} / *$ that $_{i}$ believes the claim.
b. Nobody $_{\mathrm{i}}$ is such that $\mathrm{he}_{\mathrm{i}} /$ that $_{\mathrm{i}}$ believes the claim.

Fifth, it is necessary in Sobin's theory that an element with the feature [ +WH ] be allowed to Fuse with an element with the feature [ -WH ]. Why should this be possible? Sixth, in order for the Adverb Effect to be captured by Fuse, it is necessary to assume that the C created by adjunction of the adverbial to the complementizer that counts as null. Why should addition of overt structure to an overt element make the element null? Furthermore, he requires that the structure created by the Fuse of the adverbial with the complementizer have a lexical category, C , but that the syntax not treat it as a lexical item. This means that the syntax must be somehow sensitive to the distinction between unfused heads and fused heads. How is the distinction drawn in the syntax?

## 7 Conclusions and Future Work

The CAE constraint is a simple constraint that captures a wide variety of data, including basic CAE effects, the Adverb Effect, the Relative Clause Paradox, and Embedded VP Topics. The constraint makes no reference to a representational device such as a trace that marks the position of the subject extraction, thus maintaining LFG's traceless theory of unbounded dependencies (Kaplan and Zaenen 1989). The intuition behind the constraint is that CAE is a constraint at the syntax-phonology interface, where linear precedence is a native relation, an idea that is shared by Kandybowicz (2006, 2009), although under quite different theoretical assumptions. The constraint is stated in terms of the $\succ$ metavariable, which identifies the next word's f-structure. The metavariable is stated in terms of the $\pi$ mapping from the phonologically parsed string to c-structure in LFG's Correspondence Architecture. Several alternatives to this approach were reviewed and were shown to have empirical and theoretical inadequacies. Nevertheless, three approaches were identified as close cousins of this one: the PF proposal of Kandybowicz and the constraint-based proposals of Levine and Hukari and of Falk.

A number of avenues for future work suggest themselves. It would be interesting to connect the notion of linear adjacency developed here with other LFG proposals concerning the syntax-phonology interface and string parsing, such as Butt and King (1998) and Bögel et al. (2009). This would also allow more of the insights of Kandybowicz $(2006,2009)$ to be captured. It is also important to consider the nature of the $\pi$ function in light of the theory of Lexical Sharing (Wescoat 2002, 2005, 2007, 2009);
the notions developed here and Lexical Sharing are not necessarily antithetical, but the $\pi$ function would likely have a more complex analysis.

The kind of adjacency effect observed in CAE is reminiscent of Zwicky's "shape conditions" (Zwicky 1985, 1986, Pullum and Zwicky 1988), which have been appealed to in previous constraintbased analyses of phenomena such as the an/a alternation in English, French liaison and Welsh mutation (Asudeh and Klein 2002, Tseng 2003). The an/a alternation is a simple illustration: the form an is conditioned by an immediately following vowel-initial word, no matter the structural relation between the article and the following word (e.g., an orange/a plum, an/*a ugly plum, an/*a unbelievably nice plum). Another apparently adjacency-based phenomenon is Welsh syntactic soft mutation (see Tallerman 2009 and references therein), in which a complement $\alpha$ to a head bears soft mutation if a phrase that c-commands $\alpha$ immediately precedes $\alpha$ (i.e., separates the head and the complement; Borsley 1999). The Welsh case is especially compelling, because a trace of extraction counts as a trigger for soft mutation. This constitutes an important challenge to a traceless theory of unbounded dependencies and one that could potentially be met using the metavariable introduced here.

## Appendix: Inadequacy of f-precedence

An objection to the $\succ$ metavariable may be that LFG already has the precedence relation of f-precedence (Bresnan 1984) and that, all else being equal, I should not introduce a new mechanism. All else is not equal: there are theoretical and empirical inadequacies with f-precedence compared to the relation that I propose. On the theoretical side, the constraint that I formalize below concerns a very local notion of precedence between two string elements. In contrast, in order to calculate f-precedence a potentially large number of c-structure nodes must be considered. In other words, f-precedence is a computationally inefficient operation; this is presumably partly why it is not implemented in the standard implementation of LFG, the Xerox Linguistic Environment (Crouch et al. 2009), which instead implements a more limited variant ("head precedence").

Let us consider two alternative definitions of f-precedence (Dalrymple 2001:172-174). ${ }^{9}$
(50) F-precedence (strong)

F-structure $f$ f-precedes f -structure $g\left(f<_{f_{\forall \forall}} g\right)$ if and only if for all $n_{1} \in \phi^{-1}(f)$ and for all $n_{2} \in \phi^{-1}(g), n_{1}$ c-precedes $n_{2}$.
(51) F-precedence (weak)

F-structure $f$ f-precedes f -structure $g\left(f<_{f \nexists \exists} g\right)$ if and only if for all $n_{1} \in \phi^{-1}(f)$ and for some $n_{2} \in \phi^{-1}(g), n_{1}$ c-precedes $n_{2}$.
(52) C-precedence ${ }^{10}$
(Dalrymple 2001:172)
A c-structure node $n_{1}$ c-precedes a node $n_{2}$ if and only if $n_{1}$ does not dominate $n_{2}, n_{2}$ does not dominate $n_{1}$, and all nodes that $n_{1}$ dominates precede all nodes that $n_{2}$ dominates.

Strong f-precedence is the relation introduced by Bresnan (1984) in unpublished work and defined in Kaplan (1987) and taken up by Kameyama $(1985,1989)$ and Zaenen and Kaplan (1995). Weak fprecedence is the relation discussed in Bresnan $(1994,1995,2001)$ in different terms, which are almost, but not entirely, equivalent; although it is somewhat tangential, this is a theoretically interesting point and I return to it at the end of this appendix.

In addition to two notions of f-precedence, we need to consider a positive constraint to the effect that the complementizer's f-structure must f-precede that of the subject of the complementizer's clause and a negative constraint that states that the subject of the complementizer's clause cannot f-precede

[^39]the complementizer's f-structure. This yields four constraints that could be part of a complementizer's lexical entry, where $f$ is the f -structure of the complementizer:

```
\(f<_{f \forall \forall}(f\) SUBJ \()\)
    \(f<_{f_{\forall \exists}}(f\) SUBJ \()\)
    ( \(f\) SUBJ) \(\nless f_{\forall \forall} f\)
\((f\) SUBJ \() \nless_{f \forall \exists} f\)
```

On the standard assumption that C is an f -structure co-head (Bresnan 2001, Toivonen 2003), the two positive constraints are out, because even in examples with no extraction, both constraints are false:
(57) Kim said that Sandy left.

That and left correspond to the same f-structure. It is not the case that all c-structure nodes that map to the complementizer's f-structure precede all nodes that map to the subject's f-structure (constraint 53 is false) and it is not the case that all c-structure nodes that map to the complementizer's f-structure precede some c-structure node that maps to the subject's f-structure, because left does not precede Sandy (constraint 54 is false). ${ }^{11}$

Next consider constraint (55), which is stated with strong f-precedence. This constraint captures basic CAE effects, because in that circumstance all of the c-structure correspondent of the subject f precedes all of the c-structure correspondent of the complementizer's f-structure; the constraint is thus violated and correctly blocks CAE. However, the constraint does not fare well on the Adverb Effect or Relative Clause Paradox. With respect to the Adverb Effect, insertion of the adverbial does not affect the f-precedence relation between the subject and the complementizer's f-structure, so the constraint is equally violated when an adverbial occurs after the complementizer and the subject is extracted; the adverbial examples are not generated. In order to appreciate the behaviour of the constraint with respect to the Relative Clause Paradox, it is useful to see the standard LFG analysis of a relevant relative clause example (Dalrymple 2001): ${ }^{12}$
the person that sneezed


The subject is identified with a relative pronoun at f -structure, but the relative pronoun has no c -structure correspondent; it is a null pronoun. Any null element both vacuously strongly f-precedes and is vacuously strongly f-preceded by anything else in the f-structure (Kameyama 1989, Dalrymple 2001). Therefore, by virtue of being equal to the null pronoun, the subject vacuously f-precedes the complementizer's f-structure. The constraint is thus equally violated in the relevant relative clause and there is undergeneration again, this time of a very basic phrase, (58).

[^40]Lastly，consider constraint（56），which is analogous to constraint（55），but stated with weak f－ precedence．Just like constraint（55），constraint（56）captures basic CAE effects，again because the c－structure correspondent of the subject f－precedes all of the c－structure correspondent of the comple－ mentizer＇s f－structure；the constraint is thus violated．The constraint is also the same as constraint（55） with respect to the Adverb Effect and the Relative Clause Paradox．Again，the adverbial does not affect the f－precedence relation between the subject and the complementizer＇s f－structure，so the constraint is still violated when an adverbial occurs after the complementizer．The result for constraint（56）is also the same as constraint（55）for the Relative Clause Paradox．Again，assuming the standard treatment of relative clauses in（59），the subject of the complementizer＇s f－structure weakly f－precedes the comple－ mentizer＇s f－structure because it is vacuously true that all of the subject＇s c－structure correspondent（it has none）precedes some（in fact，all）of the complementizer＇s f－structure＇s c－structure correspondent． Again，the constraint is violated by even a simple relative clause example like（58）．

The alternative f－precedence relation to strong f－precedence is typically given a different formula－ tion than the one given in（51），which I have called weak f－precedence．The standard alternative to strong f－precedence，as discussed in Dalrymple（2001：171－174），is：

## （60）F－precedence（edge－based）

F－structure $f$ f－precedes f－structure $g\left(f<_{f_{\mathrm{RR}}} g\right)$ if and only if for the rightmost $n_{1} \in \phi^{-1}(f)$ and for the rightmost $n_{2} \in \phi^{-1}(g), n_{1}$ c－precedes $n_{2}$ ．

Edge－based f－precedence cannot be satisfied by null pronominals，because no null pronominal has a rightmost node in its c－structure correspondent．

The Relative Clause Paradox therefore constitutes a case in which weak f－precedence and edge－ based f－precedence make different predictions．Constraint（56）is violated by relative clauses such as（58），as outlined above，but the equivalent constraint with edge－based f－precedence would not be violated by（58），because the null pronominal subject in fact does not f－precede the complementizer＇s f －structure，since the null pronominal has no rightmost node in c－structure．Edge－based f－precedence thus captures the Relative Clause Paradox and basic CAE effects，but not the Adverb Effect．${ }^{13}$

Table 1 provides a general overview of some differences between alternative f－precedence relations． $A$ and $B$ are $f$－structures．The symbol $\emptyset$ represents an $f$－structure with no $c$－structure correspondent；i．e． a c－structurally unrealized grammatical function． $\mathrm{A}_{1} \ldots \mathrm{~A}_{2}$ represents an f －structure that is mapped from disjoint parts of c－structure；i．e．what Bresnan（1995）calls a＂scattered constituent＂．The first two columns correspond to the situation of a null pronominal preceding its binder or vice versa，as discussed in Bresnan（2001：193－195）and Dalrymple（2001：173－174，288－289）．The second two columns cor－ respond to the situation of weak crossover with respect to a realized or null pronominal，as discussed with respect to the linear order condition on operator binding by Bresnan（1994，1995，2001）．

|  | D．．．A | A ．．．$\square^{\text {¢ }}$ | $\mathrm{A}_{1} \ldots \mathrm{~B} \ldots \mathrm{~A}_{2}$ | $\mathrm{A}_{1} \ldots \emptyset \ldots \mathrm{~A}_{2}$ |
| :---: | :---: | :---: | :---: | :---: |
| Strong f－precedence（ $<f_{\text {f＊}}$ ） | $\emptyset<_{f \forall \vdash} \mathrm{~A}$ | A $<_{f_{\forall b}} \emptyset$ | $\mathrm{A}_{1} \chi_{f 6 \forall} \mathrm{~B}$ | $\mathrm{A}_{1}<_{f v *} \emptyset$ |
|  |  |  | $\mathrm{B}<_{f ⿰ 丬} \mathrm{~A}_{2}$ | $\emptyset<_{f \forall 甘} \mathrm{~A}_{2}$ |
| Weak f－precedence（ $<_{f_{\nsucceq \exists} \text { ）}}$ | $\emptyset<_{f_{\forall \exists}} \mathrm{A}$ | A $\chi_{f_{w b} \emptyset} \emptyset$ |  | $\mathrm{A}_{1} \chi_{f_{\text {扫 }} \emptyset}$ |
|  |  |  | $\mathrm{B}<_{f_{\nvdash \exists}} \mathrm{A}_{2}$ | $\emptyset<_{f_{\nsucceq \exists}} \mathrm{A}_{2}$ |
| Edge－based f－precedence（ $<f_{\text {RR }}$ ） | $\emptyset ぬ_{f_{R R}} \mathrm{~A}$ | A $f_{f_{\text {RR }}} \emptyset$ | $\mathrm{A}_{1} \chi_{f_{\mathrm{RR}}} \mathrm{B}$ | $\mathrm{A}_{1}{\nless f_{\mathrm{RR}} \emptyset}$ |
|  |  |  | $\mathrm{B}<f_{\text {RR }} \mathrm{A}_{2}$ | $\emptyset ぬ_{f R \mathrm{R}} \mathrm{A}_{2}$ |

Table 1：Alternative definitions of f－precedence and some outcomes for unrealized grammatical func－ tions and scattered constituents

[^41]
## References

Asudeh, Ash. 2006. Direct Compositionality and the Architecture of LFG. In Miriam Butt, Mary Dalrymple, and Tracy Holloway King, eds., Intelligent Linguistic Architectures: Variations on Themes by Ronald M. Kaplan, 363-387. Stanford, CA: CSLI Publications.
Asudeh, Ash, and Ewan Klein. 2002. Shape Conditions and Phonological Context. In Frank van Eynde, Lars Hellan, and Dorothee Beermann, eds., Proceedings of the 8th International HPSG Conference, 20-30. Stanford, CA: CSLI Publications.

Bögel, Tina, Miriam Butt, Ronald M. Kaplan, Tracy Holloway King, and John T. Maxwell, III. 2009. Prosodic Phonology in LFG: A New Approach. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG09 Conference. CSLI Publications. This volume.

Borsley, Robert D. 1999. Mutation and Constituent Structure in Welsh. Lingua 109(267-300).
Bresnan, Joan. 1972. Theory of Complementation in English Syntax. Ph.D. thesis, MIT, Cambridge, MA. Reprinted as Bresnan (1979).
-. 1977. Transformations and Categories in Syntax. In Robert E. Butts and Jaakko Hintikka, eds., Basic Problems in Methodology and Linguistics. Part Three of the Proceedings of the Fifth International Congress of Logic, Methodology and Philosophy of Science.
—. 1979. Theory of Complementation in English Syntax. New York: Garland.
-. 1984. Bound Anaphora on Functional Structures. Presented at the Tenth Annual Meeting of the Berkeley Linguistics Society.
—. 1994. Linear Order vs. Syntactic Rank: Evidence from Weak Crossover. In Katie Beals, Jeannette Denton, Bob Knippen, Lynette Melnar, Hisami Suzuki, and Erika Zeinfeld, eds., CLS 30-1: Papers from the Thirtieth Regional Meeting of the Chicago Linguistic Society. Chicago: Chicago Linguistic Society.
—. 1995. Linear Order, Syntactic Rank, and Empty Categories: On Weak Crossover. In Dalrymple et al. 1995, 241-274.
—. 2001. Lexical-Functional Syntax. Oxford: Blackwell.
Browning, Marguerite A. 1996. CP Recursion and that-t Effects. Linguistic Inquiry 27(2): 237-255.
Butt, Miriam, and Tracy Holloway King. 1998. Interfacing Phonology with LFG. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG98 Conference. Stanford, CA: CSLI Publications. Carnie, Andrew. 2000. On the Definition of $\mathrm{X}^{0}$ and XP. Syntax 3(2): 59-106.
Chomsky, Noam. 1981. Lectures on Government and Binding. Dordrecht: Foris.
—. 1986. Barriers. Cambridge, MA: MIT Press.
—. 1995. The Minimalist Program. Cambridge, MA: MIT Press.
Chomsky, Noam, and Howard Lasnik. 1977. Filters and Control. Linguistic Inquiry 8: 425-504.
Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2009. XLE Documentation. Palo Alto Research Center, Palo Alto, CA.
http://www2.parc.com/istl/groups/nltt/xle/doc/xle_toc.html.
Culicover, Peter W. 1991. Topicalization, Inversion, and Complementizers in English. In Denis Delfitto, Martin Everaert, Arnold Evers, and Frits Stuurman, eds., Going Romance and Beyond, OTS Working Papers. Utrecht: University of Utrecht.
—. 1993. Evidence Against ECP Accounts of the that-t Effect. Linguistic Inquiry 24: 557-561.
Dalrymple, Mary. 2001. Lexical Functional Grammar. San Diego, CA: Academic Press.
Dalrymple, Mary, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, eds. 1995. Formal Issues in Lexical-Functional Grammar. Stanford, CA: CSLI Publications.
Emonds, Joseph. 1976. A Transformational Approach to English Syntax: Root, Structure-Preserving, and Local Transformations. New York: Academic Press.

Falk, Yehuda. 2000. Pivots and the Theory of Grammatical Functions. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG00 Conference, 122-138. Stanford, CA: CSLI Publications.
—. 2001. Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax. Stanford, CA: CSLI Publications.
—. 2006. Subjects and Universal Grammar: An Explanatory Theory. Cambridge: Cambridge University Press.
Forst, Martin, and Ronald M. Kaplan. 2006. The Importance of Precise Tokenizing for Deep Grammars. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC 2006).
Gazdar, Gerald. 1981. Unbounded Dependencies and Coordinate Structure. Linguistic Inquiry 12: 155-184.
Ginzburg, Jonathan, and Ivan A. Sag. 2000. Interrogative Investigations: The Form, Meaning and Use of English Interrogatives. Stanford, CA: CSLI Publications.
Halvorsen, Per-Kristian, and Ronald M. Kaplan. 1988. Projections and Semantic Description in Lexical-Functional Grammar. In Proceedings of the International Conference on Fifth Generation Computer Systems, 1116-1122. Institute for New Generation Systems, Tokyo. Reprinted in Dalrymple et al. (1995:279-292).
Ishii, Toru. 2004. The Phase Impenetrability Condition, the Vacuous Movement Hypothesis, and that-t Effects. Lingua 114: 183-215.
Kameyama, Megumi. 1985. Zero Anaphora: The Case of Japanese. Ph.D. thesis, Stanford University.
-. 1989. Functional Precedence Conditions on Overt and Zero Pronominals. Technical report, Microelectronics and Computer Technology Corporation.
Kandybowicz, Jason. 2006. Comp-Trace Effects Explained Away. In Donald Baumer, David Montero, and Michael Scanlon, eds., Proceedings of the 25th West Coast Conference on Formal Linguistics, 220-228. Somerville, MA: Cascadilla Proceedings Project.
—. 2009. Embracing Edges: Syntactic and Phono-Syntactic Edge Sensitivity in Nupe. Natural Language and Linguistic Theory 27: 305-344.
Kaplan, Ronald M. 1987. Three Seductions of Computational Psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson, and Paul Bennett, eds., Linguistic Theory and Computer Applications, 149-181. London: Academic Press. Reprinted in Dalrymple et al. (1995:339-367).
—. 1989. The Formal Architecture of Lexical-Functional Grammar. In Chu-Ren Huang and Keh-Jiann Chen, eds., Proceedings of ROCLING II, 3-18. Reprinted in Dalrymple et al. (1995:7-27).
Kaplan, Ronald M., and Joan Bresnan. 1982. Lexical-Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan, ed., The Mental Representation of Grammatical Relations, 173-281. Cambridge, MA: MIT Press. Reprinted in Dalrymple et al. (1995:29-135).
Kaplan, Ronald M., and Annie Zaenen. 1989. Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty. In Mark Baltin and Anthony Kroch, eds., Alternative Conceptions of Phrase Structure, 17-42. Chicago, IL: University of Chicago Press. Reprinted in Dalrymple et al. (1995:137-165).
Kayne, Richard S. 1981. ECP Extensions. Linguistic Inquiry 12(1): 93-133.
Levine, Robert D., and Thomas E. Hukari. 2006. The Unity of Unbounded Dependency Constructions. Stanford, CA: CSLI Publications.
Lohndal, Terje. 2009. Comp-t Effects: Variation in the Position and Features of C. Studia Linguistica 63(2): 204-232.
Merchant, Jason. 2001. The Syntax of Silence. Oxford: Oxford University Press.
Partee, Barbara H., Alice ter Meulen, and Robert E. Wall. 1990. Mathematical Methods in Linguistics. Dordrecht: Kluwer.

Perlmutter, David M. 1968. Deep and Surface Structure Constraints in Syntax. Ph.D. thesis, MIT, Cambridge, MA. Reprinted as Perlmutter (1971).
—. 1971. Deep and Surface Structure Constraints in Syntax. New York: Holt, Rinehart and Winston.
Pesetsky, David. 1982. Paths and Categories. Ph.D. thesis, MIT.
Pesetsky, David, and Esther Torrego. 2001. T-to-C Movement: Causes and Consequences. In Michael Kenstowicz, ed., Ken Hale: A Life in Language, 355-426. Cambridge, MA: MIT Press.
Pollard, Carl, and Ivan A. Sag. 1987. Information-Based Syntax and Semantics. CSLI Publications.
-. 1994. Head-driven Phrase Structure Grammar. Chicago, IL and Stanford, CA: The University of Chicago Press and CSLI Publications.
Pullum, Geoffrey K., and Arnold M. Zwicky. 1988. The Syntax-Phonology Interface. In Frederick J. Newmeyer, ed., Linguistics: The Cambridge Survey, vol. 1, 255-280. Cambridge: Cambridge University Press.
Rizzi, Luigi. 1990. Relativized Minimality. Cambridge, MA: MIT Press.
-. 1997. The Fine Structure of the Left Periphery. In Liliane Haegeman, ed., Elements of Grammar, 281-337. Dordrecht: Kluwer.
Ross, John R. 1967. Constraints on Variables in Syntax. Ph.D. thesis, MIT.
Shlonsky, Ur. 1988. Complementizer-Cliticization in Hebrew and the Empty Category Principle. Natural Language and Linguistic Theory 6: 191-205.
Sobin, Nicholas. 1987. The Variable Status of Comp-Trace Phenomena. Natural Language and Linguistic Theory 5: 33-60.
-. 2002. The Comp-Trace Effect, the Adverb Effect and Minimal CP. Journal of Linguistics 38: 527-560.
Tallerman, Maggie. 2009. Phrase Structure vs. Dependency: The Analysis of Welsh Syntactic Soft Mutation. Journal of Linguistics 45: 167-201.
Toivonen, Ida. 2001. The Phrase Structure of Non-Projecting Words. Ph.D. thesis, Stanford University. —. 2003. Non-Projecting Words: A Case Study of Swedish Verbal Particles. Dordrecht: Kluwer.
Tseng, Jesse. 2003. Edge Features and French Liaison. In Jong-Bok Kim and Stephen Wechsler, eds., Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar, 313-333. Stanford, CA: CSLI Publications.
Wescoat, Michael T. 2002. On Lexical Sharing. Ph.D. thesis, Stanford University.
-. 2005. English Nonsyllabic Auxiliary Contractions: An Analysis in LFG with Lexical Sharing. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG05 Conference, 468-486. Stanford, CA: CSLI Publications.
-. 2007. Preposition-Determiner Contractions: An Analysis in Optimality-Theoretic LexicalFunctional Grammar with Lexical Sharing. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG07 Conference, 439-459. Stanford, CA: CSLI Publications.
-. 2009. Udi Person Markers and Lexical Integrity. In Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG09 Conference. CSLI Publications. This volume.
Zaenen, Annie, and Ronald M. Kaplan. 1995. Formal Devices for Linguistic Generalizations: West Germanic Word Order in LFG. In Dalrymple et al. 1995, 215-239.
Zwicky, Arnold M. 1985. Rules of Allomorphy and Phonology-Syntax Interactions. Journal of Linguistics 21(2): 431-436.
—. 1986. The General Case: Basic Form versus Default Form. In Vassiliki Nikiforidou, Mary VanClay, Mary Niepokuj, and Deborah Feder, eds., Proceedings of the Berkeley Linguistics Society, vol. 12, 305-314. Berkeley: Berkeley Linguistics Society.

# EXPLOITING XLE's FINITE STATE INTERFACE IN LFG-BASED STATISTICAL MACHINE TRANSLATION 

Eleftherios Avramidis and Jonas Kuhn<br>Linguistics Department, University of Potsdam<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

We present the addition of a morphological generation component to an LFG-based Statistical Machine Translation System, taking advantage of existing morphological grammars and the FST (Finite State Transducer) processing pipeline of the XLE system. The extended syntax-driven translation system takes separate stochastic decisions for lemmata and morphological tags; the role of finite-state morphological grammars is to generate full forms out of a bundle of morphological tags produced by the translation component. This technique can lead to a more effective use of a given amount of training data from a parallel corpus, since lexical vs. morphosyntactic translation patterns can be induced independently.

The existing FST processing cascade for German, when added to the Statistical Machine Translation System, suffers from generation failures. These occur due to overgeneralisation by the syntax-driven translation process and originate from (i) the use of various underspecification tags in the morphological grammar, or (ii) erroneous assignment of certain tags to a given lemma. In order to deal with this, we add a set of replacement/correction rules on top of the cascade. The augmented FST cascade leads to an increase of generation coverage from $47.90 \%$ to $75.35 \%$. A detailed error analysis for the remaining $24.65 \%$ is given.


## 1 Introduction

In current work on Machine Translation (MT), purely data-driven, statistical approaches, based on very large corpora of sample translations, continue to lead to the best evaluation results, at least when tested on the same text domain as they were trained on (Callison-Burch et al., 2008). At the same time, it is conceptually clear that there are limitations to picking up certain generalisations (which can be easily described in linguistic terms) from unstructured training data - Zipf's law has it that the multitude of types of linguistic units occur rather infrequently in corpus data. Hence, an obvious goal for linguistically grounded natural language processing (NLP) research is to find effective combinations of the highly successful statistical techniques with insights from deep linguistic processing. This goal is considerably more challenging than one may first think: nearly all previous experiments on the straightforward ways of constraining the statistical models to apply only on linguistically warranted units have led to a drop in performance (e.g., Koehn et al. (2003a); Chiang (2005)). This is presumably so because the unconstrained system will quite often learn to produce a reasonable translation for some combination of words that does not form a linguistic unit at any level. The development of more structured statistical translation models, capable of incorporating linguistic knowledge while not suffering from a reduced amount of training data, remains a major goal for NLP research for the next years.

[^42]In this paper, we focus on the so-called "generation issues" in data-driven translation. These issues arise when translating from a morphologically poor language (e.g. English), into a language that requires complex morphosyntactic rules to be taken into account (e.g. German). The purely statistical MT systems have very limited capabilities of inducing the generalizations behind the morphosyntactic patterns of a language. This occurs as they rely on statistically trained word alignments, which were trained on a parallel corpus. For a given portion of text in the source language, the word-aligned word sequences in the target language are considered as translation candidates. The candidate word sequences are then assembled, mainly on the basis of statistical (n-gram) language models, which assign higher scores to typical word sequence patterns in the target language. Patterns involving high-frequency items will typically be reflected in the language model, but the patterns are not represented systematically and cannot be generalised to lower-frequency items.

The statistical translation approach which we build on (Galley et al., 2004; Hopkins and Kuhn, 2007b) has the ability to separate lexical from morphosyntactic effects during training, since it is driven by a rich syntactic source language analysis (c-structure augmented by features from f-structure). However, it can only produce the specific target language word forms that are included in the training data. This approach suffers to some degree from similar generation issues as the pure statistical MT approach. Given that high-quality morphological analysers exist for many languages, we can see a reasonable extension of this approach: We assume not only a syntactic analysis of the source language, but also a (disambiguated) morphological analysis of the target language. In this paper, we present such an extension building on the English and German resources from the ParGram project (Butt et al., 2002) and focus on the steps needed to ensure robustness of the resulting overall system. Specifically, the cascade of Finite State Transducers is adapted in order to fit the requirements of MT generation.

In the remainder of section 1, a more detailed motivation for morphologically informed generation in data-driven MT is given. In section 2, we first sketch the broader research framework in which our experimental statistical LFG-based translation system is situated, referring to related work and the potential of using LFG in statistical MT. We also present the translation approach that we are building on and show how morphological generation can be integrated straightforwardly in the statistical modelling and combined with standard FSTs. In Section 3 we address issues that arise in the use of specific existing finite state morphological analysers and we describe the adaptation methods employed. Section 4 presents the experiment set-up for an English-to-German translation scenario, evaluation results of system coverage and an error analysis. In section 5 , we briefly discuss future directions of our work, before closing with a short conclusion in section 6 .

### 1.1 Motivation: Morphology issues in Statistical MT

LFG is an excellent candidate for exploring sophisticated ways of combining statistics and deep linguistic analysis, thanks to the assumption of parallel correspondence across levels (Riezler and Maxwell, 2006; Hopkins and Kuhn, 2007a). In the case of the present proposal, we follow the LFG-based statistical translation approach of Hopkins and Kuhn (2007a,b), which already exploits c-structure and f -structure information in the source language. We then take advantage of the syntax-morphology interface of an LFG grammar for the target language, which acts as a reliable tool for disambiguating the options that a morphological analyser of the target language produces. ${ }^{1}$ This seems to be a useful tool for an exploitation of linguistic generalizations in data-driven MT: when translating from English into a language that is morphologically more challenging (in our case, German), a linguistically motivated morphological analyser can break down word forms into a lemma and a particular set of morphosyntactic features (e.g., \{Mann +NN . Masc . Gen . Sg\} for the form Mannes ('man's')).

Lexical generalizations can then be learned for a lemma and generalised to other forms than the ones seen in training. For instance (fig. 1), the system may have learned two things:
(a) starke Unwetter (lit. 'strong un-weathers') is a good translation for 'heavy windstorms', and
(b) an after-PP in English should be translated as a nach-PP in German, where the determiner and attributive adjectives should occur in their dative form (this generalization could have been picked up from examples like 'after a long game' $\rightarrow$ nach einem langen Spiel).

From these two patterns, the system will for instance be able to infer correct translations for the phrases 'a heavy windstorm' as ein starkes Unwetter and 'after a heavy windstorm' as nach einem starken Unwetter, even if the respective form of the adjective did not occur in the respective context in the training data - in fact even if the specific form never occurred in the training data at all.

The benefits to the system can be also interpreted as a way to make clearer translation decisions, if we consider the stochastic background of how the words in the two languages are automatically aligned: In the example above, there would be more than 5 candidates for translating the article ' $a$ ', which consist of the German indefinite article in various variations of genders and cases. As a simplified example, a pure word-to-word statistical translation model (Brown et al., 1990) would in principle handle this in the same way as a lexical ambiguity; it would create a set of translation candidates and each of these candidates would be assigned

[^43]

Figure 1: Using the a morpheme generator allows for better coverage for unknown inflections
a probability, depending on how many times it appears in the training corpus as translation of 'a' (its relative frequency). This appears conceptually weak, as it is obvious that the decision on the correctly inflected form has nothing to do with the relative frequency of this particular form, but is dependent on rules of syntax and agreement in the processed sentence. In our example, if ein appears more often, it would have a higher translation probability and therefore would be more likely to be chosen, even if it grammatically should not.

The state-of-the-art systems (following Koehn et al. (2003b)) have reduced this drawback by using multi-word units (the so-called phrases) and language models which penalise translation sequences which are non-fluent. Nevertheless, the issue can still be complex on the lexical level, as the previously mentioned candidate list may contain both candidates for lexical ambiguities and morphosyntactic inflections. For example, the translation candidates list for the English word bank would contain all noun case variations \{Bank (nominative/genitive - prob. $40 \%$ ), Flussufer (nominative - prob. 35\%), Flussufers (genitive - prob. 25\%) \}, but the probability of the most frequent lexical decision is split into two separate hypotheses. When applying the suggested idea, by adding a separate morphology layer, we would reduce this list to \{Bank (prob. 30\%), Flussufer (prob. $60 \%)\}$, and consequently make the decision on the noun case at a separate stage, with the possibility of considering syntax information, provided from a separate layer of the LFG analysis.

## 2 Building a morphologically informed system

### 2.1 Existing work

The idea of augmenting the generation process, when translating into morphologically complex languages, has already been applied to purely statistical systems. Koehn and Hoang (2007), Toutanova et al. (2008) prefer to translate on lemmata and consequently train a separate generation process by using morphological and syntactic factors/features. Other approachess include the use of information from
the source-side syntax, aiming to improve the morphology issues at the target side (Minkov et al., 2007; Avramidis and Koehn, 2008). Much research has also been motivated by the needs of agglutinative languages, such as Turkish (El-Kahlout and Oflazer, 2006; Oflazer, 2008), where integrating morphotactic knowledge at the generation stage appears to be essential for creating a fluent output.

### 2.2 The "tree labelling" approach

For reasons explained in section 1, we attempt to approach the issue in a more linguistically motivated approach, where LFG is the structural backbone of the translation process. We build on top of a statistical tree-to-string translation approach as in Hopkins and Kuhn (2007b), the "PTOLEMAIOS approach", working with the XLE system and the grammars developed in the ParGram project (Butt et al., 2002) and a parallel corpus, word-aligned with GIZA++ (Och and Ney, 2003). Information from the source language LFG analysis drives a "tree labelling" approach to translation: a cascade of statistical (discriminative) classifiers is trained, that traverses the c-structure analysis, taking into account f-structure information and all previous decisions. The new "labels" assigned to the source c-structure tree will contain target language word forms and tree re-structuring instructions (which can have the effect of changing the word order), so a particular target language string can be read off the final tree.

The training process is characterised by the following steps:
(1) Get the XLE parse of the source sentence (e.g. English), add indices for accessing the f-structure information.
(2) For every leaf node, get the corresponding target word from the word alignment with the target (e.g. German) sentence.
(3) Based on the graph structure of the resulting tree/word alignment structure, it is possible to determine a set of "frontier nodes" among the non-terminal nodes, following Galley et al. (2004). The tree/word alignment sub-graphs rooted by these frontier nodes can be used as the building blocks for syntactically informed statistical translation.
(4) Traverse the c-structure tree top-down. In training, we simulate a decision process that subsequently assigns various labels to each tree node. The labels reflect the information needed to reconstruct the full tree/word alignment structure, given only the original source language analysis and the result of previous decisions (e.g., on the mother node). Complex decisions are broken up into simple partial decisions, reflected by sub-labels on the node (For example: Should the node be in the frontier set? Should there be discontinuous parts in the resulting target string? What is the target language word that should be used as a translation for cooperation? Should the translation of the right-most daughter precede the translation of the daughter previously translated? etc.).


Figure 2: Basic training and decoding process

Each sub-step can be characterised as a discriminative classification decision, for which the training data include all the learning features and the correct outcome (label). The learning feature/label combinations for all sub-decisions are collected for the entire training corpus.
(5) The learning feature/label combinations are used for training a (large) set of specialised statistical classifiers that are able to generalise over similar situa-
tions in the top-down tree traversal process. (This involves a very sophisticated back-off technique to ensure that each classifier is based on a sufficiently large sample of evidence.) The resulting cascade of statistical classifiers represents a full tree-to-string translation model. ${ }^{2}$

In the decoding process, i.e., when the model is applied in order to translate a given source language sentence, there is obviously no target language string in the input. This means that step (1) of the training procedure is performed; steps (2) and (3) cannot be performed. This means that the top-down tree traversal of step (4) is performed as a real, cascaded labelling decision process on the nodes (not just as a simulation as in training). ${ }^{3}$ The resulting node labels can be used to determine the set of target language word (predicted by the translation model). Then, labels referring to the relative order of the graph fragments indicate the predicted word order.

### 2.3 Adding the morphology interface

It is conceptually rather simple to augment the cascade of statistical classifiers just described in order to include further labelling decisions. This makes it very straightforward to move away from generation of full word form strings on the target side. Instead, this can be replaced by a more flexible step-by-step generation of lemma information and morphological tags, as they can be used by a finitestate morphological generator. Rather than using full word forms in the tree labels, the first step is just to generate a lemma. The morphological tag specification is then added in separate classification steps, so it can take all available information into account; this information may contain agreement information, based on the analysis done on previously generated words, but may also take advantage of the syntactic analysis of the source sentence, e.g. in order to assign the proper case to the direct and indirect objects.

Whereas the original Ptolemaios approach applies a ParGram LFG grammar to the source language (in our case English) in order to perform the tree labeling, we also parse the target language (German). Since XLE incorporates finite-state transducers (FSTs) for preprocessing (tokenisation) and morphological analysis, the German parses contain a syntactically disambiguated morphological analysis for all words. This is exactly what is needed as training material for the extended tree labelling approach we just described: instead of full form like starke Unwetter, we use the following representation of the target language words to train the tree labeler: \{stark +ADJ. .Pos.MFNOnly.NA. .Pl . St \} $\{$ Unwetter +NN. Neut. NGA. . Pl \}. Here, these morphemes are syntactically disambiguated, in the sense that, even if another morphological analysis of

[^44]

Figure 3: Training and decoding process, after adding the separate morphology layers. Note that here, ich is stemmed to Sie, because Sie has been chosen by the authors of the German LFG Grammar as the citation form of the personal pronoun
the current form would be possible in different contexts, we are getting only the combination of morphemes that matches the most probable syntactic parse. ${ }^{4}$

After training is finished, in application/decoding mode, the tree labelling translator is applied to new input (a set of unseen English sentences) as follows: the English LFG parser is used to produce the most probable c-structure tree, again with information from the f-structure attached on its nodes. The cascade of statistical classifiers is then applied to add the translation labels to this tree, which are then read out to produce a string of lemmata and morphological tags. In the plain tree-to-string approach, the process was finished at this point. Now, we have to perform one last step: the string of lemmata and morphological tags is fed into the target language morphological analyser (run in reverse mode, i.e., as a morphological generator, which is straightforward in finite-state technology). ${ }^{5}$

## 3 Adapting the morphology interface

The previous section showed that in principle, the tree labelling approach can be straightforwardly extended to produce not just a string of word forms, but a sequence of lemmata and morphological tags used as input for standard FST morphological generation. However, a set of issues arises when this approach is used for a specific morphological grammar, like the one for German used as part of the German ParGram LFG grammar (based on the work by Schiller and Steffens (1990)). In this section, we present the issue and our approach to deal with it in a systematic way.

### 3.1 The compact underspecified feature format

Using a typical general-purpose morphological analyser for morphologically rich languages such as German, in a different application context than it was originally designed for, may quite naturally lead to complications. Specifically, any pipeline that includes some "soft"/machine learning component feeding the analysis level of the morphological grammar may pose systematic problems. Here we observe this type of problem regarding the set-up of the German morphological grammar, but we present a straightforward solution in the subsequent sections.

To understand the issue, it has to be noted that the feature representations used within the morphological analysers (Schiller and Steffens, 1990) rely on a compact underspecified feature format in order to avoid a proliferation of disjunctive analyses for ambiguous word forms. For instance, the form Mann ('man') can be either nominative, dative, or accusative singular (only the genitive singular differs:

[^45]Mannes). The morphological grammar assigns the following analysis to Mann: $\{$ Mann +NN. Masc. NDA . Sg \}. The case tag . NDA combines the tags for nominative, dative and accusative in one compact tag. Other singular nouns are case ambiguous for all four cases, e.g., Frau ('woman'), which is assigned the case tag . NGDA. Similar tag combinations occur for other morphosyntactic features, such as gender, number, and mood.

This compact feature representation leads to the following issue in translation: as the assignment of labels is trained from output of the morphology, the system will of course pick up generalizations that involve combined tags like . NDA. It may turn out however that the translator ends up using such a tag with a lemma that has a slightly different inflection paradigm (e.g., producing \{Frau +NN . Fem . NDA . Sg \} instead of $\{$ Frau +NN . Fem . NGDA . Sg \} ). Running the incorrect sequence through the morphological generator will result in a failure.

One may argue that we should try to improve the training so the system will learn to only produce "legal" sequences. However, even if this worked, it would unnecessarily reduce the effectiveness of the training with a given amount of data. ${ }^{6}$ It seems much more appropriate to take advantage of the available linguistic knowledge about morphological regularities in the form of morphological analysers and use this to fix the issues.

### 3.2 The "correction" module

It is relatively straightforward to augment the pre-processing FSTs used in XLE with a "correction" module: using the FST composition operation, we can map combined tags like. NDA to other, overlapping combined tags like . NGDA, operating in two stages. Hence, a new "recombination" FST is defined, by adding a set of replace rules on top of the existing deep morphology FST, without requiring any modification of the latter.

These extra replace rules could be seen as a preprocessing step for the queries that are fed to the generator. They were written manually with regard to the particular morpheme/part-of-speech categories that use a compact representation for ambiguous word forms. Accordingly, their aim is to avoid generation failures, dealing with cases when a probabilistically guessed morpheme does not exactly match the compact morpheme tag expected by the compiled morphology FST. Then they should therefore lead to at least one more compact or more generalised tag, containing the one requested, that could end up in a successful generation. In particular, this task is addressed by:
(a) explicating the combined tags towards their component features (e.g., replacing . NDA with . Nom, . Dat or .Acc, disjunctively) and then
(b) generalizing these in order to get a disjunction of all the (other) possible tag combinations that may contain them.

[^46]

Figure 4: Example of a successful generation in an enhanced Finite State Transducers cascade. Whereas Frau would previously be generated only if . NGDA was given, the replace rules can adapt an input such as .NDA so that it is accepted

This way, the desired tags would be taken into consideration, even if they are more or less explicit than the expected. Both stages are compiled out in the resulting recombination FST.

For the German grammar, apart from the noun cases, which have been explained above for means of illustration, rules were written for compact tags referring to verb persons, numbers, genders, moods, and adjective predicate markers.

With the described generalizations, the generator is essentially tuned to overgenerate, in the sense that it will produce all partial tags for a given compact tag (e.g., nominative, genitive, dative and accusative for . NGDA), even if the lemma that the tag is attached to does not have the same form for all the feature values. This is intentional since it allows for the desired degree of robustness, i.e., the cascade will typically produce at least one result even for input that would have been incompatible with the original morphological transducer. Since the preprocessing transducers are composed (or cascaded) with the actual morphological grammar transducer, the linguistic knowledge encoded in the latter will constrain the overgeneration. In almost all cases this will have the desired effect, i.e., the correct solution will be included among the solutions. However, it cannot be excluded that an unfortunate combination of overgeneration steps will lead to an incorrect result. What is quite typical is that more than one solution is produced disjunctively. There are ways for the statistical system to choose with some confidence between the alternatives at a later stage (e.g., by scoring the formed phrases with a language model that takes the left and right context in the target language into account).

### 3.3 Facing incorrect assignments

In the previous section, we addressed cases in which the use of convenient "underspecification tags" in the morphological grammar for ambiguous word forms can lead to issues in the translation-driven construction of the input to a finite-state transducer. One could argue that in the generation of a . NDA tag instead of a . NGDA tag in translation is not really a mistake, but what we see is a representational issue.

However, in some cases, the step-by-step generation of morphological tags performed in translation may lead to an incorrect assignment of unambiguous feature tags. Since the statistical system has no explicit knowledge of the gender of the nouns, but instead makes predictions based on a wide range of features, it would be possible to assign the tag . Fem to a noun that is actually masculine, e.g., leading to $\{$ Mann + NN .Fem .NDA . Sg $\}$. In such cases, even if most nouns have no flexibility in changing their gender and therefore such a specification in the generation process seems redundant, the nature of the morphology FST would lead it into


Figure 5: Example of two successful generations in an enhanced Finite State Transducers cascade, extending the one shown at Figure 4. Here, Zusammenarbeit (a feminine noun) is generated, although a tag for the masculine form of it has been incorrectly decided by the statistical system.
a generation failure. It is clearly desirable to rely on the morphological grammar for overriding such incorrect feature markings to make the system more robust.

Of course, we only want to change a feature like . Fem into .Masc in situations where the former analysis is indeed incompatible with the morphological grammar. Adding such a correction to the morphology cascade which was described in the previous section, would correct the issue concerning the nouns, but would cause problems to other parts of speech, for whom the gender information is indeed useful in order to choose within inflection options.

In order to achieve such a flexible manipulation, we actually need a model with two cascades: (a) one with the core of the correction module (section 3.2) without any gender alterations, and (b) an alternative one which takes effect only when the main one fails to generate. This gives all possible gender alternatives as alternative morphemes for the generation failures on the nouns. As mentioned, (b) should only be applied to input (a lemma/tag sequence) for which (a) fails.

### 3.4 Priority Union

The finite-state operation of priority union (Guingne et al., 2003) can be used to this effect (as a unification operation, it was proposed by Kaplan (1995)). By combining two FSTs with priority union (Figure 5), the second FST is only applied to a given input in case the input is not included in the upper side of the first FST. For instance, we may in general apply the mentioned "recombination" FST, and if this combination does not lead to a result, we prefix an additional feature correction FST:

$$
\begin{align*}
& \left(T_{\text {recomb }} \circ T_{\text {morph }}\right) \bigcup^{p}\left(T_{\text {correct }} \circ T_{\text {recomb }} \circ T_{\text {morph }}\right)  \tag{1}\\
& \equiv\left(T_{\text {recomb }} \circ T_{\text {morph }}\right) \cup\left(\neg \text { upper }\left(T_{\text {recomb }} \circ T_{\text {morph }}\right) \circ\left(T_{\text {correct }} \circ T_{\text {recomb }} \circ T_{\text {morph }}\right)\right) \tag{2}
\end{align*}
$$

where $T_{\text {morph }}$ is the existing morphology generator, $T_{\text {recomb }}$ is the tag recombination transducer and $T_{\text {correct }}$ is an FST cascade for substituting tags that may fail during the first generation.

## 4 Evaluation

The language pair that our experiments focus on is English to German. This pair, in this translation direction, is a good example for disproportional morphology, as German is much more inflected than English. In addition, XLE parsers with the desired morphology features were fully available to us for both languages.

The main focus of our evaluation was how well the morphology interface was adapted to the generation stage of our statistical system. Therefore, we had to measure the improvement in generation coverage. This can be seen as the number of the generations that succeeded, divided by the total number of generations requested. Evaluation of the full translation system will be presented in future work.

| System | Generation coverage |
| :--- | ---: |
| only morphology FST | $47.90 \%$ |
| compact tags correction | $60.41 \%$ |
| gender correction | $75.35 \%$ |

Table 1: Generation coverage, with the various adaptations of the morphology module

### 4.1 Experiment

The experiment was run on a small, simplified set of the Europarl corpus version 4 (Koehn, 2005). The training set contained 20,000 sentences which had less than 10 words, whereas the untranslated evaluation set contained 1,000 sentences of the same length. The percentages, shown in Table 1, are given based on a proportion over 5510 generation requests.

The results in table 1 show that with the techniques demonstrated, the success rate has been raised from $47.90 \%$ to $75.35 \%$, in the full system. The extra correction level for the gender correction by itself was able to improve the coverage by $14.94 \%$, which confirms that the problem was quite critical.

### 4.2 Error analysis

As the results show, there is still a considerable $24.65 \%$ of failures taking place even when all of the above corrections are applied. A first detailed evaluation was performed manually in order to further investigate the actual cause of the failures. It became clear that many failures had common reasons: a more concrete categorisation of approximately $70 \%$ of the errors has successfully been traced with regular expressions, whereas the "Wrong POS" category was estimated based on a smaller manually evaluated subset.

The outcome of the analysis is shown in Table 2, in which the percentages sum up to the $24.65 \%$ questioned. What is identified as a major cause of failures, includes:
(a) the predefined behaviour of the statistical part of the system, which does not always provide the full set of required morphemes. As this has been the most robust solution, the statistical system first decides the categories of the morphemes that a word may be assigned and then makes a decision for each morpheme value. However, FST allows the morpheme order for a small set of words to vary, especially when these words are generated by combining other smaller words. In this error category we would count complex prepositions (like gegenüber, daraus), prepositions with fused articles (zur, im), compounds (Parkordnung etc.) and some other forms which appear as articles or personal pronouns.
(b) wrong POS behaviour (e.g. when a verb lemma is requested to be inflected as

| Error type | failure perc. |
| :--- | ---: |
| fused articles (zur, im), complex prepositions (darüber) | $4.33 \%$ |
| wrong POS predicted | $3.95 \%$ |
| definite morpheme for indefinite article | $2.47 \%$ |
| proper names | $2.05 \%$ |
| ihre_attr with incompatible morphemes | $1.56 \%$ |
| die_art with incompatible morphemes | $1.38 \%$ |
| derivatives of verbs | $1.32 \%$ |
| dem with incompatible morphemes | $1.16 \%$ |
| compounds | $1.03 \%$ |
| NoGend tag required for spec. nouns (Herr, Ausschuss, Prozess) | $0.69 \%$ |
| personal pronoun derivatives | $0.58 \%$ |
| hyphens | $0.18 \%$ |
| numerical expressions | $0.18 \%$ |
| Other issues | $3.73 \%$ |

Table 2: Analysis of the persisting generation failures
a noun, or when the system requires an adjective-looking item, which is in fact a derivational form of a verb).
(c) incompatible morphemes (the definite morpheme for indefinite articles, the .NoGend morpheme for particular nouns) that should have been included in the correction layer described in section 3.3
(d) proper names, which are not known by the FST. Although the generator reports a failure, they do not consist a translation error, as they can safely be left uninflected.
(e) other issues, such as hyphens, numerical expressions etc.

Many of the issues above could be addressed with some minor machinery alterations. Point (a) above represents a large class of failures. For this case, the statistical system decision process can be adapted in order to deal with morpheme tag sets of variable width. Similarly, incompatible morphemes (b) can be addressed by adding rules as shown in Section 3.3.

## 5 Future work

Since there is still a small class of generation failures due to various issues (section 4.2 ), some effort is needed in order to guarantee robustness. We could consider a backing-off statistical model, which could perform the tree labeling process in a combined mode, for every sentence: During training, every tree node would get labels referring to both the full word form (as in the original system in section 2.2)
and the lemma+tags (as in the extended system in section 2.3). Then, during the decoding, when the morphology generator fails to produce a word from out of its morphemes, the trained full word form label would be used.

There is also ongoing work in order to efficiently handle the overgenerating phenomenon, which was explained at section 3.2. The method of $n$-best re-scoring (Och and Ney, 2002; Koehn and Knight, 2003), creates a set of alternatives forms of the whole sentence and uses an n-gram language model to re-score them, based on their fluency. That could be a useful tool for getting a more certain decision for the outcome of the generations that resulted in several alternative inflections.

Additionally, the order in which the tree nodes are being traversed has an impact on the availability of the agreement features within the sentence. Whereas the experiments were performed on a simple top-down, left-to-right tree traversal (and hence left-right in the sentence), this does not provide enough agreement features from words following the ones we examine at a certain point. For example, the determiners and the adjectives would have more hints for their gender and case, if they know the properties of the following noun. However, nouns normally get traversed and analysed afterwards, since they are to the right of their determiner and adjectival modifiers. We are considering a restructuring on the order of the traversal mechanism, so that there is better availability of such features.

## 6 Conclusion

We have explained the adaptation of a German Morphology Finite State transducer, so that it can inflect words from given morphemes, as they have been given at the final stage of a LFG-based statistical Machine Translation system. A new "recombination" transducer was formed by writing a set of replace rules on top of the existing morphology transducer. During this adaptation, two major issues were shown to be (a) the compact underspecification tags required by the FST, which would not match what was decided by the statistical system and (b) the requirements of specific POSs for morphemes that are useful for agreement, but redundant for generation. Both issues, when addressed, led to a significant improvement at the generation coverage.

## References

Avramidis, E. and Koehn, P. (2008). Enriching input for statistical machine translation. In ACL '08: Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. Comput. Linguist., 16(2):79-85.

Butt, M., Dyvik, H., King, T., Masuichi, H., and Rohrer, C. (2002). The parallel grammar project. International Conference On Computational Linguistics, pages 1-7.

Callison-Burch, C., Koehn, P., Monz, C., Schroeder, J., and Fordyce, C. S., editors (2008). Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, Columbus, Ohio.

Chiang, D. (2005). A hierarchical phrase-based model for statistical machine translation. In Proceedings of ACL, pages 263-270.

El-Kahlout, I. D. and Oflazer, K. (2006). Initial explorations in English to Turkish statistical machine translation. In Proceedings on the Workshop on Statistical Machine Translation, pages 7-14, New York City. Association for Computational Linguistics.

Galley, M., Hopkins, M., Knight, K., and Marcu, D. (2004). What's in a translation rule? In Proceedings of HLT/NAACL 2004.

Guingne, F., Nicart, F., Champarnaud, J., Karttunen, L., Gaal, T., and Kempe, A. (2003). Virtual operations on virtual networks: The priority union. International Journal of Foundations of Computer Science, 14(6):1055-1070.

Hopkins, M. and Kuhn, J. (2007a). Deep grammars in a tree labeling approach to syntax-based statistical machine translation. In Proceedings of ACL Workshop on Deep Linguistic Processing 2007.

Hopkins, M. and Kuhn, J. (2007b). Machine translation as tree labeling. In Proceedings of NAACL-HLT/AMTA Workshop on Syntax and Structure in Machine Translation 2007.

Huang, L., Knight, K., and Joshi, A. (2006). Statistical syntax-directed translation with extended domain of locality. Proc. AMTA, pages 66-73.

Kaplan, R. M. (1995). Three seductions of computational psycholinguistics. In Dalrymple, M., Kaplan, R. M., Maxwell, J. T., and Zaenen, A., editors, Formal Issues in Lexical-Functional Grammar. CSLI Publications.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. MT Summit, 5.

Koehn, P. and Hoang, H. (2007). Factored translation models. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868876.

Koehn, P. and Knight, K. (2003). Feature-rich statistical translation of noun phrases. In ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, pages 311-318, Morristown, NJ, USA. Association for Computational Linguistics.

Koehn, P., Och, F. J., and Marcu, D. (2003a). Statistical phrase-based translation. In Proceedings of the Human Language Technology Conference 2003 (HLT-NAACL 2003), Edmonton, Canada.

Koehn, P., Och, F. J., and Marcu, D. (2003b). Statistical phrase-based translation. In NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 48-54, Morristown, NJ, USA. Association for Computational Linguistics.

Minkov, E., Toutanova, K., and Suzuki, H. (2007). Generating complex morphology for machine translation. In ACL '07: Proceedings of the 45th Annual Meeting of the Association of Computational linguistics, pages 128-135, Prague, Czech Republic. Association for Computational Linguistics.

Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 295-302.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. Computational Linguistics, 29(1):19-51.

Oflazer, K. (2008). Statistical machine translation into a morphologically complex language. In CICLing, volume 4919 of Lecture Notes in Computer Science, pages 376-387. Springer.

Riezler, S. and Maxwell, J. (2006). Grammatical machine translation. In Butt, M., Dalrymple, M., and King, T. H., editors, Intelligent Linguistic Structures: Variations on themes by Ronald M. Kaplan, chapter 3, pages 35-52. CSLI Publications, Stanford, CA.

Schiller, A. and Steffens, P. (1990). A two-level environment for morphological descriptions and its application to problems of german inflectional morphology. Terminology and Knowledge Engineering, 1:318-329.

Toutanova, K., Suzuki, H., and Ruopp, A. (2008). Applying morphology generation models to machine translation. In ACL '08: Proceedings of the 46th Annual Meeting of the Association of Computational linguistics, pages 514-522. Association for Computational Linguistics.

Yamada, K. and Knight, K. (2001). A syntax-based statistical translation model. In ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pages 523-530, Morristown, NJ, USA. Association for Computational Linguistics.

# PROSODIC PHONOLOGY IN LFG: A NEW PROPOSAL 

Tina Bögel $\dagger$, Miriam Butt $\dagger$, Ronald M. Kaplan $\ddagger$, Tracy Holloway King $\ddagger$, and John T. Maxwell III*<br>Universität Konstanz $\dagger$, Microsoft Corp. $\ddagger$, PARC*<br>Proceedings of the LFG09 Conference Miriam Butt and Tracy Holloway King (Editors) 2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper we outline a new architecture for modeling the interaction between syntax and prosody. This architecture does not make use of correspondences between separate projections, but it is still consonant with the overall framework of LFG. We propose that prosodic information is developed in a component that operates independently of the syntax, thus allowing easy description of misalignment phenomena. We also propose a simple way of making prosodic information accessible to syntax, so that it is possible to condition syntactic rules and preferences on prosodic boundaries. We place the prosodic and syntactic components of the grammar in a pipeline configuration such that the terminal string of the syntactic tree is a sequence of lexical formatives intermixed with features inserted by the prosodic component. Depending on how they are distributed with respect to syntactic groupings, those features may or may not have an impact on the syntactic analysis.


## 1 Introduction

An open question in theoretical linguistics is how to characterize the interactions between the syntactic and prosodic components of a grammar. ${ }^{1}$ One approach to this question takes syntax as primary, following the tradition of proposals made by Selkirk (1981, 1984, 1986) and Nespor and Vogel (1986) and summarized by Selkirk (2001). Under this approach prosodic information is mapped directly from syntax, and prosodic units are therefore naturally aligned with syntactic constituents. It is expected that deviations from straightforward alignment will be quite unusual, and it can become very complex and unintuitive to describe exceptions when they do appear (e.g., Cinque, 1993). The co-description architecture proposed by Butt and King (1998) and Bögel et al. (2008) uses the formal mechanisms of LFG in a concrete instantiation of this approach. ${ }^{2}$

In contrast, a second school of thought assumes that syntax and prosody are typically misaligned. This idea was put forward early on by scholars like Henry Sweet, Eduard Sievers, Franz Saran and Hermann Paul (Plank, 2005, see references therein). This version of the interaction has generally had little appeal to prosodic phonologists, but recent work is undertaking a reconsideration. O’Connor (2005a) and Lahiri and Plank (2009), for example, argue that a simple correspondence between prosody and syntax is more the exception than the rule. Similarly, although Mycock (2006) works within the co-description architecture, she also assumes that there is no simple correspondence between phonology and syntax.

[^47]An example of a typical misalignment is the contrast underlined in (1), discussed by Lahiri et al. (1990, 118). The substring I talked to groups syntactically with the prepositional phrase to its right but prosodically with the preceding noun phrase. ${ }^{3}$
(1) a. [[[The man] [[I] [[talked to] [in the school] $]$ ] [is ill $]$ ]
b. $(((($ The man $) \underline{(\text { I talked to })})$ (in the school) $)$ (is ill))
(Wheeldon and Lahiri, 1997, 357-358)
Throughout this paper, syntactic bracketing is indicated by square brackets [], while prosodic bracketing is indicated by parentheses (). Also, the bracketing shown often collapses syntactic and prosodic levels when they are unimportant for showing the grouping of constituents.

In this paper we outline a new architecture for modeling the interaction between syntax and prosody. This architecture does not make use of correspondences between separate projections but is still consonant with the overall framework of LFG. We propose that prosodic information is developed in a component that operates independently of the syntax, thus allowing easy description of misalignment phenomena. We also propose a very simple way of making prosodic information accessible to syntax, so that it is possible to condition syntactic rules and preferences on prosodic boundaries. We place the prosodic and syntactic components of the grammar in a pipeline configuration such that the terminal string of the syntactic tree (the LFG c-structure) is a sequence of lexical formatives intermixed with features inserted by the prosodic component. Depending on how they are distributed with respect to syntactic groupings, those features may or may not have an impact on the syntactic analysis.

As support for our model, we explore the prosody-syntax relationship with respect to two different types of clitic phenomena. Clitics are interesting for investigation of the prosody-syntax interface since they often reflect misalignments between prosody and syntax and therefore give us insight into what kinds of mismatches need to be accounted for (Halpern, 1995; Halpern and Zwicky, 1996).

Out of the wealth of possible clitic phenomena, we look at just two in the context of this paper. We chose these two because they have recently figured in discussions either with respect to reconsidering the prosody-syntax alignment assumption or with respect to discussions around the prosody-syntax interface in LFG. We leave aside for now a discussion of second position clitics such as those

[^48]found in Serbo-Croatian (e.g., O'Connor (2005a) for an analysis within LFG) and many other languages around the world.

As our first focus we examine the prosody-syntax discrepancies posed by clitics in Germanic languages. In particular, we look at Dutch pronominal clitics, which have been discussed extensively in the (mostly phonological and psycholinguistic) literature (Berendsen, 1986; Gussenhoven, 1986; Carlos Gussenhoven, 1989; Lahiri et al., 1990; Wheeldon and Lahiri, 1997, 2002). The contrast that is important for us is is illustrated in (2) (from Lahiri et al. 1990, 118).
(2) a. [[ik] [zoek [der krant]]]]

I look for her newspaper
'I look for her newspaper.' Spoken Dutch
b. ((ik zoek der) (krant)) I look for her newspaper 'I look for her newspaper.' Spoken Dutch

As can be seen, the pronominal clitic der is incorporated into the prosodic word to its left in (2b) rather than grouping with the syntactic constituent to its right, as in (2a), thus providing an instance of a prosody-syntax mismatch. Further discussion of this mismatch is provided in section 2.1.

Urdu ezafe is the second phenomenon that we examine. Bögel et al. (2008) have argued that ezafe is a clitic whose properties follow straighforwardly if the prosodic dimension is taken into account. They present an analysis which builds on Butt and King (1998), who implemented the interaction between Bengali clitics and prosody as analyzed by Hayes and Lahiri (1991) and Lahiri and FitzpatrickCole (1999) via a p(rosodic)-projection. This p-projection follows the standard LFG architecture in that it is projected from the c-structure in parallel to the f structure and thus follows the alignment assumption of Selkirk (1981, 1984, 1986), Nespor and Vogel (1986), and Truckenbrodt (1999). However, the implementation that Bögel et al. (2008) present has some difficulties which are resolved under the alternative approach presented in this paper in section 2.2.

Finally, we point to prosodically-determined resolution of syntactic ambiguities as another source of evidence for our model. Without prosodic information the string old men and women has two different syntactic bracketings, corresponding to two different interpretations (3):
(3) Syntactic bracketing
a. [[old men] and [women]]
b. [old [men and women]]

In our proposed pipeline architecture, the first of these would be preferred given the prosody in (4a) and the second would be preferred for the pattern in (4b): ${ }^{4}$

[^49]a. (old men) (and women)
b. (old) (men and women)

We will use these examples to illustrate how information coming from the otherwise independent prosody component can influence the distribution of optimalitytheory preferences (Frank et al., 1998) and thus affect the selection of particular syntactic analyses.

The following sections first provide more detail about the phenomena under consideration. We then introduce our LFG-oriented architecture of independent components that communicate and interact through symbols on a shared string. For the sake of concreteness, we show how the syntactic aspects of our proposal can be implemented by means of the notations and formal mechanisms that already exist in the XLE computational interpreter for LFG grammars (Crouch et al., 2009), and thus we show that this approach does not require mathematical or computational extensions of LFG syntactic theory. As a separate hypothesis, we suggest that the independent prosodic component needs no more than the mathematical and computational power of regular relations and finite-state transducers, the same devices that are already used for morphological analysis within the XLE system. Since the LFG languages are closed under pipeline composition with regular relations, and since the XLE system can perform finite-state transductions, the combination of LFG syntax with a regular prosodic component fits comfortably within the formal systems that already exist.

## 2 The Interaction of Prosody and Syntax

Selkirk (1986) made a particularly straightforward proposal for the interaction of prosody and syntax. She put forth the requirement that a unit of prosodic structure must have as its terminal string the stretch of the surface syntactic structure that is demarcated by the right and left ends of selected syntactic constituents. This postulates a relation of close alignment between prosodic units (inferred by their blocking or triggering of postlexical phonological and prosodic processes) and syntactic constituents (determined by traditional arguments involving substitution, coordination, extraction, and the like). The prosodic and syntactic structures are not isomorphic under this conception, because it does not require a distinct prosodic unit for every level of the syntactic hierarchy - the prosodic structure can be flatter than the syntactic. The situation where some elements of a syntactic constituent belong to one prosodic unit and other elements of that same constituent belong to another prosodic unit is then seen as being an exceptional instance of misalignment
often written ' $n$ in representations of colloquial English. This does not change the fact that different prosodic groupings of the prosodic words that correspond to the content words prefer different syntactic structures.
or mismatching. The items der and krant in example (2) above are thus misaligned according to Selkirk's algorithm.

Butt and King (1998) showed how the co-description architecture of LFG can be used to implement this general conception. They introduced an explicit prosodic structure ( $p$-structure) that is projected from the c-structure by co-describing constraints in the same way that the f-structure is projected from the c-structure. Codescribing constraints can correlate selected syntactic categories to particular levels of the prosodic hierarchy, as suggested by Selkirk, and constraints that equate the prosodic units corresponding to other mothers and daughters allow for the flattening of prosodic structures, as desired. The architecture naturally permits a limited amount of "heightening" - the specification of intermediate prosodic levels that do not correspond directly to syntactic constituents - but that expressive power is not required to implement the Selkirk proposal.

Butt and King (1998) also observe, however, that the co-description architecture does not easily allow for prosodic units that are misaligned (in the sense defined above) with syntactic constituents. In this section we discuss two sets of linguistic data that suggest that misalignments are not atypical: Germanic, primarily Dutch, clitic placement and Urdu ezafe. Evidence of this sort is what motivates our consideration of new architectural arrangements.

### 2.1 Misalignment of Germanic Clitics

The Germanic languages are among those where phonological phrasing systematically diverges from syntactic phrasing (Lahiri and Plank, 2009). As a particular case in point, a series of psycholinguistic experiments has shown that the prosodic properties of Dutch clitics are misaligned with their conventional morphosyntactic properties (Lahiri et al., 1990; Wheeldon and Lahiri, 1997, 2002). The psycholinguistic reality of mismatches is demonstrated by the prosodic and syntactic phrasing of the Dutch article $d e$, which, being a clitic, needs to be incorporated into another prosodic word, as in (5).
(5) Ik drink de wijn
'I drink the wine'

|  | ik | drink | de | wijn |
| :--- | :--- | :--- | :--- | :--- |
|  | I | drink | the | wijn |
| Syntactic Phrasing: | $[[i \mathrm{ik}]$ | $[$ drink | [de | wijn $]]]$ |
| Phonological Phrasing: | ik | ((drink) | de) | wijn |

(Wheeldon and Lahiri, 1997, 358)
The experiments by Wheeldon and Lahiri (1997) show that the phonological phrasing in (5) is the correct one and that the Dutch definite determiner der/de is indeed a clitic that forms a prosodic word with the word to its left. In one experiment, using an experimental method whereby the speaker was offered a delayed response, they sought to determine whether the number of prosodic units or the
number of actual words account for the length of speaker preparation time for sentences such as the following:
(6) a. (ik drink de) (wijn) 'I drink the wine' ( 2 prosodic units, 4 words)
b. (ik drink) (Jans) (wijn) 'I drink Jan's wine' (3 prosodic units, 4 words)
c. (ik drink) (wijn) 'I drink wine' (2 prosodic units, 3 words)

By hypothesis, sentences (6a) and (6c) share the same number of prosodic units although they differ in the number of words. These should take the same speaker preparation time, while sentence (6b), which has one additional prosodic unit, should take longer for the speaker to prepare. Indeed, Lahiri and Wheeldon's results show that (6a) and (6c) do take the same amount of speaker preparation time, while (6b) takes the speaker significantly longer to prepare. This leads to the conclusion that it is the number of prosodic units and not the number of words that is relevant for speaker preparation time, and it supports the claim that the definite determiner $d e$ forms a prosodic word with the word to its left.

To further determine whether the definite determiner de in (6a) attached to the host on the left and could therefore be classified as a clitic, Lahiri and Wheeldon conducted a second experiment in which they required an immediate response from the speaker (thus allowing minimal or no time for planning). The idea behind the experiment was that it would be the size of the first prosodic unit that mattered for the speaker and not the number of prosodic units as in the first experiment. The result of this experiment was that sentences (6b) and (6c) took the speaker the same time to prepare, while sentence (6a) needed a significantly longer time. The conclusion drawn from this is that with the spontaneous response, the size of the first unit matters. The longer preparation time for sentence (6a) can only be explained if $d e$ is attached to the left and is therefore acting as a clitic.

Further proof for the Dutch definite determiner being a clitic and being integrated into the prosodic unit to the left comes from junctural rules. Consider voice assimilation in Dutch:
(7)

|  | expression | [gd] | [kd] | $[\mathrm{kt}]$ |
| :--- | :---: | :---: | :---: | :---: |
| Compound | $(\text { zak })_{\omega}(\text { doek })_{\omega}$ <br> 'handkerchief | + | - | - |
| Clitic | (zoek der) $)_{\omega}$ <br> 'seek her' | + | - | + |
|  |  |  |  |  |

The Dutch compound noun in (7) shows voicing assimilation across a prosodic word boundary, with the assimilation of unvoiced $k$ to voiced $g$ in the context of the voiced consonant $(d)$ starting the second prosodic word. In contrast, the clitic example in (7) argues for a single prosodic word because the voicing assimilation can also go the other way, from the final voiceless consonant of the host $(k)$ to the initial consonant of the clitic (Gussenhoven, 1986; Lahiri et al., 1990).

This evidence shows that de/der are clitics in Dutch and that they incorporate into the prosodic word on their left. As our main interest is the determination of
the interaction between prosody and syntax, the important observation for us is that function words such as the Dutch definite determiner syntactically group with various syntactic phrases to their right, but incorporate prosodically with various phrase types to their left. The same point is made by the English function words $a$ and of in the natural pronunciation of the admonition in (8), discussed by Lahiri and Plank (2009).

| (8) | Syntactic Phrasing | [Drink [[a pint] [of milk]] [a day]] |
| :--- | :--- | :--- |
|  | Phonological Phrasing | (Drink a) (pint a) (milk a) (day) |

Writing special rules to account for the variety of prosody-syntax mismatches would seriously complicate all of the prosody-syntax mapping algorithms that assume a basic prosody-syntax alignment. We therefore see the prosodic behavior of Germanic clitics as strong motivation to try for an alternative conception.

### 2.2 Misalignment of Urdu Ezafe

The South Asian language Urdu contains a construction traditionally called ezafe or izafat, which it borrowed via language contact from Persian. Persian ezafe has been analyzed in a number of papers; Samvelian (2007) provides a very thorough overview of the phenomenon and the work that has been done on it.

In Persian, the ezafe originated from a relative clause construction. Its interest for modern linguistics is that it does not respect the usual headedness patterns of the language and that it licenses complements from an unexpected position. Samvelian (2007), while providing an otherwise very lucid account of the properties of Persian ezafe, analyzes it as a morphological affix and as an instance of head-marking of grammatical relations. She does discuss the possibility of ezafe being a clitic but dismisses this possibility. Rather than explicitly integrating a full-blown prosodic module into her analysis, she integrates a reference to the prosodic properties of ezafe by including an EDGE feature by which the ezafe has to percolate to a clausal edge in the syntactic representation.

Urdu ezafe has a more restricted use than Persian ezafe. However, there is much of the same type of evidence as in Persian that it is a clitic and that it is prosodically incorporated into the right edge of the preceding phrasal constituent. Bögel et al. (2008) therefore argue against treating ezafe as an affix and in favor of treating it as a clitic. In order to model the conflicting syntactic and prosodic properties of ezafe, they build on Butt and King (1998) and integrate a prosodic projection into the analysis.

The prosodic projection proposed by Butt and King (1998) and used in the analysis of Urdu ezafe by Bögel et al. (2008) follows the standard architecture of LFG, which projects other levels of representation from the c-structure. Under this view, syntax is primary and the c-structure (and f-structure) are central to analyses in other projections. This arrangement is therefore in line with approaches that assume that prosodic and syntactic units are typically aligned.

Example (9) illustrates a typical ezafe construction. There is a syntactic dependency between the head noun sher 'lion' and a modifier to the right of that NP, panjAb 'Punjab'. 5 This dependency relation is licensed by the ezafe ( $=e$ in (9)) even though the ezafe is prosodically attached to the head of the construction. Note that the usual pattern of headedness in Urdu (as in Persian) is head-final.
(9) sher=e panjAb
lion=Ez Punjab
'a/the lion of Punjab'
Bögel et al. (2008) analyze this construction as follows. Syntactically, the ezafe is part of the modifying construction. It licenses the modifier panjAb and is therefore part of the same constituent. The category EzP is assigned to that constituent to model the idea that ezafe is a head that licenses a complement, namely the pan$j A b$ in our example.
(10)
a. C-Structure

b. F-Structure


At the c-structure level shown in (10a), the ezafe is inserted as a terminal node and is thus analyzed as a syntactic word in its own right. ${ }^{6}$ It combines with its complement to form the modifying constituent for the head noun sher. This modification relationship is expressed within the f-structure, which models the functional information and dependencies. In (10b), sher is the head of the phrase and panjAb functions as its modifier (MOD).

[^50]However, this syntactic analysis is at odds with the fact that the ezafe is prosodically incorporated into the head noun to its left. That is, we have a clear mismatch between syntactic and prosodic constituency:

| (11) | Syntactic Phrasing: | [[sher] | [e panjAb]] |
| :--- | :--- | :--- | :--- |
|  | Prosodic Phrasing: | ((sher | e) |
| panjAb) |  |  |  |

Bögel et al. (2008) make use of the Butt and King (1998) p(rosodic)-structure to address this problem and arrange for the p-structure in (12) to be assigned to this example.
(12) P-Structure for sher e panjAb
$\left[\begin{array}{l}\left\{\begin{array}{l}{[\text { DOMAIN P-WORD, P-FORM sher, CL-FORM ezafe }]}\end{array}\right\} \\ {[\text { DOMAIN P-WORD, P-FORM panjAb }]}\end{array}\right\}$
The outer unit of this p-structure corresponds to the top-level NP in (10a) and is marked as a prosodic phrase. The fact that this phrase covers the two prosodic words is formalized by the set containing the individual p -structures corresponding to those words. Crucial to this analysis, ezafe is not encoded as an independent prosodic word in the p-structure. It appears instead as a CL(itic)-FORM in the wordlevel p-structure corresponding to its prosodic host.

Bögel et al. associate the p-structure in (12) with the c-and f-structures in (10) by adding to the conventional LFG syntactic rules a set of constraints that describe the p-structure alongside the constraints that characterize the f-structure. The grammar constraints most relevant to this discussion are shown as annotations on the nodes in (13), a decorated version of the c-structure (10a). In these constraints the designators $\uparrow$ and $\downarrow$ denote the f-structures corresponding to mother and daughter nodes, as usual, and $\uparrow_{p}$ and $\downarrow_{p}$ denote the p-structure units projected from those nodes. Unless otherwise specified and as is conventional for LFG, the $\mathrm{f}-\mathrm{and} \mathrm{p}$ structures corresponding to a daughter node are assumed by default to be the same as the structures corresponding to its mother. Thus all of the p-structure constraints under the left NP must hold of a single structure; that word-level structure is a member of the set in the top-level p-structure by virtue of the $\downarrow_{p} \in \uparrow_{p}$ assertion on the N. Similarly, the collection of constraints on the N under the EzP node define its properties as another prosodic-word component of the larger prosodic phrase. The f-structure in (10b) satisfies all of the functional constraints from the left NP (by the default convention) and also includes the MOD structure by virtue of the constraints below the panjAb noun.


What stands out in this representation is that the constraint that adds the CLFORM feature to the host p-structure is attached to the N under the left NP and not to any of the nodes on the clitic side. This is because the p-structure of the leftward host is not accessible by ordinary co-description from any of the nodes on the right. Constraints on the right can make reference to the top-level p-structure, which does contain the sher structure as a set-element, but there is no co-descriptive designator by which that particular element can be picked out from other elements that might also belong to that prosodic phrase. The solution shown in this tree is to assign the CL-FORM on the left side, where the host p-structure is directly available. This requires an alternative expansion of the general NP rule that attaches the CLFORM p-structure constraint to the N head but only when the NP is part of an ezafe construction. Relying on the fact that the top-level f-structure (unlike the p-structure) is accessible on both branches, Bögel et al. impose this restriction by having ezafe register its presence by adding a special CHECK feature to the top f structure, and then testing for that feature via the $(\uparrow$ CHECK EZAFE $)={ }_{c}+$ constraint where the CL-FORM is assigned. ${ }^{7}$

Although this co-descriptive arrangement of $\mathrm{c}-, \mathrm{f}$ - and p -structures does model the properties of the Urdu ezafe, this account is unsatisfactory in several ways. The grammar under this analysis does not express the pretheoretic intuition that a clitic operates on its host and instead makes the host anticipate that it might have an attached clitic. This leads to a complicated and carefully orchestrated distribution of prosodic and syntactic constraints across both lexical entries and syntactic rules. And these constraints are special to this particular construction and make no further predictions about the interaction of prosodic structure with the rest of the grammar. Other phenomena, for example case clitics or focus clitics (Butt and King, 2004), would have to be modeled individually, on a case by case

[^51]basis. ${ }^{8}$ In contrast, the architecture we develop in the next section allows prosodic generalizations to be stated independently of syntactic ones and does not require otherwise unmotivated structures or bookkeeping features to correctly model the interactions of these linguistic subsystems.

## 3 A Pipeline Architecture

A growing body of evidence, some of which we have summarized, calls into question the hypothesis of strong alignment between prosody and syntax and thus also the LFG co-description account of the relationship between these components. The challenge is to define an architecture of components that allows a close linkage between prosodic and syntactic phenomena in some situations but still allows for independent operation in cases of misalignment.

We suggest that this challenge can be met with a pipeline arrangement of independent components that interact through a very simple channel of communication. Our proposal depends on pre-existing aspects of the LFG syntactic formalism, including the capability of expressing optimality-theoretic preferences to impose soft constraints on syntactic interpretations (Frank et al., 1998; Sells, 2001; Bresnan, 2000). Our architecture has the following key features:
(14) a. An independent prosodic component interprets various phonological properties to determine the boundaries of prosodic phrases.
b. Prosodic boundaries are made visible to the syntax as distinct symbols in the terminal string of the syntactic constituent structure.
c. Prosodic boundary symbols augment but do not disrupt syntactic patterns.
d. The syntactic component obeys a Principle of Prosodic Preference: syntactic structures with constituent boundaries that do not coincide with prosodic boundaries are dispreferred.

[^52]In this architecture the boundary-annotated output of the separate prosodic component becomes the input to the syntactic component. The input to the syntax for our Urdu ezafe example is the prosodically-bracketed string in (11) instead of the simple word-string. The syntactic component interprets its rules as allowing prosodic brackets to be freely intermixed among the other syntactically specified terminals (property (14c)), and the resulting syntactic structures have nodes and branches that cover the prosodic brackets in addition to the required syntactic formatives. This means that misaligned prosodic brackets will not interfere with the usual syntactic analysis. Thus rule (15a) is interpreted as (16a) and (15b) is interpreted as (16b) when they apply to our ezafe example (RB and LB are the lexical categories of the right/left prosodic brackets, the terminal parentheses).

$$
\text { a. } \quad \mathrm{EzP} \rightarrow \quad \text { EZ } \quad \underset{ }{\downarrow} \begin{align*}
& \mathrm{N}  \tag{15}\\
& \\
&
\end{align*}
$$

b. $\quad \mathrm{NP} \rightarrow \mathrm{N}$
a. $\quad \mathrm{EzP} \rightarrow \quad$ EZ $\quad$ RB $\quad \underset{\downarrow \in(\uparrow \text { MOD })}{\mathrm{N}}$
b. $\quad \mathrm{NP} \rightarrow \mathrm{LB} \quad \mathrm{N}$

This architecture allows a drastic simplification of the rules needed to describe an ezafe construction. In contrast to the annotations needed for the previous solution, the rules in (15) do not encode any information about prosodic properties and do not involve any CHECK-features to ensure that an ezafe clitic appears to the right of the head noun. With the extended rule interpretation in (16), the tree in (17) is the resulting c-structure.


The outer prosodic brackets are aligned with the syntactic constituents but the internal ones are not. The input is still accepted by the grammar and is assigned the f -structure in (10b). The traditional syntactic c-structure can be seen as a projection of (17) formed by systematically deleting prosodic nodes and branches.

Consider (18) as another illustration of misalignment in our pipeline. Again we assume that the prosodic component introduces phonologically-determined boundaries into the syntactic input string. We have added square brackets on top of the prosodically-bracketed input to indicate the syntactic constituents - the syntactic analysis goes through despite the confusion of prosodic boundaries.

| Phonological String | Drink a pint a milk a day |
| :--- | :---: |
| PROSODIC PARSER | (Drink a) (pint a) (milk a) (day) |
| Prosodically Bracketed | (milk $]$ [a) (day)]] |
| SYNTACTIC PARSER | [[(Drink ] [[a) (pint] [of) (mile |

We have seen how properties (14a-c) allow for arbitrary misalignments of syntactic and prosodic boundaries. The Principle of Prosodic Preference (14d) completes the architecture by introducing a soft dependency between prosody and syntax: among a competing set of syntactic structures, those with the fewest number of misaligned brackets will be selected as the correct analyses. This reflects Selkirk's original intuition of close alignment, at least in certain situations. As an immediate consequence, it also captures the fact that prosodic information can have the effect of disambiguating between several possible parses. The phrase old men and women is syntactically ambiguous in the absence of prosodic phrasing, as indicated by the following syntactic structures:
(19) Syntactic constituents
a. [[old men] and [women]]
b. [old [men and women]]

But suppose that the syntax is instead given a prosodically bracketed string, for example, the one in (20a):

## Prosodic phrasing

a. (old men) (and women)
b. (old) (men and women)

With this prosodic phrasing the analysis in (19b) is dispreferred by virtue of the bracket configurations shown in (21), and the compatible analysis in (19a) is selected. The asterisks mark the prosodic brackets that are unaligned with syntactic phrases and are therefore dispreferred. The alternative phrasing in (20b) will select the analysis (19b).
(21) a. [[(old men)] *(and [women $])]$
b. [(old [men*) *(and women])]

To summarize, this architecture for the interface between prosody and syntax allows a proper analysis of the Urdu ezafe clitic, the systematic misalignment between prosody and syntax in Drink a pint a . . . , the Dutch definite determiner clitic, and, as far as we can determine, all other clitic phenomena and other instances of misalignment. We allow misalignments as a matter of course, as suggested by one school of thought on these matters, but we also incorporate a preference for more aligned analyses, in accord with the second and more conventional school of
thought. Prosody operates as a separate component that communicates with syntax through the narrow channel of a prosodically annotated string. The syntactic interactions are governed by an extended interpretation of ordinary c-structure rules and an optimality-theoretic Principle of Prosodic Preference. Unlike previous LFG proposals, we do not incorporate a co-described prosodic projection and so avoid the detailed specifications that define its properties.

## 4 Implementation by Metarule Expansion

Our proposed architecture assigns an extended interpretation to the ordinary rules of a conventional LFG grammar. In (16) we showed that the effect of this extended interpretation for some particular rules is equivalent to including in the grammar some additional rules that are systematically related to the originals. We observe now that this is generally the case: the behavior of every syntactic rule according to our proposed architecture can be modeled by a finite expansion to a set of rules that could have been written in standard, pre-existing notations. In other words, the architectural principles in (14) can be implemented as metagrammatical operations that systematically transform the rules of a conventional grammar. As a consequence, we know that this architecture implies no changes to the mathematical and computational properties of the syntactic component.

A conventional LFG grammar contains a set of c-structure rules of the form
(22) $\mathrm{CAT} \rightarrow$ RHS
where CAT is a nonterminal category and the right-hand side RHS denotes a regular language over categories annotated with functional (or other co-describing) constraints. To implement the architectural specifications, we replace each such rule with another rule of the form

$$
\text { (23) CAT } \rightarrow
$$

$$
\text { (LB) RHS / [ LB } \mid \mathrm{RB} \quad](\mathrm{RB})
$$

Disprefer
The prosodic brackets and their lexical categories ( LB and RB ) belong to the terminal and nonterminal vocabularies of the enlarged grammar, in accordance with (14b). The right-side of the original rule is replaced by a rule expansion which allows for the parsing of prosodic brackets. The categories of the original righthand side can be optionally preceded by a left prosodic bracket (as indicated by the parentheses) and optionally followed by a right prosodic bracket. In addition, the expansion will match a daughter sequence that would match the RHS regular expression if all occurrences of either LB or RB in that sequence are ignored (the | indicates a disjunction). The / is a notation for the "Ignore" operator first introduced by Kaplan and Kay (1994); it is included in the Xerox finite-state machine calculus (Beesley and Karttunen, 2003) and in the c-structure notation of the XLE
system (Crouch et al., 2009). The effect of this use of the Ignore operator is to implement property ( 14 c ) of the architecture: it ensures that occurrences of prosodic brackets cannot disrupt otherwise valid phrase-structure expansions.

The "Disprefer" annotation implements the Principle of Prosodic Preference (14d). Whenever a prosodic bracket is ignored in the middle of the RHS, the structure is assigned a dispreference optimality mark. The effect of this is to determine a ranking over possible syntactic analyses, as described by Frank et al. (1998). The only brackets that are not dispreferred are those that match the optional LB and RB categories, the ones that appear on the edges of constituents. Replacement rules produced in this way by metagrammatical expansion thus provide dispreferences only for misaligned prosodic brackets, as required.

By way of illustration, the example in (24b) shows what results from the metarule expansion of the simple rule (24a).
a. $\mathrm{VP} \rightarrow \mathrm{V} \quad \mathrm{NP}$ $(\uparrow$ OBJ $)=\downarrow$
b. VP $\rightarrow$

| $(\mathrm{LB})$ | $[\mathrm{LB} \mid \mathrm{RB}]^{*}$ | V | $[\mathrm{LB} \mid \mathrm{RB}]^{*}$ | NP | $[\mathrm{LB} \mid \mathrm{RB}]^{*}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Disprefer |  | Disprefer | $(\uparrow$ OBJ $)=\downarrow$ | Disprefer |

The Kleene-star operators derive from the Ignore specification. They allow for misaligned prosodic brackets to appear in any position as well as for the possibility of no misalignments.

The metagrammatical implementation of our architecture can be instantiated quite directly within the XLE computational system (Crouch et al., 2009). XLE includes a metarule expansion facility whose purpose is to express generalizations over all syntactic rules or over particular subsets of them. This facility is invoked by defining a "metarule macro". The input to a metarule macro is the category and right-hand side of an existing rule, and the output is the replacement rule for that input. ${ }^{9}$ The macro definition in XLE notation in (25) is equivalent to the metagrammatical expansion in (23).
(25) METARULEMACRO(CAT, RHS $)=$
(LB) RHS / $\{\mathrm{LB} \mid \mathrm{RB}\}$ : @ (DISPREFER) (RB)
The @ (DISPREFER) annotation is an invocation of an XLE template that can be defined to add a prosodic dispreference to the collection of optimality marks associated with the c-structure.

[^53]
## 5 The Prosodic Component: Some Speculations

Our architecture postulates an independent prosodic component that recognizes prosodic phrases and marks their boundaries in the input string to the syntax. On this view, the internal properties of the prosodic component are not accessible to syntax and are not constrained by syntactic requirements, and indeed are not especially relevant to the overall architectural conception. Still, it is worthwhile to consider how the prosodic component might operate.

The prosodic component must embody knowledge about rhythmic structure of the language (trochaic/iambic), it must be able to parse tones (e.g. high/low), and it must be sensitive to part of speech information, at least enough to differentiate function vs. content words. There have been some suggestions in the literature that recursive rules may be needed for prosodic analysis (Booij, 1995, 1996; Peperkamp, 1997; Vigário, 1999, 2003). However, the notion of recursivity within phonology seems to be confined to the level of the prosodic word and mainly seems to concern clitic phenomena across languages. More recent work seems to be distancing itself from the notion of recursivity in phonology. Vogel (2009) argues that recursive power is not necessary for clitic phenomena, and Kabak and Revithiadou (2009) attribute its appearance at the prosodic level to the interaction with morphosyntax. That is, recursivity within morphosyntax is reflected within prosody but is not inherent to prosodic structures (see also Selkirk's (1984) Strict Layer Hypothesis, which legislates against recursion). O'Connor (2004) also points out that center-embedding recursion is not needed for prosodic structure.

If prosodic rules lack center-embedding recursion and have a bounded number of levels (e.g. prosodic words, prosodic phrases, intonational phrases), then the prosodic component as a whole defines a regular relation between its inputs and outputs, a relation that can be implemented by a finite-state transducer. Thus O'Connor (2005b) proposes to model prosodic information via a series of rewrite rules that apply to the representation of intonation in the AM/ToBI annotation scheme (Pierrehumbert, 1980; Goldsmith, 1976, among others). His proposal allows only bounded reapplication of these tune structure rules and so his system can describe only regular relations.

These observations have some interesting and important consequences. If the prosodic component defines only regular relations, we can characterize it using notational devices whose mathematical and computational properties are very well understood (Kaplan and Kay, 1994; Beesley and Karttunen, 2003). We also know that the LFG languages are closed under composition with regular relations, so that the formal power of the combination of components in our pipeline architecture is no greater than the formal power of the syntactic component by itself.

As another consequence, we can immediately create and experiment with a concrete implementation of our architecture. In the XLE system the terminal string of the syntactic tree is constructed by applying a pipeline sequence of finite state transductions to an original input string. Typically the input is a string of ordinary text, and the transducers perform standard transformations such as tokenization and
morphological analysis. We can reconfigure the system so that its input is a string annotated with tonal information and other prosodically-relevant features. Then the initial step in the cascade of transformations can be carried out by a transducer that introduces prosodic brackets that are consistent with the prosodic annotations. We are now experimenting with a first version of this type of transducer, constructed using the tools of the XFST finite-state calculus (Beesley and Karttunen, 2003).

It is a strong and useful hypothesis that the independent prosodic component is so limited in its computational power, but it is not a theoretical necessity of our proposed architecture.

## 6 Conclusion

We have proposed an LFG-oriented pipeline architecture that allows for misalignments between prosody and syntax in a natural manner but also still incorporates a preference for an alignment of prosodic and syntactic phrases. We postulate that an independent prosodic component delivers prosodically-bracketed strings as the input to the syntactic component and that syntax can ignore these brackets with some degree of dispreference if they are incompatible with proper syntactic analyses. This architecture provides explicit accounts of the syntactic and prosodic properties of clitics using simpler rules and representations than previous approaches have required. We have also shown how this architecture can be implemented by means of metagrammatical expansions, both conceptually and computationally, so that it adds no new formal power to the basic LFG framework. This architecture addresses the challenges coming from the two traditional schools of thought concerning the alignment of prosody and syntax by allowing both for the primacy of syntax and for rampant mismatches between syntactic and prosodic structure.

## References

Anderson, Stephen. 2005. Aspects of the Theory of Clitics. Oxford University Press.
Asudeh, Ash. 2009. Adjacency and Locality: A Constraint-based Analysis of Complementizer-Adjacent Extraction. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG09 Conference, CSLI On-line Publications.
Beesley, Kenneth and Karttunen, Lauri. 2003. Finite State Morphology. CSLI Publications.
Berendsen, Egon. 1986. The Phonology of Cliticization. Foris.
Bögel, Tina, Butt, Miriam and Sulger, Sebastian. 2008. Urdu Ezafe and the Morphology-Syntax Interface. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG08 Conference, CSLI On-line Publications.
Booij, Geert. 1995. The Phonology of Dutch. Clarendon Press.
Booij, Geert. 1996. Cliticization as Prosodic Integration: The Case of Dutch. The Linguistic Review 13, 219-242.

Bresnan, Joan. 2000. Optimal Syntax. In Joost Dekkers, Frank van der Leeuw and Jeroen van de Weijer (eds.), Optimality Theory: Phonology, Syntax, and Acquisition, pages 334-385, Oxford University Press.
Butt, Miriam and King, Tracy Holloway. 1998. Interfacing Phonology with LFG. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG98 Conference, CSLI On-line Publications.
Butt, Miriam and King, Tracy Holloway. 2004. The Status of Case. In Veneeta Dayal and Anoop Mahajan (eds.), Clause Structure in South Asian Languages, Kluwer.
Butt, Miriam and King, Tracy Holloway. 2007. Urdu in a Parallel Grammar Development Environment. Language Resources and Evaluation 41, 191-207.
Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. A Grammar Writer's Cookbook. CSLI Publications.
Carlos Gussenhoven, C. 1989. Cliticization in Dutch as Phonological Word Formation, University of Nijmegen.
Cinque, Guglielmo. 1993. A Null Theory of Phrase and Compound Stress. Linguistic Inquiry 24, 239-267.
Crouch, Dick, Dalrymple, Mary, Kaplan, Ronald M., King, Tracy, Maxwell, John and Newman, Paula. 2009. XLE Documentation. On-line documentation, Palo Alto Research Center (PARC).
Dalrymple, Mary, Kaplan, Ronald M., Maxwell III, John T. and Zaenen, Annie (eds.). 1995. Formal Issues in Lexical-Functional Grammar. CSLI Publications.
Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and Maxwell III, John T. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG98 Conference, CSLI On-line Publications.
Goldsmith, John. 1976. Autosegmental Phonology. Ph. D. thesis, Massachusetts Institute of Technology.
Gussenhoven, Carlos. 1986. Over de Fonologie van Nederlandse Clitica. Spektator 15, 18-200.
Halpern, Aaron. 1995. On the Placement and Morphology of Clitics. CSLI Publications.
Halpern, Aaron and Zwicky, Arnold (eds.). 1996. Approaching Second: Second Position Clitics and Related Phenomena. CSLI Publications.
Hayes, Bruce and Lahiri, Aditi. 1991. Bengali Intonational Phonology. Natural Language and Linguistic Theory 9, 47-96.
Kabak, Bariş and Revithiadou, Anthi. 2009. An Interface Approach to Prosodic Word Recursion. In Janet Grijzenhout and Bariş Kabak (eds.), Phonological Domains: Universals and Deviations, pages 105-133, Mouton de Gruyter.
Kaplan, Ronald M. 1987. Three Seductions of Computational Psycholinguistics. In Peter Whitelock, Mary McGee Wood, Harold L. Somers, Rod Johnson and Paul Bennett (eds.), Linguistic Theory and Computer Applications, pages 149-181, Academic Press, reprinted in Dalrymple et al. (1995:339-367).
Kaplan, Ronald M. 1989. The Formal Architecture of Lexical-Functional Gram-
mar. In Chu-Ren Huang and Keh-Jiann Chen (eds.), Proceedings of ROCLING II, pages 3-18, reprinted in Dalrymple et al. (1995:7-27).
Kaplan, Ronald M. and Kay, Martin. 1994. Regular Models of Phonological Rule Systems. Computational Linguistics 20, 331-478.
Lahiri, Aditi and Fitzpatrick-Cole, Jennifer. 1999. Emphatic Clitics and Focus Intonation in Bengali. In René Kager and Wim Zonneveld (eds.), Phrasal Phonology, Nijmegen University Press.
Lahiri, Aditi, Jongman, Allard and Sereno, Joan A. 1990. The Pronomincal Clitic [der] in Dutch: A Theoretical and Experimental Approach. Yearbook of Morphology 3, 115-127.
Lahiri, Aditi and Plank, Frans. 2009. What Linguistics Universals can be True of. In Sergio Scalise, Elisabetta Magni and Antonietta Bisetto (eds.), Universals of Language Today, pages 31-58, Springer.
Luís, Ana and Otoguro, Ryo. 2005. Morphological and Syntactic Well-formedness: The Case of European Portuguese Clitics. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG05 Conference, CSLI On-line Publications.
Miller, P. 1992. Clitics and Constituents in Phrase Structure Grammar. Garland.
Mycock, Louise. 2006. The Typology of Constituent Questions: A LexicalFunctional Grammar Analysis of wh-questions. Ph. D. thesis, University of Manchester.
Nespor, Marina and Vogel, Irene. 1986. Prosodic Phonology. Foris.
O'Connor, Rob. 2004. Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croation. Ph. D. thesis, University of Manchester.
O’Connor, Rob. 2005a. Clitics in LFG: Prosodic Structure and Phrasal Affixation. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG05 Conference, CSLI On-line Publications.
O'Connor, Rob. 2005b. Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croation. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG05 Conference, CSLI On-line Publications.
Peperkamp, Sharon. 1997. Prosodic Words. Ph. D. thesis, University of Amsterdam.
Pierrehumbert, Janet. 1980. The Phonology and Phonetics of English Intonation. Ph. D. thesis, Massachusetts Institute of Technology.
Plank, Frans. 2005. ((FRISCH ge) (WAGT ist)) ((HALB ge) (WON nen)), unpublished manuscript, Universität Konstanz.
Sadler, Louisa and Spencer, Andrew. 2000. Workshop on Morphology and LFG. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG00 Conference, CSLI On-line Publications.
Samvelian, Pollet. 2007. A (Phrasal) Affix Analysis of the Persian Ezafe. Journal of Linguistics 43, 605-645.
Selkirk, Elisabeth. 1981. On Prosodic Structure and its Relation to Syntactic structure. In Thorstein Fretheim (ed.), Nordic Prosody II, pages 111-140, Tapir.

Selkirk, Elisabeth. 1984. Phonology and Syntax: The Relation between Sound and Structure. MIT Press.
Selkirk, Elisabeth. 1986. On Derived Domains in Sentence Phonology. Phonology Yearbook 3, 371-405.
Selkirk, Elisabeth. 1995. The prosodic structure of function words. In Jill Beckman, Laura Walsh Dickey and Suzanne Urbanczyk (eds.), Papers in Optimality Theory, pages 439-70, GLSA Publications, also in Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition, ed. James L. Morgan and Katherine Demuth, 187-214: Lawrence Erlbaum Associates.
Selkirk, Elisabeth. 2001. The syntax-phonology interface. In N.J. Smelser and Paul B. Baltes (eds.), International Encyclopedia of the Social and Behavioral Sciences, pages 15407-15412, Pergamon.
Sells, Peter (ed.). 2001. Formal and Empirical Issues in Optimality Theoretic Syntax. CSLI Publications.
Truckenbrodt, Hubert. 1999. A Null Theory of Phrase and Compound Stress. Linguistic Inquiry 30, 219-255.
van der Leeuw, Frank. 1997. Clitics: Prosodic Studies. Holland Academic Graphics.
Vigário, Marina. 1999. On the Prosodic Status of Stress Function Words in European Portuguese. In Tracy Alan Hall and Ursula Kleinhenz (eds.), Studies on the Phonological Word, pages 253-293, John Benjamins.
Vigário, Marina. 2003. The Prosodic Word in European Portuegese. Mouton de Gruyter.
Vogel, Irene. 2009. The Status of the Clitic Group. In Janet Grijzenhout and Bariş Kabak (eds.), Phonological Domains: Universals and Deviations, Mouton de Gruyter.
Wheeldon, Linda and Lahiri, Aditi. 1997. Prosodic Units in Speech Production. Journal of Memory and Language 37, 356-381.
Wheeldon, Linda and Lahiri, Aditi. 2002. The Minimal Unit of Phonological Encoding: Prosodic or Lexical Word. Cognition 85, B31-B41.
Zwicky, Arnold. 1987. Suppressing the Z's. Journal of Linguistics 23(1), 133-148.

# ON THE SPLIT NATURE OF THE DUTCH LATEN-CAUSATIVE 

Gerlof Bouma and Jonas Kuhn<br>Universität Potsdam

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

We present a discussion of the laten-perception-causative construction in Dutch, formed by combining laten 'let' and a non-volitional perception/cognition verb like zien 'see'. On the basis of partly novel argument alternation and binding data, we show that these causatives cannot be captured by a standard raising analysis. This sets these data apart from other uses of causative laten. The laten-perception-causative is best analyzed as a complex predicate. Interestingly, although its behaviour cannot be explained from the syntactic combinatory rules, the observed argument alternations and binding effects are predictable from Dutch syntax if we were to assume that the laten-perception-causative is a complex member of a semantically coherent group of communication verbs.

After having introduced the monoclausal complex predicate analysis, we approach the question of how the intermediate status of laten-perceptioncausatives between idiomatic constructions and transparent phrasal combinations could be formally captured in an extension of the LFG architecture. We believe that the construction sheds some interesting light on long-standing issues from the complex predicate formation debate, especially when viewed from the perspective of recent considerations about the use of a template hierarchy as a theoretically motivated device in LFG.


## 1 Introduction

The Dutch verbs laten 'let', doen 'do' and dwingen 'force' all take verbal complements and allow one to express causation of the embedded event by the matrix subject. Examples are given in (1). ${ }^{1}$
(1) a. Ik laat mijn man de uien snijden!
ik let my husband the onions cut
'I have my husband cut the onions.' ${ }^{2}$
b. Maar een glas wijn doet de gemoederen steeds weer bedaren.
but a glass wine does the moods always calm
'But a glass of wine will always calm people down.'3
c. Economische crisis dwingt Hongarije wereldwijd ambassades te sluiten. economic crisis forces Hungary worldwide embassies to close 'Economic crisis forces Hungary to close embassies worldwide. ${ }^{4}$

[^54]Causatives in Dutch, and then especially the laten-causative (1a) and the doencausative (1b), have been extensively studied, for instance in the Dutch functionalist literature (Dik, 1980; Verhagen, 1997; Verhagen and Kemmer, 1997; Loewenthal, 2003). As AcI verbs and (possible) cases of object-control, they have been the topic of countless papers in formal syntax, of which the LFG treatment of Dutch verbal clusters presented in Kaplan and Zaenen (2003) is only one.

In this paper, we will restrict our attention to the laten-causatives. In particular, we investigate the combination of laten with a class of perception/cognition verbs. As an example of the construction under investigation, consider (2a), in which we have replaced snijden 'cut' in (1a) with zien 'see'. On the face of it, the result is analogous to (1a). However, there is a contrast between the two verbs when embedded under laten: unlike the agent of snijden, the experiencer of zien can be realized in a PP headed by aan 'to'. This is unexpected, since this option does not appear in the regular argument realizations of zien 'see'.
(2) a. Ik laat mijn man de uien zien.

I let my husband the onions see
'I show the onions to my husband.'
b. Ik laat de uien aan mijn man zien / *snijden.

I let the onions to my husband see / cut.
In fact, as we will see, combinations like laten zien 'let see' differ from combinations like laten snijden 'let cut' in terms of binding, interpretation and argument realization. On the basis of those arguments, we show we can distinguish the laten-perception-causative (LPC). We will argue that the LPC should be modelled as a complex predicate in LFG. Complex predicate analyses for Dutch causatives and/or constructions with laten have been proposed before with differing levels of formalization (Coopmans and Everaert, 1988; Verhagen, 1997; Booij, 2002). However, to our knowledge, there is no detailed description of the LPC as a construction with a highly internally coherent behaviour.

In this paper, we start by describing the laten-causative in section 2 . We discuss a standard analysis of laten as a raising verb. Then we show why this raising model is unsuitable for capturing the LPC subset of laten-causatives. Instead, we follow previously made proposals to treat the LPC as a monoclausal construction. In section 3, we incorporate insights from the functional literature, which lead us to conclude that the LPC seen as a whole is a complex example of a communication verb. This in turn allows us to explain the syntactic behaviour that is anomalous under a raising analysis. Additional binding data for the monoclausal analysis is given in section 4 . Section 5 discusses the base verbs that participate in the LPC in more detail. We make no attempt to provide a fully formalized analysis of the monoclausal LPC in this paper, but the picture that we arrive at puts the LPC in the gray area between a regular syntactic entity and an idiosyncratic, lexically specified, idiomatic construction. In section 6, we therefore discuss some interesting consequences of our analysis when put in the standard LFG architecture as it is currently conceived.

## 2 The Dutch laten-Causative

The verb laten 'let' is, like the perception verbs zien 'see' and horen 'hear' (amongst others), an accusativus cum infinitivo (AcI) verb: it combines with an object NP and an unmarked infinitival VP.
(3) Jan laat / hoort [np hem] [vp een liedje zingen].

Jan lets hears him a song sing
'Jan lets/hears him sing a song.'
Although an AcI with laten 'let' is referred to as the 'laten-causative', there is a range of meanings that the construction can have. Examples of different readings are given in (4): in (a), the embedded event is requested by the matrix subject; in (b), there is coercion by an authority; sentence (c) expresses that the matrix subject will not intervene in the embedded event; and (d) describes a case of mechanical causation.
(4) a. Ik laat een makelaar het huis taxeren.

I let an estate agent the house appraise
'I'm having an estate agent appraise the house.'
b. De professor laat zijn medewerkers al het onderwijs geven.

The professor lets his employees all the education give
'The professor makes his assistants do all the teaching.'
c. Ik laat de baby nog even slapen.

I let the baby a bit longer sleep
'I will let the baby sleep a bit longer.' (='I will not wake the baby yet.')
d. Belg laat het regenen in de woestijn.

Belgian lets it rain in the desert
'A Belgian (citizen) lets it rain in the desert.' ${ }^{5}$
A systematic investigation of the interpretation of Dutch causatives can for instance be found in Verhagen and Kemmer (1997) and Loewenthal (2003). Henceforth, we shall refer to the embedded verb as the 'base verb' (regenen in 4d).

A well-known observation about the laten-causative is that it allows the so called laten-passive. In this construction, the base verb subject is suppressed or demoted to a door-PP (5a). This is also observed for German lassen (Reis, 1976; Gunkel, 1999; Müller, 2002, a.o.). Although more marked, Dutch also allows passive-like AcIs with verbs like horen (5b).
(5) a. Ik laat het huis taxeren (door een makelaar).

I let the house appraise by an estate agent
'I let someone / an estate agent appraise the house.'
b. Ik hoor een lied zingen.

I hear a song sing
'I hear someone sing a song.'

[^55]The laten-passive is passive-like in several ways. The base verb's agent is suppressed or demoted. Just like in the regular passive, demoted agents are marked with door. The laten-passive can be used with many verbs that allow regular passivization, including - to some extent - intransitive verbs that allow an impersonal passive. In fact, it appears that the set of possible laten-passive base verbs forms a subset of the verbs that allow normal passivization (Müller, 2002, makes this claim for German). To illustrate, the verb houden 'keep in possession' does not allow a regular passive. It cannot appear in the laten-passive either.
(6) a. *De lamp wordt (door Jan) gehouden. the lamp is by Jan kept
b. *Hij laat de lamp houden (door Jan). he lets the lamp keep by Jan
Thus far, the data could in principle receive an object-control analysis. In their LFG analysis of word order in the Dutch verb cluster, Kaplan and Zaenen (2003) assume that laten is a raising-to-object verb, that is, it is lexically specified with an $(\uparrow$ OBJ $)=(\uparrow$ XCOMP SUBJ $)$ annotation. A grammar fragment based on Kaplan and Zaenen's grammar is given in (7). An analysis of a simple laten-causative under this grammar is given in (8).
(7) a. $\quad \mathrm{CP} \rightarrow \mathrm{C} \quad \mathrm{VP}$

$$
\begin{aligned}
& \mathrm{VP} \rightarrow \quad \mathrm{NP}^{*} \quad \mathrm{~V}^{\prime} \quad(\mathrm{VP} \mid \mathrm{CP}) \\
& (\uparrow \text { XCOMP* SUBJ } \mid \text { OBJ })=\downarrow \quad \uparrow=\downarrow \quad\left(\uparrow \text { хСОмP* }{ }^{*} \text { COMP }\right)=\downarrow \\
& \mathrm{V}^{\prime} \rightarrow \mathrm{V} \quad\left(\mathrm{~V}^{\prime}\right) \\
& \uparrow=\downarrow \quad(\uparrow \text { хСОмР })=\downarrow \\
& \left(\uparrow \mathrm{XCOMP}^{+} \text {SUBJ } \mid \mathrm{OBJ}\right) ~ \neg<_{\mathrm{f}}(\uparrow \mathrm{NGF})
\end{aligned}
$$

b. laat $\quad \mathrm{V} \quad \mathrm{PRED}={ }^{\prime} \operatorname{let}\langle(\uparrow \mathrm{SUBJ})(\uparrow \mathrm{XCOMP})\rangle(\uparrow \mathrm{OBJ})$,
$(\uparrow$ OBJ $)=(\uparrow$ XCOMP SUBJ $)$
bekijken $\mathrm{V} \quad$ PRED $={ }^{\prime}$ look-at $\langle(\uparrow$ SUBJ $)(\uparrow$ OBJ $)\rangle$ '
(8)


The analysis of the laten-causative is biclausal. The two verbal c-structure nodes each head different f-structures; the only f-structure to be immediately contained in both clauses is the one corresponding to the raised constituent. Special provision
would still have to be made for the passive-like cases. For German, Reis (1976) assumes a kind of unmarked infinitival passive, whereas Gunkel (1999) offers a HPSG solution based on argument inheritance. For Dutch, Everaert (1991) argues on the basis of binding data for a complex predicate solution in a GB context. In the interest of keeping this paper focused on the perception causatives, we reserve investigating the implementation of the laten-passive in LFG for future work.

When the base verb is one out of a small group of perception/cognition verbs the situation changes. As mentioned in the introduction, the combination laten zien allows one to optionally drop the experiencer of zien or to realize it as a aan-PP. Note that this is a bit like the laten-passive, but with a different preposition to head the PP of the demoted argument. Used independently, zien does not allow this neither in the active ( 9 b ) nor the passive ( 9 c ). In addition, in the context of laten zien, realization of the subject of zien in a door-PP is marginal at best (9d).
(9) a. Ik heb het boek (aan iemand) laten zien.

I have the book to somebody see
'I let someone see the book.'
b. *Het boek ziet (aan iemand)
the book sees to somebody
c. Het boek wordt (*aan iemand) gezien the book is to someone seen
d. ??Ik laat het boek door iemand zien.

I let the book by someone see
This pattern can be observed for the following nine verbs: zien 'to see', horen 'hear', ruiken 'smell', proeven 'taste', voelen 'feel', lezen 'read', weten 'know', merken 'notice', blijken 'become clear'. ${ }^{6}$ Verhagen and Kemmer (1997) speculate that laten lezen 'let read' is the only combination that readily allows realizing the base verb's experiencer/agent as either an aan-PP or door-PP. We add that laten proeven 'let taste' also offers this possibility. Because of the type of base verbs that allow this alternation, we will refer to these laten-causatives as 'laten-perception-causatives', or 'LPCs' for short.

The LPC thus presents us with a puzzle: where does the possibility to demote to an aan-PP come from and why do we not have the possibility to demote to a doorPP? From the syntax of raising in Dutch, we do not expect such a thing to happen. Nor does the existence of the laten-passive predict the LPC. The LPC (although not thus-named) has received some attention in the literature on Dutch. Coopmans and Everaert (1988) consider a great number of idiomatic constructions with laten, including LPC instances. They propose laten be analyzed in the GB framework as a causative bound morpheme that 'internalizes the external argument' of its base. The combination laten zien is thus a complex $\mathrm{V}^{0}$ and its newly acquired internal argument (zien's experiencer) may be left unrealized (experiencer suppression) or

[^56]realized as an aan- PP (experiencer demotion). The argument alternations seen with the LPC could then be lexically stipulated for the complex lexical entries with laten. Like other bound morphemes, laten acts as the syntactic head: Of the two verbs in the $\mathrm{V}^{0}$, it is laten that inflects for tense and agreement and moves to the second position in a declarative main clause.

LFG neither forces nor allows us to assume that laten zien is one complex lexical node in c-structure. Complex predicate analyses in LFG have put the monoclausality at f-structure (Butt, 1995; Alsina, 1996; Frank, 1996). Thus, an LPC could receive the analysis as in (10).


The two verbal nodes laten zien now project to one f-structure, with a PRED $=$ 'lpc-see $\left\langle(\uparrow\right.$ SUBJ $)\left(\uparrow \mathrm{OBJ}_{\mathrm{exp}}\right)(\uparrow$ OBJ $\left.)\right\rangle$ '. Cases in which the base verb subject has been demoted to an aan-PP would be captured by specifying PRED $=$
'lpc-see $\left\langle(\uparrow\right.$ SUBJ $)\left(\uparrow\right.$ OBL $\left._{\text {exp }}\right)(\uparrow$ OBJ $\left.)\right\rangle$ ' as one of the alternative LPC annotations. The c-structure in (10) is identical to the analysis of the non-LPC laten-causative in (8). The precise details of the c-structure tree are likely to depend on where one decides to place the complex predicate formation - in the lexicon or in syntax. In particular the subtree below the highest $\mathrm{V}^{\prime}$-node will depend on this. In this paper, we remain relatively agnostic about the technical question of complex predicate formation, but see section 6 for a bit more discussion.

Coopmans and Everaert (1988) deal with idioms. Their aim is to show that complex idioms/idiom-like constructions formed with laten behave as if they where idioms headed by a single verb and to show that one can give combinations with laten an analysis as lexical entries. This, however, means that they do not pay much attention to the systematic argument realization possibilities that exist for both the base verb's experiencer and its theme in the LPC. In our sketch of an LFG analysis, we have not made any provisions for this either, treating the LPC alternations as lexical idiosyncrasies. Before we turn to why these the argument realization possibilities in an LPC are not idiosyncratic in Dutch syntax, let us conclude this section with a more precise overview of these possibilities in the LPC.

Suppression/demotion of the experiencer Our introduction of the LPC started with the observation that a group of base verbs allow demotion of their subject to a
aan-PP. There is more to be said about this, however. First, note that laten on its own does not offer this possibility in general: the laten-passive is formed with door, only the LPC base verbs allow aan.

Secondly, as mentioned before, it appears that the verbs that cannot be used in a regular passive cannot appear in a laten-passive either. There is some evidence that there is no such link between passivizability and suppression/demotion in an LPC. Intuitively, the verbs horen and zien in their readings as non-volitional perception verbs are marked in a passive, but they are fine in an LPC with a demoted experiencer. More concretely, the verb weten 'know' is clearly not passivizable (11a). Again, suppressing or demoting its experiencer in an LPC is unproblematic (11b).
(11) a.
*[Dat de lamp van Jan is] wordt geweten. that it is Jan's lamp is known
b. Hij laat (aan iedereen) weten [dat de lamp van Jan is]. he lets to everybody know that it is Jan's lamp 'He says / tells everybody that the lamp belongs to Jan.'
Thirdly and finally, the base verb blijken 'become clear' is special, in that its normal subject is not the experiencer but the theme. Because argument realization with blijken is not like it is with the other base verbs, we shall return to blijken in section 5.2. Here, we just note that in the LPC, it is the experiencer that can be suppressed or demoted to an aan-PP and not the theme (see also Verhagen, 1997).
(12) a. Ik laat mijn ongenoegen (aan Jan) blijken.

I let my discontent to Jan become clear
'I show Jan my discontent.'
b. *Ik laat Jan (aan mijn ongenoegen) blijken.

The proper generalization is then at the level of the thematic role of experiencer, not the grammatical role of subject: With all of its base verbs, the LPC allows realizing the experiencer a) as an NP; b) as an aan-PP; or c) dropping it altogether. These possibilities cannot be explained from the syntactic properties of raising-laten or of the base verbs alone.

Realization of the theme The LPC base verbs can all realize a propositional theme as a CP complement. The LPC inherits this ability. Examples with horen 'hear' are given in (13ab). A little noted fact about the LPC, however, is that the propositional theme can be realized as a te-marked infinitival VP, too. ${ }^{7}$ The understood subject of this VP may be interpreted as either one of the other LPC participants. A constructed example with the base verb horen 'hear' is in (13d). Used independently, horen cannot realize a propositional theme as a te-VP (13c).
(13) a. Ik hoor dat ik boos ben.

I hear that I angry am
'I can hear that I am angry.'

[^57]b. Ik laat horen dat ik boos ben.

I let hear that I angry am
c. *Ik hoor ${ }_{t e-\mathrm{VP}}$ boos te zijn].

I hear angry to be
d. Ik laat horen $\left[_{t e-V P}\right.$ boos te zijn].

I let hear angry to be
The construction is intuitively more common with the cognition verbs, for instance with weten 'know' in example (14a). The verb weten as used in the LPC does not allow a $t e$-VP theme when used independently. ${ }^{8}$ Examples with a perception verb like horen 'hear' are also attested (14b).
(14) a. het CDA had eerder al laten weten ${ }_{t e-V P}$ onthutst te zijn] the CDA had earlier PART let know upset to be 'The CDA had already said that they were very upset.' ${ }^{9}$
b. Piepend liet het ons horen $[t e-\mathrm{VP}$ klaar te zijn voor de grote wereld.] squeaking let it us hear ready to be for the big world 'Squeaking, [the chick] let us know it was ready for the big world. ${ }^{10}$
To summarize, we see that the LPC offers argument realization possibilities for both the base verb's experiencer and theme that are not offered by the base verb itself. This would be very hard to explain if we were to give the LPC a biclausal raising-toobject analysis. If laten were analyzed as a raising verb in the LPC, we would have to allow it to: a) raise to object or raise to oblique, b) target the embedded subject or the embedded complement, and, most radically, c) be able to target both embedded arguments at the same time (as in (14b), where both the embedded subject and the complement appear in forms not normally allowed by the base verb). This would be at odds with what we know about the rest of Dutch syntax and we suspect it to be highly unexpected from a cross-linguistic perspective, too.

## 3 The LPC as a verb of transfer of a message

So far we have seen that an analysis of the LPC along the lines of Kaplan and Zaenen's (2003) treatment of AcI verbs would fail. It may thus seem attractive to analyze LPCs as built around an idiomatic complex verb (cf. Coopmans and Everaert, 1988; and essentially also in the computational grammar of the Alpino parser, Van Noord, 2006). However, as we shall see in this section and the next, the conclusion that the construction is strictly an idiom, meaning that the various argument realization options have to be explicitly listed as lexical idiosyncrasies, is not warranted. In fact, it would miss important generalizations.

[^58]Functionalist descriptions of Dutch doen- 'do' and laten- 'let' causatives (Dik, 1980; Kemmer and Verhagen, 1994; Verhagen and Kemmer, 1997; Verhagen, 1997; Loewenthal, 2003) have taken the perspective that such causatives should not be treated as complex clausal structures with independent syntactic domains. Instead they have treated them as a whole, as ditransitive constructions. With regards to the LPC, it has been additionally pointed out in this literature that the verb combinations in LPCd do not seem to have one of the 'regular' causal interpretations in the range given in (4). With the cognition verbs this is particularly clear: the combinations laten weten, laten blijken, laten merken do not involve any kind of exerted authority over someone else. Instead, they are statements of something being communicated. Appropriate English translations would for instance be 'tell', 'signal' or 'communicate'. LPCs with a perception base verb and a propositional theme, such as the one in (14b), can also be understood as such. The perception base verb then indicates the means by which the communicated information is perceived. We therefore propose to see the LPC as belonging to the Dutch counterpart of Levin's (1993) class of verbs of transfer of a message (VTM).

This holistic view of the LPC as a VTM sheds immediate light on the argument alternations we see with the construction. Let us take the Dutch VTM vertellen 'tell'. This ditransitive verb takes an agent, a goal and a theme. It allows one to realize the goal as an NP or aan-PP, or to drop the goal completely. The propositional theme can be realized as a pronominal $\mathrm{NP}, \mathrm{a} \mathrm{CP}$ or as a $t e-\mathrm{VP}$. All combinations are possible.
(15) a. Jan vertelt het Piet.
(NP-goal, NP-theme)
Jan tells it Piet
'Jan tells it to Piet.'
b. Jan vertelt Piet dat hij ontslagen is.

Jan tells Piet that he fired is
'Jan tells Piet that he is fired.'
c. Jan vertelt Piet ontslagen te zijn. Jan tells Piet fired to be 'Jan tells Piet that he is fired.'
d. Jan vertelt het aan Piet.
e. Jan vertelt aan Piet dat hij ontslagen is.
f. Jan vertelt aan Piet ontslagen te zijn.
g. Jan vertelt het.
h. Jan vertelt dat hij ontslagen is.
i. Jan vertelt ontslagen te zijn.
(NP-goal, CP-theme)
(NP-goal, VP-theme)

These are the same alternations we see with the LPC. The parallel between a causative construction such as the LPC and VTMs in general is not surprising if we consider decompositions of VTMs into their conceptual components. For instance, tell can (pre-theoretically) be seen as a speaking event that causes a knowing event. The LPC appears to do not much more than spell these two parts out. What is remarkable about the LPC, however, is that by virtue of having a VTM-like meaning,
it inherits VTM argument realization syntax. We can contrast this with the otherwise highly similar German lassen-causative. The German verb zeigen 'show' and the causative combination sehen lassen 'let see' can both describe a situation in which a seeing event is caused. But unlike the Dutch LPC, this conceptual similarity is not enough for the German lassen-causative to behave like zeigen. The ditransitive zeigen takes a dative and an accusative argument (16a) whereas sehen lassen has two accusative arguments - arguably one matrix object and one embedded object (16b):
(16) a. Er zeigt ihr den Riesenmammutbaum.
he shows her.DAT the.ACC giant sequoia
'He shows her the giant sequoia.'
b. Er lässt sie den Riesenmammutbaum sehen.

He lets het.ACC the.ACC giant sequoia see
To conclude the discussion thus far, we have seen that the argument alternations observed in the LPC are hard to explain if we were to assume that the LPC has a biclausal structure in which laten functions as a raising verb. We propose, however, that one can go beyond a monoclausal analysis in which the LPC is treated as an idiom: The LPC is a monoclausal construction formed around a complex instance of a verb of transfer of a message. We emphasize that our argumentation should be taken to apply the LPC alone. For the regular laten-causative, we continue to assume a default biclausal raising analysis.

## 4 Evidence from binding

We take the argument alternation evidence from the previous section as conclusive for a monoclausal analysis. However, it is reasonable to expect that a contrast in clausality between the LPC and the regular laten-causative has effects on binding. The binding data is presented here separately, since its force as evidence for monoclausality co-depends on ones theory of binding, which we will mostly leave implicit here. Preliminary GB-couched discussion of binding data in laten-causatives, including the LPC and the laten-passive, can be found in Everaert (1991).

In (17), we give minimal pairs of laten-causatives with a reflexive pronoun as the complement of the base verb. The contrast is between the LPC base verb zien 'see' and the semantically related but volitional bekijken 'look at', which does not participate in the LPC. When the base verb is bekijken, its reflexive marked theme can only refer to the subject of the base verb. Co-reference with the matrix subject is not possible. In the LPC (17b), however, the reflexive theme of zien can be understood as referring to its experiencer or to the matrix subject.
(17) a. $\quad \operatorname{Jan}_{j}$ laat Piet $_{p}$ zichzelf $_{p / * j}$ bekijken

Jan lets Piet himself look at
'Jan lets Piet ${ }_{p}$ look at himself ${ }_{p}$ '
b. $\mathrm{Jan}_{j}$ laat $\operatorname{Piet}_{p}$ zichzelf $_{j / p}$ zien

Jan lets Piet himself see
${ }^{\prime} \mathrm{Jan}_{j}$ shows $\operatorname{Piet}_{p}$ himself $_{j / p}$.'

```
c. \(\mathrm{Jan}_{j}\) toont Piet \(_{p}\) zichzelf \(_{j / p}\)
```

    Jan shows Piet himself
    The binding facts of (17b) parallel those of (17c), which is a sentence headed by the simplex ditransitive verb tonen 'show'. ${ }^{11}$ To assume that (17a) is biclausal and that (17bc) are monoclausal would put us in a position to explain these data by taking the immediately containing f-structure clause as the domain within which the reflexive has to be bound. Even though these data should be taken as merely suggestive - one could, for instance, formulate a binding theory where the volitionality of the base verb subject is important - they are in line with a monoclausal analysis of the LPC.

## 5 Remarks about the LPC base verbs

In this section we discuss the LPC base verbs in a bit more detail. We will talk about the similarities between the base verbs and their contributions to the LPC. The base verbs all take an experiencer and a theme, but the base verb blijken 'become clear' is a bit different because it assigns the subject function to the theme and by itself already shows some of the argument alternations that we claimed distinguish the LPC from other laten-causatives. We therefore review the evidence for assuming that we have an LPC with the base verb blijken in subsection 5.2.

### 5.1 The role of the base verbs in the LPC

The base verbs that participate in the LPC share that they have readings as nonvolitional verbs of perception or cognition. In terms of thematic roles, they take an experiencer and a theme. The verbs can be divided into two groups as follows:
(18) Perception: zien 'see'; horen 'hear'; ruiken 'smell'; proeven 'taste'; voelen 'feel'; lezen 'read'
Cognition: weten 'know'; merken 'notice'; blijken 'become clear'
Each of these verbs can take a propositional theme, either in the form of an (pronominal) NP or as a subordinate clause (19a). They may also take NPs that denote what is being perceived or cognized. For perception verbs, these may be abstract objects, but also concrete or even animate objects (19b). ${ }^{12}$ Cognition verb themes can in general only be abstract (19c). Exceptions to the latter generalization require a context that is rich enough to allow coercion from reference to an object to reference to the proposition that the object is present, available, etc.

[^59](19) a. Ik hoor / weet het / dat je verdrietig bent.

I hear know it that you sad are
'I can hear / know it / that you are sad.'
b. Ik hoor Jan.

I hear Jan.
'I can hear Jan.'
c. *Ik weet Jan.

I know Jan.
These selection restrictions and semantic properties of the base verbs are transferred to the LPC. For instance, sentence (20a) is perfectly natural with the base verb zien. However, if we use the base verb hear, it is only interpretable if we allow more far fetched readings in which we make noise with the paper the drawings are on. For another instance, the selection restrictions of weten explain why sentence (20b) is out.
(20) a. Ik laat de bouwtekening aan Jan zien / \#horen.

I let the construction drawings to Jan see / hear
'I let John see/\#hear the construction drawings.'
b. *Ik laat Jan aan Piet weten.

I let Jan to Piet know
Although the LPC can be said to be syntactically anomalous - as it shows the alternations of a group of simplex verbs even though it is complex - in terms of its semantics, it appears to behave compositionally. Note that treatment as an idiom would not only trivialize the VTM-like behaviour of the LPC, it would ignore this aspect completely, too.

The non-volitionality of the base verb is a rather strong constraint. For instance, ruiken 'smell' and proeven 'taste' have PP-selecting variants that are volitional ( $\sim$ 'sniff' and 'sample', respectively). See (21).
(21) a. Jan ruikt aan de bloemen.

Jan smells at the flowers
'Jan sniffs the flowers.'
b. Jan proeft van de soep.

Jan tastes of the soup
'Jan samples the soup.'
These variants cannot appear in the LPC. This is illustrated in (22), where marking the theme with a P renders the sentences ungrammatical. Note that in the ungrammatical versions, there are two PPs: an aan-PP realizing the experiencer (only possible in an LPC) and another PP realizing the theme.
...laat (*van) de soep aan Jan proeven let of the soup to Jan taste
'... let Jan taste/*sample the soup.'
b. ...laat (*aan) de bloemen aan Jan ruiken let at the flowers to Jan smell '... let Jan smell/*sniff the flowers.'

We get the same effects with volitional counterparts to see (bekijken 'watch'), hear (beluisteren 'listen to'), and feel (voelen aan 'feel at/touch'). The verb lezen 'read' may be the exception to this non-volitionality constraint. Although seeing lezen as a perception verb is plausible, it is harder to understand it as a non-volitional verb.

## 5.2 blijken is an LPC base verb

The verb blijken 'become clear' has a propositional theme subject (23a) and can optionally take an experiencer in the form of an NP (23b) or an aan-PP (23c). In addition, blijken can function as a raising-to-subject verb, in which case it selects for a non-thematic subject and a $t e-\mathrm{VP}(23 \mathrm{~d})$.
(23) a. [nP Dat] blijkt.
that becomes clear
'That's clear.'
b. Op Sicilië blijkt [ ${ }_{\mathrm{NP}}$ hem] [ ${ }_{\mathrm{CP}}$ dat hij van koninklijke geboorte is]. on Sicily b. clear him that he of royal birth is 'On Sicily, he finds out he is of royal descent.' ${ }^{13}$
c. Blijkt [aan-PP aan de ambtenaar] [CP ${ }_{\text {dat }}$ de leerplichtige becomes clear to the civil servant that the ADJ jongere [...] het onderwijs [...] niet volgt], ...
youth the education not follows 'Should it become clear to the civil servant that the child subject to compulsory education is not attending school, ... ${ }^{14}$
d. $\left.\quad{ }_{\mathrm{NP}} \mathrm{Hij}\right]$ blijkt $\quad{ }_{t e-\mathrm{VP}}$ daar toch wel een neus voor te hebben]. he becomes clear there PART a nose for to have 'It turns out he does have quite a talent for recognizing that.' ${ }^{15}$
We see that blijken on its own offers the realization options for the experiencer and the theme that we have argued distinguish the LPC from raising. If blijken gives us these alternations, what reason do have to think that laten blijken is an LPC and not a case of raising? The answer lies in the observation that blijken alone does not readily allow the combination of raising-to-subject and realizing the experiencer (24a). That is, blijken does not select for a te-VP theme and another thematic argument. In the LPC, however, realizing the theme as a $t e-\mathrm{VP}$ and realizing the experiencer is fine ( 24 bc ).

> a. ??[$\left.{ }_{\mathrm{NP}} \mathrm{Hij}\right]$ bleek $\left.\quad{ }_{\mathrm{NP}} \mathrm{me}\right]\left[\begin{array}{l}t-\mathrm{VP} \\ \text { ardig te zijn]. }\end{array}\right.$
> he became clear me nice to be

Intended: 'I learnt he was a nice person.'

[^60]b. Chirac liet [aan-PP aan Kok] blijken $\left[_{t e-\mathrm{VP}}\right.$ open te staan voor onze Chirac let to Kok b. clear open to stand for our voorstellen]. proposals
'Chirac made it clear to Kok that he was open to our proposals. ${ }^{16}$
c. Daarvóór liet ze [ ${ }_{\mathrm{NP}} \mathrm{me}$ ] nooit blijken [te-vP zulke fantasieënte hebben]. before let she me never b. clear such fantasies to have 'She had never made clear to me she had such fantasies before. ${ }^{17}$
The combination of arguments in (24bc) cannot be easily explained from a raising analysis and the syntax of blijken, but it is unremarkable for an LPC.

The corroborating binding evidence for monoclausality presented in section 4 cannot be applied to laten blijken since it cannot take an animate theme. Still, binding in laten blijken suggests that a raising analysis is wrong, too. As said, the experiencer of blijken is not the subject. In a raising analysis, this would mean that the experiencer does not become a clause-mate of the matrix subject. Still, the matrix subject can bind a reciprocal experiencer in laten blijken.
(25) a. We boften, lieten we ${ }_{i}$ elkaar $_{i}$ woordeloos blijken [.] we were lucky let we each other word-less become clear 'We were lucky, we conveyed to each other without words.' ${ }^{18}$
b. We lieten aan elkaar blijken dat we steeds meer gevoelens kregen. we let to each other b. clear that we ever more feelings got 'We made clear to each other that we were developing affections. ${ }^{19}$
An important consequence of the fact that blijken is an LPC base verb is that the proper generalization of LPC behaviour cannot be stated in terms of grammatical functions alone (Verhagen, 1997, for points to the same effect), it needs to refer to argument structure/semantics/thematic roles of the base verbs. This means either that such information needs to be made available in syntax (if we place LPC formation in syntax) or that LPC formation is lexical.

### 5.3 Summary

In the previous three sections, we have argued that we should distinguish a special subgroup of laten-causatives, the laten-perception-causative or LPC. Although the LPC on the surface contains a embedded verbal constituent, we cannot explain LPC behaviour with the standard Dutch syntax for such embeddings (in this case: through raising), which assumes two clausal domains. A monoclausal analysis is to be preferred, in which the fact that the combination of laten and a specific group of

[^61]base verbs receives special treatment has to be stipulated. However, the conclusion that the LPC is a pure idiom with an idiosyncratic syntax and a non-compositional semantics would be incorrect. Rather, the LPC is special because it takes on the syntactic behaviour - in terms of argument alternations and in terms of binding of the group of simplex verbs of transfer of a message that it overlaps in meaning with. At construction level, LPC syntax is determined by its (largely) compositional semantics, not by the syntax of its parts. This presents us with a special kind of construction that sits between the extreme points of following regular syntactic patterns and needing full lexical specification.

## 6 Consequences for Formalization and LFG Architecture

### 6.1 Complex predicate formation in LFG

The formalization of complex predicate constructions has received considerable attention in the LFG literature, especially in the 1990's (amongst others Butt, 1995; Alsina, 1996; Frank, 1996; Ackermann and Webelhuth, 1998; Andrews and Manning, 1999; Butt et al., 2003; Wedekind and Ørsnes, 2003; a recent overview is provided in Butt and Seiss, 2009). The intuitive modelling goal of monoclausality can be stated very clearly in the LFG framework: a single f- and a-structure corresponds to both lexical parts of the complex predicate, both of which contribute to the resulting interpretation. The technical ways of achieving this goal are more controversial, as some default assumptions of LFG have to be modified for an effective and satisfactory implementation. Normally, lexical items that do not have a vacuous semantics introduce their own semantic form under PRED. Hence, unification of the two lexical parts of a complex predicate is formally not an option, since it will lead to a clash of the semantic forms. Most authors agree that a special formal operation is required to model the merger of the subparts of a complex predicate into a new semantic predicate. In (26), the effect of this operation is sketched for an LPC example (following Alsina, 1996; Frank, 1996).


Opinions differ on where to locate this merger operation. Butt (1995) and Alsina (1996) assume it to be part of the c-structure rule annotations for verb complex formation. Frank (1996) ultimately argues for a lexicon-based approach, proposing a lexical rule that affects two word forms at the same time, resulting in two 'modified' versions of their lexicon entries which are constrained such that they can only enter an analysis in tandem (Frank, 1996, sec. 4.3). A massively simplified sketch of an application of such a rule to form an LPC is given in (27).

```
laten \(_{1} \quad \mathrm{~V} \quad(\uparrow\) PRED \()={ }^{\prime} \operatorname{let}\langle(\uparrow\) SUBJ \(),(\uparrow\) COMP \()\rangle\),
zien \(_{1} \mathrm{~V} \quad(\uparrow\) PRED \()=' \operatorname{see}\langle(\uparrow\) SUBJ \(),(\uparrow\) OBJ \()\rangle\) '
\(=[\) LPC formation \(] \Rightarrow\)
laten \(_{2} \quad \mathrm{~V} \quad(\uparrow \mathrm{PRED})={ }^{\prime} \mathrm{lpc}-\mathrm{see}\left\langle(\uparrow \mathrm{SUBJ}),\left(\uparrow \mathrm{OBJ}_{\theta}\right),(\uparrow \mathrm{OBJ})\right\rangle\) '
    \((\uparrow\) COPRED \()={ }_{c}\) see
zien \(_{2} \quad \mathrm{~V} \quad(\uparrow\) COPRED \()=\) see
```

Both formalization options require some amount of technical effort to make them work. In a syntax-based approach, standard assumptions about the c-structure-tof -structure mapping have to be modified. Traditionally, the identity relation holds between the f-structures of an X-bar-projected node (by $\uparrow=\downarrow$ ). With the restriction operator (Kaplan and Wedekind, 1993; Butt et al., 2003), more differentiated constraints about the two f-structures can be expressed in a general way. In particular, it can be stated that all but the PRED information (and additional information subject to the predicate merger) is systematically 'projected' in f-structure, but the two f-structures are no longer identical (see also Ackermann and Webelhuth, 1998; Andrews and Manning, 1999). Conceptually, the question arises what the limitations on such an operation in syntax are, given the principle of Lexical Integrity. In fact, on the whole, the syntax-based approach seems to require that many mapping phenomena that have traditionally been assumed to be lexical, like passivization or dative alternation, be technically treated in syntax.

The lexicon-based approach requires the standard LFG apparatus to be augmented with lexical rules that operate on two items simultaneously. To our knowledge, this has not been implemented. However, there do not seem to be any principled problems: In the XLE system, for instance, lexicon entries are not read into the system at the time of grammar compilation, but 'on demand' at parsing time, as particular forms are found in the input string. It is conceivable that dyadic lexical rules, which are triggered by two forms in the string, can be implemented in this step without much additional complexity. Similarly, at a more cognitive level, it seems reasonable to assume a model in which a word triggers activation of particular lexical entries of associated material.

Independent of the formalization of complex predicate formation chosen, it would be straightforward to implement an analysis of the LPC with the monoclausal analysis shown in (10) above. The fact that the alternation patterns for the LPC correspond to the alternations found for VTMs (section 3) would follow from the fact that the complex predicate resulting from the merger includes the same list of thematic roles as a simplex VTM. Any Lexical Mapping Theory account predicting the VTM alternations will generalize to the LPC.

Rather than spell out a particular formal account in full detail, we will use the remainder of this paper to discuss how the intermediate status of LPC between idiosyncratic phrasal combinations and fully general, compositional syntactic patterns might be captured in the LFG framework.

### 6.2 What can the evidence from the LPC tell us?

Before discussing implications of the LPC for the formal LFG architecture, let us point out that we do not intend to necessarily make a point about complex predicates in general. As the data discussion showed, the lexical entries participating in the LPC are fairly restricted. In prototypical complex predicate constructions, such as light verb constructions, one of the participating items is much less restricted. Nevertheless, LPC is a systematic construction that applies to laten and a semantically coherent class of base verbs, and displays monoclausal properties. Hence, we are in need of a systematic monoclausal analysis for this set of semantically non-vacuous items, that is, a complex predicate analysis. ${ }^{20}$ Quite clearly, there is a range of more and less restricted complex predicate constructions - cross-linguistically (as for instance Frank, 1996 points out in her discussion of complex predicates in French vs Italian), but presumably also within the same language.

The question we would like to raise here is how the systematic semantic grouping of LPC base verbs could be formally captured, while at the same time allowing for idiosyncratic exceptions, like the exclusion of near-synonymous verbs. Note that neither syntax-based nor lexicon-based complex predicate formation provides an uncontroversial handle for this. While the observed restrictions may intuitively have a more natural place in a lexicon-based account, there is nothing in a plain classical LFG lexicon that would make this systematicity explicit. The use of template hierarchies to these ends has been discussed in some more recent LFG work (Dalrymple et al., 2004; King et al., 2005; Asudeh et al., 2008), inspired by inheritance hierarchies in the HPSG tradition (Flickinger, 1987) and from Construction Grammar. As the work by Wechsler (1995) and Davis (2001) shows, the combination of systematic and idiosyncratic effects can indeed be captured in such a hierarchy.

With an eye on the complex predicate formation debate, it is worth noting that the discussions of template hierarchies by Dalrymple et al. (2004) and Asudeh et al. (2008) address their application in f -annotations in both the lexicon and syntactic rules. Complex predicate formation in syntax could in principle be augmented with an template inheritance model just like the lexical approach. We take this as an indication that the question of complex predicate formation in syntax vs the lexicon may simply not be the most pressing one. What may be more interesting is the degree of restriction and systematicity of the items that undergo complex predicate formation. A continuum of possibilities could in principle be captured by a template inheritance account. This continuum could apply at any level of the lexicality-phrasality dimension, that is, to individual lexical items, to multi-word units, and to constructions. What we see is that two independent dimensions are separated out which tended to be interleaved in classical considerations, where syntax is the place for systematic effects and the lexicon the place for idiosyncratic

[^62]listings (even though LFG has been more differentiated with regard to the lexicon from the beginning). This separation of dimensions is sketched in (28), where the wide ranges for constructions, etc., indicate that there is a whole spectrum of variants in each case, ranging rather systematic to highly idiosyncratic. ${ }^{21}$

|  | Systematicity |  |  |
| :---: | :--- | :--- | :--- |
| Location | fully systematic | $\longleftrightarrow$ | idiosyncratic |
| grammar <br> $\downarrow$ | syntactic rules |  | constructions <br> lexicon |
|  |  | $\longmapsto$multi-word expressions <br> lexical entries |  |
|  | general (default) <br> descriptions / <br> templates |  | item-specific <br> descriptions / <br> templates |

In order to capture the effect that the LPC ends up with the properties of a VTM because the complex predicate meaning matches it, an inheritance mechanism would be required. This may go beyond the original intention of templates for organizing descriptions in the lexicon and the grammar. To effectively generate two PRED values for the two parts of the LPC, the template mechanism has to be able to apply to two items simultaneously, very much like Frank's (1996) lexical rule. If such a mechanism is in place, it would again seem to be applicable both to a syntax-based and to a lexicon-based account. Such a dyadic template mechanism yet needs to be defined, but following the separation of dimensions sketched in (28), it may be exactly what is required to provide a compact formal way of talking about the systematicity dimension not just for individual lexical items and opaque multi-word combinations, but also for syntactically complex, semi-transparent units. A tentative sketch of the use of a dyadic template for capturing the place of LPCs in an inheritance hierarchy is given in (29).


[^63]
## 7 Conclusion

We argued that a subgroup of laten-causatives in Dutch, the laten-perceptioncausative or LPC should receive a monoclausal analysis while other laten-causatives are appropriately modelled with a biclausal raising analysis. In an LPC, laten and the embedded perception/cognition verb form a complex predicate that is not only semantically similar to the verbs of transfer of a message, but also has assimilated to them syntactically. This explains the otherwise unexpected argument alternation behaviour of the LPC. The construction is an interesting case of a phrasal combination situated halfway between a fully idiosyncratic construction and a transparent phrasal combination of words.

The technical implementation of complex predicate formation in LFG has been debated for a fairly long time. We addressed the option of a formation account in the syntax and one in the lexicon. The intermediate status of LPC seems to justify a re-evaluation of certain aspects of this debate, in particular when viewed from the perspective of relatively recent LFG work on a theoretically motivated use of template inheritance hierarchies for capturing the organization of systematic versus idiosyncratic knowledge.

## References

Ackermann, Farrell and Webelhuth, Gert. 1998. A Theory of Predicates. Stanford, CA: CSLI Publications.
Alsina, Alex. 1996. The Role of Argument Structure in Grammar. Stanford, CA: CSLI Publications.
Andrews, Avery D. and Manning, Christopher D. 1999. Complex Predicates and Information Spreading in LFG. Stanford, CA: CSLI Publications.
Asudeh, Ash, Dalrymple, Mary and Toivonen, Ida. 2008. Constructions with Lexical Integrity: Templates as the Lexicon-Syntax Interface. In Miriam Butt and Tracy H. King (eds.), Proceedings of the LFG08 Conference, Stanford, CA: CSLI Publications Online.
Booij, Geert. 2002. Constructional Idioms, Morphology, and the Dutch Lexicon. Journal of Germanic Linguistics 14(4), 301-329.
Butt, Miriam. 1995. The Structure of Complex Predicates in Urdu. Stanford, CA: CSLI Publications.
Butt, Miriam, King, Tracy Holloway and III, John T. Maxwell. 2003. Complex Predicates via Restriction. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG03 Conference, pages 92-104, Stanford, CA: CSLI Publications.
Butt, Miriam and Seiss, Melanie. 2009. Complex Predicates: An Overview, Ms, University of Konstanz.
Coopmans, Peter and Everaert, Martin. 1988. The Simplex Structure of Complex Idioms: The Morphological Status of laten. In M. Everaert, A. Evers, M.A.C. Huybregts and M. Trommelen (eds.), Morphology \& Modularity. In Honour of Henk Schultink, pages 75-104, Dordrecht: Foris Publications.
Dalrymple, Mary, Kaplan, Ronald M. and King, Tracy Holloway. 2004. Linguistic Generalizations over Descriptions. In Miriam Butt and Tracy H. King (eds.), Proceedings of the LFG04 Conference, Stanford, CA: CSLI Publications Online.

Davis, Anthony R. 2001. Linking by Types in the Hierarchical Lexicon. Stanford, CA: CSLI Publications.
Dik, Simon. 1980. Studies in functional grammar. London/New York: Academic Press.
Everaert, Martin. 1991. Contextual determination of the anaphor/pronominal distinction. In Jan Koster and Eric Reuland (eds.), Long Distance Anaphora, pages 77-119, Cambridge: Cambridge University Press.
Flickinger, Daniel. 1987. Lexical Rules in the Hierarchical Lexicon. Ph. D.thesis, Stanford University, Stanford, CA.
Frank, Anette. 1996. A note on complex predicate formation: Evidence from auxiliary selection, reflexivization, and past participle agreement in French and Italian. In Miriam Butt and Tracy H. King (eds.), Proceedings of the LFG96 Conference, Stanford, CA: CSLI Publications Online.
Gunkel, Lutz. 1999. Causatives in German. Theoretical Linguistics 25, 133-159.
Kaplan, Ronald and Wedekind, Jürgen. 1993. Restriction and correspondence-based translation. In Proceedings of the Sixth European Conference of the Association for Computational Linguistics, pages 193-202.
Kaplan, Ronald and Zaenen, Annie. 2003. West-germanic verb clusters in LFG. In P. Seuren and G. Kempen (eds.), Verb Constructions in German and Dutch, pages 127-150, Amsterdam: John Benjamins.
Kemmer, Suzanne and Verhagen, Arie. 1994. The grammar of causatives and the conceptual structure of events. Cognitive Linguistics 5(2), 115-156.
King, Tracy H., Forst, Martin, Kuhn, Jonas and Butt, Miriam. 2005. The feature space in parallel grammar writing. Research on Language and Compuation 3, 139-163.
Levin, Beth. 1993. English Verb Classes And Alternations: A Preliminary Investigation. Chicago/London: The University of Chicago Press.
Loewenthal, Judith. 2003. Meaning and use of causeeless causative constructions with laten in Dutch. In A. Verhagen and J. van de Weijer (eds.), Usage-Based Approaches to Dutch, Lot Occasional Series, No. 1, pages 97-130, Utrecht: LOT.
Müller, Stefan. 2002. Complex predicates: verbal complexes, resultative constructions and particle verbs in German. Stanford, CA: CSLI Publications.
Reis, Marga. 1976. Reflexierung in deutschen A.c.I.-Konstructionen. Ein transformationsgrammatisches Dilemma. Papiere zur Linguistik 9, 5-82.
van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister and Patrick Watrin (eds.), TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles, pages 20-42.
Verhagen, Arie. 1997. Causatieve en di-transitieve constructies. In Els Elffers van Ketel, Joop van der Horst, and Wim Klooster (eds.), Grammaticaal Spektakel. Artikelen aangeboden aan Ina Schermer-Vermeer bij haar afscheid van de vakgroep Nederlandse Taalkunde aan de Universiteit van Amsterdam, pages 231-235, Amsterdam: UvA.
Verhagen, Arie and Kemmer, Suzanne. 1997. Interaction and causation: Causative constructions in modern standard Dutch. Journal of Pragmatics 27, 61-82.
Wechsler, Steven. 1995. The Semantic Basis of Argument Structure. Stanford, CA: CSLI Publications.
Wedekind, Jürgen and Ørsnes, Bjarne. 2003. Restriction and Verbal Complexes in LFG: A Case Study for Danish. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG03 Conference, pages 424-450, CSLI Publications.

# CLITICS AS CALCIFIED PROCESSING STRATEGIES 

Miriam Bouzouita Stergios Chatzikyriakidis<br>Universitat de les Illes Balears<br>King's College London

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

In this paper we examine clitic placement in Medieval Spanish (MedSp) and Renaissance Spanish (RenSp) as well as the Person Case Constraint (PCC) in Modern Spanish (ModSp), arguing that a natural explanation for these phenomena can be given once we assume clitics to be the encoding of calcified processing strategies of an earlier freer word order system ( Bouzouita 2008a, 2008b, 2008c; Kempson et al. 2008; Kempson \& Cann 2007; Kempson \& Chatzikyriakidis 2009; Chatzikyriakidis forthcoming). We show that the availability of different parsing strategies being possible for one and the same string, led to cases where reanalysis in terms of the parser gave rise to syntactic change. Assuming that each clitic in effect matches one of the four different parsing strategies of the earlier Latin scrambling system, the PCC facts are straightforwardly accounted for. Assuming that syncretized and dative clitics involve the projection of an unfixed node with no form of update, any combination of $1 \mathrm{st} / 2$ nd clitics or a 3 rd dative plus a $1 \mathrm{st} / 2 \mathrm{nd}$ clitic is predicted to be illicit by a very general constraint on tree-growth, the fact that no more than one unfixed node with the same underspecified address can be present in the tree structure, since by definition these two will collapse into one by means of tree-node identity.


## 1 Introduction

In this section, we will give an overview of the data that will later on be examined from a Dynamic Syntax (DS) perspective. Firstly, we shall discuss the diachrony of Spanish clitic placement with respect to the finite verb from 13th century Medieval Spanish (MedSp) to 16th century Renaissance Spanish (RenSp). Subsequently, the Person Case Constraint (PCC) phenomenon will be examined and exemplified.

### 1.1 The Diachrony of Spanish Clitic Placement

In order to discuss the historic development of Spanish clitic placement, we shall examine both the clitic systems found in 13th century MedSp and in 16th century RenSp -the terminus a quo and ad quem of the study reported in this section- from a synchronic point of view. ${ }^{1}$ In this paper, we limit our attention to clitic placement with respect to the finite verbs. Furthermore, we will focus above all on the root clause environments as this is the locus where diachronic changes can be perceived.

### 1.1.1 Medieval Spanish

Whereas in non-root clauses the predominant clitic position is the preverbal (anteposition) one, in root contexts it is the postverbal (postposition) one. The study reported in Bouzouita (2008b: 238), for instance, records only 25\% (507/2026) of anteposition for 13th century root clauses. Notwithstanding this, some syntactic environments only license preverbal clitics, as illustrated below: ${ }^{2}$

[^64]| $\frac{\text { Que }}{}$ | vos | dara |
| :--- | :--- | :--- |
| what | CL-DAT | give-will.3SG |

'What will he give you?' (Faz.: 104)

| No | $\boldsymbol{l o}$ | matemos |
| :--- | :--- | :--- |
| not | CL-ACC | kill.1PL |

'Let's not kill him.' (Faz.: 51)

| mayor mal | les | fizo |
| :--- | :--- | :--- |
|  | CL-DAT | made.3SG |

'More pain he caused them.' (Faz.: 65)
$e$ ael se aplegaran pueblos and to him CL-REFL move-closer-will.3PL people 'and to him people will flock.' (Faz.: 210)

| Grief | me | es |
| :--- | :--- | :--- |
| hard | CL-DAT | is.3SG |

'It is hard for me.' (Faz.: 106)

Root clauses in which a wh-element, a negation adverb, a non-coreferential complement, or a prepositional or a predicative complement appears at the left-periphery, as illustrated by examples (1) to (5) respectively, all invariably display preverbal clitics. Moreover, unlike in Modern Spanish (ModSp), the mood of the verb does not influence the clitic's position in MedSp: clitics appearing with indicative, imperative, subjunctive or optative verbs will always exhibit anteposition in these syntactic environments (for an extensive discussion see Bouzouita 2008a: chapters 2-3). The same is also true for those syntactic contexts that always appear with postposition or those that license variation. Examples (6) to (9) illustrate the former. Postposition is the only clitic position observed for root clauses in which the verb appears sentence-initially, paratactic ones, or those with a contrastive coordination conjunction such as mas 'but' or those with a vocative, as exemplified below. The restriction of clitics appearing sentence/clause-initially has also been called the Tobler-Mussafia Law.
Bendixolos en es dia
blessed.3SG-CL-ACC on that day
'He blessed them on that day.' (Faz.: 59)

| Esta aqui, e yré yo; | encontrarmé | yo | con | mio | Sennor |
| :--- | :--- | :--- | :--- | :--- | :--- |
| is.3SG here and go-will.1SG I | meet-CL-will.1SG | I | with | my | Lord |


| $\frac{\text { mas }}{}$ | dales | comer | $e$ | $a$ | bever |
| :--- | :--- | :--- | :--- | :--- | :--- |
| but | give. 2 SG-CL-DAT to eat | and | to | drink |  |


| Selo | $\frac{\text { mio fijo }}{}$ | selo |
| :--- | :--- | :--- |
| know.1SG-CL-ACC | my son | know.1 SG-CL-ACC |
| 'I know, my son, I know.' (Faz.: 59) |  |  |

Notice that example (7) contains a so-called mesoclitic pronoun in the analytic future encontrarmé. As has been shown in great detail in Bouzouita (forthcoming), such analytic futures (and conditionals) appear in the same
syntactic environments as postverbal clitics with other tenses and can therefore be considered to be postposition cases.

Besides the exclusively preverbal and postverbal cases given above, there are also syntactic environments in which synchronic variation between preverbal and postverbal can be discerned. This is for instance the case following left-peripheral subjects, and coreferential objects, as shown in examples (10)-(11) and (12)-(13) respectively.
e Dios tornolo en bien
and god turned.3SG-CL-ACC in good
'and god turned it into good.' (Faz.: 61)
$e$ Dios te quiere demostrar [...]
and god CL-DAT wants.3SG show
'and god wants to show you [what he has to tell you].' (Faz.: 54)
Zebee e Salmana descabeçolos Gedeon
Zebee and Salmana decapitated.3SG-CL-ACC Gideon
'Zebee and Salmana, Gideon decapitated them.' (Faz.: 110)
$\begin{array}{lll}\frac{\text { todo }}{\text { everything }} \quad \text { lo } & \text { CL-ACC } & \text { quemo } \\ \text { burned.3SG } \\ \text { 'Everything he burned it.' } & (\text { Faz.: 111) }\end{array}$

The minimal pair in examples (10) and (11), in which the subject Dios 'god' appears at the left-periphery, clearly illustrates that one and the same syntactic environment can license both anteposition and postposition. It has been reported that subjects preceding preverbal clitics seem to be emphatic (e.g. Bouzouita 2008a: 88-99, 2008b, 2008c; Castillo Lluch 1996). By contextual analyses for subject cases found in notarial and other prose texts, Granberg (1988: 200), for instance, demonstrated convincingly that there exists a relationship between the presence/absence of emphasis on the left-peripheral subjects and the placement of the following clitic(s). Despite such contextual analyses being subjective, we shall assume this hypothesis to be correct as other Iberian-Romance languages, such as Modern Galician and Modern Asturian, have a clitic placement system the underlying principles of which are similar, if not identical, to the MedSp one (Alvarez Blanco et al. 1986: 184; Campos 1989: 22; Xove Ferreiro 1986: 522-523; Academia de la Llingua Asturiana 2001: 366-367; D’Andrés Díaz 1993: 36-43; González i Planas, 2007; Sánchez Vicente \& Rubiera Tuya 1985: 77).

Coreferential objects (Clitic Left Dislocation/Hanging Topic Left Dislocation, CLLD/HTLD) contexts also exhibit variation, as shown in (12)-(13). Unlike the subject environment which allows variation following a whole range of different types of subjects (Bouzouita 2008a: 63-74), anteposition in CLLD/HTLD cases has only been observed following the quantifiers $\operatorname{todo}(s)$ 'all' and $\operatorname{am}(b) o s(s)$ 'both'. At first sight, the variation in these latter cases seems thus more restricted. However, it has been proposed that in these environments too the left-peripheral constituents are emphatic when co-occurring with preverbal clitics (Bouzouita 2008a: 108). Moreover, cross-linguistic evidence from Modern Galician supports (once again) this claim: Xove Ferreiro (1986: 527-528) states in this respect that both preverbal and postverbal placement is possible following left-peripheral quantifiers, depending on whether they are focussed or not. He also notes that, because certain quantifiers appear overwhelmingly with a certain clitic position, some grammars claim that only anteposition is found following todos. Nonetheless, postposition is also an option, although very rare. Remarkably, this parallelism between Modern Galician and MedSp seems to have gone unnoticed in the current literature, as far as we are aware. Variation between anteposition and
postposition has also been attested following left-peripheral adverbials, coordination contexts with $e(t) / y$ 'and' and preceding non-root/absolute clauses. Due to space limitations, we shall not dwell on the details of this syntactic variation and we refer the readers to Bouzouita (2008a: chapters 2-3; 2008b) for an extensive discussion.

### 1.1.2 Renaissance Spanish

From the beginning of the 15th century onwards, diachronic changes can be observed (Arias Alvarez 1995; Eberenz 2000: 133; Nieuwenhuijsen 1999: chapter 5 inter alia). The Tobler-Mussafia Law, for instance, starts eroding as exemplified in example (14), in which the clitic se precedes the finite verb dize without there being a left-peripheral constituent present. In fact, all those syntactic environments that in MedSp only permitted the use of postverbal clitics have acquired by the 16th century the possibility of occurring with preverbal clitics as well, as illustrated schematically in Table 1 which gives a side-by-side overview of both MedSp and RenSp clitic placement.
Se dize publicamente que [...]
'Publicly it is being said that [...].' (DLNE: 1529.9)

| al señor su marido | $\boldsymbol{l e}$ | dé | mis | besamanos |
| :--- | :--- | :--- | :--- | :--- |
| to-the lord your husband | CL | give.3SG | my | regards |

'Give my regards [lit.: hand-kiss] to your husband.' (DLNE: 1529.7)

Table 1: Clitic Placement in Root Clauses

| Syntactic Environment | Medieval Spanish |  | Renaissance Spanish |  |
| :---: | :---: | :---: | :---: | :---: |
|  | Antepos. | Postpos. | Antepos. | Postpos. |
| Wh-element or si/sy | X | - | X | - |
| Negation | X | - | X | - |
| Non-coreferential object complement | X | - | X | - |
| Prepositional complement | X | - | X | - |
| Predicative complement | X | - | X | - |
| Verb in 1P | - | X | X | X |
| Paratactic root clause | - | X | X | X |
| Contrastive coordination pero/mas 'but' | - | X | X | X |
| Vocative | - | X | X | X |
| Subject | X | X | X | X |
| Adverbial | X | X | X | X |
| Coreferential object (CLLD/HTLD) | X | X | X | X |
| Coordination $e(t) / y$ 'and' | X | X | X | X |
| Non-root/absolute clause | X | X | X | X |

Apart from sentence-initial cases such as example (14), the novel anteposition environments include those which involve a preceding paratactic root clause, a contrastive coordination conjunction pero/mas 'but' or a vocative,
as shown in Table 1. Furthermore, not only the exclusively postverbal environments are affected by diachronic changes but also the variation ones (see also Bouzouita 2008a: chapter 7, 2008b: 241-245; Keniston 1937: 93). Recall that we saw in section 1.1.1 that for certain syntactic environments preverbal placement was observed under specific conditions. In RenSp, however, these restricting conditions no longer seem to apply. Consider for instance the CLLD/HTLD example in (15): as can be seen, there is a preverbal clitic present despite the fact that no quantifiers appear at the left-periphery (cf. example 13). Similar loss of restrictions is also observed for the other variation contexts (see Bouzouita 2008a: chapter 7 for more details). It is also important to note that in RenSp, as in MedSp but unlike ModSp, mood does not affect clitic placement.

In sum, preverbal placement has become more widespread in $\operatorname{RenSp}$ as (i) the environments that were previously only licensing postposition also acquired the possibility of appearing with preverbal clitics and (ii) the conditions under which anteposition was licensed in the MedSp variation environments no longer apply. As we shall see in section 2.2 , the loss of these restrictions has been implemented in the account given here as the lexical simplification of the lexical entry of the clitic. ${ }^{3}$

Now that the MedSp and RenSp data have been broadly discussed, we shall turn our attention to the Person Case Constraint.

### 1.2 The Person Case Constraint

The Person Case Constraint (PCC) is a clitic co-occurrence restriction which basically bans certain combinations of clitics from co-occurring. There are a number of different versions of the constraint. The most restrictive one, exhibited by languages like French, Greek, Spanish, Italian among others, states that 1st/2nd person accusative clitics are precluded in the presence of a dative clitic: ${ }^{4}$

| *Elle | me | lui | $a$ | donné |
| :---: | :--- | :--- | :--- | :--- |
| she | me.CL | him.CL-DAT | has | given |

'She has given me to him.' [French]

| *Gli | $\boldsymbol{m i}$ | ha | dato |
| :--- | :--- | :--- | :--- |
| them |  |  |  |

'He/She has given me to them.' [Italian]

| *Le | $\boldsymbol{m e}$ | $h a$ | dado |
| :--- | :--- | :--- | :--- |
| him.CL-DAT | me.CL-ACC | has | given |

'He/She has given him to me.' [Spanish]

| *Mu | se | exi |
| :--- | :--- | :--- |
| me.CL-DAT | you.CL-ACC | has |
| given |  |  |

There are a number of other variants of the PCC, namely the weak PCC (Bonet 1991, 1994; Anagnostopoulou 2005; Bianchi 2006) and the Romanian PCC (Sǎvescu 2007; Nevins 2007), but we will not deal with them here

[^65]for reasons of space. The interested reader is referred to Kempson \& Chatzikyriakidis (2009) for a full analysis of the PCC including these two variants. In this paper, we limit ourselves to the discussion of the strong PCC.

What makes the PCC rather peculiar is the fact that the restriction does not seem to be semantic at all, with paraphrases where one of the clitics has been substituted by a strong pronoun being well-formed: ${ }^{5}$

| Me | sistisan | se sena |
| :--- | :--- | :--- |
| me.CL-ACC | introduced | to you.ACC |
| 'They introduced you to me.' | $[$ Greek $]$ |  |

Elle m’ a donné a lui
she me.CL has given to him
'She has given me to him.' [French]

A number of approaches have been proposed in the literature. There are functional accounts such as Haspelmath (2002), arguing that the ungrammaticality of the PCC combinations is due to the inability of these clitic clusters to get grammaticalized, assuming low frequency rates of the latter. Furthermore, there are templatic approaches arguing for a separate level of morphology (Bonet 1991, 1994; Heap 2003). Lastly, pure syntactic accounts of the phenomenon also exist, mainly within the minimalist framework (Rezac 2003, 2008; Anagnostopoulou 2003, 2005; Adger \& Harbour 2007; Nevins 2007; Sǎvescu 2007, 2009 inter alia). The common denominator in all the latter analyses is that the PCC is argued to arise from a feature checking failure against a functional head. There are different formalizations of the latter assumption, e.g. the assumptions made as regards the nature of the agreeing head with respect to checking (multiple vs. single Agree) or the stipulated features that different kinds of clitics are argued to have (no person feature for 3rd person accusative clitics in Anagnostopoulou 2003, 2005 vs. the presence of a person feature for the same clitic, although negatively specified [-], in Nevins 2007), but the core proposal effectively shares the same intuitions. It is not our intention to give a full review of the PCC literature (see Kempson \& Chatzikyriakidis (2009) for such a review) and the problems that do arise in these existing analyses. However, we should note that a number of stipulatory decisions must be made in all these analyses as regards the feature specifications of each clitic, in order for all these analyses to work. Notable examples include Adger \& Harbour's assumption that all indirect objects carry a [+Participant] feature, independently disputed by Bonet (2007) and Kempson \& Chatzikyriakidis (2009), and Anagnostopoulou's $(2003,2005)$ assumption that 3rd person dative clitics bear a person feature, albeit specified as minus [-], whereas 3rd person accusative clitics bear no such feature at all. There are a number of other similar examples plus a number of other stipulations used in these analyses but we will not go through them. The interested reader is referred to Kempson \& Chatzikyriakidis (2009) for an extensive critique of such approaches.

In this paper, we will argue that the PCC can be naturally explained as a processing constraint, according to which no more than one underspecified relation can be used simultaneously. The PCC will then be argued to be no more than a restriction on underspecification, a hard-wired constraint of the tree-logic language underpinning the Dynamic Syntax framework.

[^66]
## 2 Dynamic Syntax Analyses

### 2.1 Availability of Different Processing Strategies

The accounts to be given for the different clitic phenomena adopt the Dynamic Syntax framework (DS; Cann et al. 2005; Kempson et al. 2001). DS is a grammar formalism that reflects the dynamics of parsing, with syntax defined as the incremental growth of semantic trees following the time-linear parsing/production process. These semantic trees represent a possible interpretation of the natural language string. Once the processing process is completed, the top node of the tree is decorated with a propositional formula and each daughter node with some sub-term of that formula, representing a predicate-argument structure. Various processing strategies i.e. different ways of building up semantic content for a natural language string, are made available. More specifically, DS licenses the construction of (i) fixed nodes, (ii) unfixed nodes, which represent structural underspecification (which is similar to functional uncertainty in LFG) and which can be constructed locally or non-locally, and (iii) linked structures, i.e. trees that are hooked together and often share semantic content. Moreover, as a set of strategies for parsing, the grammar standardly makes available more than one sequence of strategies for parsing a string with little or no difference in content associated with the distinct output structures. For example, in parsing a pro-drop language with case such as Latin, there are three strategies available for the parsing of a subject expression (Kempson et al. 2008; Kempson \& Chatzikyriakidis 2009). It may be parsed following the strategy available for parsing all argument expressions, which is to construct an unfixed node merely indicating argumenthood (step 1), then decorate it as indicated by the nominal (step 2), and eventually use case to fix the structural relation as that of a subject (step 3). This process is illustrated for the parse of the subject of the Latin example given in (22). As shown in (23), first the locally unfixed node is projected (step 1). The NP is then parsed as annotating that locally unfixed node (step 2). Finally, constructive case fixes the locally unfixed node's address into that of the subject (step 3):

| Catullus | Lesbiam | amavit |
| :--- | :--- | :--- |
| Catullus.NOM | Lesbia.ACC | loved.3SG |
| 'Catullus loved Lesbia.' |  |  |



The second strategy is to take that subject expression as providing a context relative to which the reminder is interpreted, i.e., in DS terms to build a linked structure decorated solely with information provided by the subject expression and use that structure as the point of departure for constructing an independent tree containing a proposition with subject agreement indicating the identification of that term with the already presented context.
(24) Parsing Catullus in Catullus Lesbiam amavit as a linked structure


Finally, there is also the possibility of taking the subject expression to decorate a node initially constructed as unfixed that is not immediately updated, but rather is identified as subject only subsequent to parsing the verb; this decision to fully determine its role in the propositional structure at only this very late stage is a means of achieving a non-backgrounding/contrastive effect.
(25) Just before MERGE of the unfixed node in the parse of Catullus Lesbiam amavit


The availability of these different processing strategies will be crucial to the analysis of MedSp clitic placement as well as the PCC. In what follows, we will show the latter claim starting with MedSp and moving on to the PCC.

### 2.2 Clitic Placement in Spanish

As has been argued elsewhere (Bouzouita 2007, 2008a, 2008b, 2008c; Bouzouita \& Kempson 2006), MedSp and RenSp clitic placement seem to be regulated by the processing strategies used for the left-peripheral constituents that precede the clitics. More specifically, in MedSp, preverbal placement is found after a disjunct set of triggers, to wit, whenever a negation marker, a tense marker, or a constituent that can be represented as structurally (syntactically) underspecified i.e., an expression decorating an unfixed node, precedes the clitic pronoun. ${ }^{6}$ The left-peripheral constituents of most syntactic environments that trigger anteposition can be parsed/produced using this unfixed node strategy. This is, for instance, the case for $w h$-questions, non-coreferential object complements, and prepositional and predicative complements. This is also the case for those variation examples that appear with preverbal clitics, e.g. subjects. Furthermore, the left-peripheral constituents in these variation cases can not only be analyzed as involving an unfixed node but also as decorating a linked structure (or a fixed node). This explains in turn why variation is observed: whenever the unfixed node strategy is used anteposition will appear whereas in the absence of this processing environment postverbal clitics will be used.

As shown in the lexical entry given in (26), the various anteposition triggers are stored as part of the clitic pronoun's lexical specification. The lexical incorporation of these processing environments is said to be due to a routinization process (see Bouzouita 2007, 2008a, 2008b for more details).

Lexical entry of Medieval Spanish accusative clitic lo


As shown, preverbal placement in MedSp has three possible environment triggers. However, as we shall see shortly, these anteposition restrictions have been lost for the RenSp clitic. Postverbal placement, on the other hand, appears in the absence of these anteposition-triggering processing environments (both in MedSp and RenSp). There is thus a complementary cluster of restrictions. Further, it should be noted that both preverbal and postverbal accusative clitics are taken to annotate fixed object nodes. As we shall see later in more detail, not all clitics involve

[^67]the construction of a fixed argument node. In leísta dialects, for instance, the clitic $l e$ will be taken to introduce and annotate a locally unfixed node due to its case ambiguity. This is also the case for the dative $l e$. Further, the nodes decorated by the postverbal clitics have been introduced by the lexical specifications of the preceding verb. Those being annotated by preverbal clitics, on the other hand, have been constructed by the lexical entry of the clitic pronoun itself due to the lexical calcification of the accusative case in Old Romance. The self-evident complexity of the disjunctive form is what then gets progressively simplified, as we can now see with a characterization of the RenSp lo in (27).

Lexical entry of Renaissance Spanish accusative clitic lo

```
IF \(\quad ? T y(t)\)
THEN make \(\left(\left\langle\downarrow_{1}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{1}\right\rangle\right)\);
    make \(\left(\left\langle\downarrow_{0}\right\rangle\right)\); go \(\left(\left\langle\downarrow_{0}\right\rangle\right)\);
    \(\operatorname{put}(F o(\mathbf{U}), T y(e), ? \exists x . F o(x)\),
    \(\left.[\downarrow] \perp, ?\left\langle\uparrow_{0}\right\rangle T y(e \rightarrow t)\right)\)
ELSE IF ?Ty(e), \(\langle\uparrow\rangle \top\)
    THEN IF \(\left\langle\uparrow_{0} \uparrow_{1}^{*}\right\rangle^{+}(? T y(t),[N E G+], \exists x . T n s(x)) \vee\)
    \(\left\langle\uparrow_{0} \uparrow_{1}^{*}\right\rangle^{+}\left(? T y(t),\left\langle\downarrow_{*}\right\rangle(F o(\alpha), ? \exists x \cdot T n(x))\right.\),
    \(\exists x . T n s(x)) \vee\)
    \(\left\langle\uparrow_{0} \uparrow_{1}^{*}\right\rangle^{+}(? T y(t), ? \exists x . \operatorname{Tns}(x))\)
    THEN abort
    \(\operatorname{ELSE} \quad \operatorname{put}(F o(\mathbf{U}), T y(e), ? \exists x . F o(x)\),
    \(\left.[\downarrow] \perp, ?\left\langle\uparrow_{0}\right\rangle T y(e \rightarrow t)\right)\)
ELSE abort
```

Recall that in RenSp anteposition is much more widespread (see Table 1). This can be faithfully reflected in the DS characterization by the loss of the restrictions on the appearance of preverbal clitics, as shown in example (27). Notice that the disjunctive specification constraining postverbal placement is retained. This lexical entry thus directly reflects the fact that all MedSp postverbal environments in the intervening period acquired the possibility of also licensing preverbal pronouns (see Table 1). As this specification shows, this was due to a relatively small change in the lexical entry of the clitic pronoun: the so-called anteposition triggers that were present in MedSp (the presence of a negation marker, an unfixed node or a tense requirement) are dropped from the RenSp characterization, as shown in example (27). The immediate result of the loss of these triggers is the occurrence of preverbal clitcs in substantially more environments: RenSp clitics can appear preverbally as long as there is a ?Ty $(t)$-requirement. Note however that the same does not apply to the appearance of postposition, as these restrictions remain unchanged. Further, the effect of this lexical 'simplification' is not a more simplified distribution, as what emerges is a greater number of environments in RenSp that exhibit syntactic variation.

There remains the question why this simplification in the lexical entry occurred. Recall that DS regularly makes available more than one strategy for interpretation: for certain variation environments, for instance, (i) the strategy of building a pair of linked structures, with the left-peripheral NP decorating that first linked tree as an independent structure, and, in addition, (ii) the strategy of inducing the construction of an unfixed node for that left-peripheral expression to decorate. Additionally, once routinisation took place in MedSp, the original pragmatic motivation underpinning clitic placement gradually disappeared, as it had become a short-circuited. With no pragmatic basis or intonation cues present, there is then nothing to determine which of these two processing strategies to select. Accordingly, a processing mismatch between speaker and hearer is then plausible for these variation environments. The left-peripheral subject in a sentence containing a preverbal clitic, for instance, can be produced relative to a
strategy for building and annotating an unfixed node. The hearer, on the other hand, can parse this subject as annotating a ?Ty(e)-linked structure. Once the preverbal clitic has been heard, the hearer has two processing choices: (i) they can access the lexical entry for MedSp clitics and notice that the left-peripheral subject should have been parsed as an unfixed node due to the occurrence of this preverbal pronoun and consequently choose to parse this subject as an unfixed node instead or (ii) they can ignore this MedSp lexical entry and infer that preverbal pronouns are allowed after linked structures since that is how they just parsed the left-peripheral subject. In the latter option, the hearer will have effectively reanalyzed the lexical entry for the weak pronoun as given in (27). In other words, a production-parsing mismatch in the variation environments could accordingly have led to the inference that there are no conditions on the occurrence of preverbal pronouns. Once the hearer has made such a move, and indeed has done so on a recurrent basis, this reanalysis could be used as the basis for a production decision, thereby confirming a shift of analysis in the system itself. Notice further that this production-parsing mismatch, restricted to taking place in variation environments only, led to the reanalysis of the clitic's lexical entry, hence affecting all the other environments as well.

### 2.3 The Person Case Constraint

We have argued that one of the reasons behind diachronic change in the case of the Spanish clitic system is the availability of different parsing strategies for a given string. These different parsing possibilities are not only possible with subjects, but have been argued to be the mechanism that derives the Latin scrambling effects ( Kempson \& Cann 2007; Kempson et al. 2008; Kempson \& Chatzikyriakidis 2009). What we are going to argue is that the clitic system of MedSp, leaving aside the positioning restrictions and concentrating on the actual actions induced by the clitics, can be seen as a calcification of the parsing strategies of an earlier 'freer' in terms of word order system, namely the Latin scrambling system. The Latin case system made use of four parsing strategies that effectively allowed scrambling. Three of them have been mentioned previously in our discussion of the different parsing strategies being available for the same word, such as for the parsing of a subject. To recap, the first strategy involves parsing of an NP as decorating a locally unfixed node, with subsequent fixing of that node via case information provided by the morphology on the NP (constructive case) (see section 2.1):
(28) Parsing Lesbiam in Catullus Lesbiam amavit (see example (22)) via constructive case


In the above example, Lesbiam is parsed on a locally unfixed node. Then the accusative case information uniquely identifies the unfixed node as the direct object node, since only the latter node will satisfy the requirement $?\left\langle\uparrow_{0}\right\rangle T y(e \rightarrow t)$ within the range of the underspecified modality $\left\langle\uparrow_{0}\right\rangle\left\langle\uparrow_{1}^{*}\right\rangle$.

The second parsing strategy assumes case to function not fully constructively as above, but rather as a filter on output. In that sense, the potential fixing sites of the node get reduced by means of a case filter but the node still remains unfixed, since more than one structural position is available for fixing the node. This parsing strategy is exemplified in (30), showing an intermediate step in the parse of the sentence given in (29):
Stercilinum magnum stude ut habeas
dunghill.ACC big.ACC ensure.2SG-IMP that have.2SG
'Ensure that you have a large dunghill.'
(30) Before MERGE has applied in the parse of Stercilinum magnum stude ut habeas


In the above tree structure, stercilinum is parsed as decorating an unfixed node (not locally unfixed). The case filter the NP stercilinum carries, i.e., ? $\left\langle\uparrow_{0}\right\rangle T y(e \rightarrow t)$, although restricting the potential fixing sites of the node (to all direct object nodes no matter the level of embedding), cannot however fix the node itself. Fixing of the node is done later via MERGE, as soon as a fixed node with no conflicting formula, type, address or any other value or requirement exists.

The third parsing strategy involves a much weaker structural relation encoded as a linked structure, i.e. a separate tree structure peripheral to the main tree. Linked structures are in general used in DS for relative clauses and topic constructions. In the case of these two constructions, a requirement for a shared term between the linked tree and the main tree is posited. However, the strategy we present here makes use of the link relation but unlike the topic or CLLD/HTLD case no requirement for a shared term is posited. Such a strategy is used to encode weak non-argumental structural relations, traditionally called 'ethical datives', as the one found in the example (31). The use of this strategy is illustrated in the parse given in (32):
Quid mihi Celsus agit?
what me.DAT Celsus.NOM.SG do.3SG
'How, pray, is Celsus?' (Lit. 'What to me Celsus does?')


The additional parsing strategy we will introduce shares with the first two strategies described the effect that parsing of the NP is on an unfixed node. However, the difference lies in the fact that no case information is projected on the unfixed node. In that sense, neither fixing of the node nor filtering of the potential fixing sites of the unfixed node is possible. A natural candidate for the use of this strategy will be a highly syncretized noun, e.g. a fourth declension Latin neuter noun, such as cornus 'cornel tree': ${ }^{7}$
(33) Parsing cornus


What we are going to argue is that these four parsing strategies used in Latin scrambling became lexicalized and even though some of them got lost as general language strategies in Spanish (via loss of case-marking), most of them got encoded lexically in the entries for clitic pronouns, in which case marking is still active (albeit highly syncretic). According to this story, each clitic matches one or more of the described strategies. The constructive case strategy has already been exemplified, although implicitly, by the entries given for the MedSp and RenSp 3rd person accusative clitic $l o$ in (26) and (27). According to these lexical entries, the clitic builds and decorates the direct object node, in effect producing a fixed structure. This action is nothing else than the outcome of constructive

[^68]case use, in which case information eventually fixes the node's address. On the other hand, $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics in Spanish being syncretized are assumed to project a locally unfixed node without any such form of update, in effect matching the third parsing strategy presented. Furthermore, 3rd person dative clitics, even though nonsyncretic, are also assumed to project locally unfixed nodes, since their function is still underspecified (indirect and direct objects, possessives, benefactives/malefactives). Putting all these assumptions together, we end up with the following entries for ModSp lo, le and me/te respectively:
(34) Lexical entry for Modern Spanish accusative clitic $l o$

```
IF \(\quad ? T y(t)\)
THEN IF \(\left[\downarrow_{1}^{+}\right] ? T y(x) \quad \vee\)
    [NON - FINITE+]
    THEN make \(\left(\left\langle\downarrow_{1}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{1}\right\rangle\right)\);
    make( \(\left.\left\langle\downarrow_{0}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{0}\right\rangle\right)\);
    \(\operatorname{put}\left(T y(e), F o\left(\mathbf{U}_{M a l e} /\right.\right.\) Neuter \(\left.), ? \exists x . F o(x)\right)\);
    gofirst(?Ty(t))
ELSE abort
```

(35) Lexical entry for Modern Spanish dative clitic le

| IF | $? T y(t)$ |  |
| :--- | :--- | :--- |
| THEN | IF | $\left[\downarrow_{1}^{+}\right] ? T y(x) \quad \vee$ |
|  |  | $[N O N-F I N I T E+]$ |
|  | THEN | make $\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ; \operatorname{go}\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ;$ |
|  |  | make $\left(\left\langle\downarrow_{0}\right\rangle\right) ; \operatorname{go}\left(\left\langle\downarrow_{0}\right\rangle\right) ;$ |
|  |  | $\operatorname{put}\left(\left\langle\uparrow_{0}\right\rangle\left\langle\uparrow_{1}^{*}\right\rangle ? T y(t), T y(e), F o\left(\mathbf{U}_{\text {Male } / \text { Neuter }}\right)\right.$, |
|  |  | $? \exists x . F o(x), ? \exists x . \operatorname{Tn}(x)) ;$ |
|  |  | gofirst $(? T y(t))$ |

Lexical entry for Modern Spanish syncretic clitics melte

```
IF \(\quad\) ?Ty \((t)\)
THEN IF
\(\left[\downarrow_{1}^{+}\right] ? T y(x) \quad \vee\)
    [NON - FINITE+]
    THEN make \(\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ; \operatorname{go}\left(\left\langle\downarrow_{1}^{*}\right\rangle\right)\);
    make \(\left(\left\langle\downarrow_{0}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{0}\right\rangle\right)\);
    \(\operatorname{put}\left(\left\langle\uparrow_{0}\right\rangle\left\langle\uparrow_{1}^{*}\right\rangle ? T y(t)\right), T y(e), F o\left(\mathbf{U}_{\text {Speaker } / \text { Hearer }}\right)\),
    ? \(\exists x . F o(x), ? \exists x \cdot T n(x))\);
    gofirst(?Ty(t))
ELSE abort
```

Given the above lexical entries, the PCC is straightforwardly accounted via a hard-wired processing constraint, namely the fact that no more than one unfixed node with the same underspecified address will ever be possible, since by definition two such nodes will collapse into one by means of tree-node identity. Assuming that 1 st/2nd person and dative clitics project locally unfixed nodes, no combinations of these clitics will ever be made possible,
thus the PCC. Let us go into more detail. Assuming that we want to parse the illicit me te combination, we first parse the 1st person clitic me. This will project a locally unfixed node along with its type value and metavariable specifications. However, as soon as te gets parsed and projects another locally unfixed node, these two nodes collapse into one. What we are left with is one node carrying incompatible formula metavariables, as shown in red in the second step below:

Parsing the clitic cluster me te


Under the present account the PCC is derived by an entirely general processing constraint, a restriction on underspecification. A further welcomed result is that the current account predicts that substituting one of the clitics of the illicit combinations with a strong pronoun will render the sentence grammatical. This is because strong pronouns on a par with full NPs are analyzed in DS as involving a ?Ty (e)-trigger, provided either by the verb itself or via an unfixed but not a locally unfixed node. The reason strong pronouns are parsed as decorating an unfixed but not a locally unfixed node is that strong pronouns can also appear outside their domain of interpretation (left dislocation), in which case the rule of LOCAL *ADJUNCTION would produce the wrong results. Two (or more) underspecified nodes, as long as these are encoded by different underspecified modalities, are predicted to be possible according to the system. This is what happens in the case of a preverbal strong pronoun plus a syncretized clitic. Within this line of reasoning, combinations of a $1 \mathrm{st} / 2 \mathrm{nd}$ person clitic and a strong pronoun are predicted to be grammatical by our account, which is indeed the case. ${ }^{8}$

The analysis provided for Spanish extends naturally to other PCC languages like French or Italian. However, strong PCC languages like Greek seem to pose a problem for such an analysis. This is because in Greek 1st/2nd person singular clitics are non-syncretized. In that respect, $1 \mathrm{st} / 2 \mathrm{nd}$ person accusative clitics will have to be encoded as fixed, some might argue. Before we abandon the account proposed as inadequate for Greek, let us first look at the nature of non-syncretism with $1 \mathrm{st} / 2$ nd person clitics. First of all, non-syncretism is only partial for $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics in Greek, since their plural counterparts are syncretized with respect to case (mas.1PL, sas.2PL), in contrast to 3 rd person clitics that are non-syncretic across the board. Assuming a unitary analysis of $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics in Greek, we will either have to encode plural clitics as projecting fixed nodes or singular clitics as projecting unfixed nodes. The first option is clearly on the wrong track, since it will predict that plural $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics

[^69]can be interpreted as either direct or indirect objects but not both. On the other hand, the second option can be naturally encoded given what we have said, assuming that 1 st/2nd person accusative clitics, even though unfixed, they do project a case requirement that acts as a filter on output but nevertheless does not fix the node itself. Under such an analysis, the PCC is captured within the same mechanisms in both the Greek and the Spanish case. The lexical entry for $1 \mathrm{st} / 2$ nd person accusative clitics in Greek is shown below:

Lexical entry for Greek 1st/2nd accusative clitics

```
IF \(\quad ? T y(t)\)
THEN IF \(\left[\downarrow_{1}^{+}\right] ? T y(x) \quad \vee\)
    Mood(Imp)
    THEN make \(\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{1}^{*}\right\rangle\right)\);
    make \(\left(\left\langle\downarrow_{0}\right\rangle\right) ;\) go \(\left(\left\langle\downarrow_{0}\right\rangle\right)\);
    \(\operatorname{put}\left(\left\langle\uparrow_{0}\right\rangle\left\langle\uparrow_{1}^{*}\right\rangle ? T y(t), T y(e), F o(\mathbf{U})\right.\),
    \(\left.? \exists x . F o(x), ? \exists x . \operatorname{Tn}(x), ?\left\langle\uparrow_{0}\right\rangle(T y(e \rightarrow t))\right)\);
    gofirst(?Ty(t))
ELSE abort
```

The difference then between 1st/2nd person clitics in Spanish and $1 \mathrm{st} / 2 \mathrm{nd}$ person accusative clitics in Greek is that the latter, even though they also project a locally unfixed node, further project a statement $\left(?\left\langle\uparrow_{0}\right\rangle(T y(e \rightarrow t))\right)$ which acts as a filter on output, ensuring that the locally unfixed node will be updated into the direct object node at some point. On the other hand, dative clitics in Greek, even though non-syncretized, do not need such a case filter, since their interpretation is still not determined given that dative clitics can also function as direct objects, possessives, and malefactive/benefactive datives as well as indirect objects (see Kempson \& Chatzikyriakidis (2009) for the relevant examples and argumentation). In that sense, positing an indirect case filter will exclude all the latter interpretations. Thus, no such case filter is needed for datives in Greek:

Lexical entry for Greek dative clitics

| IF | $? T y(t)$ |  |
| :--- | :--- | :--- |
| THEN | IF | $\left[\downarrow_{1}^{+}\right] ? T y(x) \quad \vee$ |
|  |  | $[N O N-F I N I T E+]$ |
|  | THEN | make $\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ; \operatorname{go}\left(\left\langle\downarrow_{1}^{*}\right\rangle\right) ;$ |
|  |  | make $\left(\left\langle\downarrow_{0}\right\rangle\right) ; \operatorname{go}\left(\left\langle\downarrow_{0}\right\rangle\right) ;$ |
|  |  | $\operatorname{put}\left(\left\langle\uparrow_{0}\right\rangle\left\langle\uparrow_{1}^{*}\right\rangle ? T y(t), T y(e), F o(\mathbf{U})\right.$, |
|  |  | $? \exists x . F o(x), ? \exists x . T n(x)) ;$ |
|  |  | gofirst $(? T y(t))$ |

To recap, we have argued that clitics encode parsing strategies for scrambling of an earlier system. Each of the clitics encodes one or more of these strategies. In effect, the clitics in Spanish and Greek can be seen as vestiges of the earlier freer scrambling system (Latin and Ancient Greek respectively). Assuming that $1 \mathrm{st} / 2 \mathrm{nd}$ person and dative clitics project locally unfixed nodes, the PCC is directly explained as a processing constraint in which no more than one unfixed node with the same underspecified modality can be present. The PCC is thus reduced to a hard-wired general constraint and no feature stipulations or added framework machinery is needed.

## 3 Conclusions

We have argued that the clitic systems of MedSp and RenSp as well as the PCC can be accounted for assuming that clitics constitute calcified processing strategies of the Latin scrambling system. In the case of RenSp positioning, it was shown that reanalysis of the different parsing strategies by the parser gives rise to syntactic change. In the same vein, we have argued that the actual actions induced by the clitics match one or more of these strategies of the earlier Latin system (or Ancient Greek one in the case of Greek). Then, we have argued that $1 \mathrm{st} / 2 \mathrm{nd}$ clitics as well as dative clitics project locally unfixed nodes with no immediate form of update or case filter, while 3rd person accusative clitics, on the other hand, project fixed structure. Given that no more than one unfixed node with the same underspecified modality is possible, combinations of $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics with a dative are predicted to be impossible within such an account, thus the PCC.

## References

Academia de la Llingua Asturiana, 2001. Gramática de la llingua asturiana. Oviedo: Academia de la Llingua Asturiana.

Arias Alvarez, B., 1995. Colocación de los pronombres átonos en El libro de los gatos. In: González, A., L. von der Walde and C. Company (eds), Palabra e imagen en la Edad Media. Mexico City: Universidad Nacional Autónoma de México, 113-125.

Adger, D. and Harbour, D., 2007. Syntax and Syncretisms of the Person Case Constraint. Syntax 10, 2-37.
Alvarez Blanco, M.R., Reguiera, X.L. and Monteagudo, H., 1986. Gramática galega. Vigo: Galaxia.
Anagnostopoulou, E., 2003. The Syntax of Ditransitives: Evidence from Clitics. Berlin/New York: Mouton de Gruyter.

Anagnostopoulou, E., 2005. Strong and Weak Person Restrictions: A Feature Checking Analysis. In: Heggie, L. and Ordonez, F. (eds), Clitics and Affixation. Amsterdam/Philadelphia: John Benjamins, 199-235.

Blackburn, P. and Meyer-Viol, W., 1994. Linguistics, Logic and Finite Trees. Bulletin of Interest Group of Pure and Applied Logics 2, 2-39.

Bianchi, V., 2006. On the Syntax of Personal Arguments. Lingua 116, 2023-2067.
Bonet, E., 1991. Morphology after Syntax: Pronominal Clitics in Romance. PhD Dissertation, Cambridge, MA: MIT.

Bonet, E., 1994. The Person Case Constraint: A Morphological Approach. In: Harley, H. and Philips, C. (eds), The Morphology Syntax Connection. Cambridge, MA: MIT, 33-52.

Bonet, E., 2007. The Person Case Constraint and Repair Strategies. In: Alessandro, R., Fischer, S. and Hrafnbjanson, G. H. (eds), Person Restrictions. Berlin/New York: Mouton de Gruyter, 103-128.

Bouzouita, M., 2007. Processing Factors in Syntactic Variation and Change: Clitics in Medieval and Renaissance Spanish. In: Salmons, J. C. and Dubenion-Smith S. (eds.), Historical Linguistics 2005: Selected papers from the 17th International Conference on Historical Linguistics, Madison, 31 July-5 August 2005. Amsterdam/Philadelphia: John Benjamins, 51-71.

Bouzouita, M., 2008a. The Diachronic Development of Spanish Clitic Placement, PhD Dissertation, London: King's College London.

Bouzouita, M., 2008b. At the Syntax-Pragmatics Interface: Clitics in the History of Spanish. In: Cooper, R. and Kempson, R. (eds), Language Evolution and Change. London: College Publications, 223-265.

Bouzouita, M., 2008c. Modelling Syntactic Variation. Diálogo de la lengua 1: 15-39.
Bouzouita, M., forthcoming. Future Constructions in Medieval Spanish: Mesoclisis Uncovered. In: Gregoromichelaki, E. and Kempson, R. (eds), At the Lexicon-Syntax Interface. Stanford: CSLI.

Bouzouita, M. and Kempson, R., 2006. Clitic Placement in Old and Modern Spanish: A Dynamic Account. In: Thomsen Nedergaard, O. (ed.), Competing Models of Linguistic Change. Amsterdam/Philadelphia: John Benjamins, 253-268.

Campos, H., 1989. Clitic Position in Modern Gallegan. Lingua 77: 13-16.
Cann, R., Kempson, R. and Marten, L., 2005. The Dynamics of Language. Oxford: Elsevier.
Castillo Lluch, M., 1996. La posición del pronombre átono en la prosa hispánica medieval. PhD Dissertation, Madrid: Universidad Autónoma de Madrid.

Chatzikyriakidis, S., forthcoming. Clitics in Four Dialects of Modern Greek: A Dynamic Account. PhD Dissertation, London: King's College London.

D'Andres Diaz, R., 1993. Allugiamentu de los pronomes átonos col verbu n'asturianu. Oviedo: Universidad de Oviedo.

Eberenz, R., 2000. El español en el otoño de la Edad Media. Sobre el artículo y los pronombres. Madrid: Gredos.
Granberg, R.A., 1988. Object Pronoun Position in Early and Medieval Spanish. PhD Dissertation, Los Angeles, CA: University of California Los Angeles.

González i Planas, F., 2007. Sintaxis de los clíticos pronominales en asturleonés. Ianua. Revista Philologica Romanica 7: 15-35.

Haspelmath, M., 2004. Explaining the Ditransitive Person - Role Constraint. Electronic form article. Available at: http://email.eva.mpg.de/ haspelmt/2004dpr.pdf.

Heap, D. 2005. Constraining Optimality: Clitic Sequences and Feature Geometry. In: Heggie, L. and Ordóñez, F. (eds.), Clitic and Affix Combinations. Amsterdam/Philadelphia: John Benjamins, 81-102.

Kempson, R., Meyer-Viol, W. and Gabbay, D., 2001. Dynamic Syntax: The Flow of Language Understanding. London: Blackwell Publishing.

Kempson, R. and Cann, R., 2007. Production Pressures, Syntactic Change and the Emergence of Clitic Pronouns. Ms., King's College London.

Kempson, R. Cann, R. and Marten, L., 2008. Treegrowth Dynamics. In C. Cristi (ed.) (title tba).
Kempson, R. and Chatzikyriakidis, S., 2009. The Person Case Constraint as a Treegrowth Property. Ms., King's College London.

Keniston, H., 1937. The Syntax of Castilian Prose: the Sixteenth Century. Chicago: University of Chicago Press.
Nevins, A., 2007. The Representation of Third Person and its Consequences for the Person-Case Constraint, Natural Language and Linguistic Theory 25.2: 273-313.

Nieuwenhuijsen, D., 1999. Cambios en la colocación de los pronombres átonos. Phd Dissertation, Groningen: University of Groningen.

Ormazabal, J. and Romero, R., 2007. The Object Agreement Constraint. Natural Language and Linguistic Theory 25, 315-347.

Rezac, M. and Bejar, S., 2003. Person Licensing and the Derivation of PCC Effects. In: Perez-Leroux A. and Roberge, Y., (eds.), Romance Linguistics: Theory and Acquisition, Amsterdam/Philadelpia: John Benjamins, 49-62.

Rezac, M., 2007. The Syntax of Eccentric Agreement: The Person Case Constraint and Absolutive Displacement in Basque. Natural Language and Linguistic Theory, 61-106.

Rezac, M., 2008. Phi across Modules. Ms, University of Nantes.
Sánchez, X. and Rubiera Tuya, C., 1985. Enclisis, proclisis y pronomes átonos n’asturianu. ALLAS, Lletres Asturianes 18: 71-81.

Sǎvescu, O., 2007. Challenging the Person Case Constraint: Evidence from Romanian. In: Camacho, J., FloresFerran, N., Sanchez, L., Deprez, V., Jose Cabrera, M. (eds), Romance Linguistics 2006. Selected Papers from the 36th Linguistic Symposium on Romance Languages (LSRL), New Brunswick, March-April 2006. John Benjamins, 261-274.

Sǎvescu, O., 2009. A Syntactic Analysis of Pronominal Clitic Clusters in Romance: The View from Romanian. Phd Dissertation, New York: New York University.

Xove Ferreiro, X, 1986. Sobre a posición do pronome persoal átono en galego: eu o sei e eu seino. In: Actes du XXVIIIe Congrès International de Linguistique et Philologie Romanes. Tübingen: Max Niemeyer 2.18: 515-529.

## A Sources and Texts Consulted

DLNE (year, document number) = Documentos lingüísticos de la Nueva España: Altiplano central. C. Company Company, Mexico City: Universidad Nacional Autónoma de México, 1994.

Faz. (page) $=$ La Fazienda de Ultra Mar. Edition of M. Lazar, Salamanca: Universidad de Salamanca, 1963.

# USING TRI-LEXICAL DEPENDENCIES IN LFG PARSE DISAMBIGUATION 

Aoife Cahill, Marion Weller, Christian Rohrer and Ulrich Heid<br>IMS, Universität Stuttgart, Germany

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper we describe experiments where we automatically extract large lists of tri-lexical dependencies. We investigate how effective they are in PP attachment disambiguation by integrating them into a log-linear model for parse disambiguation. We show that we achieve a statistically significant improvement in parse accuracy with the new model that incorporates the trilexical dependencies.


## 1 Introduction

The use of lexical dependencies in parse disambiguation is not a new idea. For example, it is one of the key ideas behind the Collins (1999) parser. In that parser, bi-lexical dependencies are used, but in later work Bikel (2004) showed that the bi-lexical dependencies in fact had little or no impact on the parser decisions. One of the reasons these bi-lexical dependencies are thought to be ineffective is because of sparse data. In treebank-derived parsers, there are simply not enough instances of the dependencies for them to have a significant impact.

Intuitively, on the other hand, it seems that including lexical dependencies should help. In this paper, we consider tri-lexical dependencies between verbs, prepositions and nouns, automatically extracted from a very large corpus, independent from any training data the parser uses. Consider the Example (1) from German with the tri-lexical dependency between the verb stehen ('to stand'), the preposition $z u$ ('to'), and the noun Verfügung ('disposition'), which correspond to the dependency 'to be available' in English:
(1) Dass das Geld zur Verfügung steht, wird angenommen.

That the money to the disposition stands, is assumed.
'The fact that the money is available is assumed.'
For a parser, there is ambiguity here about where to attach the PP zur Verfügung, either to the DP das Geld or to the verb stehen as shown in Figure 1. We observe that for the extracted dependencies with high log-likelihood values the PP almost never attaches to the DP. ${ }^{1}$

In this paper we attempt to incorporate this observation (and test its validity) in an LFG parse disambiguation scenario.

## 2 Automatically Extracting Tri-Lexical Dependencies

We used parsed text to extract multiword expressions with their morphosyntactic features, focusing on preposition-noun-verb triples (PNV) and verb-object combinations. These basic patterns can be expanded to include further components like

[^70]


Figure 1: PP-Attachment ambiguity for German example (1) parsed with the LFG parser
adjectives or additional objects. The extracted PNV-triples were then ranked according to their log-likelihood values.

The data from which we automatically extracted the PNV-triples consists of 230 million tokens of parsed newspaper text. ${ }^{2}$ We use the fspar parser (Schiehlen, 2003) to create dependency structures for each sentence, where ambiguities remain unresolved. This is especially the case with PP-attachment and case marking on nouns, as can be seen in the example in Table 1.

### 2.1 Extraction

The extraction of PNV-triples begins with first identifying the full verb of a sentence and then systematically 'collecting' every relevant item annotated as a dependent of this verb. Extraction steps are illustrated in Table 1.

The verb legen ('to put') in line 8 and the auxiliaries in lines 2 and 9 form a verbal complex (war... gelegt worden) which is referred to by the prepositions nach - 'after' (line 3) and auf - 'on' (line 6). The ambiguous attachment of nach to the noun in line 5 is ignored. Triples are extracted by combining the verb, the preposition and head noun of the object of the preposition. The first PNV-triple we extract is built from liegen, auf and Eis; the second is built from liegen, nach and

[^71]| line | token | gloss | POS | lemma | features | dependencies |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| 0 | Dieses | this | PDAT | dies |  | 1 |
| 1 | Abkommen | agreement | NN | Abkommen | Nom:N:Sg | $8 / 9 / 2$ |
| 2 | war | was | VAFIN | seinP | $1 \mid 3: S g: P a s t: I n d$ | -1 |
| 3 | nach | after | APPR | nach | Dat | $8 / 9 / 2$ |
| 4 | dem | the | ART | d |  | 5 |
| 5 | U-Boot- | submarine | NN | U-Boot- | Dat:M:Sg | 3 |
|  | Zwischenfall | incident |  | Zwischenfall |  |  |
| 6 | auf | on | APPR | auf | Dat $\mid A k k$ | $8 / 9 / 2\| \| 5$ |
| 7 | Eis | ice | NN | Eis | Dat:N:Sg $\mid$ Akk:N:Sg | 6 |
| 8 | gelegt | put | VVPP | legen | PPart | $8 / 9 / 2$ |
| 9 | worden | passive | VAPP | werdenP | PPart | $8 / 9 / 2$ |
| 10 | . | . | $\$$. | . |  | -1 |

Table 1: Slightly simplified example of a sentence parsed with fspar

U-Boot-Zwischenfall. Of the two identified triples, only auf Eis legen (lit. 'to put on ice': to put something on hold) is a valid, idiomatic multiword expression, while nach $U$-Boot-Zwischenfall legen is a random combination of a preposition, noun and verb. For each one of the extracted triples, we will compute a log-likelihood association score to identify associated word combinations. The higher the loglikelihood association, the more likely the triple is to be an idiomatic multiword expression.

A list of 760 manually checked triples that was created as part of the B 3 project of the Stuttgart SFB732 research project is used to handle the ambiguous attachment of a preposition to two or more verbs: if one of the possible combinations is known to be valid, the remaining ones can be discarded; this leads to a lower number of trivial triples.

### 2.2 Log-Likelihood

In order to distinguish (highly) associated PNV-triples from random cooccurrences, triples were ranked according to their log-likelihood-scores using the UCS-toolkit ${ }^{3}$ (Evert, 2004) to compute the scores. Log-likelihood is based on the cooccurrences and individual occurrences of word pairs. For this reason, the extracted triples needed to be reduced to pairs: We tried two different settings by adding the preposition to the noun or to the verb resulting in N-PV and NP-V pairs (cf. Heid et al. (2008)). Table 2 gives a sample of the dependencies extracted, together with their log-likelihood values.

## 3 Parsing System

In our experiments we use the handcrafted German LFG of Rohrer and Forst (2006) coupled with a log-linear disambiguation component (Riezler et al., 2002; Forst, 2007). This is a robust large-scale grammar that has been implemented within the XLE system and achieves complete spanning parses for around $80 \%$ of newspaper text.

For both training and evaluation of the log-linear disambiguation models described in this paper, we use data constructed with the help of the TIGER Treebank (Brants et al., 2002). Our training data consists of 11,504 pairs of labelled and unlabelled packed representations of c - and f -structures which have been produced by our grammar. The labelled representations were constructed by matching the f -structure part of the unlabelled representations produced by the grammar against packed f-structure representations that were derived from the original TIGER Treebank graphs (Forst, 2003). Only sentences for which a proper subset of the readings is compatible with the treebank annotations (and can be determined as such in a reasonable amount of time) were included in the training data, since only these

[^72]| PV | Log-likelihood | NP-V | Log-likelihood |
| :---: | :---: | :---: | :---: |
| Verfügung - zu stehen 'to be available' | 7.1814156e+04 | Verfügung zu - stehen 'to be available' | $3.9224998 \mathrm{e}+04$ |
| Verfügung - zu stellen 'to make available' | 5.1476639e+04 | Verfügung zu - stellen 'to make available' | $3.1667310 \mathrm{e}+04$ |
| Leben - um kommen 'to die' | $4.5470218 \mathrm{e}+04$ | Leben um - kommen 'to die' | $2.9221682 \mathrm{e}+04$ |
| Mittelpunkt - in stehen 'to be central' | $3.7863808 \mathrm{e}+04$ | Mittelpunkt in - stehen 'to be central | $2.7858338 \mathrm{e}+04$ |
| Anspruch - in nehmen 'to make use of' | $2.8404239 \mathrm{e}+04$ | Anspruch in - nehm 'to make use of' | $2.5917412 \mathrm{e}+0$ |
| Vordergrund - in stehen 'to be to the fore' | $2.4039711 \mathrm{e}+04$ | Grenze in - halten (refl) <br> 'to keep within reasonable limits' | $1.8137896 \mathrm{e}+04$ |

Table 2: Example PNv-triples and their log-likelihoods

| Feature Type | Sample feature | Log-likelihood |
| :---: | :--- | ---: |
| A | VERB_PP_ATTACH_DP | -1.1977411824394557094 |
|  | VERB_PP_ATTACH_NODP | -0.365429809372071146 |
| B | NOUN_PP_ATTACH_DP | -3.6529562031401150435 |
|  | NOUN_PP_ATTACH_NODP | 4.4348141926322766082 |
| C | ACC_PP_ATTACH_DP | 0.26424665120489210235 |
| D | VERB_PP_ATTACH_NODP_zu | 7.4110203067279920575 |
|  | NOUN_PP_ATTACH_NODP_ab | 17.000320083423069661 |

Table 3: Example of each feature type with log-likelihoods after training
are useful for discriminative training. For evaluation, we use the TiGer Dependency Bank (TiGer DB) (Forst et al., 2004), a dependency-based gold standard for German parsers.

## 4 Experiments

We carry out a number of experiments to test the effectiveness of the tri-lexical dependencies in parse disambiguation. There are a number of ways to integrate the tri-lexical dependencies into the log-linear model. We design four features to achieve this and experiment with various combinations of the features. Given an ambiguous PP attachment decision, where the PP headed by preposition prep and with object head noun noun can either attach to the VP headed by verb or the $\mathrm{DP},{ }^{4}$ we design the following four feature types:

A PP attached to VP or DP and log-likelihood of N-PV
B PP attached to VP or DP and log-likelihood of NP-V
C If the NP in the PP is in accusative case, does it attach to DP or VP?
D PP headed by prep attached to VP or DP and log-likelihood of N-PV or NP-V
Table 3 gives an example of each feature type along with the log-likelihood it is assigned after training.

We automatically extract these feature types using the 40,000 most-likely PNV dependencies, in addition to the standard parse disambiguation features described in Forst (2007) and train a standard log-linear model. We tune the parameters of the log-linear model on a development set of 362 sentences and carry out the final testing on 1451 sentences. Table 4 gives the results for various combinations of feature types. ${ }^{5}$ The results show that combining all four types of features results in

[^73]| Features | F-Score | Features | F-Score |
| :--- | ---: | :--- | ---: |
| None | 79.54 | BC | 79.54 |
| A | 79.53 | BD | 79.47 |
| B | 79.54 | ABC | 79.54 |
| D | 79.47 | ABD | 79.47 |
| AB | 79.54 | ACD | 79.47 |
| AC | 79.54 | BCD | 79.46 |
| AD | 79.47 | ABCD | $\mathbf{7 9 . 9 9}$ |

Table 4: Initial Results of incorporating tri-lexical dependencies into parse disambiguation
the biggest improvement over the baseline system. Although the difference is small (0.45), the Approximate Randomization significance test (Noreen, 1989) shows that it is statistically significantly better. Some feature combinations in particular cause the results to degrade. In particular, feature type $D$ that attempts to learn attachment preferences for each preposition does badly. However, in combination with all other feature types, it leads to an improvement.

An example sentence where the new model performs better than the baseline is given in (2). The most-probable solution according to the baseline model attached the PP 'on Tuesday' to the NP 'the Federal Constitutional Court', whereas the most probable solution according to the new model (ABCD) correctly attaches the PP to the verb 'to decide'.
(2) Das entschied das Bundesverfassungsgericht (BVG) am Dienstag. That decided the Federal Constitutional Court (BVG) on Tuesday.
'The Federal Constitutional Court decided that on Tuesday'

### 4.1 How Much Data?

An arbitrary decision was made to use the first 40,000 tri-lexical dependencies in the experiments above. However, an interesting question is how many tri-lexical dependencies do you need? The log-likelihood values are an indication of the reliability of the dependencies, and so the more that are used, the more noise that is introduced. The log-linear model does take the log-likelihood into account, however at some point, one would expect the noise to drown out the reliable and useful tri-lexical dependencies. We carry out an ablation experiment to test the effect of the number of tri-lexical dependencies used on the overall f-score. The results are presented in Figure 2.

The graph shows that almost the same f-score can be achieved with only 10,000 tri-lexical dependencies: it is even slightly higher. Table 5 gives the p-values from applying the approximate randomization test to each pair of feature sets. It shows that using 10,000 dependencies is not statistically significantly better than using


Figure 2: Increasing the number of tri-lexical dependencies

|  | 10 k | 25 k | 40 k | 50 k | 75 k | 250 k |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| 10 k |  | 0.019 | 0.131 | 0.724 | 0.001 | $<0.0001$ |
| 25 k |  |  | 0.131 | 0.243 | 0.252 | $<0.0001$ |
| 40 k |  |  |  | 0.975 | 0.019 | $<0.0001$ |
| 50 k |  |  |  |  | 0.057 | $<0.0001$ |
| 75 k |  |  |  |  |  | $<0.0001$ |

Table 5: P-Values for testing significance between pairs of feature lists

40,000 or 50,000 ; however is it statistically significantly better than $25,000,75,000$ and 250,000 . We see very few significant differences in the table. What is noteworthy, however, is that all feature lists perform significantly better than the list with 250,000 features. We conclude from this that by adding in so many dependencies, we have introduced too much noise into the model and this is causing the model to degrade. It seems sensible therefore to choose the model with 10,000 features, since this performs best (albeit only slightly) and is the most effective in terms of performance. The dip in performance between 10,000 and 25,000 is due to 9 sentences with PP-attachment ambiguity that receive an improved f -score but 11 sentences with worse f-scores.

In total, there are only 27 sentences in the larger test set where there is a trilexical dependency involved in an ambiguous PP attachment decision. These sentences alone make up a very small test set, and further evaluation on a larger, more targeted test suite would be required for complete evaluation. In the meantime, however, we look at the performance of each of the models on only these 27 sen-

|  | \# Sentences <br> Model |  | PP Attachment changes  <br> incr. f-score decr. f-score |  |
| :--- | ---: | ---: | ---: | ---: |
| correct | incorrect |  |  |  |
| A | 0 | 0 | 0 | 0 |
| B | 0 | 0 | 0 | 0 |
| D | 2 | 0 | 1 | 0 |
| AB | 0 | 0 | 0 | 0 |
| AC | 0 | 0 | 0 | 0 |
| AD | 2 | 0 | 1 | 0 |
| BC | 0 | 0 | 0 | 0 |
| BD | 2 | 0 | 1 | 0 |
| ABC | 0 | 0 | 0 | 0 |
| ABD | 2 | 0 | 1 | 0 |
| ACD | 2 | 0 | 1 | 0 |
| BCD | 2 | 0 | 1 | 0 |
| ABCD | 11 | 5 | 9 | 0 |

Table 6: The numbers of sentences with tri-lexical dependencies involved in ambiguous attachment decisions
tences, and show how many of the sentences have a higher f-score than the baseline system and how many have a lower f-score. We also show of the sentences with a changed PP attachment, how many of these are now correct and how many are incorrect. Table 6 shows the results.

These results are very interesting. They show a different pattern to the results presented in Table 4, where most feature combinations involving the $D$-type features performed worse. Here we see now that these feature combinations are actually performing better than the baseline on the sentences with ambiguous PP attachment, while other feature combinations perform at the same overall level. On closer inspection, we see that this improvement is due to improved PP attachment in only one of those sentences. It is still clear, that the combination of all feature types is what gives the most improvement, also in terms of PP attachment. Of the 11 sentences with improved f-score, nine of these are due to now correct PP attachments. In the five sentences with lower f-scores than the baseline, three of these are due to the new model now incorrectly attaching the PP.

## 5 Generalising the dependencies

Although we achieved a statistically significant improvement in overall accuracy of the most probable f-structure, we wondered if we could increase performance even more. One of the reasons often cited for the relatively unsuccessful performance of lexical dependencies in parse disambiguation is sparse data. Many of the lexical dependencies extracted are simply too infrequent to be useful in most cases. Our

| number of possible categories | 2 | 3 | 4 | 5 | 6 |
| :--- | :--- | :--- | :--- | :--- | :--- |
| number of words with $n$ possible categories | 2243 | 278 | 42 | 18 | 3 |

Table 7: Ambiguous assignment to GermaNet-classes
attempt to combat this issue was to backoff from the lexical level to a more general level. We used the German version of WordNet, GermaNet to do the backoff.

### 5.1 GermaNet

GermaNet defines 23 categories, which we used to generalise the head nouns in our dependency lists. We also introduced an additional category to account for nouns not included in GermaNet. As the heads of nouns are specified in the parse output, words that could not be found in the GermaNet-lists could be searched for with their head-nouns. Unknown words ending with -ist (as in Linguist) were tagged as Mensch ('human').

Another obvious problem is word sense disambiguation. It would be impossible to choose the correct category in isolation given several alternatives (e.g. ice $\rightarrow$ nutrition, substance); we randomly chose an assignment to one of the possible categories (e.g. ice $\rightarrow$ nutrition). Table 7 shows the amount of ambiguity for the GermaNet categories.

Table 8 contains the GermaNet-classes, the number of entries in each class and its most frequent word. While most of the example entries seem reasonable for their respective classes, some are not very intuitive: The word Mark can mean 'bone-marrow' (and thus qualify for the category body) or the currency Mark which would be the intended meaning in most cases. The example entry in the REST class is an ambiguous lemma that can mean either cabinet or a sort of vine.

### 5.2 GermaNet Experiments

We carry out the same experiments as in Section 4, choosing the feature set that performs best, ABCD . We evaluate the most probable parse against the same test set and achieve an f -score of 79.83 . This is 0.16 f -score points lower than the previous results, although it is not statistically significant. This result is disappointing. We had hoped that by backing off to a more general level, our results would have improved. We also combine the two models and achieve an f-score of 79.78, even lower than the GermaNet model alone.

## 6 Conclusions

In this paper we presented a method for automatically extracting tri-lexical dependencies and ranking them using log-likelihood. In order to evaluate how effective these dependencies were for the PP attachment disambiguation, we carried out a

| frequency | Measure Class |  | most frequent word |  |  |
| :---: | :--- | :--- | :---: | :--- | :--- |
| 1755565 | Artefakt | artefact | 43364 | Tag | day |
| 1730719 | Geschehen | event | 32502 | Fall | case/fall |
| 826886 | Zeit | time | 311257 | Jahr | year |
| 784217 | Gruppe | group | 23170 | Welt | world |
| 642875 | Kommunikation | communication | 29837 | Frage | question |
| 543302 | Ort | location | 23074 | Bereich | area |
| 533880 | Mensch | human | 13053 | Mensch | human |
| 528944 | Besitz | property | 51407 | Land | land |
| 517856 | Attribut | attribute | 18453 | Weise | way |
| 511247 | Kognition | cognition | 13768 | Ansicht | view |
| 466077 | REST | REST | 2012 | Kabinett | Cabinet |
|  |  |  |  | quantity | 154686 |
| Prozent | percent |  |  |  |  |
| 419586 | Menge | body | 102769 | Mark | Mark |
| 251135 | Koerper | feeling | 10722 | Sinn | sense |
| 155851 | Gefuehl | form | 9610 | Kreis | circle |
| 136961 | Form | relation | 9073 | Gegensatz | contrast |
| 98682 | Relation | natural | 8269 | Tod | death |
| 71626 | natPhaenomen |  |  |  |  |
|  |  | phenomenon | 4650 | Wasser | water |
| 49975 | Nahrung | nutrition | 5034 | Luft | air |
| 48201 | Substanz | substance | motive | 20584 | Leben |
| 46696 | Motiv | natural | 3688 | Erde | life |
| 17711 | natGegenstand | narth |  |  |  |
|  |  | object | animal | 1910 | Tier |

Table 8: List of GermaNet-categories and the number of entries as well as the most frequent word in each category.
number of parse-disambiguation experiments. We integrated these dependencies into the log-linear disambiguation model by means of four different feature types and achieved a statistically significant improvement over a baseline when all four feature types were combined. We experimented with backing off the individual word dependencies to try and tackle the sparse data problem. We used classification from GermaNet as a backoff: however we found that this did not improve results. We also combined both models, which also did not lead to an improvement. We found that in our experiments, the 10,000 most likely dependencies contributed most to the improved f-score, and that 250,000 introduced too much noise.

Our initial results were encouraging: we achieved a small improvement in f-
score without using any backoff. However, the results with the GermaNet backoff were disappointing. They are consistent, however, with other experiments using WordNet/GermaNet resources as a backoff for lexical dependencies (Bikel, 2004).

## References

Bikel, Daniel M. 2004. A Distributional Analysis of a Lexicalized Statistical Parsing Model. In D. Lin and D. Wu (eds.), Proceedings of EMNLP 2004, pages 182-189, Barcelona, Spain: Association for Computational Linguistics.

Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang and Smith, George. 2002. The TIGER Treebank. In E. Hinrichs and K. Simov (eds.), Proceedings of the first Workshop on Treebanks and Linguistic Theories (TLT'02), pages 24-41, Sozopol, Bulgaria.

Collins, Michael. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph. D. thesis, MIT.

Evert, Stefan. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. Ph. D. thesis, University of Stuttgart.

Forst, Martin. 2003. Treebank Conversion - Creating a German f-structure bank from the TIGER Corpus. In Proceedings of the LFG03 Conference, pages 205216, Saratoga Springs, NY: CSLI Publications.

Forst, Martin. 2007. Disambiguation for a Linguistically Precise German Parser. Ph. D. thesis, University of Stuttgart.

Forst, Martin, Bertomeu, Núria, Crysmann, Berthold, Fouvry, Frederik, HansenSchirra, Silvia and Kordoni, Valia. 2004. Towards a dependency-based gold standard for German parsers - The TiGer Dependency Bank. In Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC '04), Geneva, Switzerland.

Heid, Ulrich, Fritzinger, Fabienne, Hauptmann, Susanne, Weidenkaff, Julia and Weller, Marion. 2008. Providing Corpus Data for a Dictionary for German Juridical Phraseology. In A. Storrer, A. Geyken, A. Siebert and K. Wrzner (eds.), Proceedings of the 9th Conference on Natural Language Processing, KONVENS 2008, Berlin, Germany: Mouton de Gruyter.

Noreen, Eric W. 1989. Computer Intensive Methods for Testing Hypotheses: An Introduction. New York: Wiley.

Riezler, Stefan, King, Tracy H., Kaplan, Ronald M., Crouch, Richard, Maxwell, John T. III and Johnson, Mark. 2002. Parsing the Wall Street Journal using a

Lexical-Functional Grammar and Discriminative Estimation Techniques. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 271-278, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Rohrer, Christian and Forst, Martin. 2006. Improving Coverage and Parsing Quality of a Large-scale LFG for German. In Proceedings of the Language Resources and Evaluation Conference (LREC-2006), Genoa, Italy.

Schiehlen, Michael. 2003. A Cascaded Finite-State Parser for German. In Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest.

# OBJ $_{\text {® }}$ WITHOUT OBJ: A TYPOLOGY OF MESKWAKI OBJECTS 

Amy Dahlstrom<br>University of Chicago<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu


#### Abstract

Meskwaki exhibits a typologically unusual valence pattern in which certain two-place verbs subcategorize for a subject and an OBJ $_{\ominus}$ but no OBJ. Verbs with the valence pattern of interest here are tested to show that their non-subject argument is $\mathrm{OBJ}_{\theta}$, not unrestricted OBJ, nor OBL. A brief survey of recent work on similar phenomena is presented in order to place Meskwaki in typological perspective.


## 1 Introduction

The Algonquian language Meskwaki (also known as Fox) exhibits a typologically unusual valence pattern in which certain two-place verbs subcategorize for a subject and an $\mathrm{OBJ}_{\ominus}$ but no OBJ. ${ }^{1}$ The structure of the paper is as follows: I first give some background information on Meskwaki, necessary to understand the arguments which follow. I then examine ditransitive verbs in order to establish diagnostics for OBJ and $\mathrm{OBJ}_{\odot}$. Verbs with the valence pattern of interest here are tested to show that their non-subject argument is OBJ ${ }_{\theta}$, not unrestricted OBJ, nor OBL. In the final sections of the paper I consider the range of thematic roles associated with $\mathrm{OBJ}_{\theta}$, ask whether one can predict which verbs will display this pattern, and compare the Meskwaki phenomenon with other languages in which $\mathrm{OBJ}_{\ominus}$ can appear with no OBJ.

### 1.1 Background on Meskwaki: verb inflection

Meskwaki and the other Algonquian languages are almost entirely headmarking in the sense of Nichols (1986): nouns are case-marked only for a locative case; verbs are inflected for subject and object; verbs in relative clauses bear an additional inflection for the head of the relative clause. First and second person inflection always functions as incorporated pronouns; third person inflection may be either pronominal or agreement with a lexical subject or object. There are 26 inflectional paradigms for verbs, sensitive to syntactic, semantic, and pragmatic factors.

The agreement categories are person, number, gender (+/- animate), and OBVIATION. Obviation is a discourse-based opposition within third person: unmarked PROXIMATE forms refer to the third person most central to the discourse; marked OBVIATIVE forms are used for more peripheral third persons. Animate gender includes not only humans and animals but also some notionally inanimate items (e.g. drum, pipe, sun, fingernail, kidney, raspberry...). Inanimate is the unmarked member of the gender opposition, containing most

[^74]body parts, most plants, and most natural and manufactured items (e.g. medicine, fire, blood, heart, strawberry...).

### 1.2 Verb stem classes

Verb stems are specialized for the gender of their OBJ if transitive or SUBJ otherwise:

| $\begin{align*} & \text { amw- 'eat }<\text { S O }>\text { ' }  \tag{1}\\ & (\uparrow \text { OBJ GEND })={ }_{c} \text { ANIM } \end{align*}$ | $\begin{aligned} & m i \cdot c \check{c} \text { - 'eat <S O>' } \\ & (\uparrow \text { OBJ GEND })={ }_{\mathrm{c}} \text { INAN } \end{aligned}$ |
| :---: | :---: |
| meškosi-' 'be red $<\mathrm{S}>$ ' <br> ( $\uparrow$ SUBJ GEND) $=$ ${ }_{c}$ ANIM | meškwa'- 'be red <S> <br> $\left(\uparrow\right.$ SUBJ GEND) $=_{\mathrm{c}}$ INAN |

The Algonquianist labels for these stem classes are:
(2) Transitive Animate (TA) Transitive Inanimate (TI)

Animate Intransitive (AI) Inanimate Intransitive (II)

### 1.3 Stem-internal components

Since the discussion below touches upon questions of stem-internal structure it should be noted that most Algonquian verb stems are bipartite, consisting of an INITIAL and a FINAL. In the paired verb stems in (1), the suppletive pairing of 'eat' is exceptional; the norm is to have an initial like meškw- 'red' combine with a final. It is the final morpheme which bears valence information and constrains the gender of OBJ or SUBJ, in addition to the semantic information it contributes, as can be seen below with -esi- stative (AI) vs. $-a \cdot$ - stative (II).

```
meškw- -esi- meškw- -a`-
red- STATIVE <S> red- STATIVE <S>
        (\uparrowSUBJ GEND) =
                                ( }\uparrow\mathrm{ SUBJ GEND) = = INAN
```

(4) lists a few pairs of transitive stems with the initial pan- 'miss' combined with various instrumental finals:


The finals exemplified in (4) are -en/-en 'by hand', -am/-at 'by mouth', and -eškaw/-ešk 'by foot'.

Another stem-internal component is MEDIAL, consisting of incorporated nouns and classifiers. An incorporated body-part noun is controlled by the OBJ, if present, otherwise by SUBJ:
(5) mešketone $n$ - 'open OBJ's mouth by hand'
mešk-etone--en
open-mouth-by.hand
(6) mešketone $\cdot k w a \cdot m$ -
mešk-etone•-ekwa•m
open-mouth-sleep
'open OBJ's mouth by hand'
$(\uparrow$ OBJ GEND $)={ }_{\mathrm{c}}$ ANIM
'sleep with one's mouth open'
$(\uparrow$ SUBJ GEND $)={ }_{c}$ ANIM

### 1.4 Inventory of GFs

Meskwaki permits athematic SUBJ and OBJ, as expected with the semantically unrestricted GFs. Athematic arguments are inanimate gender and are never expressed by an independent pronoun.

| (7) | kemiya- | $\begin{aligned} & \text { 'rain }<>\text { S' } \\ & \text { ( } \uparrow \text { SUBJ GEND })={ }_{\mathrm{c}} \text { INAN } \end{aligned}$ |
| :---: | :---: | :---: |
| (8) | a•hkwamat- | $\begin{aligned} & \text { 'be sick }<\mathrm{S}>\mathrm{O} \text { ' } \\ & (\uparrow \text { OBJ GEND })={ }_{\mathrm{c}} \text { INAN } \end{aligned}$ |

Besides SUBJ, OBJ, and of course $\mathrm{OBJ}_{\theta}$, the focus of the present paper, Meskwaki also exhibits OBLs of numerous types. OBLS in Meskwaki are often associated with specific morphemes appearing in stem-initial position or as a preverb (a phonologically separate word compounded with the verb stem). For example, the morpheme for $\mathrm{OBL}_{\text {source }}$ is ot-, realized as an initial in (9) and as a preverb in (10).
9) očiwen-
'take O from $<\mathrm{SOOOBL}_{\text {source }}>$ '
$\left(\uparrow\right.$ OBJ GEND) $=_{\mathrm{c}}$ ANIM
(10) oči nowi-
'go out from $<$ S OBL $_{\text {source }}>$ '
$(\uparrow$ SUBJ GEND $)={ }_{\mathrm{c}}$ ANIM
The sense of "source" here is the starting point of a path of motion, or the cause of an event. Human sources, as in 'steal from', are expressed as OBJ, as will be seen below in (17a).

The inventory of grammatical functions in Meskwaki includes COMP:
(11)
anwači•-
'be willing to $<$ S COMP $>$ '
( $\uparrow$ SUBJ GEND) $=_{\mathrm{c}}$ ANIM

There are no nonfinite forms of verbs in Meskwaki, so propositional arguments of verbs like 'be willing to' are always COMP and not XCOMP. Meskwaki does, however, have XCOMPs incorporated into a verb stem, in initial position:

$$
\begin{array}{ll}
\text {-e•nem- } & \text { 'consider }<\text { S O XCOMP }>'  \tag{12}\\
\text { e.g, nepwa•hka•we•nem- 'consider O smart' }
\end{array}
$$

See Dahlstrom (2000) for discussion of incorporated XCOMPs.

### 1.5 Word order

The order of elements within the clause is sensitive to the following template:


Obliques appear to the left of the verb; the unmarked position for all other arguments is to the right of the verb, unless the NP is put in topic or focus position.

## 2 Ditransitive verbs

### 2.1 Basic ditransitives

With the above background on Meskwaki we can now examine ditransitive verbs, both basic stems and those derived by valence-increasing processes, in order to establish diagnostics distinguishing OBJ from $\mathrm{OBJ}_{\ominus}$ in Meskwaki. As in many languages, the verb 'give' is a prototypical ditransitive verb. The first object (OBJ) of 'give' is the recipient and the second object ( $\mathrm{OBJ}_{\ominus}$ ) is the theme argument, the item which is given. If the objects are expressed by NPs their unmarked position is to the right of the verb. If both are NPs OBJ nearly always precedes $\mathrm{OBJ}_{\theta}$, as seen in the following textual example: ${ }^{2}$

[^75]| nemi $\cdot$ na $\cdot$ waki nešise $\cdot$ haki me:šomakini |  |  |
| :--- | :--- | :--- |
| ne-mi $\cdot \mathrm{n}-\mathrm{a} \cdot$ waki | ne-šise $\cdot$ h-aki | IC-mešw-emakini |
| 1 -give- $1>3 \mathrm{P} / \mathrm{IND}$ | 1 -uncle-PL | IC-shoot-1>3'/PART/3'head |
|  | $\mathbf{O}$ | $\mathbf{O}_{\Theta}$ |

'I gave my uncles the [game] which I shot.'
Note that Meskwaki, unlike English, does not have an alternation of two different structures for dative expressions - there is no way of expressing the recipient as an oblique, something like "I gave the game which I shot to my uncles." The double object construction is the only possibility.
(15) lists a few ditransitive verb stems with their subcategorizational requirements.
(15) Basic ditransitives

| a. | mi'n- | 'give $<$ S O O O $_{\ominus}>$ |
| :--- | :--- | :--- |
| b. | ašam- | 'feed $<$ S O O O $_{\ominus}>$ |
| c. | manih- | 'rob O of O O |$<$ S O O O $_{\ominus}>$ ',

Besides mi•n- 'give' other basic ditransitives include ašam- 'feed', where the recipient is first object and the theme, the food, is $\mathrm{OBJ}_{\ominus}$, manih- 'rob', where the robbery victim, here a source argument, is OBJ and the thing stolen, the theme, is $\mathrm{OBJ}_{\theta}$, and $a \cdot$ sim- 'urge', where the addressee is OBJ and the thing or person urged is $\mathrm{OBJ}^{\mathrm{e}}$. The OBJ of such verbs is always animate and nearly always human. (The constraint equations have been omitted for readability.) The $\mathrm{OBJ}_{\boldsymbol{\Theta}}$ may be grammatically animate or inanimate, and typically bears the thematic role of theme.

### 2.2 Applicatives

Ditransitive verbs may also be the result of derviational processes. Applicative formation, for example, adds a new OBJ to a verb's argument structure; the old OBJ of the input form gets demoted to $\mathrm{OBJ}_{\ominus}$. Applicatives may add a beneficiary, as in the textual example in (16), where the grandmother is OBJ.
(16) nehtamawi ko•hkomese•hena•na ma•hani ki•hče•wani nehtamaw-i ko•hkomese•hena•na [ma•hani ki•hče•w-ani] kill. ${ }_{\theta}$.for-2<3/IMP our.g.mother this.ANIM.OBV turkey-OBV $\mathbf{O} \quad \mathbf{O}_{\boldsymbol{\theta}}$
'Kill this turkey for our grandmother.'
A few more applicative forms are listed in (17). In 17a the OBJ has the thematic role of 'source' (who you accept the $\mathrm{OBJ}_{\ominus}$ from), while the forms in b and c have beneficiary OBJs.
(17) a. nahkonamaw- 'accept $\mathrm{O}_{\Theta}$ from $\mathrm{O}<\mathrm{S} \mathrm{O} \mathrm{O}_{\Theta}>$ '
b. mi•winehkamaw- 'chase $\mathrm{O}_{\Theta}$ away for $\mathrm{O}<\mathrm{S} \mathrm{O} \mathrm{O}_{\Theta}>$ '
c. aka•wa•tamaw- 'desire $\mathrm{O}_{\ominus}$ for $\mathrm{O}<\mathrm{S} \mathrm{O} \mathrm{O}_{\ominus}>{ }^{\prime}$

Compare the monotransitive stems nahkon- 'accept', mi•winehk- 'chase', and aka•wa•t- 'desire'.

### 2.3 Causative

Ditransitive stems may also result from adding a causative suffix to a monotransitive verb stem, as seen in (18). Causative adds a new argument, the causer, as a SUBJ, demoting the old SUBJ to OBJ and the old OBJ to OBJ ${ }_{\ominus}$.

| a. | kehke $\cdot$ netamwih- | 'make O know $\mathrm{O}_{\Theta}<\mathrm{SOO} \mathrm{O}_{\Theta}>$ ' |
| :--- | :--- | :--- |
| b. | awata $\mathrm{h}-$ | 'make O take $\mathrm{O}_{\Theta}<\mathrm{SO} \mathrm{O}_{\Theta}>'$ |
| c. | awih- | 'lend $<\mathrm{SOOO}_{\Theta}>$ ' |

Compare the monotransitive stems kehke•net- 'know', awat- 'take', and awi'have'.

### 2.4 Possessor Raising

A final type of derived ditransitive is possessor raising. If the OBJ of a monotransitive verb is a possessed noun, speakers will often express the possessor as the OBJ of the verb. As a consequence, the possessed item gets demoted to $\mathrm{OBJ}_{\ominus}$. The morphology of the verb stem reflects that it is a three place verb, as can be seen by comparing (19a) and (b).
(19) a. ne.t- 'see $<\mathrm{S} \mathrm{O}>$ '
b. ne•tamaw- 'see $\mathrm{O}^{\prime} \mathrm{s} \mathrm{O}_{\Theta}<\mathrm{SOO}_{\Theta}>$ '

As with the basic ditransitives, the OBJ of applicative, causative, and possessor raising derived ditransitives is always animate in gender.

## 3 OBJ-suppressing processes applied to ditransitives

We now turn to a consideration of the Meskwaki verbs which I claim have a subject and an $\mathrm{OBJ}_{\ominus}$ but no OBJ.

One way in which the subcategorizational pattern of interest can arise is if a ditransitive verb undergoes a valence-reducing process which suppresses the OBJ. An example of a process that suppresses the OBJ is antipassive, as in (20a). Here the ditransitive verb 'give' has had the recipient argument suppressed. The verb takes a subject and a theme argument, that which is given, but the recipient is left unspecified. My claim is that the theme argument remains an $\mathrm{OBJ}_{\Theta}$ and does not advance to OBJ.
a. mi•šiwe•-
'give $\mathrm{O}_{\ominus}$ away $<\mathrm{SO}_{\ominus}>$ ' [antipassive]
b. ašameti•- 'feed each other $\mathrm{O}_{\Theta}<\mathrm{S} \mathrm{O}_{\Theta}>$ ' [reciprocal]
c. aka•wa•tama•tiso- 'desire $\mathrm{O}_{\ominus}$ for oneself $\quad<\mathrm{S} \mathrm{O}_{\Theta}>$ ' [reflexive]

Other ways in which an OBJ can be suppressed include reciprocal formation, as in (20b), where the ditransitive verb 'feed' becomes 'feed each other', and reflexive formation as in (20c), where a reflexive suffix has been added to the applicative stem seen above in (17c). In 20b the recipient argument is suppressed but the theme argument remains; likewise, the beneficiary argument of (20c) is suppressed but the theme argument is unaffected.

## 4 Verbs inherently subcategorized for SUBJ and OBJ $\boldsymbol{O}_{\boldsymbol{e}}$

The subcategorization frame of subject and $\mathrm{OBJ}_{\ominus}$ is also found with stems which are inherently specified for that valence; that is, they are not derived from a more basic ditransitive stem. Some examples are listed in (21). ${ }^{3}$

| a. | we•pa•hke-- |
| :--- | :--- |
| b. | meno- |
| c. | ata•-we•- |
| d. | wani•hke-- |
| e. | wača•ho- |
| f. | ahčike•- |
| g. | kemot- |
| h. | kehekwi- |

```
'throw < S O
'drink <S O O}>
'sell, trade < S O
'forget < S O
'cook < S O O
'plant <S O
    'steal < \S O
    ''O}\mp@subsup{O}{\Theta}{}\mathrm{ gives S the slip <S O}\mp@subsup{\textrm{O}}{\Theta}{}>\mathrm{ ''
```

One can see that this valence pattern is found with some of the most basic verbs in the language, such as 'throw' and 'drink'. The verb in (21h), however, is unusual: kehekwi- is used for a hunter losing his prey, or a warrior having a captive escape.

The forms in (21) do not display any recurring morphological elements, but consider the forms in (22), with initials of ahp- 'on', takw- 'together with', or kek- 'having'.

| a. | ahpe nemo- | 'depend on $<\mathrm{S} \mathrm{O}_{\ominus}>$ ' |
| :---: | :---: | :---: |
| b. | ahpapi- | 'sit on $<\mathrm{S} \mathrm{O}_{\ominus}>$ ' |
| c. | ahpeka-- | 'dance on $<\mathrm{S} \mathrm{O}_{\Theta}>$ ' |
| d. | takwi-- | 'join $<\mathrm{SO}_{\ominus}>$ ' |
| e. | takwisen- | 'lie together with $<\mathrm{SO}_{\Theta}>$ ' [INAN. SUBJ] |
| f. | kekišin- | 'lie having, be buried with $<\mathrm{SO}_{\ominus}>$ ' |
| g. | kekate•mo- | 'weep holding $\mathrm{O}_{\Theta}<\mathrm{S} \mathrm{O}_{\Theta}>$ ' |

Perhaps the most commonly used verbs subcategorized for a subject and $\mathrm{OBJ}_{\ominus}$ are those derived from kinship terms and other possessed nouns. A few such verbs are listed in (23).
(23) a. oki- 'have $\mathrm{O}_{\ominus}$ as a mother $<\mathrm{S} \mathrm{O}_{\Theta}>$ '
b. owi•wi- 'have $\mathrm{O}_{\ominus}$ as wife, marry $\mathrm{O}_{\ominus}<\mathrm{S} \mathrm{O}_{\ominus}>$ '

[^76]c. owi•hka•ni- 'have $\mathrm{O}_{\Theta}$ as a friend $<\mathrm{SO}_{\Theta}>$ '
d. owi ${ }^{\text {či }}$-škwe•hi- 'have $\mathrm{O}_{\Theta}$ as an enemy $<\mathrm{S} \mathrm{O}_{\Theta}>$ '

The forms in (23) would be more idiomatically glossed in English as ' $\mathrm{O}_{\ominus}$ is S 's mother', etc.

## 5 Behavior of $\mathrm{OBJ}_{\Theta}$ vs. OBJ

In order to argue that the verbs in (21-23) take subject and $\mathrm{OBJ}_{\theta}$, rather than subject and OBJ, we must discover what the properties of the two types of object are for Meskwaki.

### 5.1 Valence-decreasing processes

Meskwaki ditransitives are of the asymmetric type (Bresnan and Moshi 1990), with the syntactic behavior of $\mathrm{OBJ}_{\ominus}$ differing from that of OBJ in several respects. As we have already seen in (20), an OBJ may undergo lexical processes which suppress the object, e.g. antipassive, reciprocal, and the verbal reflexive; $\mathrm{OBJ}_{\ominus}$ cannot be the target of these processes. Such processes apply to the sole object of a monotransitive verb and to the first object of a ditransitive verb. They cannot apply to the second object of a ditransitive, nor can they apply to the non-subject argument of the verbs in (21-23), the two place verbs which I claim take a subject and an $\mathrm{OBJ}_{\ominus}$.

## $5.2[+/-r]$

It was shown above (7-8) that Meskwaki permits athematic SUBJ and OBJ, as expected for the two GFs associated with the [-r] feature: semantically unrestricted. In contrast, there are no athematic secondary objects of ditransitives, nor any athematic arguments of the verb class under examination here.

### 5.3 Gender

Third, as seen above in (1), in Algonquian languages verb stems come in pairs, specialized for the gender of one of the verb's arguments. Transitive verb stems are sensitive to the gender of OBJ. (24) gives some further examples of monotransitive stem pairs, with the inanimate object form on the left and the animate object form on the right.


An $\mathrm{OBJ}_{\mathrm{e}}$, on the other hand, may be either animate or inanimate without affecting the form of the verb. This can be seen by looking at the ditransitive form of 'bring, bring for' in (25a). Here the OBJ (the recipient or beneficiary argument)
must be animate. But the $\mathrm{OBJ}_{\Theta}$, the thing brought, may be grammatically animate or inanimate, with no change in the shape of the verb stem. Compare the monotransitive forms of 'bring' in (24c), where bringing an inanimate object such as 'rattle' requires a different form of the verb stem from bringing an animate object such as 'drum'.
a. pye•tahw - 'bring $\mathrm{O} \mathrm{O}_{\Theta}{ }^{\prime}$
[OBJ must be animate]
b. ne-pye•tahw-a•wa te•we•hikan-ani 'I brought him a drum'
1-bring.for-1>3/IND drum-ANIM.OBV.SG
c. ne-pye•tahw-a•wa ši•ši $\cdot$ kwan-i 'I brought him a rattle'
1-bring.for-1>3/IND rattle-INAN.SG

Now consider one of the two place verbs of interest here, ahpe-nemo'depend on, rely on'. One can depend upon a human being, as in (26b), or upon an inanimate object such as medicine, as in (26c). In either situation, the form of the verb stem is the same. The absence of paired stem morphology is another way in which the non-subject argument of verbs like ahpe-nemo- 'depend on' patterns with the second objects $\left(\mathrm{OBJ}_{\Theta}\right)$ of ditransitives.

```
a. ahpe·nemo- 'depend on, rely on O}\mp@subsup{O}{\Theta}{}\mp@subsup{}{}{\prime
b. ahpe-nemo-wa o-si}\cdotme\cdoth-ani
    depend.on-3/IND his-younger.sibling-ANIM.OBV.SG
    'He relies on his younger brother.'
c. ahpe\cdotnemo-wa na\cdottawino\cdotn-i
    depend.on-3/IND medicine-INAN.SG
    'He relies on the medicine.'
```


### 5.4 Pronominal OBJ and $\mathrm{OBJ}_{\boldsymbol{e}}$

A further difference between the two types of Meskwaki objects is that ditransitive verbs are inflected for OBJ but not for $\mathrm{OBJ}_{\theta}$. Two place verbs like ahpe-nemo- 'depend on' likewise do not bear inflection for their non-subject argument. The verbal inflection for OBJ may function pronominally in the absence of a full NP argument, as can be seen in the ditransitives of ( 25 b and c) above, where the recipient of pye•tahw - 'bring' is understood to be a singular third person.

The question then arises, how is a pronominal $\mathrm{OBJ}_{\Theta}$ expressed, since there is no verbal inflection for $\mathrm{OBJ}_{\Theta}$ ? A third person pronominal second object is nearly always expressed by zero anaphora:
ne-pye $\cdot$ tahw-a $\cdot$ wa 'I brought it (animate or inanimate) for him.'
1-bring.for-1>3/IND

A first or second person pronominal $\mathrm{OBJ}_{\Theta}$ is expressed by an independent personal pronoun - a grammaticalized possessed form of the inalienably possessed noun stem -i'yaw- 'body':

```
netahpe:nemo ki\cdotyawi
    ne-t-ahpe-nemo-Ø ki`yawi
    1-EP-depend.on-1/IND you [literally, 'your body']
    'I depend on you.'
```

An interesting fact about the usage of the 'body' pronouns is that third person pronominal $\mathrm{OBJ}_{\Theta} \mathrm{S}$ are expressed by a 'body' pronoun when $\mathrm{OBJ}_{\ominus}$ is proximate and the subject or OBJ is obviative. (29) and (30) are textual examples showing this usage:

| e-h-ahpe-nemo-niči | mehtose•neniw-ahi | owi•yawi |
| :--- | :--- | :--- |
| AOR-depend.on-3'/AOR | person-OBV.PL | him |

'The people (obviative) depended on him (proximate).'

| nekoti | aša $\cdot h$-ani | e-h-pye-tahw-emeči | owi ${ }^{\text {yawi }}$ |
| :---: | :---: | :---: | :---: |
| one | Sioux-OBV | AOR-bring. $\mathrm{O}_{\odot}$.to- $\mathrm{X}>3 \mathrm{l} /$ AOR | her |

'They (unspecified) brought her (proximate) to a certain Sioux (obviative).'

In other words, the appearance of an independent third person pronoun for an $\mathrm{OBJ}_{\Theta}$ is analogous to the inverse forms of inflectional morphology on monotransitive verbs: a marked formal option signaling the pragmatically marked situation of the proximate third person outranked syntactically by an obviative third person.

What is important for our purposes here, however, is that the third person 'body' pronouns appear both with the $\mathrm{OBJ}_{\Theta}$ of a ditransitive like pye•tahw'bring $\mathrm{O}_{\Theta} \mathrm{O}_{\Theta}$ ', as in (30), and with the non-subject argument of verbs like ahpe-nemo- 'depend on', in (29). Again, this is evidence that the non-subject argument in (29) bears the same grammatical function as the $\mathrm{OBJ}_{\Theta}$ of a ditransitive verb.

### 5.5 Reflexive $\mathrm{OBJ}_{\boldsymbol{\ominus}}$

Although $\mathrm{OBJ}_{\Theta}$ cannot undergo the verbal reflexive strategy seen above in (20c), in which a reflexive suffix attaches to the verb stem and decreases the valence of the verb, it is in fact possible to express a reflexive $\mathrm{OBJ}_{\theta}$. This is accomplished by using the 'body' series of independent pronouns, exemplified in the previous section. With these independent reflexive pronouns, we can see another asymmetry between OBJ and $\mathrm{OBJ}_{\ominus}$ : an OBJ can be the antecedent of an $\mathrm{OBJ}_{\ominus}$ reflexive, as in (31), but not vice versa.

| ne-wa-pato $\cdot$ n-a $\cdot$ wa | apeno $\cdot$ h-a |  |
| :--- | :--- | :--- |
| child-SG | owi | owawi <br> her(self) |
| 'show- $1>3 /$ IND |  |  |

### 5.6 Noun incorporation

A further difference between OBJ and $\mathrm{OBJ}_{\ominus}$ concerns noun incorporation. Recall that an incorporated body part noun is construed with the object of a transitive verb, as in (5), repeated below:
(5) mešketone•n- 'open OBJ's mouth by hand'
mešk-etone--en
open-mouth-by.hand
Verbs subcategorized for SUBJ and $\mathrm{OBJ}_{\Theta}$, on the other hand, always have the SUBJ as controller of the incorporated noun, not $\mathrm{OBJ}_{\ominus}$ :

```
ahpanasite}\cdot\textrm{ka}\cdot\textrm{pa}\cdot-\quad 'stand with one's feet on OBJ , ,'
ahp-anasite-ika`par- [not "stand on OBJ ''s feet"]
```

on-foot-stand

To sum up the results of this section: using the criteria for distinguishing OBJ from $\mathrm{OBJ}_{\ominus}$, we must analyze some two-place verbs as being subcategorized for a subject and an $\mathrm{OBJ}_{\Theta}$, not an OBJ. That is, the nonsubject argument of such verbs cannot be the target of antipassive, reflexive or reciprocal verb formation, it is never an athematic object, it may be either animate or inanimate without changing the form of the verb stem, it does not trigger agreement on the verb, it may be expressed by pronouns from the 'body' series or by zero anaphora, and it cannot be construed with an incorporated noun, all characteristic of $\mathrm{OBJ}_{\ominus}$ as opposed to OBJ .

## 6 Distinguishing $\mathrm{OBJ}_{\boldsymbol{\Theta}}$ from OBL

Before concluding that the non-subject argument of a verb like ahpe-nemo'depend on' is an $\mathrm{OBJ}_{\theta}$, it is necessary to also investigate the possibility that the relevant grammatical function borne by the non-subject argument is instead OBL. There is, after all, nothing unusual about a given two-place verb being subcategorized for a subject and an oblique (e.g. English depend (on)). In Meskwaki, however, obliques exhibit well-defined syntactic behavior and it is clear that the arguments of interest here do not pattern with obliques.

### 6.1 Word order

Let us first consider word order patterns. As mentioned above, obliques in Meskwaki nearly always appear immediately to the left of the verb, as seen in (33 and 34). The verb in (33) requires an oblique expressing stationary location, expressed here with the locative pronoun $i \cdot n a h i$ 'there'. The verb in (34)
requires a goal oblique, here expressed by the phrase manahka si•po•ki 'yonder river'.
(33) i•nahi netapihapi
i•nahi ne-t-apih-api-Ø
there 1-EP-REDUP-sit-1/IND
$\mathrm{OBL}_{\text {loc }}$
'I was sitting there'
(34)
manahka si•po•ki neta•pi•ha
[manahka si•po•w-eki] ne-t-a•pi•ha•-Ø
yonder river-LOC 1-EP-go.thither.\&.return-1/IND
OBL $_{\text {goal }}$
'I have been to yonder river'
$\mathrm{OBJ}_{\Theta}$, in contrast, appears to the right of the verb, as seen in (35), with a ditransitive verb. The non-subject argument of verbs like ahpe-nemo-'depend on' likewise appears to the right of the verb as its unmarked position, as in (26b), repeated below:
ata hpenamaw-ihko ne-ši ši $\cdot$ kwan-i
take.hold.of.O ${ }_{\ominus}$.for 2-1/IMP my-rattle-INAN.SG
$\mathbf{O}_{\boldsymbol{\ominus}}$
'Get my rattle for me!'

| ahpe-nemo-wa | o-si•me•h-ani |
| :--- | :--- |
| depend.on-3/IND | his-younger.sibling-ANIM.OBV.SG |
| $\mathbf{O}_{\boldsymbol{\Theta}}$ |  |

'He relies on his younger brother.'

### 6.2 Case-marking

Another difference between obliques and $\mathrm{OBJ}_{\ominus}$ has to do with case morphology. Some obliques take a locative case ending, as seen in (34) on $s i \cdot p o \cdot k i$ 'river'. Locative case never appears on an $\mathrm{OBJ}_{\Theta}$ of ditransitives or on the putative $\mathrm{OBJ}_{\Theta}$ argument of verbs like ahpe-nemo- 'depend on'.

### 6.3 Relative clause formation

Another syntactic difference between $\mathrm{OBJ}_{\Theta}$ and obliques can be seen in the formation of participles, the verb forms used in relative clauses. Participles bear an additional inflectional suffix on the right edge of the verb agreeing with the head of the relative clause.

If the head of a relative clause is a subject, object, or $\mathrm{OBJ}_{\ominus}$ in the lower clause, the participle is inflected with a suffix agreeing in gender, number, and obviation with the head of the relative clause. For example, in (36), the participle bears the suffix $-a$, indicating that the head is animate proximate singular. The head of the relative clause is coreferential with the non-subject
argument of ahpe•nemo- 'depend on', the class of argument I am claiming is an $\mathrm{OBJ}_{\ominus}$. The fact that the rightmost suffix on the participle expresses gender, number, and obviation information about the head is consistent with my analysis of this argument being an $\mathrm{OBJ}_{\ominus}$.
$e \cdot h p e \cdot n e m o y a \cdot n a$
IC-ahpe•nemo-ya na
IC-depend.on-1/PART/3.HEAD
'the one whom I depend on'
(final $-a=$ animate proximate singular head of rel.cl.)
In (37) the head of the relative clause is 'tobacco', coreferential to the $\mathrm{OBJ}_{\ominus}$ associated with the preverb keki- 'having'. The final suffix on the participle is -ini, indicating that the head is (grammatically) animate and obviative singular. Again, this morphosyntactic behavior is what we would expect for an $\mathrm{OBJ}_{6}$.

| nese•ma'wani wi•hkeki-nowi•wa-čini |  |
| :---: | :---: |
| se•ma•w-ani | IC-wi•h-keki-nowi-wa-čini |
| bacco-OBV | IC-FUT-having. $\mathrm{O}^{\text {e }}$-go.out-3P/PART/3'.H |
| 'tobacco for them to take out with them' (Goddard 1987:110) |  |
| (final -ini $=$ ani | ate obviative singular head of relative clause) |

Obliques, on the other hand, behave differently in relative clause formation. If the head of the relative clause is an oblique in the lower clause, the participle is simply suffixed with $-i$, even if the head refers to an animate third person:
(38) wi•nwa•wa we-či-mehtose•neniwiyani
wi•nwa $\cdot$ wa $\quad$ IC-oči-mehtose $\cdot$ neniwi-yani
they IC-from-be.person-2/PART/OBL.HEAD
'They [your parents] are why you are alive.'
(final $-i=$ oblique head of rel.cl.)
If the non-subject arguments of the verbs in (36) and (37) were obliques, we would expect to see the participle forms suffixed with $-i$, not with $-a$ or with -ini. This test provides further evidence for the syntactic status of the nonsubject argument of verbs like ahpe•nemo- 'depend on'.

## 7 Thematic roles mapping to $\mathrm{OBJ}_{\boldsymbol{\Theta}}$

As is well known, the motivation for labeling as $\mathrm{OBJ}_{\Theta}$ the second object of a ditransitive verb or an applicative in Bantu is that such objects are restricted with regard to the type of thematic role associated with the grammatical function. It is therefore important to ask what sort of thematic roles are associated with the Meskwaki OBJ $_{\Theta}$. We certainly find themes as the OBJ $_{\Theta}$ of 'give' and other ditransitives, as well as with verbs like 'throw' and the kinship verbs listed in (23). The verbs beginning with the initial ahp-, listed in (22a-c), show that
locative arguments may also map onto $\mathrm{OBJ}_{\ominus}$. (39) and (40) informally present sample argument structures:
$\mathrm{OBJ}_{\ominus}$ in ditransitives: always THEME/PATIENT

$$
\begin{array}{clll}
m i \cdot n- & \text { 'give }<\text { agent } & \text { recip } & \text { theme }>' \\
\mathrm{~S} & \mathrm{O} & \mathrm{O}_{\ominus}> \tag{40}
\end{array}
$$

$\mathrm{OBJ}_{\mathrm{e}}$ in two place verbs:
THEME/PATIENT

| we•pa•hke-- | 'throw <agent theme>' |
| :--- | :--- |
| wani $\cdot h k e$-- | 'forget <experiencer theme>' |

LOCATIVE (verbs with initial ahp- ‘on')
ahpeka'- 'dance on <agent locative>'
ahpe•nemo- 'depend on <experiencer? locative?>
Verbs beginning with takw- 'together with' (22d-e) also take an $\mathrm{OBJ}_{\theta}$; these verbs seem to require a comitative, if that is to be recognized as a distinct thematic role. ${ }^{4}$

## 8 Can the marked valence pattern be predicted?

As stated above, the thematic role most frequently associated with $\mathrm{OBJ}_{\Theta}$ is theme/patient, but obviously not all themes and patients map onto $\mathrm{OBJ}_{\ominus}$. This can be clearly seen by comparing the Meskwaki verbs for 'eat' and 'drink': 'eat' takes an ordinary OBJ, as we saw in (1), repeated below, while 'drink' requires an $\mathrm{OBJ}_{\ominus}$.
(1) $a m w$ - 'eat $<$ S O $>$ ' $m i \cdot c \check{c} i$ - 'eat $<\mathrm{S} \mathrm{O}>$ '

$$
(\uparrow \text { OBJ GEND })==_{\mathrm{c}} \text { ANIM } \quad(\uparrow \text { OBJ GEND })={ }_{\mathrm{c}} \text { INAN }
$$

(41) meno- 'drink $<\mathrm{S} \mathrm{O}_{\ominus}>$ '

Similar observations may be made for the cases of locative $\mathrm{OBJ}_{\Theta}$ vs. locative OBLS, as in (33).

Other languages have been described in recent work (in LFG and in other frameworks) as having a similar valence pattern, in which an $\mathrm{OBJ}_{\ominus}$ appears without an OBJ. In this section I will briefly survey a few such works to place the Meskwaki phenomenon in typological perspective.

First, we may observe that the Meskwaki valence pattern under consideration here is akin to the Differential Object Marking analyzed by Aissen

[^77](2003). But note that the Meskwaki verbs under consideration here are unlike the DOM facts treated by Aissen: the linking to $\mathrm{OBJ}_{\Theta}$ occurs without regard to the definiteness, specificity, or animacy of the non-subject argument of these verbs.

Butt (1998), analyzing Urdu, proposes a modification to Lexical Mapping Theory in which themes may be intrinsically either [+r] or [-r]; the [+r] themes are mapped to $\mathrm{OBJ}_{\ominus}$. The [-r] feature is "aspectually inert" while the [ +r ] feature is associated with specificity. In causatives, the [ +r ] feature on causees results in a reading of affectedness at s-structure. However, in the Meskwaki case we find differences of neither affectedness nor aspect associated with the distinction between OBJ and $\mathrm{OBJ}_{\ominus}$. The non-subject arguments of 'eat' and 'drink' would seem to be equally affected.

Nor can information structure be appealed to, as an explanation for the unusual linking pattern. Unlike Northern Ostyak (Dalrymple and Nikolaeva 2005), there is no correlation between (secondary) topic and OBJ for Meskwaki, nor between focus and $\mathrm{OBJ}_{\odot}$.

Perhaps the closest analog to the Meskwaki pattern is found in Turkish (Çetinoğlu and Butt 2008). Turkish has an alternation in objects tied to specificity; in addition, certain verbs always take non-canonical objects (dative or ablative case), which Çetinoğlu and Butt analyze as $\mathrm{OBJ}_{\theta}$. The latter group of verbs includes psych verbs plus others (e.g. 'ride' and 'help').

It seems that in Meskwaki, as in Turkish, we must simply list certain two-place verbs as taking an $\mathrm{OBJ}_{\Theta}$ argument. In fact, because of the complex stem morphology of Algonquian languages, in Meskwaki the association with $\mathrm{OBJ}_{\ominus}$ must be made not only with full stems but also with certain initials and finals. We have already seen the initial ahp-takw- and kek- associated with $\mathrm{OBJ}_{\ominus}$; the final $-a \cdot h k e \cdot-$ 'throw' likewise always takes its theme argument as $\mathrm{OBJ}_{\ominus}$ :
(42) initial/preverb elements
a. ahp- 'on $\mathrm{OBJ}_{\Theta}{ }^{\prime}$
b. takw- 'together with $\mathrm{OBJ}_{\Theta}{ }^{\prime}$
c. kek- 'having $\mathrm{OBJ}_{\Theta}$ '
$-\mathrm{a} \cdot \mathrm{hke} \cdot \mathrm{-}$ 'throw, fling $\mathrm{OBJ}_{\ominus}$ ' (final)
a. we•pa•hke-- 'throw OBJ $_{\ominus}{ }^{\text {a }}$ [=(21a) above]
b. ina'hke•- 'fling OBJ thither'
c. ni•sa'hke•-fling $\mathrm{OBJ}_{\ominus}$ down'
d. nowa•hke•- 'fling $\mathrm{OBJ}_{\ominus}$ out'
[requires an OBL goal]

## 9 Conclusion

Recent years have seen several in-depth investigations of ditransitives, such as Maling (2001) and Kibort (2008). One recurring theme has been the observation that the properties of OBJ and $\mathrm{OBJ}_{\ominus}$ are not always so clearly distinguished from each other as standard treatments assume. Moreover, Börjars and Vincent (2008), in a critical appraisal of the OBJ function, raise the possibility that theme should be eliminated as a distinct theta-role, instead allowing the semantics of an individual verb to determine the content of the argument mapping onto OBJ. As
a result, they say "the standard distinction between OBJ and $\mathrm{OBJ}_{\ominus}$ disappears, in some sense all objects are OBJ $_{e}$."

Meskwaki, however, provides evidence in the opposite direction, in favor of retaining a distinction between OBJ and $\mathrm{OBJ}_{\ominus}$. Given the complications of ditransitive constructions, perhaps it is in constructions like the Meskwaki twoplace verbs where $\mathrm{OBJ}_{\Theta}$ occurs with no object co-argument that the properties of $\mathrm{OBJ}_{\Theta}$ can be most clearly seen.

## References

Aissen, Judith. 2003. Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory 21, 435-483.
Alsina, Alex, Mohanan, Tara, and Mohanan, KP 2005. How to Get Rid of the COMP. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG05 Conference, 21-41. Stanford: CSLI Publications.
Bloomfield, Leonard. 1962. The Menomini Language. New Haven: Yale University Press.
Börjars, Kersti and Vincent, Nigel. 2008. Objects and OBJ. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG08 Conference, 150-168. Stanford: CSLI Publications.
Bresnan, Joan and Moshi, Lioba. 1990. Object Asymmetries in Comparative Bantu Syntax. Linguistic Inquiry 21, 147-185.
Butt, Miriam. 1998. Constraining Argument Merger through Aspect. In Erhard Hinrichs, Andreas Kathol, and Tsuneko Nakazawa (eds.), Complex Predicates in Nonderivational Syntax, 73-113. New York: Academic Press.
Çetinoğlu, Özlem and Butt, Miriam. 2008. Turkish Non-canonical Objects. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG08 Conference, 214-234. Stanford: CSLI Publications.
Dahlstrom, Amy. 1991. Plains Cree Morphosyntax. New York: Garland.
Dahlstrom, Amy. 2000. Morphosyntactic Mismatches in Algonquian: Affixal Predicates and Discontinuous Verbs. In Arika Okrent and John Boyle (eds.), Proceedings from the Panels of the Chicago Linguistic Society's Thirty-sixth Meeting, 63-87. Chicago: Chicago Linguistic Society
Dalrymple, Mary and Nikolaeva, Irina. 2005. Nonsubject Agreement and Discourse Roles. In Anna McNay (ed.), Oxford Working Papers in Linguistics, Philology and Phonetics, 71-91.
Goddard, Ives. 1987. Fox Participles. In Paul D. Kroeber and Robert E. Moore (eds.), Native American Languages and Grammatical Typology, 105118. Bloomington, IN: Indiana University Linguistics Club.

Kibort, Anna. 2008. On the Syntax of Ditransitive Constructions. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG08 Conference, 312-332. Stanford: CSLI Publications.
Maling, Joan. 2001. Dative: the Heterogeneity of the Mapping among Morphological Case, Grammatical Functions, and Thematic Roles. Lingua 111:419-464.

Rhodes, Richard. 1991. Secondary Objects in Ojibwe. In Katarzyna Dziwirek, Patrick Farrell and Errapel Mejías-Bikandi (eds.), Grammatical Relations: A Cross-theoretical Perspective, 401-414. Stanford: CSLI Publications.

# A TALE OF TWO TAQS: <br> THE OT-LFG OF PLURALS AND DISTRIBUTIVES IN K'ICHEE' MAYAN 

Lachlan Duncan<br>University at Albany, SUNY

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

## CSLI Publications

http://csli-publications.stanford.edu/


#### Abstract

This paper investigates the distributive pluralizer taq (PL) of K'ichee' Mayan. As a nominal pluralizer, the non-bound morpheme taq barely registers in the Mayanist literature, while the distributive taq (DISTR) is virtually non-existent. Semantically the distributive pluralizer taq pluralizes nominals that are ambiguous between collective and distributive readings. Morphosyntactically the distributive pluralizer taq is a phrasal particle that (left) adjoins to stringadjacent constituents. This contrasts with the morphosyntax of the distributive taq that I argue elsewhere is a non-projecting particle that (right) head-adjoins to verbs only. Using Optimality Theoretic Lexical-Functional Grammar (otLFG), the complex phrasal distribution of the distributive pluralizer taq, which is unaccountable using phrase-structure rules alone, can be straightforwardly modeled using a modest number of universal constraints.


This paper investigates the distributive pluralizer taq (PL) of K'ichee' Mayan. ${ }^{1,2}$ While little has been said about the non-bound morpheme taq as a nominal pluralizer in the grammars and dictionaries of the K'ichee' an language family, virtually nothing has been said about its use as a distributive (DISTR). ${ }^{3}$

The only substantive description of the morpheme taq is in Willson (2004, 2005), where it is interpreted as a distributive and a pluralizer. As a distributive, $t a q$ associates with verbs. As a pluralizer, taq follows adjectives, possessed nouns, relational nouns, prepositions, and 'splits' compound nouns. Judgment is reserved about whether taq is one morpheme with two uses, or two morphemes each with its own use. As for word type, Willson provisionally interprets taq as a clitic.

Employing a variety of data and linguistic constructions, I demonstrate conventional use of the distributive pluralizer taq and show the categories of words that it associates with and the positions that it occupies in the phrase. As a nominal pluralizer (PL), I indicate that taq is used with $w h$-interrogatives, NPs, (possessive) DPs, relational nouns, QPs, PPs, and non-verbal predicates. I propose that the distributive pluralizer taq pluralizes nominals that are semantically ambiguous between collective and distributive readings. I argue that the distributive pluralizer taq is a phrasal particle that (left) adjoins to string-adjacent constituents.

[^78]The complex phrasal distribution of the distributive pluralizer taq, which is unaccountable using phrase-structure rules alone, can be straightforwardly modeled using Optimality Theoretic Lexical-Functional Grammar (OT-LFG) (Bresnan 2000, et al.) and a modest number of universal constraints. Data on the distributive pluralizer taq is shown in section 1, and the OT-LFG analysis in section $2 .{ }^{4}$

## 1 K'ichee' data

Nominals Inanimate entities, like lee leej 'the tortilla(s),' are ambiguous between singular and plural readings. Structurally the distributive pluralizer taq in (1) cannot precede the determiner of the DP, nor can it be placed inside the DP, between the determiner and the head noun. The distributive pluralizer taq cannot immediately follow the noun that it pluralizes and, at the same time, be phrase-final:

| Lee leej | *Taq lee leej | *Lee taq leej | *Lee leej taq |
| :--- | :--- | :--- | :--- |
| DET tortilla | PL DET tortilla | DET PL tortilla | DET tortilla PL |
| 'The tortilla(s)' | ('The tortillas') | ('The tortillas') | ('The tortillas') |

Negation The negation of a singular and a plural bare NP is shown in (2). The negation word $m a$ 'no' and the non-projecting irrealis word $t a(j)$ (IRR) typically frame the negated constituent. The distributive pluralizer taq cannot be negated:

| Ma leej taj | *Taq leej taj | *Ma taq leej taj | *Ma taq taj |
| :--- | :--- | :--- | :--- |
| NEG tortilla IRR | PL tortilla IRR | NEG PL tortilla IRR | NEG PL IRR |
| 'No tortilla' | ('No tortillas') | ('No tortillas') | ('No (PL)') |

Attributive adjectives In a DP with an attributive adjective pre-head modifier, as in (3), the distributive pluralizer taq must follow the adjective:

| a. Lee q'an-a leej | Lee q'an-a taq leej |  |
| :--- | :--- | :--- |
| DET yellow-ATT tortilla | DET yellow-ATT PL tortilla |  |
| 'The yellow tortilla' | 'The yellow tortillas' |  |
| b. *Lee taq q'an-a leej | *Lee q'an-a leej taq |  |
|  | DET PL yellow-ATT tortilla | DET yellow-ATT tortilla PL |
| ('The yellow tortillas') | ('The yellow tortillas') |  |

The distributive pluralizer taq in (4a) follows the first adjective nim 'big.' Following the second adjective $q$ 'eq 'black' in (4b) is not preferred. Although not ungrammatical, using taq after multiple attributive adjectives is always avoided (4b). The overwhelming preference, then, is for the distributive pluralizer taq to follow the left-most attributive adjective, and to be used once per clause:

[^79](4) a. Lee nim-a q'eq-a ab'aj Lee nim-aq taq q'eq-a ab'aj DET big-ATT black-ATT rock DET big-PL PL black-ATT rock 'The big black rock' 'The big black rocks'
b. ?Lee nim-aq q'eq-a $\boldsymbol{t a q} a b$ 'aj Lee nim-aq $\boldsymbol{t a q} q$ 'eq-a $\boldsymbol{t a q} a b \prime a j$ D big-PL black-ATT PL rock D big-PL PL black-AT PL rock 'The big black rocks' 'The big black rocks'

Numerals A DP modified by a cardinal cannot be pluralized with the distributive pluralizer $\operatorname{taq}$ (5). If used with a distributive pluralizer, cardinals could be confused with a distributive numeral, for example, jo'taq 'by fives, five-by-five':

| Lee jo'-ob' ab'aj | *Lee jo'-ob' taq ab'aj | *Lee taq jo'-ob' ab'aj |
| :--- | :--- | :--- |
| DET five-PL rock | DET five-PL PL rock | DET PL five-PL rock |
| 'The five rocks' | ('The five rocks') | ('The five rocks') |

Possessives - morphological The noun ja' 'water' in (6) is possessed by the inanimate noun tinamit 'town,' and $m e$ 's 'cat' by the animate noun $a k$ 'aal 'child':

| Lee u-ja' lee tinamit | Lee u-me's lee ak'aal |
| :--- | :--- |
| DET 3SPOS-water DET town | DET 3SPoS-cat DET child |
| 'The town's water' | 'The child's cat' |

The data in (7), with and without the distributive pluralizer taq, are the pluralized forms of the singular inanimate possessor nominal tinamiit 'town' from (6). These data show that the two possessive phrases are semantically identical:

| Lee u-ja' taq lee tinamit | Lee ki-ja' lee tinamit |
| :--- | :--- |
| DET 3SPOS-water PL DET town | DET 3PLPOS-water DET town |
| 'The towns' water' | 'The towns' water' |

The data in (8-9) are the pluralized forms of the singular animate possessor ak'aal in (6). The morphological plural form using plural agreement ki- without the distributive pluralizer taq is shown in (8). Because the animate possessor is morphologically marked as plural, the possessed noun must agree in number. Nominals without morphological plurals do not automatically trigger number agreement. Because the two phrases in (8) are semantically equivalent, it follows that no exclusive distributive reading exists using the distributive pluralizer taq with plural nominals:

$$
\begin{array}{ll}
\text { Lee ki-me's lee ee ak'al-aab' } & \text { Lee ki-me's taq lee ee ak'al-aab' }  \tag{8}\\
\text { DET 3PLPOS-cat D PLU child-PL } & \text { DET 3PLPos PL D PLU child-PL } \\
\text { 'The children's cat' } & \\
& \text { 'The children's cat' }
\end{array}
$$

Although the distributive pluralizer taq can also be used in data with plural agreement, such as in (7-8), there appears to be a distinct preference against this by my consultants. Agreement on the possessed nouns in (9) is a mismatch:
(9) *Lee u-me's lee ee ak'al-aab' *Lee u-me's taq lee ee ak'al-aab’ DET 3sPos-cat DET PLU child-PL DET 3SPOS PL D PLU child-PL ('The children's cat')
('The children's cat')

Possessives - lexical An alternate method for indicating possession exists using the inflecting relational noun $-e e(c h)$ 'of, possession.' The distributive pluralizer taq in (10b) follows the relational noun ree 'of (it)' to pluralize its possessor DP lee tinamit 'the towns.' When the unpossessed noun lee ee $t z$ ' $i$ ' 'the dogs' in (10c) is immediately followed by the distributive pluralizer taq and then a PP, the use of taq to pluralize the PP's complement DP lee tinamit 'the town' is not permitted:
a. Lee ee tz'i' r-ee lee tinamit

DET PLU dog 3sPos-Poss DET town
'The town's dogs / The dogs of the town'
b. Lee ee tz'i' r-ee taq lee tinamit

DET PLU dog 3SPos-Poss PL DET town
'The towns' dogs / The dogs of the towns'
c. *Lee ee tz'i' taq r-ee lee tinamit

DET PLU dog PL 3SPos-Poss DET town
('The towns' dogs / The dogs of the towns')
If the distributive pluralizer taq follows an unpossessed DP and is itself then followed by a PP, taq cannot be used to pluralize the PP's DP complement (11c):
a. Ee k'oo k'a'n-a tz'i' pa lee tinamit

3PLABS exist mean-ATT dog PREP DET town
'There are mean dogs in the town.'
b. Ee k'oo k'a'n-a tz'i' pa taq lee tinamit

3PLABS exist mean-ATT dog PREP PL DET town
'There are mean dogs in the towns.'
c. *Ee k'oo k'a'n-a tz'i' taq pa lee tinamit 3PLABS exist mean-ATT dog PL PREP DET town ('There are mean dogs in the town(s).')

Phrasal compounds The distributive pluralizer taq can pluralize phrasal compounds. The latter consist of two separate words that act as a single lexical unit. The phrasal compounds in (12a) are inanimate [Adjective Noun] and animate [Noun Noun]. The pluralized versions of the inanimate and animate phrasal compounds are shown in (12b). The only position the distributive pluralizer taq can occupy in (12) is preceding the second noun of the phrasal compound:
a. K'im-a jaa
thatch-ATT house
'Thatched house'

Lee ati't ak'
DET female chicken
'The hen (La gallina)'
b. K'im-a taq jaa thatch-ATT PL house 'Thatched houses'

Lee ee ati't-ab' taq ak'
DET PLU female-PL PL chicken
'The hens (Las gallinas)'

Consider a DP headed by an adjective-noun [A N] phrasal compound with a cardinal and attributive adjective. The attributive marker $-a$ on the pre-head word k'im in (13b) indicates that the word $k^{\prime}$ im is a modifying adjective, and that, lexically, it is part of the phrasal compound k'ima jaa 'thatched house.' The distributive pluralizer taq follows the attributive adjective niitz' in (13a) and the adjective $k^{\prime} i m$ in (13b). The alternation indicates ideolectical or dialectical microvariation:
a. Lee niitz' k'im-a jaa Lee jo'ob' niitz' k'im-a taq jaa D little thatch-ATT house $D$ five little thatch-ATT PL house 'The little thatched house' 'The five little thatched houses'
b. Lee jo'-ob' niitz' taq k'im-a jaa DET five-PL little PL thatch-ATT house 'The five little thatched houses'
c. ??Lee jo'-ob' niitz' taq k'im-a taq jaa DET five-PL little PL thatch-ATT PL house ('The five little thatched houses')

Prepositional phrases The complex preposition puwi' 'above' in the second part of (14) agrees in number and person with the preposition's morphologically singular (but semantically plural) DP complement lee chee' 'the tree'. ${ }^{5}$

P-u-wi' lee chee'
PREP-3SPos-head DET tree
'Above the tree.'

P-u-wi’ taq lee chee'
PREP-3SPOS-head PL DET tree 'Above the trees.'

The distributive pluralizer taq cannot 'split' a PP's unmodified DP complement:
*P-u-wi' lee taq chee' *Pa-ki-wi' lee taq chee'
PREP-3SPOS-head DET PL tree PREP-3PLPos-head DET PL tree ('Above the trees.') ('Above the trees.')

If a pre-head attributive adjective modifies the head noun of the PP's DP complement, the distributive pluralizer taq must follow the DP's attributive adjective: ${ }^{6}$
a. P-u-wi lee rax-a taq chee'

PREP-3SPOS-head DET green-ATT PL tree
'Above the green trees.'

[^80]> b. Pa-ki-wi’ lee rax-a taq chee' PREP-3PLPOS-head DET green-ATT PL tree 'Above the green trees.'

When an attributive adjective modifies the nominal head of the DP complement, the distributive pluralizer taq cannot immediately follow the preposition (17): ${ }^{7}$

| a. | *P-u-wi' taq lee rax-a chee' |
| ---: | :--- |
|  | PREP-3SPOS-head PL DET green-ATT tree |
|  | ('Above the green trees.') |
| b. | Pa-ki-wi' taq lee rax-a chee' |
|  | PREP-3PLPos-head PL DET green-ATT tree |
| ('Above the green trees.') |  |

If a cardinal quantifies the head noun of a DP complement, the distributive pluralizer must follow the preposition (18b), not the cardinal (18c):
a. Ch-u-paam taq lee tinamit PREP-3SPOS-stomach PL DET town 'Inside the towns.'
b. Ch-u-paam (taq) lee ox-ib' tinamit PREP-3SPOS-stomach PL DET three-PL town 'Inside the three towns.'
c. ??Ch-u-paam lee ox-ib' taq tinamit
PREP-3SPOS-stomach DET three-PL PL town ('Inside the three towns.')

If a cardinal is followed by a pre-head attributive adjective, the distributive pluralizer taq follows the adjective, not the cardinal (19a). Clearly plural cardinals do not behave like attributive adjectives. In the configuration of pre-head modifiers in (19c), the distributive pluralizer taq cannot follow the preposition:


The restriction on the pluralization of cardinals by taq might be due to possible confusion with distributive numerals, like waqitaq 'six by six,' for example (20):

[^81]| Waq-iib' | *Waq-ib' taq | Waq-i-taq |
| :--- | :---: | :--- |
| six-PL | six-PL | PL | | six-ATT-DISTR |
| :--- |
| 'Six' |
| ('Six') |

Phrasal compound DP complement When the preposition's DP complement is a phrasal compound, pluralization is somewhat more involved. The phrasal compound, tiox jaa 'church' consists of two juxtaposed nominal heads, tiox 'Dios' and jaa 'house.' The distributive pluralizer taq in (21a) follows the PP's (prepositional) head. Pluralizing the phrasal compound tiox jaa 'church' in (21b) with the distributive pluralizer taq is questionable at best: ${ }^{8}$

| a. | Ch-u-wach lee tiox jaa | Ch-u-wach taq lee tiox jaa |
| :--- | :--- | :--- |
|  | PREP-3sPos-face D god house | P-3sPos-face PL D god house |
|  | 'In front of the church.' | 'In front of the churches.' |

b. ??Ch-u-wach lee tiox taq jaa PREP-3sPos-face DET god PL house 'In front of the churches.'
c. Ch-u-wach taq lee tiox taq jaa PREP-3SPos-face PL DET god PL house 'In front of the churches.'

If an attributive adjective is used as a DP complement's pre-head modifier, the adjective seems to strongly 'attract' the distributive pluralizer taq. The pluralizer taq in (22b) directly follows the attributive adjective $q$ 'el 'old.' The pluralization of the phrasal compound in (22c) by the distributive pluralizer taq is not preferred. Alternatively when the attributive adjective in (22e) modifies the phrasal compound, the distributive pluralizer taq is not permitted to follow the preposition chuwach:
a. Ch-u-wach lee q'el-a tiox jaa PREP-3SPOS-face DET old-ATT god house 'In front of the old church.'
b. Ch-u-wach lee q'el-a taq tiox jaa PREP-3SPOS-face DET old-ATT PL god house 'In front of the old churches.'
c. ?Ch-u-wach lee q'el-a tiox taq jaa PREP-3SPOS-face DET old-ATT god PL house 'In front of the old churches.'
d. ??Ch-u-wach lee q'el-a taq tiox taq jaa PREP-3SPos-face DET old-ATT PL god PL house 'In front of the old churches.'
e. *Ch-u-wach taq lee q'el-a tiox jaa PREP-3SPOS-face PL DET old-ATT god house ('In front of the old churches.')

[^82]Interrogatives Interrogative operators can be pluralized in two ways. When referencing an animate argument, an interrogative can be pluralized with the animate pluralizer $e e$, and the distributive pluralizer taq optionally (23a). An interrogative can also be pluralized with the distributive pluralizer taq alone, particularly when the operator references an inanimate entity (23b): ${ }^{9}$

> a. Ee jachin (taq) k-ee-b'ii-n la' ch-aw-ee? PLU INT PL INC-3PLABS-say-AF DEM PREP-2SPOS-Poss 'Who (PL) said that to you?'
> b. Jachin taq k-ee-b' an-ow la' ch-k-ee?
> INT PL INC-3PLABS-make-AF DEM PREP-3PLPOS-Poss 'What (PL) did that to them?'

Ambiguity of plural descriptives I argue that standard plural nominals in $\mathrm{K}^{\prime} \mathrm{i}$ chee' are semantically ambiguous between collective and distributive readings. DP complements pluralized with the distributive pluralizer taq are not interpreted as having exclusive distributive readings. Both collective and distributive readings remain available, but the collective reading is the default.

The PP in (24) with a plural DP complement has at least two interpretations; a collective reading, which is the default (24a), and a distributive reading (24b):

```
Pa taq lee juyub,
PREP PL DET mountain
a. 'In all of the mountains.' (Collective reading)
b. 'In each of the mountains.' (Distributive reading)
```

Temporal events can be expressed with PPs. Because of the use of the distributive pluralizer taq, the PPs in (25) appear to have a distinctly distributive reading:

| Pa saq'iij | Pa taq saq'iij | Pa martes | Pa taq martes |
| :--- | :--- | :--- | :--- |
| PREP summer | PREP PL summer | PREP T. | PREP PL T. |
| 'In summer.' | 'Every summer.' | 'On Tuesday.' | 'Every Tuesday.' |

When used following prepositions, taq is typically a pluralizer with a collective reading. Yet in (25) the distributive reading seems more appropriate. The data support my contention that plural nominals in K'ichee' are semantically ambiguous between collective (default) and distributive (marked) readings.

Non-verbal predicates The distributive pluralizer taq is also used in non-verbal predicates, which can provide additional insight about the morpheme taq. Let us consider in particular the pluralization of subjects and the hosting of the pluralizer.

Subject pluralization The glosses in (26b) indicate clearly that the subject nominals of the non-verbal predicates are targeted for pluralization by the distributive

[^83]pluralizer taq. In this form of clausal arrangement, the distributive pluralizer taq cannot be used 'inside' the subject nominal in order to pluralize it (26c):
a. Saq lee jaa Q'or lee ala
white DET house 'The house is white.'
b. Saq taq lee jaa white PL DET house
'The houses are white.'
c. $*$ Saq lee $\boldsymbol{t a q}$ jaa white DET PL house
('The houses are white.')
lazy DET boy 'The boy is lazy.' Ee q'or-ib' taq lee alab'oom 3PLABS lazy-PL PL DET boy:PL 'The boys are lazy.'
*Ee q'or-ib' lee taq alab'oom 3PLABS lazy-PL DET PL boy:PL ('The boys are lazy.')

Contrarily if an attributive adjective modifies the head noun as in (27), the distributive pluralizer taq must immediately follow the attributive adjective:

Saq lee q'el-a-laj taq jaa Ee q'or-ib' lee alaj taq alab'oom white D old-ATT-INT PL house 3PLA lazy-PL D small PL boy:PL 'The very old houses are white.' 'The small boys are lazy.'

And if the distributive pluralizer taq instead follows the predicative adjective and not the attributive adjective, the clause is ungrammatical (28): $:^{10}$
*Saq taq lee q'el-a jaa *Ee q'or-ib' taq lee alaj alab'oom white PL DET old-ATT house 3PLA lazy-PL PL DET small boy:PL ('The old houses are white.') ('The small boys are lazy.')

Pluralization in the non-verbal predicates using the distributive pluralizer taq is syntactically similar to pluralization in PPs and QPs. But the distributive pluralizer $t a q$ is not a distributive in non-verbal predicates because the latter are not verbs. Rather non-verbals are non-eventives, non-dynamic statives that can never distribute over sorting keys as distributive shares.

Pluralizer host The data in (29) illustrate that the distributive pluralizer taq precedes the plural subject DP lee $t z$ ' $i$ ' 'the dogs' in (29a), but does not precede it in (29b). In the former, lee $t z ' i$ ' follows the predicate as grammatical subject, whereas, in the latter, lee $t z^{\prime} '$ ' is in sentence-initial position, in this case as external topic. Crucially the distributive pluralizer taq remains in situ in (29b) when the subject DP extracts to external topic position. Example (29) includes the antipassive voiced verb keeti'onik 'they bite' used here as a restrictive relative clause:
a. Ee k'a'n taq lee tz'i' k-ee-ti'o-n-ik

3PLABS mean PL DET dog INC-3PLABS-bite-AP-IPF
'The dogs that bite are mean.'

[^84]b. Lee tz'i' ee k'a'n taq k-ee-ti'o-n-ik

DET dog 3PLABS mean PL INC-3PLABS-bite-AP-IPF
'The dogs that bite are mean.'
The distributive pluralizer taq cannot extract with the subject it pluralizes to sen-tence-initial position (30a). Even if the extracted subject is not sentence-initial (30b), the sentence is ill-formed. If the distributive pluralizer taq extracts along with the subject, the sentence is ill-formed (30c). It is obvious from (29-30) that the pluralizer taq does not necessarily attach to the DP that it pluralizes:

> a. *Taq lee tz'i' ee k'a'n k-ee-ti'o-n-ik PL DET dog 3PLABS mean INC-3PLABS-bite-AP-IPF ('The dogs that bite are mean.')
> b. *Ojeer taq lee tz'i' ee k'a'n k-ee-ti'o-n-ik before PL DET dog 3PLABS mean INC-3PLABS-bite-AP-IPF ('In the past, the dogs that bite were mean.')
> c. *Ojeer lee tz'i' taq ee k'a'n k-ee-ti'o-n-ik before DET dog PL 3PLABS mean INC-3PLABS-bite-AP-IPF ('In the past the dogs that bite were mean.')

Non-projecting word or phrase? The category and word type of the distributive pluralizer taq have not yet been established. I argue elsewhere that the distributive taq (DISTR), used exclusively in verbal predicates, is a non-projecting word. So is the distributive pluralizer taq also a non-projecting word? Let us first consider a DP with a coordinated attributive adjectival modifier. As we know, the distributive pluralizer taq preferentially follows the left-most pre-head attributive adjective (4b). One could conclude that the distributive pluralizer taq would follow the left-most adjective in a coordinated phrase. This assumes that the pluralizer taq is a nonprojecting word because it head-adjoins to its host, and as such, does not respect phrasal boundaries. Thus in a coordinated phrase, a non-projecting word would be predicted to follow the left-most adjective. Nonetheless it is clear that the distributive pluralizer taq in (31) follows the entire coordinated adjectival phrase q'eqa chi'l saqa, not the first attributive adjective $q$ 'eqa 'black.' Because a non-projecting word can penetrate the phrasal boundaries of any phrase, the distributive pluralizer taq, as a hypothesized non-projecting word, should be able to immediately follow the DP's left-most adjective, $q$ 'eq 'black.' But as (31c) demonstrates, it does not:

[^85]The non-projecting adverb $\operatorname{chi}(k)$ can precede or follow the head noun kape: ${ }^{11}$

```
Jun q'eq-a chi kape
DET black-ATT again coffee
'Another black coffee'
Jun q'eq-a kape chik
DET black-ATT coffee again
'Another black coffee'
```

I propose elsewhere that the non-projecting adverb chik and the distributive taq (DISTR) can order freely after the verb complex. If the distributive pluralizer taq were a non-projecting word like the non-projecting adverb chik, then the two words should similarly be able to order freely after the pre-head attributive adjective. The data in (33) clearly show that the two words do not order freely. This surprising result suggests that the distributive pluralizer taq may not be a non-projecting word:

$$
\begin{array}{lll}
\text { a. Jujun q'eq-a taq kape } & \text { Jujun q'eq-a taq chi kape }  \tag{33}\\
& \text { DISTR black-ATT PL coffee } & \text { DISTR black-ATT PL again coffee } \\
& \text { 'Some black coffees' } & \text { 'Some more black coffees' } \\
\text { b. Jujun q'eq-a taq kape chik } & \text { *Jujun q'eq-a chi taq kape } \\
& \text { DISTR black-A PL coffee again } & \text { DISTR black-A again PL coffee } \\
& \text { 'Some more black coffees' } & \text { ('Some more black coffees') }
\end{array}
$$

Let us consider PPs that include the distributive pluralizer taq and directionals. The distributive pluralizer taq is used to pluralize the PP's object complement (34b):
a. Ee k'oo lee kyeej pa lee saq'umb'al 3PLABS exist DET horse PREP DET field 'The horses are in the field.'
b. Ee k'oo lee kyeej pa taq lee saq'umb'al 3PLAbS exist DET horse PREP PL DET field 'The horses are in the fields.'

Directionals can be used in a PP immediately following the preposition (35a). But the distributive pluralizer $t a q$ and directional $a q$ 'an 'above' can not be used together following the preposition in a PP irrespective of their order (35b-c): ${ }^{12}$
a. Ee k'oo lee kyeej pa aq'an lee saq'umb'al 3PLABS exist DET horse PREP DIR DET field 'The horses are up above the field.'
b. *Ee k'oo lee kyeej pa aq'an taq lee saq'umb'al 3PLABS exist DET horse PREP DIR PL DET field ('The horses are up above the fields.')
c. *Ee k'oo lee kyeej pa taq aq'an lee saq'umb'al 3PLABS exist DET horse PREP PL DIR DET field ('The horses are up above the fields.')

[^86]In addition, it is possible to gap the head of a PP whose object complement has been pluralized by the distributive pluralizer taq. The preposition in (36) following the conjunctive adverb chuq(e) 'also, and' in the sentence-final PP has been gapped: ${ }^{13}$
(36) Lee siink aree ka-chooman taq lee chaak pa lee tinamit DET syndicate 3SPRO INC-organize DISTR DET work PREP DET town
xuq pa taq juyub', k'ayb'al, chuqe taq lee b'eh CONJ PREP PL aldea market CONJ PL DET road
'El síndico, es él que arregla los trabajos en el pueblo, las aldeas, los mercados, y las carreteras (Ajpacajá Tum et al. 2005:361).'
'The syndicate organizes every job in the town, and in the aldeas, markets, and the roadways. ${ }^{14}$

Because the preposition has been elided in (36), the distributive pluralizer taq cannot head-adjoin to it. In sum, the data support the proposal that the distributive pluralizer taq is a phrase, not a non-projecting word. In that case, the distributive pluralizer taq adjoins to whichever constituent is right string-adjacent. ${ }^{15}$

## 2 The ot-LFG of the distributive pluralizer taq

In brief, I argue that the K'ichee' morpheme taq denotes two grammaticized concepts: plurality (PL) and distributivity (DISTR), and represents two word types: phrase and non-projecting word. To indicate the plurality of nominals, the phrase taq follows attributive adjectives, interrogatives, prepositions, non-numerical quantifiers, the heads of possessive constructions, and non-verbal predicates. Restrictions on the phrasal distribution of taq are substantial: taq can never be phrase-initial or phrase-final, can never follow determiners, cardinals, or unpossessed nouns, and can only follow a phrasal compound's initial word. Preferred usage of taq is one per clause. To indicate distributivity (DISTR), the non-projecting word taq immediately follows finite verbs only, freely ordering with other non-projecting words, like the adverb chik and the directionals, for example. As regards category, I suggest that both forms of taq are non-phonologically dependent particles. ${ }^{16}$

The lexical entries of the non-bound morpheme taq are shown in (37):

$$
\begin{array}{lll}
\operatorname{taq} & \mathrm{Pt}^{0} & (\uparrow \mathrm{NUM})=\{\text { DISTRIBUTIVE } \mid \text { PLURAL }\}  \tag{37}\\
\operatorname{taq} & \widehat{\mathrm{Pt}} & (\uparrow \mathrm{NUM})=\text { DISTRIBUTIVE }
\end{array}
$$

Constraints Phrase-structure rules are, of course, indispensable in that they license the phrasal organization of constituent categories. But unordered PS rules

[^87]account only for dominance relations of phrasal constituents, not their linear order. Some have proposed a limited set of generalized ordering rules to account for linear word order in the clause. ${ }^{17}$ It has been suggested, however, that a more representative method of explaining linear word order can be captured using ot (Prince and Smolensky 1993) or OT-LFG (Bresnan 2000). Let us consider the constraints.

The constraint in (38a-b) penalizes the placement of the distributive pluralizer taq initially in a $[-\mathrm{V}]$ constituent (NP, DP, PP). The constraint in (38c-d) penalizes placing the distributive pluralizer taq finally in a -V constituent (NP, DP, PP). Let us propose, then in (38e-f), to unify the two 'edge' constraints as Avoid(Edge):
(38) a. Distributive pluralizer taq may not be initial in [-V] constituent
b. *INitial $(t a q) \Rightarrow$ *Initial
c. Distributive pluralizer taq may not be final in a $[-\mathrm{V}]$ constituent
d. *FINAL $(t a q) \Rightarrow$ FINAL
e. Unify *INITIAL and *FINAL so phrasal boundaries are penalized
f. *Initial $\cup *$ Final $\Rightarrow$ *Edge

When all the candidates badly violate ranked constraints, no output is generated resulting in ineffability. To account for ineffability, the constraint MPARSE (Prince and Smolensky 2004) is used because it penalizes no output. MPARSE resolves the tableau by satisfying all candidates except the null parse candidate ' $\varnothing$ ':

Ineffability: use null parse candidate $\emptyset$, and the constraint MPARSE
The distributive pluralizer taq displays strong preferences for following attributive adjectives. Formalizing this preference is straightforward: always penalize a phrase in which the distributive pluralizer taq does not abut an adjective (40):
(40) a. Align left edge of distr. pluralizer taq with right edge of an adjective
b. Align(taq, L, Adj, R) $\Rightarrow$ Align-ADJ

Several types of phrasal compound occur in K'ichee' (e.g., A N, N N). The distributive pluralizer taq must be constrained so that it only follows the phrasal compound's initial word. The necessary constraint must also penalize the distributive pluralizer taq for not following adjectives, interrogatives, possessed nouns, prepositions, quantifiers, and so on. Therefore the constraint in (41) requires the distributive pluralizer taq to be placed immediately after a lexical category ( $\mathrm{N}, \mathrm{A}, \mathrm{P}, \mathrm{Q}$ ):
(41) a. Align left edge of taq with right edge of a $[-\mathrm{V}]$ lexical category
b. $\quad \operatorname{Align}\left(t a q, L, X_{[+ \text {lexical }]}, R\right) \quad \Rightarrow \quad$ ALIGN-LEX

Constraint ranking The constraints, *Edge, Mparse, Align-adj, AlignLEX, are ranked according to the hierarchy in (42):
(42) *Edge $\gg$ Mparse $\gg$ Align-adj $\gg$ Align-Lex

[^88]Determiner phrases The PS rules in (43) license a DP configured as 'Det N ': ${ }^{18}$

$$
\begin{equation*}
\mathrm{DP} \rightarrow \underset{\uparrow=\downarrow}{\mathrm{D}^{0}} \quad, \quad \underset{\uparrow=\downarrow}{\mathrm{NP}} \tag{43}
\end{equation*}
$$

$$
\mathrm{NP} \rightarrow \underset{\substack{\mathrm{~N}^{0} \\ \uparrow=\downarrow}}{\mathrm{N}^{2}}
$$

The OT-LFG of the distributive pluralizer taq in the DP in (1) is shown in tableau 1. But tableau 1 is suboptimal because it produces no optimal or winning candidate.

Tableau $1 \mathrm{DP} \Rightarrow \operatorname{Det} \mathrm{N}+\operatorname{taq}$ (PL)

| $t a q$ | Det N | *EDGE | ALIGN-ADJ | Align-LEX |
| :---: | :---: | :---: | :---: | :---: |
| a. | $\operatorname{taq} \operatorname{Det} \mathrm{N}$ | $*!$ | $*$ | $*$ |
| b. | Det $\operatorname{taq} \mathrm{N}$ |  | $*!$ | $*$ |
| c. | Det $\mathrm{N} \operatorname{taq}$ | $*!$ | $*$ |  |

Ineffability Ineffability occurs when the candidates violate the constraints so egregiously that no optimal output is produced. In tableau 1, which shows DP $\Rightarrow$ Det N, no output is optimal, and the result is ineffability. But ineffability can be accounted for using Prince and Smolenski's (2004) constraint MPARSE, which penalizes no output. Essentially all candidates compete with the null parse candidate ' $\emptyset$,' which satisfies all constraints, except for the constraint MPARSE.

An OT-LFG account of the $\mathrm{DP} \Rightarrow$ Det N in (1) pluralized with the distributive pluralizer taq is shown with the constraint MPARSE in tableau 2. Tableau 2 indicates that the optimal candidate is candidate (d), which represents the null parse candidate $\emptyset$. Therefore the output is null. Nonetheless tableau 2 remains well-formed with an optimal output, unlike tableau 1 , which is ineffable.

Tableau $2 \mathrm{DP} \Rightarrow \operatorname{Det} \mathrm{N}+$ taq (PL)

| $t a q \quad$ Det N | *EDGE | MPARSE | ALIGN-ADJ | ALIGN-LEX |
| :---: | :---: | :---: | :---: | :---: |
| a. $\operatorname{taq} \operatorname{Det~} \mathrm{N}$ | $*!$ |  | $*$ | $*$ |
| b. $\operatorname{Det} \operatorname{taq} \mathrm{N}$ |  |  | $*!$ | $*$ |
| c. Det $\mathrm{N} \operatorname{taq}$ | $*!$ |  | $*$ |  |
| d. $\emptyset$ |  | $*$ |  |  |

Consider the DP in (3) configured as 'Det Adj N.' The phrase structure rules in (43) added to (44) license 'Det Adj $N$ ' pluralized by the distributive pluralizer taq:

$$
\begin{equation*}
\mathrm{NP} \rightarrow \underset{\substack{\mathrm{NP} \\ \mathrm{NP}}}{ }, \quad \mathrm{AP} \tag{44}
\end{equation*}
$$

$$
\begin{aligned}
& \mathrm{PLP} \rightarrow \mathrm{PL}^{0} \\
& \uparrow=\downarrow
\end{aligned}
$$

[^89]

Figure $1 \mathrm{DP} \Rightarrow$ Det Adj N

$$
\mathrm{NP} \rightarrow \underset{\substack{\mathrm{NP} \\
\uparrow=\downarrow}}{ } \quad \begin{aligned}
& \mathrm{PLP} \\
& \\
& \\
& \\
& \\
& (\uparrow=\downarrow \\
& \\
& \uparrow \text { NUM })=\mathrm{PL}
\end{aligned}
$$

$$
\mathrm{AP} \rightarrow \mathrm{~A}^{0}
$$

$$
\uparrow=\downarrow
$$

An OT-LFG account of $\mathrm{DP} \Rightarrow$ Det Adj N pluralized by taq is shown in tableau 3.
Tableau 3 DP $\Rightarrow$ Det Adj $\mathrm{N}+\operatorname{taq}$ (PL)

| taq $\quad$ Det Adj N | *EDGE | MPARSE | AlIGN-ADJ | AlIGN-LEX |
| ---: | :---: | :---: | :---: | :---: |
| a. $\quad$ taq Det Adj N | $*!$ |  | $*$ | $*$ |
| b. Det taq Adj N |  |  | $*!$ | $*$ |
| c. Det Adj taq N |  |  |  |  |
| d. Det Adj N taq | $*!$ |  | $*$ |  |
| e. $\quad \varnothing$ |  | $*!$ |  |  |

The c-structure and f-structure in figure 1 show candidate (c) of tableau 3.

Possessive DPs (N [DP Pos $]$ ) Consider the possessor DP in (7) pluralized by the distributive pluralizer taq. The possessed DP (Possessum/POSM) is the entity possessed, and is the head of the possessive construction. The semantic role possessor (syntactic genitive) is the entity that possesses the possessum. The genitive posses-
sor is designated as $\mathrm{DP}_{\text {Pos }} .{ }^{19}$ Using prefixed 'set A' possessive morphology, the possessum agrees with the number and person of the possessor.

The phrase-structure rules in (45) license possessor DPs. The possessor DP itself is functionally annotated with ( $\uparrow$ POSS) $=\downarrow$ :
$(\uparrow$ NUM $)=P L$
An OT-LFG account of the possessor $\mathrm{DP}\left(\mathrm{DP} \Rightarrow \mathrm{NDP}_{\mathrm{Pos}}\right)$ pluralized by the distributive pluralizer taq, is shown in tableau 4.

Tableau $4 \mathrm{DP} \Rightarrow \mathrm{N} \mathrm{DP}_{\text {Pos }}+t a q$ (PL)

| $t a q \quad \mathrm{~N} \mathrm{DP}_{\mathrm{Pos}}$ | *EDGE | MPARSE | ALIGN-ADJ | ALIGN-LEX |
| :---: | :---: | :---: | :---: | :---: |
| a. $\operatorname{taq} \mathrm{N} \mathrm{DP}_{\mathrm{Pos}}$ | $*!$ |  | $*$ | $*$ |
| b. N taq $\mathrm{DP}_{\mathrm{Pos}}$ |  |  | $*$ |  |
| c. $\mathrm{N} \mathrm{DP}_{\mathrm{Pos}} t a q$ | $*!$ |  | $*$ |  |
| d. $\emptyset$ |  | $*!$ |  |  |

The c-structure in figure 2 shows the optimal candidate (b) in tableau 4 of the possessor DP pluralized by the distributive pluralizer taq.

DP phrasal compound Phrasal compounds include [A N], where the initial word is a restricting adjective (see (13)). An OT-LFG account of the DP $\Rightarrow$ Det Adj [A N ] pluralized by taq is shown in tableau 5. Although candidate (c) is the winner in tableau 5, candidate (d) does also account for well-formed data. The alternation probably represents another ideolect or dialect, or stylistic variation.

Tableau $5 \mathrm{DP} \Rightarrow \operatorname{Det} \operatorname{Adj}[\mathrm{A} \mathrm{N}]+\operatorname{taq}(\mathrm{PL})$

| taq | Det Adj [A N] | *EDGE | MPARS | AlIGN-ADJ |
| ---: | :---: | :---: | :---: | :---: |
| ALIGN-LEX |  |  |  |  |
| a. $\quad$ taq Det Adj [A N] | $*!$ |  | $*$ | $*$ |
| b. Det taq Adj [A N] |  |  | $*!$ | $*$ |
| c. Det Adj taq [A N] |  |  |  |  |
| d. Det Adj [A taq N] |  |  |  |  |
| e. Det Adj [A N] taq | $*!$ |  | $*$ |  |
| f. $\quad \varnothing$ |  | $*!$ |  |  |

[^90]
$\mathrm{N}^{0} \quad \mathrm{DP}_{1}$
uja'

PLP $\quad \mathrm{DP}_{2}$
$\uparrow=\downarrow \quad \uparrow=\downarrow \quad \uparrow=\downarrow$
$\mathrm{PL}^{0} \quad \mathrm{D}^{0} \quad \mathrm{NP}$
taq lee
$$
\uparrow=\downarrow
$$
$$
\mathrm{N}^{0}
$$
tinamiit



Figure 3 Complement DP and taq:
'Above the trees'

Figure 2 Possessive DP and taq: 'The water of the towns'

Prepositional phrases The phrase structure rules in (46) license the PP in (14) with a DP complement pluralized by the distributive pluralizer taq: ${ }^{20}$

$$
\mathrm{PP} \rightarrow \underset{(\uparrow \text { PCASE })=\downarrow}{\mathrm{P}^{0}}, \quad \begin{gather*}
\mathrm{DP}  \tag{46}\\
\uparrow=\downarrow
\end{gather*}
$$

The PP can pluralize its object complement DP by placing the distributive pluralizer taq immediately after the preposition. The PP in (14) has a DP configured as 'Det $\mathrm{N}^{\prime}$ without an attributive adjective. An OT-LFG account of the $\mathrm{PP} \Rightarrow \mathrm{P}$ Det N whose object complement is pluralized by taq is shown in tableau 6.

The optimal or winning candidate, candidate (b), can also be presented in a constituent structure, which encodes the phrase structure's constituency and its ID rules. The c -structure in figure 3 shows candidate (b) of tableau 6.

In tableau 6, the object complement of a PP can be pluralized by placing the distributive pluralizer taq after the preposition. The DP complement in (16) is pluralized by immediately placing taq after the attributive adjective. So in (16) for example, the pluralizer can follow both the preposition and the attributive adjective

[^91]Tableau 6 PP $\Rightarrow$ P Det $\mathrm{N}+\operatorname{taq}$ (PL)

| $\operatorname{taq} \quad$ P Det N | *EDGE | MPARSE | AlIGN-ADJ | AlIGN-LEX |
| ---: | :---: | :---: | :---: | :---: |
| a. $\quad \operatorname{taq}$ P Det N | $*!$ |  | $*$ | $*$ |
| b. P taq Det N |  |  | $*$ |  |
| c. P Det taq N |  |  | $*!$ | $*$ |
| d. P Det N taq | $*!$ |  | $*$ |  |
| e. $\quad \varnothing$ |  | $*!$ |  |  |

or just the attributive adjective. But the distributive pluralizer taq cannot only follow the preposition if there is an attributive adjective modifying the DP complement's nominal head. The PP in (16) has an object complement with an attributive adjective and is configured as ' $P$ Det Adj N.' An OT-LFG account of the PP $\Rightarrow P \operatorname{Det} \operatorname{Adj} N$ pluralized by the distributive pluralizer taq is shown in tableau 7.

Tableau 7 PP $\Rightarrow$ P Det Adj $\mathrm{N}+$ taq (PL)

| taq $\quad$ P Det Adj N | *EDGE | MPARSE | ALIGN-ADJ | ALIGN-LEX |
| ---: | :---: | :---: | :---: | :---: |
| a. $\quad$ taq P Det Adj N | $*!$ |  | $*$ | $*$ |
| b. $\quad$ P taq Det Adj N |  |  | $*!$ |  |
| c. P Det taq Adj N |  |  | $*!$ | $*$ |
| d. P Det Adj taq N |  |  |  |  |
| e. P Det Adj N taq | $*!$ |  | $*$ |  |
| f. $\quad \varnothing$ |  | $*!$ |  |  |

PP phrasal compound The PP's object complement in (21) whose head is a phrasal compound can be pluralized with taq. The phrasal compound is composed of two nouns [ $\mathrm{N} N$ ], typed and ordered. An OT-LFG account of the PP $\Rightarrow$ P Det [ N N ] pluralized by the distributive pluralizer taq is shown in tableau 8. Nonetheless tableau 8 is somewhat problematic because although candidate (b) is supported empirically, candidate (d) is not (see (21b)).

The object complement of the PP in (22) whose head is a phrasal compound modified by an attributive adjective can also be pluralized by the distributive pluralizer taq. The phrasal compound is composed of two nouns [ N N$]$ modified by a pre-head attributive adjective. An OT-LFG account of the PP $\Rightarrow$ P Det Adj [N N] pluralized by the distributive pluralizer taq is shown in tableau 9 .

Non-verbal predicates To pluralize the non-verbal predicate's subject in (26) with taq, the non-verbal predicate must immediately be followed by the distributive

Tableau $\mathbf{8}$ PP $\Rightarrow$ P Det $[\mathrm{N} N]+\operatorname{taq}(\mathrm{PL})$

| $t a q$ | P Det [N N] | *EDGE | MPARSE | AlIGN-ADJ | ALIGN-LEX |
| ---: | ---: | :---: | :---: | :---: | :---: |
| a. $\quad \operatorname{taq}$ P Det [N N] | $*!$ |  | $*$ | $*$ |  |
| b. $\quad$ P taq Det [N N] |  |  | $*$ |  |  |
| c. | P Det $\operatorname{taq}[\mathrm{N} \mathrm{N]}$ |  |  | $*!$ | $*$ |
| d. | P Det [N taq N] |  |  | $*$ |  |
| e. P Det [N N] taq | $*!$ |  | $*$ |  |  |
| f. $\quad \emptyset$ |  | $*!$ |  |  |  |

Tableau 9 PP $\Rightarrow$ P Det Adj [N N] + taq (PL)

| taq $\quad$ P Det Adj [N N] | *EDGE | Mpar | Align-ADJ | ALIGN-LEX |
| ---: | :---: | :---: | :---: | :---: |
| a. $\quad$ taq P Det Adj [N N] | $*!$ |  | $*!$ | $*$ |
| b. $\quad$ P taq Det Adj [N N] |  |  | $*!$ |  |
| c. $\quad$ P Det taq Adj [N N] |  |  | $*!$ | $*$ |
| d. $\quad$ P Det Adj taq [N N] |  |  |  |  |
| e. $\quad$ P Det Adj [N taq N] |  |  | $*!$ |  |
| f. $\quad$ P Det Adj [N N] taq | $*!$ |  | $*$ |  |
| g. $\quad \emptyset$ |  | $*!$ |  |  |

pluralizer taq. An OT-LFG account of the non-verbal predicate $\Rightarrow$ Pred Det N whose subject is pluralized by the distributive pluralizer taq is shown in tableau 10.

Tableau 10 Non-verbal predicate $\Rightarrow$ Pred Det $\mathrm{N}+\operatorname{taq}$ (PL)

| $t a q$ | Pred Det N | *EDGE | MPARSE | AlIGN-ADJ |
| ---: | :---: | :---: | :---: | :---: |
| ALIGN-LEX |  |  |  |  |
| a. $\quad \operatorname{taq}$ Pred Det N | $*!$ |  | $*$ | $*$ |
| b. Pred taq Det N |  |  | $*$ |  |
| c. Pred Det taq N |  |  | $*!$ | $*$ |
| d. Pred Det N taq | $*!$ |  | $*$ |  |
| e. $\emptyset$ |  | $*!$ |  |  |

If the non-verbal predicate's subject is modified by a pre-head attributive adjective, the subject can be pluralized by the distributive pluralizer taq (27). However the distributive pluralizer taq must follow the pre-head attributive adjective, not the non-verbal predicate. An OT-LFG account of the non-verbal predicate $\Rightarrow$ Pred Det Adj N whose grammatical subject is pluralized by taq is shown in tableau 11.

Tableau 11 Non-verbal predicate $\Rightarrow$ Pred Det Adj N + taq (PL)

| $t a q \quad$ Pred Det Adj N | *Edge | MPAR | ALIGN-ADJ | Align-LEX |
| ---: | :---: | :---: | :---: | :---: |
| a. $\quad$ taq Pred Det Adj N | $*!$ |  | $*$ | $*$ |
| b. $\quad$ Pred taq Det Adj N |  |  | $*!$ |  |
| c. Pred Det taq Adj N |  |  | $*!$ | $*$ |
| d. $\quad$ Pred Det Adj taq N |  |  |  |  |
| e. $\quad$ Pred Det Adj N taq | $*!$ |  | $*$ |  |
| f. $\quad \emptyset$ |  | $*!$ |  |  |

## 3 Conclusion

This paper has investigated the distributive pluralizer taq (PL) of K'ichee' Mayan. While little has been said about the non-bound morpheme taq as a nominal pluralizer in the Mayanist literature, virtually nothing has been said about its use as a distributive (DISTR). Employing a variety of data and linguistic constructions, I demonstrate conventional usage of the distributive pluralizer taq and show the categories of words that it associates with and the positions that it occupies in phrases. I argue that the distributive pluralizer taq is a phrasal particle that (left) adjoins to string-adjacent constituents. This contrasts with the distributive taq (DISTR), which I contend elsewhere is a non-projecting particle that head-adjoins to verbs only. The complex phrasal distribution of the distributive pluralizer taq, which remains unaccountable using phrase-structure rules alone, can be straightforwardly modeled using OT-LFG and a modest number of universal constraints.

## References

Ajpacajá Túm, Pedro Florentino et al. 1996. Diccionario del Idioma K’iche’. La Antigua, Guatemala: PLFM.
Bresnan, Joan. 2000. Optimal syntax. In Optimality Theory: Phonology, Syntax, and Acquisition. Eds. Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer, 334-385. Oxford, New York: Oxford University Press.
Falk, Yehuda N. 2001. Lexical-Functional Grammar: An Introduction to Parallel Cons-traint-Based Syntax. Stanford, Ca.: CSLI Publications.
García Hernández, Abraham and Yac Sam, Santiago. 1980. Diccionario Quiche-Español. Guatemala: Instituto Lingüístico de Verano.
King, Tracy Holloway. 1995. Configuring Topic and Focus in Russian. Stanford, Ca.: CSLI Publications.
Prince, Alan and Paul Smolensky. 2004. Optimality Theory: Constraint Interaction in Generative Grammar. Malden, Ma.: Blackwell.
Toivonen, Ida. 2003. Non-projecting Words: A Case Study of Swedish Verbal Particles. Dordrecht, The Netherlands: Kluwer Academic Publishers.
Willson, Heather. 2004. The Kiche Taq: Distributivity \& Plurality. Unpublished Man.
Willson, Heather. 2005. The Kiche morpheme(s) täq. Handout for Workshop on Kiche Maya at UCLA.

# ISLANDS: A MIXED ANALYSIS 

Yehuda N. Falk<br>The Hebrew University of Jerusalem

Proceedings of the LFG09 Conference
Miriam Butt and Tracy Holloway King (Editors) 2009

## CSLI Publications

http://csli-publications.stanford.edu/


#### Abstract

Analyses of island constraints have categorized them either as pragmatically driven or syntactically driven. It is argued here that most islands have both an extrasyntactic (pragmatics, processing) component and a syntactic component. On the formal side, it is proposed that islands are the result of an off-path constraint referring to a WHPATH feature, rather than the specification of grammatical functions permitted on the path..


## 1 Overview: Syntax vs. Pragmatics

Ever since Ross (1967), one of the topics that has dominated the literature on wh (long-distance dependency or LDD) constructions is that of "islands," regions of sentences which are opaque to wh dependencies.
(1) a. *This book, I know a student who read. [Complex NP]
b. *What did you eat bagels and? [Coordinate Structure]
c. *It was the Mets that I traveled to New York before I watched. [Adjunct]
d. *Star Trek, to watch is important. [Sentential Subject]

While Ross simply listed types of structures which display this opacity, subsequent literature has attempted to discover unifying principles to explain islandhood. Despite these efforts, none has been entirely successful. The purpose of this paper ${ }^{1}$ is to discuss the source of island constraints, and to propose an LFG account of islands which differs somewhat from the standard account in Kaplan and Zaenen (1989).

The literature on islands can basically be split into two groups: those that provide a syntactic explanation and those that provide a pragmatic explanation. For example, consider the ill-formedness of (2).
(2) *Which word processor did you hear the rumor that Bill Gates uses?

In the transformational literature going back to Chomsky (1977), the ungrammaticality of this sentence is attributed to the inability of the wh phrase to move in local steps, due to the intervention of an NP (or DP) node. That is to say, it is seen as a technical structural limitation, purely syntactic in nature. Similarly, other islands are taken to be the result of the failure of local movement, and thus purely syntactic in nature.

Syntactic accounts of the ill-formedness of (2) are not limited to the transformational literature: one finds them also in LFG. The standard LFG

[^92]account of the ill-formedness of (2), due to Kaplan and Zaenen (1989), traces it to constraints on the path between the two grammatical functions borne by the wh element: in English, the path is limited to COMP, XCOMP, and OBL ${ }_{\theta}$, but in this case it would have to go through OBJ. There are important differences between the transformational account and the LFG account. In LFG, islands are based on grammatical functions rather than on structural configurations, and they are the result of a language specific description of the path rather than inherent properties of the wh construction. This latter point means that different languages can have different islands. For example, as noted by Kroeger (1993) while English treats sentential subjects as islands, Tagalog allows wh constructions to cross sentential subjects, ${ }^{2}$ and disallows sentential non-subjects on the path. However, while the transformational view assumes more cross-linguistic uniformity than is justified, the LFG approach is based on the idea that islands are essentially arbitrary and can display infinite variation. There is more crosslinguistic uniformity than is suggested by the LFG approach.

On the other side, it has been observed that the ill-formedness of (2) is not a fact to be viewed in isolation. It is correlated with the inability of the fronted element to be pragmatically prominent in the clause in which it appears. Viewed from this perspective, islands are not arbitrary constraints imposed by the syntax, but rather a consequence of the informational content of sentences.

The exact nature of the pragmatic prominence is a little vague in the literature. One formulation is that the main clause has to be a comment on the fronted element, and in (2) it cannot be so interpreted (Kuno 1976, 1987):
(3) a. In a discussion of different word processors: A: What about TextMangler? B: \#I heard the rumor that Bill Gates uses it.
b. *As for TextMangler, I heard the rumor that Bill Gates uses it.

A slightly different version of this account is that of Erteschik-Shir and Lappin (1979), under which the crucial concept is "dominance," the property of being the item to which the speaker intends to draw the attention of the hearer. This can be tested by making it the topic of further discourse.
(4) a. John said: "I heard the rumor that Bill Gates won't use TextMangler."
b. \#... which is a lie, he will.

It has also been proposed that fronted elements must be associated with clauses that are focal (i.e. present new information), and in (2) the subordinate clause

[^93]is presupposed (Van Valin and LaPolla 1997, Erteschik-Shir 1997). For present purposes, the differences between these are not important, and the tests provided by Kuno and by Erteschik-Shir and Lappin typically give the same result: the crucial point is pragmatic prominence. For the most part, this paper will use tests consistent with Kuno's formulation. ${ }^{3}$

In this study, I will examine the nature of islands in light of the tension between the syntactic and pragmatic approaches. I will conclude that neither is adequate on its own: rather, (some) islands are grammaticalizations (i.e. syntactic expressions) of the pragmatic "aboutness" constraint. It will also transpire that pragmatics is not the only extra-syntactic factor involved in islands. ${ }^{4}$

## 2 Test Case: The Complex NP Constraint

### 2.1 Evidence for Pragmatics

Returning to (2), which is a case of the Complex NP Constraint (CNPC), the idea is that it is ungrammatical because of a pragmatic prominence constraint. The reason that (2) is ill-formed is that you heard the rumor that Bill Gates uses (it) cannot be construed as being about which word processor. If we consider the f-structure of the sentence, we can see why there might be such a constraint.
(5)


[^94]Which word processor is a functional element of the main clause as well as of the subordinate clause. In this respect, it is similar to a raising nominal, which is also a functional element of two distinct clauses.
(6) a. Julius Caesar struck me as honest.
b. I found Julius Caesar (to be) boring.

As noted by Postal (1974), these sentences are odd because there is an implication of direct perception of Julius Caesar by the speaker. ${ }^{5}$ In other words, although Julius Caesar is not a thematic argument of the verb, it functions as an argument (SUBJ in (6a) and OBJ in (6b)), and thus displays argument-like properties. In the present case, which word processor functions in the main clause as a discourse-prominent element (FOCUS). Its inability to be discourse prominent in the main clause naturally causes the sentence to be ill-formed. In the case of long-distance dependency constructions, as in the case of raising, the syntactic designation of an element as having a grammatical function in a higher clause results in semantic/pragmatic properties appropriate to that grammatical function.

The pragmatic analysis of the CNPC is supported by the observation that changing the sentence in order to make a pragmatically prominent reading for the fronted element more plausible improves the wh construction. For example, replacing the definite article with the indefinite article improves the ability of the fronted element to be discourse prominent in the main clause.
a. In a discussion of different word processors:

A: What about TextMangler?
B: ?I heard a rumor that Bill Gates uses it.
b. ?As for TextMangler, I heard a rumor that Bill Gates uses it.

The wh question is also improved:
(8) ?Which word processor did you hear a rumor that Bill Gates uses?

The amelioration when the indefinite article is used is a result of the fact that the definite article carries with it more semantic/pragmatic content (in the form of a claim of familiarity or identifiability) and is therefore less conducive to making something else pragmatically prominent.

In a related vein, the following contrast has often been noted:
a. *Which word processor did you hear the claim that Bill Gates
${ }^{5}$ For related discussion, see Bolinger (1967) and Borkin (1973), among others. On similarities between Raising and wh constructions, see Alsina (2008).
uses?
b. Which word processor did you make the claim that Bill Gates uses?

Here again, the contrast is matched by the possibilities of pragmatic prominence.
(10) a. A: What about TextMangler?

B: \#I heard the claim that Bill Gates uses it.
b. A: What about TextMangler?

B: The newspaper made the claim that Bill Gates uses it.
This contrast appears to be related to the fact that make the claim is synonymous with claim, and thus has less semantic weight to it than a sequence like hear the claim.

Outside of the Complex NP Condition, this correlation between semantic/pragmatic content and the ability of embedded elements to be pragmatically prominent can be seen most clearly in the distinction between bridge verbs and non-bridge verbs.
(11) a. A: What about TextMangler?

B: My friend said that Bill Gates uses it.
b. Which word processor did your friend say that Bill Gates uses?
a. A: What about TextMangler?

B: \#My friend whispered that Bill Gates uses it.
b. *Which word processor did your friend whisper that Bill Gates uses?

In (12) the use of the verb whisper makes the manner of speaking pragmatically prominent, and thus blocks the fronted element from assuming a prominent status. It is this pragmatic status that results in the inability of whisper to be a bridge verb.

An alternative explanation that has been proposed for CNPC effects is that they are the result of processing difficulties (e.g. Sag, Hofmeister, and Snider 2007). Teasing apart the effects of pragmatics and processing is difficult, not in the least because it is plausible that pragmatic infelicity itself impedes processing. The experiment reported by Sag, Hofmeister, and Snider provides mixed results. It also should be noted that the existence of CNPC effects in insitu constructions, discussed in the next section, casts doubt on the basic explanation offered by the processing account, as there is no "filler" that needs to be kept in memory. At the very least, studies of the processing of in-situ wh constructions would be needed.

To conclude, I propose that the cases I have discussed in this section are accounted for by the following pragmatic prominence constraint on longdistance dependency constructions:
(13) If an element $f$ bears a grammaticized discourse function in a nucleus $n$, it must be interpretable as pragmatically prominent in $n$.

### 2.2 The Return of Syntax, Part 1

The previous section establishes a role for pragmatics in island constraints. However, the pragmatic condition is a consequence of the syntactic analysis; an item must have a pragmatic relation with the main clause because it bears a discourse-related grammatical function in that clause. Returning for a moment to the analogy with raising constructions, the "direct perception" property (and analogous properties for other raising verbs) are not present in non-raising versions of the construction, in which the "raising nominal" is not part of the higher clause. The sentences in (14) do not share the oddness of (6) above, precisely because of the difference in the syntax.
(14) a. It struck me that Julius Caesar was honest.
b. I found (that) Julius Caesar was boring.

Thus, syntax retains a role as well.
The syntactic aspect of the pragmatic prominence condition is relevant when considering the analysis of in-situ wh constructions. Consider the following from Iraqi Arabic (Wahba 1991: 255):
a. Mona itmannat tištiri šeno?

Mona hoped to.buy what
'What did Mona hope to buy?'
b. Mona nasat tištiri šeno?

Mona forgot to.buy what
'What did Mona forget to buy?'
(also: ‘Mona forgot what to buy.')
The question is whether 'what' is syntactically a functional element of the main clause in these sentences. The crucial point is that this construction obeys the CNPC.
(16) *Mona furfit il- bint illi ištarat šeno?

Mona knew the- girl who bought what 'What did Mona know the girl who bought?'

The ungrammaticality of this sentence is a result of 'what' being functionally a part of the main clause.

A similar set of facts can be seen in Kikuyu (Bergvall 1983). In Kikuyu, both in-situ and ex-situ constructions are possible:

| a. | Oywefiiria | Goye oiyire mahzirg kenani o:? |
| :--- | :--- | :--- | :--- | :--- |
|  | you.think Ngũgĩ said they.gave crab who |  |
| 'Who do you think Ngũgĩ said they gave a crab to?' |  |  |

As in Iraqi Arabic, the Kikuyu in-situ construction obeys the CNPC:

a. | Mornire mo:do orea otiniris mahao mareko? |
| :--- |
| they.saw person DEM cut flowers which |
| 'Which flowers did they see the person who cut?' |
| b. $\quad$ *Kamau osire mo:do orea oringire o? |
|  |
| Kamau saw person DEM hit |
|  |
| 'Who did Kamao see the person who hit?' who |

Not all in-situ constructions are island sensitive. For example, in Egyptian Arabic (Kenstowicz and Wahba 1983) more conservative speakers obey islands and more progressive speakers do not. Thus, the following is grammatical for some speakers but not for others.
(19) \%Fariid simif išaa@it inn Mona yimkin titgawwiz miin? Fariid heard rumor that Mona might marry who 'Who did Fariid hear a rumor that Mona might marry?'

This means that in the more progressive variety, there is no FOCUS in the f -structure and the question is not a functional-uncertainty construction. In English, in-situ questions are not island-sensitive. Ginzburg and Sag (2000) observe that non-echo in-situ questions are possible in English given a situation where the question has pre-existent discourse accessibility. The following example, an actual utterance, is a CNPC violation:
(20) Talk show host Michael Krasny, addressing a guest who has not said anything yet, about the interim chief of the US Attorney's office: This is a position that is HOW IMPORTANT in your judgment, Rory?

I conclude from the foregoing that in-situ $w h$ constructions sometimes involve multiclausal multifunctionality, as in "wh movement" constructions, and sometimes not. In the examples we have seen here, Iraqi Arabic, Kikuyu
and conservative Egyptian Arabic are multifunctional constructions (i.e. long distance dependencies), while English and progressive Egyptian Arabic are not. I thus disagree with Mycock $(2005,2006)$, who takes the position that there is no functional uncertainty in in-situ constructions. ${ }^{6}$

Obviously, there is much more to be said about in-situ wh constructions. However, consideration of islands provides one piece of the puzzle concerning the analysis of in-situ constructions.

### 2.3 The Return of Syntax, Part 2

In addition to inducing the pragmatic prominence condition, syntax plays a more direct role in some islands. In this section I will discuss the Complex NP Condition facts in English when the wh element is part of a relative clause rather than a complement clause.

Complex NP Condition violations with relative clauses are usually crashingly bad.
(21) a. *Which word processor have you made fun of people who like?
b. *TextMangler is the word processor that I have made fun of people who like.

This unacceptability corresponds, not surprisingly, to an inability to bear pragmatic prominence.
(22) \#As for TextMangler, I have made fun of many people who like it.

As with complement clauses, it is possible to ameliorate the pragmatic effect by reducing the semantic content of the main clause.
(23) a. As for TextMangler, there are many people who like it.
b. As for TextMangler, I know many people who like it.

In Danish (Erteschik-Shir and Lappin 1979), the judgments on CNPC violations match the pragmatic status:
a. Det har jeg drillet mange der har gort.
that have I made.fun.of many who have done
'That, I have made fun of many who have done.'

[^95]b. Det er der mange der kan lide. that are there many who like 'That, there are many who like.'
c. Det kender jeg mange der kan lide. that know I many who like 'That, I know many who like.'

However, in English the facts are not quite so congenial. With there, there is some amelioration, although speakers disagree on how much.
(25) a. ?Which word processor are there many people who like?
b. ?TextMangler is a word processor that there are many people who like.

However, in other cases, there is little or no amelioration.
(26) a. *Which word processor have you found many people who like?
b. *TextMangler is a word processor that I have found many people who like.

This lack of amelioration is unexpected given the pragmatics-based analysis.
I propose to account for this difference between CNPC with complement clauses and CNPC with relative clauses by hypothesizing that in the latter case, the pragmatically based constraint has been grammaticalized, and the syntax disallows LDD constructions into relative clauses. That is to say, the CNPC for relative clauses is, in English, a syntactic constraint which, while it has roots in pragmatics, is no longer dependent on pragmatics.

I thus reject the idea which is implicit in much of the literature that all islands must be accounted for by the same mechanisms. While it is sometimes claimed that the theory is simpler if islands are a unified phenomenon, this is only true if the islands in question display similar properties. What we have seen here is that the CNPC itself is not a uniform phenomenon. While all CNPC effects have their roots in pragmatics, and specifically in the ability of the LDD element to bear pragmatic prominence in the higher clause, the CNPC as it applies to relative clauses has undergone grammaticalization in English (but not in Danish).

## 3 Other Islands

### 3.1 Adjunct Condition

One of the most puzzling constraints on LDD constructions is the inability of an LDD construction to go into an adjunct (in transformational terms, the
inability to extract from an adjunct). The basic facts are clear:
(27) a. *Which astronaut did you get to the moon [before]?
b. *Which book did you cancel your library card [without reading]?
c. $\quad$ Which cubicle did you read the file [in]? (cf. Which cubicle did you put the file in?)

However, closer investigation reveals that things are not so simple. Bouma, Malouf and Sag (2001) question the very existence of the Adjunct Condition, on the basis of examples like:
(28) a. Which student is Roger capable of working [independently of]?
b. Which people can Robin run [nearly as fast as]?
c. Who does Kim write letters [more frequently than]?

Even more puzzling are contrasts such as:
(29) a. *Which book did you go to the library [in order to read]?
b. Which book did you go to the library [to read]?

Pragmatics does not help us with these grammaticality contrasts. In general, elements of adjuncts are difficult but not impossible to interpret as pragmatically prominent.
(30) a. A: What about Neil Armstrong?

B: ?\#I got to the moon before him.
b. A: Tell me about the Olympic running team.

B: ?\#Robin can run nearly as fast as them.

This relative inability of elements of adjuncts to be pragmatically prominent is presumably a consequence of the looser connection that an adjunct has to its clause, as opposed to arguments. (This is reflected, for example, in the representation of adjuncts in the theory of Van Valin and LaPolla 1997, in which adjuncts are in the periphery.) On the other hand, the pragmatic status of elements of adjuncts is not as straightforward as that of elements of complex NPs; the correct context can make the discourses in (30) better. This is why extraction from adjuncts is (at least sometimes) essentially acceptable. However, there is a grammaticality contrast which is not the result of the pragmatic facts.

A closer look reveals that the clearest cases of ungrammatical extraction from adjuncts involve adjuncts which are PPs (following Jackendoff's 1973 analysis of PP, under which such words as before and after are prepositions). I propose that the syntax designates adjunct PPs as islands, but not other
adjuncts. ${ }^{7}$ This singling out of PP adjuncts may be related to the fact that PPs can, and commonly do, function as both arguments and adjuncts with no superficial distinction. The syntactic designation of PP adjuncts as islands, based on the pragmatic status of adjuncts, may be a formal distinction between adjunct and argument PPs.

### 3.2 Sentential Subject Condition

I turn next to the Sentential Subject Condition. Unlike the cases I have dealt with earlier, pragmatic prominence does not appear to be a factor here. Constituents of sentential subjects can be pragmatically prominent. ${ }^{8}$ Note the following:
(31) Concerning that book on nuclear physics, reading it was really an eye opener.
(32) A: What about Star Trek?

B: Watching it can teach you a lot about outer space.
The islandhood of sentential subjects thus cannot be attributed to pragmatics.
A promising alternative has been proposed by Kuno (1973) and modified by Grosu (1981). According to the Kuno-Grosu proposal, the Sentential Subject Condition is a result of a perceptual problem rather than a pragmatic one: extracting from a subject results in an incomplete constituent clause-internally. Under this proposal, there is a syntactic condition, the function of which is to prevent difficult-to-parse structures

That completeness of constituents is implicated is suggested by Kuno's observation that (for some speakers at least) ${ }^{9}$ pied piping a whole PP is better than having a stranded preposition. (Grammaticality markings are those in Kuno.)

[^96](33) a. Of which cars were the hoods damaged in the explosion?
b. *Which cars were the hoods of damaged by the explosion?
a. Which cars did the explosion damage the hoods of?

Kuno and Grosu note that for some speakers the constraint applies to non-subjects as well, as long as they are not final constituents.
(34) a. *Here is something which doing strikes Mary as repulsive.
b. *Here is something which Mary considers doing repulsive.
c. Here is something which Mary has objected to doing (on numerous occasions).

And this is also true (although less sharply) with non-clauses:
(35) a. John handed a picture of Mary to Bill.
b. ??Who did John hand a picture of to Bill?
(36) a. ?*Here is someone who the parents of have been murdered by assassins.
b. ?*Here is someone who John has turned the parents of down.
c. Here is someone who John dislikes the parents of (intensely).

The idea that the Sentential Subject Condition is the result of an incomplete non-final constituent is supported by its behavior in Spanish and Catalan (Alex Alsina, personal communication). In these languages, subjects can either be sentence-initial or sentence-final. Extraction is possible out of sentence-final subjects, but not sentence-initial subjects, as shown in the following Catalan sentences.

> a. Quines ulleres et sorprèn [que porti __]? which glasses you surprise that I.wear
> 'Which glasses does it surprise you that I wear?'
> $\begin{array}{cccllll}\text { b. } & \text { *Quines } & \text { ulleres } & \text { [que } & \text { porti } & \text { et } & \text { sorprèn? } \\ \text { which } & \text { glasses } & \text { that } & \text { I.wear } & & \text { you } & \text { surprise }\end{array}$

Finally, a suggestive piece of evidence, cited by Grosu, is that SOV languages, in which all constituents are non-final, typically lack the Sentential Subject Condition. Grosu suggests that this is because they would have to rule out extraction from everything. An exception to this is Navajo, in which some speakers obey the Sentential Subject Condition and others do not (Platero
1974). ${ }^{10}$


However, as one would expect given the Kuno/Grosu analysis, the constraint is not limited to subjects in Navajo.

```
(39)
?Łééchąą'í nishxash-
dog 2.PERF.3.bite- NMNL PFCTV.1.hear- REL nahał' in.
IMPFC.3.bark
'The dog that I heard bit you is barking.'
```

I conclude that the Sentential Subject Condition (in its various forms) is a syntactic constraint that prevents the creation of sentences that are difficult to parse. Like the (relative-clause) Complex NP Constraint and the Adjunct Condition, which prevent the creation of pragmatically awkward sentences, the Sentential Subject Condition is the grammaticalization of a non-syntactic property.

## 4 Formal Expression of Islands

### 4.1 Preliminaries

Given that (some) islands are enforced syntactically, the question is how this is expressed formally. The standard LFG approach, originally proposed by Kaplan and Zaenen (1989), is to restrict the grammatical functions on the wh path. For example, Kaplan and Zaenen incorporate the Adjunct Condition into the whpaths of Icelandic and English as follows:

$$
\begin{array}{lll}
\text { (40) } & \text { Icelandic: } & (\uparrow \text { TOPIC })=\left(\uparrow(\mathrm{GF}-\mathrm{ADJ})^{*} \mathrm{GF}\right) \\
& \text { English: } & (\uparrow \text { TOPIC })=(\uparrow\{\mathrm{COMP}, \mathrm{XCOMP}\} *(\mathrm{GF}-\mathrm{COMP}))
\end{array}
$$

[^97]In both of these cases, the LDD path is restricted so as not to include the function ADJ, thus resulting (inter alia) in the Adjunct Condition. However, as we have seen, islands such as the Adjunct Condition are more complicated than this, and a simple enumeration of possible grammatical functions on the path is inadequate. The path specified for English, for example, will rule out grammatical cases of extraction from adjuncts as well as ungrammatical ones. In addition, the specification of the path in terms of grammatical functions is inadequate. For example, Dalrymple (2001) specifies the path in English as (41):


Unlike Kaplan and Zaenen, Dalrymple enhances the specification of the path with off-path constraints. The result is a rather complicated specification in which every possible grammatical function on the path is specified individually, and most of them with additional off-path constraints.

I propose to dispense with the inclusion of specific grammatical functions in the path, and instead use off-path conditions exclusively to impose island constraints. The key is an f-structure feature originally proposed by Zaenen (1983), and named [BND] there and [LDD] in Dalrymple (2001) (as in (41)). In this study, I will refer to the feature as [WHPATH].

### 4.2 The Feature [WHPATH]

The [WHPATH] feature was originally introduced to provide an account of phenomena which occur along the wh path (sometimes called morphological signaling; cf. Dalrymple 2001). As generally presented, it ranges over the values ' + ' and '-': an f-structure contained within a wh path is [WHPATH + ], while one which is not is [WHPATH -]. Path phenomena, such as Kikuyu downstep deletion and Irish complementizer selection (Zaenen 1983) are realizations of the feature [WHPATH + ]. However, a closer look reveals that the [WHPATH] feature needs to be more complicated. In some languages, wh path phenomena are sensitive to where in the $w h$ path an f -structure is.

As an example, consider the third-person SUBJ pronoun in Ewe (Collins 1994). If it is not on a wh path, it is $\dot{e}$.
a. É/*Wo fo Kosi.
he hit Kosi
'He hit Kosi.'
b. Kofi gblo be é/*wo fo Kosi.

Kofi said that he hit Kosi
'Kofi said that he hit Kosi.'

If it is the top（or outermost）clause of a wh path，it is wo．${ }^{11}$
（43）Kofi bie «be lamata＊wo／é fo Kosi．』
Kofi asked that why he hit Kosi
＇Kofi asked why he hit Kosi．＇
If it is an embedded clause in the wh path，either form is grammatical．
（44）【 Kofi $\varepsilon$ me gblo［be é／wo fo］』． Kofi FOC I said that he hit
＇It was Kofi that I said that he hit．＇
Of course，below the wh path，only $e$ is grammatical．
（45）【 Kofi $\varepsilon$ me gblo na】 be é／＊wo fo Kosi．
Kofi FOC I said to that he hit Kosi
＇It was Kofi that I told that he hit Kosi．＇
Consider also Duala（Epée 1976）．In fronting constructions，the particle no（glossed here as WHPATH）is inserted after the first verbal element in the clause．This includes topicalization，relativization，and wh questions when the $w h$ is fronted，but not in situ．
（46）a．$\llbracket \mathrm{Nu}$ moto nde Kuo a bodi no kalati kiele．】 that man FOC Kuo he give WHPATH book yesterday ＇It＇s that man Kuo gave a book to yesterday．＇
b．Muto 【na tondi no】 a si tondi mba． woman I love WHPATH she not love me ＇The woman I love doesn＇t love me．＇
c．（i）Kuo a po njika ponda． Kuo he come WH time ＇At what time will Kuo arrive？＇
（ii）【 Njika ponda Kuo a po no？』 WH time Kuo he come WHPATH ＇At what time will Kuo arrive？＇
d．$\quad \mathrm{Na}$ si bi $\llbracket$ nga wenge nde Kuo a ben no I not know if today FOC Kuo he have WHPATH kekise．】
exam
＇I don＇t know if it＇s today that Kuo has an exam．＇

[^98]e．Baise Kuo 【nje a pula no】． ask Kuo what he want WHPATH ＇Ask Kuo what he wants．＇

This marking only appears in the top clause of a wh construction．Other clauses do not have the no marking．
a． $\mathbb{I} \mathrm{Ni}$ kalati nde na ta no na kwalane Kuo that book FOC I PST WHPATH I tell Kuo ［na a－angamene wana］．1］ that he－must bring ＇That＇s the book I told Kuo he should bring．＇
b．【 Njika ponda o mende no pula［na Kuo a WH time you FUT WHPATH want that Kuo he keke wanea wa mo］？】
try bring you it
＇When will you want Kuo to try to bring it to you？＇
c．Buña $\llbracket$ na si ta no $n$－ongele［na Kuo a day I not PST WHPATH I－think that Kuo he po］］，a poi．
come he come
＇The day I was not expecting Kuo to arrive，he did arrive．＇
d．【Kuo nde o kwadi no［na a po wenge］？］ Kuo FOC you say WHPATH that he come today ＇Is it Kuo you said would arrive today？＇

The Ewe and Duala facts require us to distinguish between the top clause of the wh path and the rest of the path．In addition，Duala（along with other languages， such as Kikuyu）shows us that we need to distinguish between specification of the wh path in in－situ constructions from that in ex－situ constructions；in－situ constructions do not trigger path phenomena．

I propose that the［WHPATH］feature has a value consisting of at least one subfeature：$[ \pm \mathrm{TOP}]$ ．I leave open the possibility that there is a second subfeature：$[ \pm$ BOTTOM $],{ }^{12}$ but will not be using it here．I further propose that

[^99]there are two versions of the wh path feature, one for ex-situ constructions and one for in-situ constructions. We can refer to these features as [ $\mathrm{WHPATH}_{\text {High }}$ ] and [ $\mathrm{WHPATH}_{\text {Low }}$ ], with the high and low referring to the position in which the wh element is overtly expressed. The path phenomena of Ewe and Duala can be handled straightforwardly using this feature:

| a. | Ewe 3rd person Subj pronoun: |  |
| :---: | :---: | :---: |
|  | é | $\left(\uparrow\right.$ wHPATH $\left._{\text {High }}\right)=[$ |
|  | wo | $\left(\uparrow\right.$ whPATH $_{\text {High }}$ ) |
| b. |  |  |

I conjecture that wh path phenomena are a kind of aid to parsing wh constructions. When the fronted wh element is encountered, it needs to be kept in memory and an argument function needs to be found for it. Since in-situ constructions are parsed differently - the wh element is in its argument position, which thus does not need to be found, and there is no fronted element to be kept in memory - they do not signal the presence of a wh path overtly. Formally, this is accomplished by the two varieties of the [WHPATH] feature. For the purpose of imposing island constraints, there is (usually) no difference between the two varieties.

## 4.3 [WHPATH] and Islands

Given the [WHPATH] feature, the structure of $w h$ constructions can be uniformly defined with the following structure.

$$
\begin{equation*}
\underset{(\rightarrow \text { WHPATH }=[+\mathrm{T}])}{\mathrm{GF}_{2}} \stackrel{\mathrm{GF}^{*}}{(\rightarrow \text { WHPATH }=[-\mathrm{T}])} \tag{48}
\end{equation*}
$$

This structure, under which the off-path constraints refer only to the WHPATH feature, will be incorporated into the functional uncertainty equations that license wh constructions.

Under such a system, syntactic island constraints are imposed by manipulating the value of the [WHPATH] feature. The phrase structure rules of English, for example, will include functional annotations such as the following:


$$
\text { c. } \quad \begin{array}{ll}
\left.\left.\mathrm{S} \rightarrow \begin{array}{cc}
\text { NP } \\
(\uparrow \text { SUBJ }) & =\downarrow
\end{array} \begin{array}{c}
\text { VP } \\
\\
\\
(\downarrow=\downarrow
\end{array}\right] \text { WHPATH }\right) \neq[-\mathrm{T}]
\end{array}
$$

These equations prevent island f-structures from being the continuation (i.e. non-top) of a wh construction

In-situ and ex-situ constructions generally behave the same with regard to island constraints. However, since the facts of wh path phenomena force us to distinguish between [ $\mathrm{WHPATH}_{\text {High }}$ ] and [ $\mathrm{WHPATH}_{\text {Low }}$ ], the possibility exists of a distinction in island constraints. One language in which such a difference exists is Iraqi Arabic (Ouhalla 1996: 678): the CNPC constraint with regard to relative clauses is only imposed syntactically on the in-situ construction: ${ }^{13}$
(50) a. *Gurfut Mona il- bint illi ištarat šeno ? knew Mona the- girl who bought what 'What did Mona know the girl who bought?'
b. ??Šeno furfut Mona il- bint illi ištarat ? What knew Mona the- girl who bought 'What did Mona know the girl who bought?'

In the framework proposed here, Iraqi Arabic has phrase-structure rules such as the following.

$$
\begin{array}{cc}
\mathrm{NP} \rightarrow \begin{array}{c}
\text { NP } \\
\uparrow=\downarrow \\
\\
\\
\\
\left(\downarrow \text { WHPATH }{ }_{\text {Low }}\right) \neq[-\mathrm{T}]
\end{array} & \begin{array}{c}
\text { CP } \\
\downarrow \in(\uparrow \mathrm{ADJ})
\end{array} \tag{51}
\end{array}
$$

It is instructive that in Iraqi Arabic, it is the in-situ construction that has the syntactic constraint. It shows that syntactic island constraints are not constraints on extraction, or the result of increased processing complexity that results from trying to find a gap in which to place a filler.

## 5 Conclusion

Island effects are primarily the result of non-syntactic properties of constructions: pragmatics, prosody, processing, etc. In some cases, they become grammaticalized as constraints on the wh path, through the [WHPATH] feature.

[^100]
## References

Alsina, Alex (2008) "A Theory of Structure-Sharing: Focusing on LongDistance Dependencies and Parasitic Gaps." in Miriam Butt and Tracy Holloway King, eds., Proceedings of LFG08. On-line.: CSLI Publications. http://csli-publications.stanford.edu/LFG/13/lfg08.html.
Bergvall, Victoria L. (1983) "Wh-Questions and Island Constraints in Kikuyu: A Reanalysis." in Jonathan Kaye, Hilda Koopman, Dominique Sportiche, and André Dugas, eds., Current Approaches to African Linguistics. Vol. 2. Dordrecht: Foris. 245-60.
Bolinger, Dwight (1967) "Apparent Constituents in Surface Structure." Word 23: 47-56.
Borkin, Ann (1973) "To Be or Not To Be." CLS 9: 44-56.
Bouma, Gosse, Robert Malouf and Ivan A. Sag (2001) "Satisfying Constraints on Extraction and Adjunction." Natural Language and Linguistic Theory 19: 1-65.
Chomsky, Noam (1977) "On Wh-Movement." in Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, eds., Formal Syntax. New York: Academic Press. 71-132.
Collins, Chris (1994) "Economy of Derivation and the Generalized Proper Binding Condition." Linguistic Inquiry 25: 45-61.
Dalrymple, Mary (2001) Lexical-Functional Grammar. New York: Academic Press.
Dixon, R. M. W. (1994) Ergativity. Cambridge: Cambridge University Press.
Epée, Roger (1976) "On Some Rules That Are Not Successive Cyclic in Duala." Linguistic Inquiry 7: 193-8.
Erteschik-Shir, Nomi and Shalom Lappin (1979) "Dominance and the Functional Explanation of Island Phenomena." Theoretical Linguistics 6: 41-86.
Erteschik-Shir, Nomi (1997) The Dynamics of Focus Structure. Cambridge: Cambridge University Press.
Falk, Yehuda N. (2006) Subjects and Universal Grammar: An Explanatory Theory. Cambridge: Cambridge University Press.
Ginzburg, Jonathan and Ivan A. Sag (2000) Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives. Stanford, Calif.: CSLI Publications.
Grosu, Alexander (1981) Approaches to Island Phenomena. Amsterdam: North-Holland.
Jackendoff, Ray (1973) "The Base Rules for Prepositional Phrases." in Stephen R. Anderson and Paul Kiparsky, eds., A Festschrift for Morris Halle. New York: Holt Rinehart and Winston. 345-56.
Kaplan, Ronald M. and Annie Zaenen (1989) "Long-Distance Dependencies, Constituent Structure, and Functional Uncertainty." in Mark R. Baltin
and Anthony S. Kroch, eds., Alternative Conceptions of Phrase Structure. Chicago: University of Chicago Press. 17-42.
Kenstowicz, Michael and Wafaa Wahba (1983) "Wh-In-Situ Constructions in Egyptian Arabic." in Jonathan Kaye, Hilda Koopman, Dominique Sportiche, and André Dugas, eds., Current Approaches to African Linguistics. Vol. 2. Dordrecht: Foris. 261-81.
Kroeger, Paul (1993) Phrase Structure and Grammatical Relations in Tagalog. Stanford, Calif.: CSLI Publications.
Kuno, Susumu (1973) "Constraints on Internal Clauses and Sentential Subjects." Linguistic Inquiry 4: 363-85.
Kuno, Susumu (1976) "Subject, Theme, and the Speaker's Empathy-A Reexamination of Relativization Phenomena." in Charles N. Li, ed., Subject and Topic. 417-44.
Kuno, Susumu (1987) Functional Syntax: Anaphora, Discourse and Empathy. Chicago: University of Chicago Press.
Mycock, Louise (2005) "'Wh'-In-Situ in Constituent Questions." in Miriam Butt and Tracy Holloway King, eds., Proceedings of the LFG 05 Conference, University of Bergen. On-line.: CSLI Publications. 313-33. http://cslipublications.stanford.edu/LFG/10/lfg05.html.
Mycock, Louise (2006) A Typology of Constituent Questions: A LexicalFunctional Grammar Analysis of 'Wh'-Questions. Ph.D. dissertation, University of Manchester.
Ouhalla, Jamal (1996) "Remarks on the Binding Properties of Wh-Pronouns." Linguistic Inquiry 27: 676-707.
Platero, Paul R. (1974) "The Navajo Relative Clause." International Journal of Americal Linguistics 40: 202-46.
Postal, Paul M. (1974) On Raising: One Rule of English Grammar and Its Theoretical Implications. Cambridge, Mass.: MIT Press.
Ross, John Robert (1967) Constraints on Variables in Syntax. Ph.D. dissertation, MIT.
Sag, Ivan A., Philip Hofmeister, and Neal Snider (2007) "Processing Complexity in Subjacency Violations: The Complex Noun Phrase Constraint." CLS 34.
Torrego, Esther (1984) "On Inversion in Spanish and Some of Its Effects." Linguistic Inquiry 15: 103-29.
Van Valin, Robert D. and Randy J. LaPolla (1997) Syntax: Structure, Meaning, and Function. Cambridge: Cambridge University Press.
Wahba, Wafaa Abdel-Faheem Batran (1991) "LF Movement in Iraqi Arabic." in C.-T. James Huang and Robert May, eds., Logical Structure and Linguistic Structure: Cross-Linguistic Perspectives. Dordrecht: Kluwer. 253-76.
Zaenen, Annie (1983) "On Syntactic Binding." Linguistic Inquiry 14: 469-504.

# CONSTRAINING DISJUNCTIVE AGREEMENT IN MODERN GREEK 

Maria Flouraki and Despina Kazana SOAS University of Essex<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

Singular disjunctive coordinate nouns in Modern Greek (MG) and in a number of other languages are interesting since the verb can show either singular or plural agreement. This variation is seen as the result of an analysis of the disjunction or either as 'inclusive' or 'exclusive' in truth-conditional semantics. We will support that the variation in verb agreement is contextually motivated and therefore it is immediately related to the contextual interpretation of the disjunctive coordinate phrase either as an 'exclusive' (singular) interpretation or as an 'and-coordinate' (plural) interpretation. Our proposal will predict both verb agreement forms in singular disjunctive nouns taking into account the various discourse conditions, and combining $\lambda$ calculus and the DRT theory, known as $\lambda$-DRT.


## 1 Introduction

Although disjunction was discussed by a number of scholars, it has not been studied as extensively as conjunction. More recently, there has been some interest by a number of linguists including Morgan (1972, 1984, 1985), Peterson (1986), Jennings (1994), Eggert (2002), who have focused on two major issues related to disjunction. The first one concerns the semantics of disjunction. In truth-conditional semantics, logicians agree on the existence of the primary logical connective or represented as $\vee$ and known as the 'inclusive' or. There is disagreement, however, as to whether a secondary logical connective represented as $\underline{V}$ and known as the 'exclusive' or exists. Most logicians resort to propositional logic in order to account for the existence or not of the 'exclusive' or (Jennings, 1966; Barret and Stanner, 1971; Hurford, 1974; Pelletier, 1978).

The second major issue concerns the agreement of the verb with the disjunctive coordinate nouns. Some of the scholars argue that verb agreement with disjunctive coordinate nouns is seen as the result of various speaker strategies, such as the PROXIMITY, the PLURAL WINS or the FIRST CONJUNCT WINS strategies (Peterson, 1986). Others claim that the actual interpretation of the disjunctive coordinate phrase is the determining factor to verb agreement (Morgan, 1985). Thus, when the disjunctive coordinate phrase has an exclusive sense then singular verb agreement is more likely, as in examples (1) and (2):
(1) John or Bill is/???are going to win the race.
(2) John or Bill is/are going to come tonight.
(Morgan, 1985)
In the first example, the singular verb agreement form is preferable since in any world there must be only one winner. The plural verb agreement form is not excluded if we consider that the phrase is asserting something about a group of two

[^101]individuals, meaning that one of the two is going to win the race (Morgan, 1985, 72). The same reasoning seems to follow in example (2). If the interpretation supposes that only one of the two people will visit the speaker tonight, then singular verb agreement is prefered. If, however, the same phrase is interpreted as claiming something about the group of individuals, then the plural verb form is assumed.

When the disjunctive coordinate phrase has an inclusive (and-coordinate) sense then the plural verb agreement is more likely (Morgan, 1985, 73), as in examples (3):
(3) I don't think that John or Bill are/*is going to win the race. ${ }^{1}$
(Morgan, 1985, 72)
In the example above, the most likely interpretation is that both John and Bill will lose, which motivates a plural verb form (Morgan, 1985, 73).

We will argue that MG native speakers adopt a distinction similar to the one presented above and therefore the choice of the verb agreement form follows from the interpretation they assign to the coordinate phrase.

The data analysis is based on a questionnaire that was developed in order to work towards the prevalent verb agreement form in disjunctive structures. The main aim is to consider the different interpretations assigned to the coordinate structures by MG native speakers. The questionnaire consists of 20 declarative and interrogative sentences with singular disjuncts of the same or different gender and of the same or different person.

The questionnaire was issued to 15 native speakers who are all university graduates from different areas in Greece. The participants were asked to make a choice between two possible verb forms and were asked to consider both verb forms whenever they found it appropriate. Their choice depends on acceptability judgements on the basis of what they thought they would say and not on grammaticality judgements. Private conversations followed as to what motivated the choice of a singular or a plural verb agreement form in order to confirm the initial intuitions we had concerning the interpretation of the disjunctive coordinate phrase.

The rest of the paper is organised as follows. In part 2 , we will present a series of MG data that illustrate cases of disjunctive coordinate nouns interpreted as 'and-coordinate' (plural) or as 'exclusive' (singular). In part 3, we will present the proposal developed by Eggert (2002) couched within the DRT framework and, in part 4, we will present an analysis of our own proposal, which follows Dalrymple (2001) and Kokkonidis (2005) within the theory of $\lambda$-DRT.

## 2 The Modern Greek Data

The disjunctive coordinate phrases that will be discussed in MG include conjoined singular nouns. The first group of data presents examples with singular verb agree-

[^102]ment and the second group of data focuses on phrases with plural verb agreement.
In the first group of examples, the majority of native speakers assigned an exclusive interpretation in the coordinate phrase and therefore chose a singular verb agreement form, as seen below:
(4) O Kostas i i Maria tha me parei me to The.SG Kostas.SG or the.SG Maria.SG will me pick-up with the autokinito.
car
'Kostas or Maria will pick me up with the car'
(5) O Kostas i o Giorgos ine xadelfos tis Marias. The.SG Kostas.SG or the.SG Giorgos.SG is.SG cousin of-the Maria 'Kostas or Giorgos is Maria's cousin'
(6) O adelfos su i i adelfi su irthe exthes. The.SG brother.SG your or the.SG sister.SG your arrived.SG yesterday 'Your brother or your sister arrived yesterday'

In the above examples, the informants showed that they seem to conceive the action as performed by the individuals separately and this is why they prefer the exclusive interpretation. Thus, in example (4) the informants assume that only one of the people mentioned is able to pick-up the speaker and not both. Also, in example (5), one of the two people must be Maria's cousin otherwise they are all related to each other, which is not the intended meaning. Similarly, in example (6) the speaker confirms that it must be only one of the two "the brother or the sister" who arrived yesterday and not both.

The exclusive interpretation is confirmed by the presence of modifiers, such as separately, individually, only, etc. that make the exclusive sense even stronger in the sentence. Thus, examples (4), (5) and (6) can be written as follows:
(7) O Kostas i i Maria tha me parei me to The.SG Kostas.SG or the.SG Maria.SG will me pick-up with the autokinito, ohi kai i dio.
car, not and the both
'Kostas or Maria will pick me up with the car, not both'
(8) O Kostas i o Giorgos ine xadelfostis Marias, ohi The.SG Kostas.SG or the.SG Giorgos.SG is.SG cousin of-the Maria, not kai i dio. and the both
'Kostas or Giorgos is Maria's cousin, not both'
(9) O adelfos su i i adelfi su irthe exthes, ohi The.SG brother.SG your or the.SG sister.SG your arrived.SG yesterday, not kai i dio. and the both
'Your brother or your sister arrived yesterday, not both'
Therefore, in the above cases MG native speakers have a tendency towards an exclusive interpretation and that interpretation leads to the choice of a singular verb agreement form.

In the second group of data, native speakers showed a preference towards an and-coordinate interpretation in the disjunctively conjoined nouns and therefore the verb agreement form that they chose was the plural one. Some examples are illustrated below:
(10) O giatros i o odontiatros mporoun na grapsoun The.SG doctor.SG or the.SG dentist.SG can.PL to write farmaka.
prescriptions
'The doctor or dentist can write prescriptions'
(11) Kafes i tsai servirontai dorean meta to geuma. Coffee.SG or tea.SG are-served.PL for-free after the dinner
'Coffee or tea are served for free after the dinner'
(12) I eggios gineka $i$ to pedi hriazonte to embolio The pregnant.SG woman.SG or the child.SG need.PL the immunisation kata tis neas gripis. against the swine flu
'The pregnant woman or child need immunisation against swine flu'
All the examples are perceived by most native speakers as actions carried out or applied to both conjuncts and thus are cases of and-coordinate interpretation. In example (10), both the doctor and the dentist are able to write prescriptions and therefore the plural verb agreement form is chosen by the speakers. In example (11), the emphasis is on the fact that both drinks are offered with the meal and not necessarily on the fact that a choice needs to be made. Finally, in the third example, both conjuncts are perceived as high risk groups that need the immunisation and therefore the interpretation is the and-coordinate.

The above examples can be paraphrased with a collective meaning, as follows:
(13) O giatros ke o odontiatros mporoun na grapsoun The.SG doctor.SG and the.SG dentist.SG can.PL to write farmaka.
prescriptions
‘The doctor and dentist can write prescriptions’
(14) Kafes ke tsai servirontai dorean meta to geuma. Coffee.SG and tea.SG are-served.PL for-free after the dinner
'Coffee and tea are served for free after the dinner'
(15) I eggios gineka ke to pedi hriazonte to embolio The pregnant.SG woman.SG and the child.SG need.PL the immunisation kata tis neas gripis.
against the swine flu
'The pregnant woman and child need immunisation against swine flu'
Such examples, where the sense is clearly as and-coordinate, do not accept easily modifiers which assign distributivity, such as separately or individually, since the meaning conveyed with the presence of distributive modifiers is different from the one perceived by the native speakers and does not necessarily correspond to a true situation. The following examples include modifiers of distributivity:
(16) ??? O giatros i o odontiatros mporoun na grapsoun The.SG doctor.SG or the.SG dentist.SG can.PL to write
farmaka, ohi kai i dio.
prescriptions, not and the both
'The doctor or dentist can write prescriptions, not both'
(17) ??? Kafes i tsai servirontai dorean meta to geuma, ohi kai Coffee.SG or tea.SG are-served.PL for-free after the dinner, not and
ta dio.
the both
'Coffee or tea are served for free after the dinner, not both'
(18) ??? I eggios gineka $i$ to pedi hriazonte to

The pregnant.SG woman.SG or the child.SG need.PL the
embolio kata tis neas gripis, ohi kai i dio.
immunisation against the swine flu, not and the both
'The pregnant woman or child will need immunisation against swine flu, not both'

Example (16) does not really apply to reality since both a doctor and a dentist can write prescriptions. Similarly, in example (17) the statement is not true if one wants to convey the meaning that there are two available drinks after dinner. Finally, example (18) cannot be true considering the fact that both groups are of high risk and need to be immunised.

In both groups of data, however, we argue that there is a preference towards a specific interpretation either 'exclusive' or 'and-coordinate'. Different discourse conditions may result in a different interpretation of the same example, as in the case of "Coffee or tea is/are served after dinner", which results in an 'exclusive' (singular) sense when a choice is made and in an 'and-coordinate' (plural) when the availability is what matters.

To conclude, the MG data showed that native speakers assume two different interpretations for the disjunctive coordinate noun phrases, either as 'and-coordinate' where verb agreement is plural or 'exclusive' where verb agreement is singular.

This forms the central notion for the theory that we will adopt and the analysis that we will propose in the following sections.

## 3 Eggert's theory of disjunction

Eggert (2002) formulates an analysis for coordination, whose main characteristic is that it accounts for agreement phenomena. This analysis has the following advantages:

- accounts for sub-propositional coordination
- takes discourse factors into consideration when determining an argument's semantic number
- represents semantic number of coordinative arguments in a straightforward way

Traditionally the logical connectors and and or are treated as boolean meet and join (or in some cases set-intersection and set-union) and are propositional operators. This treatment, however, faces problems when sub-propositional coordination needs to be accounted for where non-distributive conjunctions are involved. In (19), there is coordination between the propositions Grant ran and Abigail ran but not in (20), where we can not infer the reading Grant met and Abigail met.
(19) Grant and Abigail ran.
(20) Grant and Abigail met.

Similarly with disjunction the interpretation of (21) is not The environment is a depressing choice or the economy is a depressing choice.
(21) The environment or the economy is a depressing choice.

To overcome these problems Eggert (2002) favours a unified analysis for each of the two connectors and and or that allows them to operate over conjunct sets of any type and of any number and not as binary propositional connectors, as has been done in the past.

To achieve this he supports that:

1. or is a subset function that has common characterisitcs with the existentials
2. distributivity and collectivity factors are contextually explained.

He treats and and or as quantifiers based on the observation that both operators get involved in the same types of scopal ambiguities as quantified NPs since the conjunction and and the disjunction or resemble semantically the quantifier all and the existentials, respectively.

As is well known, all and a interact in sentences and produce scope ambiguities, such as the ones below:
b. 'all of his friends are band members'
c. 'there is a band that all of his friends belong to' (Eggert, 2002, 78)

Similar, interaction takes place in example (23a) with and and a, being ambiguous between (23b) and (23c):
(23) a. Grant and Jacob are members of a band.
b. 'Grant and Jacob are band members'
c. 'there is a band that Grant and Jacob are members of'' (Eggert, 2002, 79)

The same takes place with all and or:
(24) a. All of his friends are members of Sunset Valley or Sketchy Afterdeal.
b. 'All of his friends belong to one of the two bands'
c. 'All of his friends belong to Sunset Valley or all of his friends belong to Sketchy Afterdeal'
(Eggert, 2002, 79)
In coordination, however, discourse factors also seem to play an important role in the determination of the semantic number triggered by the conjuncts.

### 3.1 Discourse factors in Agreement

Eggert (2002) shows that apart from the strict semantic factors, there are also discourse factors involved when determining the 'semantic number' of an argument in coordinate phrases. Thus, he claims that "'semantic agreement' should be analysed as a discourse phenomenon which is informed by the semantics" (Eggert, 2002, 92). Characteristic cases of discourse factors mediating in coordinate NPs and determining verb agreement are cases of "appositive conjunction" (Hoeksema, 1988, 36) and "deferred reference" (Nunberg, 1995, 115).

In example (25), which is a case of appositive conjunction, the conjuncts denote the same referent and they trigger singular agreement. A semantic theory can capture that. Appositive conjunction is allowed with descriptions as in example (25), but not with proper nouns (Hoeksema, 1988, 36), as in (26):
(25) Grant's former wife and his mother's present girlfriend was on the Jerry Springer show.
(26) ?? Cassius Clay and Muhammed Ali has/have always been my father's favorite boxer.
(Eggert, 2002, 93)

However, Hoeksema (1988) records cases with proper nouns that can occassionally be conjoined; in these the verb admits plural verb agreement since the phrase introduces two distinct individuals that correspond to the same real-world entity. These cases cannot be captured by semantic theories since intentionality determines verb agreement. A characteristic example is below:
(27) Cassius Clay and Muhammed Ali are the same person.

In Hungarian, there is also evidence of discourse factors in cases where agreement seems to be mediated by some peculiar conception of the referent. For instance cross-linguistically there are cases where the conjunctive subject triggers singular or plural agreement according to how high in the animacy hierarchy the conjuncts are. There is singular agreement triggered when the conjuncts are nonhuman in (28) and plural when they are human in (29).
(28) Csilla es Gabor itt van-nak.
csilla and gabor here be-3.pl
'Csilla and Gabor are here'
(29) A konyv-ed es a kalap-od itt van. the book-your and the hat-your here be 'your book and your hat are here' (Hungarian)

Also, in cases of "deferred reference" agreement is only with the intended referent. Thus, demonstratives or verbs show singular or plural agreement depending on the intended referent, not the demonstratum:
(30) [In a restaurant:]

That/*those french fries is/*are getting impatient.
A similar case is found in (31), where there is agreement with the intended referent: wedding and not with the subject conjunct: Grant and Gertrude.
(31) There have been lots of weddings recently. Grant and Gertrude makes the fifth one this year.

Therefore, Eggert (2002) shows that such cases cannot be handled by purely semantic theories and thus a discourse based account should be considered.

### 3.2 An analysis for disjunction

To capture the discourse and semantic factors in agreement, Eggert (2002) proposes an analysis of coordinate phrases based on Discourse Representation Theory (Kamp and Reyle, 1993). The main advantage of DRT is that it uses a discourse structure that is mapped off of grammatical structure. Such a discourse structure allows the effective incorporation of any discourse-pragmatic features into the meaning of sentences when the latter are uttered in a particular discourse context. DRT
analysis is advantageous since it takes both discourse and semantic factors into consideration when determining the semantic number of an argument. Thus, considering a given argument, its semantic number is determined by whatever discourse conditions apply to the discourse referent that corresponds to the argument.

Eggert (2002) develops a uniform theory for and and or in order to capture the wide range of data and also introduce a straightforward definition for plurality. He treats and as a type-specific operator meaning that it is a generic operator that is identified with "whatever operator is defined for the domain of the conjoined terms" (Eggert, 2002, 92) and not with meet per se, the Boolean approach proposed by Keenan and Faltz (1985). ${ }^{2}$

In his analysis of or, he clearly shows that or is not and's dual but rather it must be treated as a subset function, which means a function that moves from sets to subsets. In other words, the subset function picks up either one (i.e. resulting in SG agreement) or both (i.e. resulting in PL agreement) members of the set, being polysemous between two meanings. Thus, he completely rejects the possibility of analysing or as the Boolean or clausal, and therefore having the same meaning as distibutive and.

Formally the subset function is "a function $f$ such that for any non-empty set $A$, $f(A)$ is a non-empty subset of $A$.

$$
\begin{equation*}
\mathbf{S U B}=\{\mathbf{f}:(\forall \mathbf{X}: \mathbf{X} \neq \varnothing) \mathbf{f}(\mathbf{X}) \subset \mathbf{X} \wedge \mathbf{f}(\mathbf{X}) \neq \varnothing\} \quad(\text { Eggert, 2002, 110) } \tag{32}
\end{equation*}
$$

Eggert (2002) presents a proposal for verb agreement with disjunctive coordinate nouns, formulated within DRT where agreement phenomena can be incorporated, and following a similar analysis to the DRT analysis of existentials. An important difference, however, between disjunction and existentials is that disjunction does not introduce a discourse referent as the existentials do. This becomes clear from examples like the following, where the continuation of the discourse does not assume that a discourse referent is introduced, i.e. she.
(33) Gertrude or Abigail is singing tonight.
?She might dance too.
He rather supports that the disjuncts in (33) form a set and or selects a member of that set i.e. either Gertrude or Abigail. There is no introduction, however, for a referent for Gertude or Abigail in the DRS but rather the predicate sing combines directly with the function on the set that consists of the discourse referents standing for Gertrude and Abigail. The denotation of this disjunction is the union of the subset of the disjuncts.

[^103]| $x, y, f$ |
| :---: |
| gertrude $(x)$ |
| abigail $(y)$ |
| $\operatorname{SU} B(f)$ |
| $\operatorname{sing}(x, f \cup\{x, y\})$ |

In the following example there are two interpretations: the 'exclusive' or interpretation (e.g. 'Grant is taller than Abigail or Grant is taller than Gertrude' and the 'and-coordinate' or interpretation, which states that for "all choice functions $f$ Grant is taller than $f\{$ Gertrude,Abigail\}" (Eggert, 2002, 111).
(35) Grant is taller than Abigail or Gertrude. (Eggert, 2002, 110)

Both of these interepretations have the following DRS:
(36)

| $x, y, z, f$ |
| :---: |
| $\operatorname{grant}(x)$ |
| abigail $(y)$ |
| gertrude $(z)$ |
| $S U B(f)$ |
| taller $(x, f \cup\{y, z\})$ |

(Eggert, 2002, 135)
Depending on the assignment of $f(y, z)$, the DRS can be interepreted either way. The first interpretation follows from:

$$
\begin{align*}
& \mathrm{f}(\mathrm{y}, \mathrm{z})=\{\{\text { abigail }\}\}  \tag{37}\\
& \mathrm{f}(\mathrm{y}, \mathrm{z})=\{\{\text { gertrude }\}\}
\end{align*}
$$

which are both subsets of $f(y, z)=\{\{$ abigail $\}$, \{gertrude $\}\}$. The second interpretation follows from $\mathrm{f}(\mathrm{y}, \mathrm{z})=\{\{$ abigail $\}$, $\{$ gertrude $\}\}$, which is a possibility since $\{\{$ abigail $\},\{$ gertrude $\}\} \subseteq\{\{$ abigail $\},\{$ gertrude $\}\}$.

The difference in interpretations is achieved by adopting the partition analysis of plurals (Schwarzschild, 1996), which claims that the collective and distributive semantic difference in sentences comes from a contextually determined variable, a partition on the universe of discourse (Schwarzschild, 1996). In the example above, the variable partitions abigail and gertrude into one cell in which case we get the collective reading in the second case, whereas in the first reading the variable partitions abigail and gertrude into two separate cells.

## 4 Analysing Verb Agreement in Disjunctive Coordinate Nouns

We formalise the above concepts in lambda-DRT and Glue Semantics, following Dalrymple (2001) and Kokkonidis (2005). The different $\lambda$-DRT expressions which correspond to the meaning parts will be combined together using the glue language.

We take the following simple example which shows either singular or plural verb agreement:
(38) Jane or Mary is/are singing.

We treat or as a subset function, which ranges over the set of disjuncts. Or is represented with the complex type $e \rightarrow(e \rightarrow e)$ since it functions over individuals and has the following lexical entry with the relevant DRS.

```
( \(\uparrow\) CONJ \()=\) 'or'
or: \(\mathrm{e}_{\uparrow s u b j} \multimap\left[\mathrm{e}_{\uparrow s u b j} \multimap \mathrm{e}_{\uparrow s u b j}\right]\)
```

$$
\lambda x . \lambda y \begin{array}{|c}
\hline \begin{array}{c}
f, x, y \\
S U B(f) \\
f \cup\{x, y\}
\end{array} \\
\hline
\end{array} \cup y \cup x
$$

The important remark in the DRS side is that there is no introduction of a new referent introduced by $f\{x, y\}$ but only the subset function is introduced. The glue side states that two semantic resources are required of type e, which are members of the set and they are represented by the $(\uparrow \in)_{\sigma<e>}$ symbol which corresponds to each argument. Once these are found, they are consumed and therefore we can deduce the semantic resource of the whole coordinate phrase, represented as $\uparrow_{\sigma<e>}$.

The lexical entries for the proper names, which are also of type $e$, and the verbal one-place predicate, which is of type $\mathrm{e} \rightarrow t$, are the following:
(40) $\quad(\uparrow$ PRED $)=$ 'jane'
jane: $\mathrm{e}_{\uparrow \text { subj }}$

$\lambda x .$|  |
| :---: |
| jane $(x)$ |

(41) ( $\uparrow$ PRED) $=$ 'mary’
mary: $\mathrm{e}_{\uparrow s u b j}$
$\lambda y$.
(42) ( $\uparrow$ PRED $)=$ 'sing' sing: $\mathrm{e}_{(\uparrow S U B J)} \multimap \mathrm{t}_{\uparrow}$

$$
\lambda x^{\prime} \cdot \square
$$

If we do the union of or with Jane we get:
(43)

$$
\begin{aligned}
& \lambda y . \begin{array}{c}
f, x \\
\begin{array}{c}
S U B(f) \\
f \cup\{x, y\}
\end{array} \\
: \mathrm{e}_{\uparrow \text { label }} \multimap \mathrm{e}_{\uparrow \text { label }}
\end{array}
\end{aligned}
$$

If we do the union of or Jane with Mary we get:
(44)

$$
\begin{aligned}
& \begin{array}{|c|}
\hline \frac{f, x, y}{} \begin{array}{l}
S U B(f) \\
f \cup\{x, y\} \\
\hline
\end{array} \cup \frac{y}{\operatorname{mary}(y)} \\
\hline \text { jane }(x) \\
\hline \mathrm{e}_{\uparrow \text { label }}
\end{array}
\end{aligned}
$$

If we do the DRS unions, we get:
(45)

| $f, x, y$ |
| :---: |
| $S U B(f)$ |
| $f \cup\{x, y\}$ |
| jane $(x)$ |
| mary $(y)$ |

: $\mathrm{e}_{\uparrow \text { label }}$
Next we need to apply the verb sing to the disjunctive coordinate phrase to get the desired result:
(46)

| $f, x, y$ |
| :---: |
| $S U B(f)$ |
| $f \cup\{x, y\}$ |
| $j a n e(x)$ |
| $\operatorname{mary}(y)$ |
| $\operatorname{sing}(f \cup\{x, y\})$ |

:t

The above DRS represents the whole sentence which is of type $t$.
This simple approach, which uses $\lambda$-DRT, accounts for a simple disjunctive phrase which results in singular or plural verb agreement and which treats individuals of type e necessarily. Further work is required to account for disjunctive phrases with more than two disjuncts and include other types.

## 5 Conclusion

The current paper has presented an analysis of verb agreement in disjunctive coordinate nouns. The analysis has focused on the hypothesis that verb agreement in such phrases depends strictly on interpretation factors and this is why agreement is relatively unpredictable. Our field work shows that MG native speakers interpret disjunctive coordinate phrases in two ways, as 'exclusive', admitting singular verb agreement, and as 'and-coordinate', admitting plural agreement. The same hypothesis applies to other languages too. A similar assumption is found in Eggert (2002), who supports that there are discourse factors in determining the 'semantic number' of an argument in a coordinate phrase. Thus, following Eggert (2002), we also assume that or is a subset function which means that it is a function from sets to subsets, and we propose an analysis on $\lambda$-DRT, a discourse-based framework. Although the current analysis captures disjunctive coordinate nouns of type $e$, there is room for more research in order to extend the analysis to other types (i.e. nouns with in/definite determiners or predicates) and also to disjunctive phrases with more than two disjuncts.

## References

Barret, Robert B. and Alfred J. Stanner. 1971. The Myth of the Exclusive 'Or'. Mind pages 116-121.

Dalrymple, Mary. 2001. Lexical Functional Grammar. Academic Press.
Eggert, Randall. 2002. Disconcordance: The syntax, semantics and pragmatics of or-agreement. Ph.D. thesis, University of Chicago, Chicago: Illinois.

Hoeksema, Jack. 1988. The semantics of non-Boolean "AND". Journal of Semantics 6(1):19-40.

Hurford, James R. 1974. Exclusive or inclusive disjunction. Foundations of Language 11:409-411.

Jennings, Ray E. 1966. Or. Analysis 26:181-184.
Jennings, Ray E. 1994. The Genealogy of Disjunction. Oxford University Press.
Kamp, Hans and Uwe Reyle. 1993. From Dircourse to Logic, vol. 1, 2. Boston: Kluwer.

Keenan, Edward L. and Leonard M. Faltz. 1985. Boolean Semantics for Natural Language. Boston: D. Reidel.

Kokkonidis, Miltiadis. 2005. Why glue a donkey to an f-structure when you can constraint it and bind it instead? In M. Butt and T. H. King, eds., Proceedings of the LFG05 Conference, pages 238-252. CSLI Publications.

Morgan, Jerry. 1972. Verb agreement as a rule of English. In J. L. Paul Peranteau and G. Phares, eds., Papers from the Eighth Regional Meeting of the Chicago Linguistic Society, Chicago Linguistic Society, pages 278-286. Chicago: Chicago Linguistic Society.

Morgan, Jerry. 1984. Some problems of agreement in English and Albanian. In Proceeding of the Berkeley Linguistic Society, vol. 10.

Morgan, Jerry. 1985. Some problems of determination in English number agreement. In B. B. Gloria Alvarez and T. McCoy, eds., Proceeding of the First Eastern States Conference on Linguistics. Columbus: Ohio State University.

Nunberg, Geoffrey. 1995. Transfer of meaning. Journal of Semantics (12):109132.

Pelletier, Francis J. 1978. Or. Theoretical Linguistics pages 61-74.
Peterson, Peter. 1986. Establishing verb agreement with disjunctively conjoined subjects: Strategies vs. Principles. Australian Journal of Linguistics (6):231249.

Schwarzschild, Roger. 1996. Pluralities. Dordrecht: Kluwer.

# PROBLEMS OF GERMAN VP COORDINATION 

Martin Forst<br>Christian Rohrer<br>Natural Language Engineering Institut für Maschinelle Sprachverarbeitung<br>Powerset/Microsoft<br>Universität Stuttgart

## Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper, we present the implementation in the German ParGram LFG of verb phrase (VP) coordinations involving conjunction reduction and/or right node raising. We show how the computationally expensive approach proposed by Maxwell \& Manning (1996) can be adopted for VP coordinations in a computationally efficient way so that many of these coordinations, which previously did not receive a correct analysis, are now analyzed soundly. We also show that the new rules obviate the need for a recursive right-branching VP rule and make it possible to define a flat VP rule instead. This is desirable for a number of reasons, including the definition of both hard and soft constraints on constituent order in the VP domain (needed in particular for generation).


## 1 Introduction

Traditionally, generative grammarians working on German assume a binary right-branching verb phrase (VP) that ends with a binary left-branching verb complex (VC) (Haider 2003, Berman 2003, Dipper 2003). While this approach can account for most of the data including coordination data, a non-negligible portion of coordinations involving verbs with arguments or adjuncts cannot be accounted for in this view -or only by recurring to notions like "left deletion".

From a grammar development perspective, this is problematic. Mechanisms like "left deletion" being beyond the scope of current LFG parsing systems (and any other practical parsing system for that matter), coverage is of course negatively affected when certain phenomena are systematically not covered and perhaps most problematically- a broad-coverage grammar is likely to analyze sentences that contain those phenomena in grossly erroneous ways, due to unrelated, potentially dispreferred, rules that can cover the sentence.

One appealing and formally well-defined approach to the non-constituent coordination phenomena of right node raising and conjunction reduction is the proposal by Maxwell \& Manning (1996). However, although their theory is formally well-defined, the ParGram grammars' efficiency would be affected very adversely if it were implemented in the LFG grammar development and processing tool XLE (Crouch et al. 2009) on a general level. We therefore propose a limited implementation of their approach, targeted at capturing certain VP coordinations that involve non-constituent coordination.

In order to demonstrate that these kinds of non-constituent coordination are actually not only of theoretical interest, but that their treatment has very practical consequences for the coverage and accuracy of a computational grammar, we also provide figures and examples from corpus data that we obtained from the TIGER Corpus (Brants et al. 2002, 2003) using the treebank query tool TIGERSearch (König et al. 2003, Lezius 2002).

The remainder of this paper is organized as follows: In Section 2, we will present the various types of VP coordination that our proposal is able to capture and exemplify them with examples from the TIGER Corpus. In Section 3, we will discuss the various possible analyses, including the right-branching recursive VP rule formerly implemented in the German ParGram LFG, the potential abandonment of the verb complex, and the use of special coordination rules inspired by the Maxwell \& Manning (1996) approach. Section 4 then discusses why our new treatment of VP coordinations makes it possible to define a flat rather than a recursive VP rule and how such a flat VP rule benefits other aspects of grammar development. Finally, Section 5 concludes.

## 2 Various Types of German VP Coordinations

Coordinations of VPs that can be covered by the "traditional" account are illustrated in examples (1) and (2). What characterizes them is that two VPs, participial ones in (1) and infinitival ones in (2), are
coordinated, but an additional constituent occurring to the left of the coordination is an argument of both conjuncts and hence somehow has to be distributed into them.
(1) Nach Angaben der Polizei hatte er eine Stewardess mit einem Messer According to statements of the police had he a stewardess with a knife bedroht und politisches Asyl verlangt. threatened and political asylum demanded.
'According to the police, he had threatened a stewardess with a knife and demanded political asylum.' (TIGER sentence \# 13672)
(2) Insofern soll meine Präsenz vor allem in der Öffentlichkeit wirken und Insofar shall my presence above all in the public have an effect and die Einsatzleitung bestärken. the operation controllers encourage.
'Insofar, my presence is primarily supposed to have an effect in public and to encourage the operation controllers.' (TIGER sentence \# 4878)

Coordinations of this kind are rather frequent in German newspaper corpora. Among the 48,470 TIGER Corpus sentences that are not used for evaluation (sentences 8,001 through 10,000 are commonly set aside for that purpose), 360 sentences match the TIGERSearch (König et al. 2003, Lezius 2002) query below, which describes this sort of VP coordination according to the TIGER annotation scheme (Brants \& Hansen 2002, Brants et al. 2002, 2003). Note that, overall, there are only 1,503 sentences that contain CVP (co-ordinated verb phrase) constituents, so that the 360 matching sentences are indeed a considerable portion of these.

```
#s >OC #vp:[cat = "CVP"] &
#s >HD #shd &
#vp >CJ #vp1:[cat = "VP"] &
#vp >CJ #vp2:[cat = "VP"] &
#vp1 .* #vp2 &
#vp1 >HD [T & pos != "VVIZU"] & // head of #vp1 is a terminal, but not a `zu' inf.
#vp2 >HD [T & pos != "VVIZU"] & // head of #vp2 is a terminal, but not a `zu' inf.
tokenarity(#vp1, 2, 200) & // #vp1 spans between 2 and 200 terminals
tokenarity(#vp2, 2, 200) & // #vp2 spans between 2 and 200 terminals
#vp2 >@r #r & // #r is the rightmost terminal of #vp2
#r .* #shd // #r precedes #shd
```

Another type of VP coordination can be found in examples (3) and (4). Characteristic of these coordinations is a verbal element at the right edge of the coordination that has to be distributed over the conjuncts to make the first conjunct "complete". In (3), this is the auxiliary infinitive zu haben; in (4), it is the modal $m u \beta$. Note that the two distribution phenomena can co-occur, i.e., in addition to verbal elements at the right edge, there can be argument constituents to the left of the coordination that have to be distributed over the conjuncts to make the second conjunct complete. This is the case of the NP constituent er in (4).
(3) Die Regierung [...] wirft ihm vor, [...] $z u$ dem Massenmord aufgerufen und The goverment [...] accuses him [...] for the mass murder called and Massaker organisiert zu haben. massacres organized to have.
'The government [...] accuses him of having called for the mass murder [...] and of having organized massacres.' (TIGER sentence \# 44943)
(4) Doch Lafontaine weiß, daß er Schröder einbinden und seine Talente nutzen mu $\beta$. But Lafontaine knows that he Schröder involve and his talents use must. 'But Lafontaine knows that he must involve Schröder and capitalize on his talents.' (TIGER sentence \# 30327)

While these coordinations are less frequent than the first type, they are not negligible in number in newspaper corpora. 171 sentences in the non-evaluation part of the TIGER Corpus match the TIGERSearch query below, which describes coordinated VPs with right-node-raised material.

```
#s >OC #vp:[cat = "CVP"] &
#s >HD #shd &
#vp >CJ #vp1:[cat = "VP"] &
#vp >CJ #vp2:[cat = "VP"] &
#vp1 > #n:[] &
#vp1 .* #vp2 & // #vp1 precedes #vp2
#vp1 >HD [T & pos != "VVIZU"] & // head of #vpl is a terminal, but not a `zu' inf.
#vp2 >HD [T & pos != "VVIZU"] & // head of #vp2 is a terminal, but not a 'zu' inf.
tokenarity(#vp1, 2, 200) & // #vp1 spans between 2 and 200 terminals
tokenarity(#vp2, 2, 200) & // #vp2 spans between 2 and 200 terminals
((#s >SB #ssb & // Either #s has a subject, which we'll call #sb, and
    #shd .* #ssb) | // #shd precedes that subject
    (#n >~ #vp2 & // or there is a secondary edge from #n to #vp2
    #n !>~}\textrm{CP}#\textrm{Vp}2)) // but the label of that secondary edge is not CP
```

Finally, there is a substantial number of coordinations in German newspaper corpora that involve conjunction reduction or argument cluster coordination. Examples (5) and (6) illustrate this non-constituent coordination phenomenon, the latter being an instance where the distribution of an argument (es) from the left, the distribution of verbal elements (geben wird) from the right (or right node raising), and argument/adjunct cluster coordination all interact.
(5) In den ersten vier Monaten stiegen die Exporte um 8,4 und die Importe um sechs In the first four months rose the exports by 8.4 and the imports by six Prozent.
percent.
'In the first four months, the exports rose by 8.4 percent, and the imports, by six.' (TIGER sentence \# 39150)
(6) [..] daß es im Kampf gegen die PKK keine Kompromisse und für die [...] that it in the struggle against the PKK no compromises and for the kurdischen Bürgerrechtler keine Zugeständnisse geben wird. Kurdish civil right activists no concessions give will.
'[...] that there will be no compromises in the struggle against the PKK and no concessions for the Kurdish civil rights activists.' (TIGER sentence \# 24221)

To estimate the frequency of argument/adjunct cluster coordination, we formulated the TIGERSearch query below. It matches 161 sentences in the non-evaluation part of the TIGER Corpus.

```
#cs:[cat = ("CS"|"CVP")] & // #cs is a coord. clause or a coord. VP
#CS >CJ #s1:[cat = ("S"|"VP")] & // clause or VP #s1 functions as a conjunct of #cs
#cs >CJ #s2:[cat = ("S"|"VP")] & // clause or VP #s2 functions as a conjunct of #cs
#s1 >HD #s1hd:[pos = ("VVFIN"|"VVIMP"|"VVINF"|"VVIZU"|"VVPP")] &
    // main verb #slhd functions as the head of #s1
#s1hd > ND #s2 & // A secondary edge indicates that #s1hd also
    // functions as the head of #s2
((#s1 .* #s2 & // Either #s1 precedes #s2 and
    #s1 >@r #slr & // #slr is the rightmost terminal of #s1 and
    #slr !> ~SVP #s2 & // #slr is not distributed into #s2 as a verb particle and
    #slr !> ~HD #s2 & // #slr is not distributed into #s as a head and
    #s1 > #slfirst & // #slfirst is some daughter of #s1 and
    #slfirst .* #slhd & // #slfirst precedes #slhd and
    #slfirst >~ #s2) | // #s1first is distributed into #s2 via a secondary edge
    (#s2 .* #s1 & // or #s2 precedes #s1 and
    #s1 >@r #s1hd)) // #slhd is the rightmost terminal of #s1
```


## 3 Possible Analyses

In this section, we discuss several possible analyses of the VP coordination phenomena introduced above. First, we briefly present the analysis implemented in previous versions of the German ParGram LFG. Subsequently, we present two alternatives.

### 3.1 Recursive right-branching VP and left-branching VC (verb complex)

The coordination facts exemplified in examples (1) and (2) are traditionally used as the strongest argument for assuming a recursive right-branching VP rule. For a long time, the German ParGram LFG therefore had a recursive VPx rule, which produced c-structures as the ones shown in Figures 1 and 2. ${ }^{1}$

The advantage of assuming such a recursive right-branching VPx rule is that the coordinations in (1) and (2) can be treated as same-category coordination under this analysis. Its disadvantage, however, is that it cannot capture coordinations that involve right-node-raised verbal elements because the verbal elements that have to be distributed into the first conjunct end up too low in the tree for this to happen. Similarly, conjunction reduction (or argument/adjunct cluster coordination) cannot be captured by the recursive right-branching VPx rule. Given that the latter two phenomena are almost as frequent as the distribution of arguments/adjuncts from the left, we consider this disadvantage serious enough to search for alternative analyses. This is particularly true since, with the "traditional" VP implementation, these sentences are either not associated with an analysis spanning the entire sentence or only analyzed erroneously. The latter is actually worse sometimes because, as a consequence of a bad analysis of the

[^104]

Figure 1: C-structure with recursive VPx corresponding to (1)
According to the police, he had threatened a stewardess with a knife and demanded political asylum.
coordinated verbs, large parts of the remainder of the sentence are typically analyzed erroneously, too, and the resulting analysis contains blatantly wrong predicate-argument triples.

### 3.2 Abandoning the verb complex (VC)

One possible way to resolve the "conflict" between argument/adjunct distribution from the left as illustrated in (1), (2), and (4) and verbal element distribution from the right as illustrated in (3) and (4) is to do away with the distinction between VPs and verbal complexes and simply assume a right-branching recursion that allows for the introduction of an argument/adjunct and a left-branching recursion for the introduction of an auxiliary or modal. The VP rule would then look as follows, where the VPconst macro expands to any major category that can appear in a VP.

```
VP --> { @(VPconst) VP
    | VP Vaux
    | ... }.
```

This solution, however, encounters two important problems: From a theoretical point of view, it is problematic that, with this set of VP rule variants, there is no way at all to constrain the placement of arguments such as predicative phrases, which, unlike most other kinds of arguments in German, clearly have to be adjacent to the verb. From the point of view of grammar engineering, the problem with this solution is that, as soon as a clause has more than one element in clause-final position, the rules can produce several c-structures for identical f-structures, which is undesirable for efficiency reasons and with respect to ambiguity management.

Finally, it should be noted that this solution only addresses the problem of VP coordinations involving right node raising, but not conjunction reduction. Apart from being problematic for independent reasons, it would hence only be a partial solution anyway.

### 3.3 Special coordination rules in the verb complex

An effective and efficient way to capture the distribution of arguments/adjuncts from the left, right node raising and conjunction reduction consists in allowing for the verb complex to expand to a coordination of VPs. In other words, a special rule for a coordinated construction is introduced, and this rule in fact produces c-structures and f-structures very similar to the ones proposed by Maxwell \& Manning (1996). The main difference is that instead of allowing the coordination of partial constituents "across the board", we do this very selectively, namely for VPs, so that the effects on efficiency remain reasonable. ${ }^{2}$

In the German ParGram LFG, the relevant rule part looks as follows:

```
VC --> { @(VP-COORD)
    | VC Vaux
    | ... }.
```

The grammar has always contained a macro for the coordination of VPs named VP-COORD, so this part is nothing new. The macro was introduced to account for idiosyncrasies of VP coordination whose treatment was not possible with the general macro for same-category coordination. It used to be called by the VP rule and covered sentences like (3) and (4) without undesirable vacuous ambiguities. By simply calling it in the VC rule rather than the VP rule, we make it possible for verbal elements to be distributed from the right over both conjuncts of a VP coordination that would traditionally be analyzed as involving "left deletion" or right node raising.

Although it may seem somewhat counterintuitive to attach a VP coordination under a VC, the rule gives rise to c-structures which, apart from this little oddity, look intuitive and from which linguistically sound f-structures are projected. Figures 3 and 4 show the c- and f-structures of (1) and (2) under the revised analysis, and Figures 5 and 6 illustrate how (3) and (4) are accounted for by the rule in (8).

[^105]
"Insofern soll meine Präsenz vor allem in der Öffentlichkeit wirken und die Einsatzleitung bestärken


Figure 2: C-structure with recursive VPx corresponding to (2) and f-structure projected from it. Insofar, my presence is primarily supposed to have an effect in public and to encourage the operation controllers.

"Nach Angaben der Polizei hatte er eine Stewardess mit einem Messer bedroht und politisches Asyl verlang


Figure 3: C-structure with VP-COORD under VC corresponding to (1) and f-structure projected from it According to the police, he had threatened a stewardess with a knife and demanded political asylum.

"Insofern soll meine Präsenz vor allem in der Öffentlichkeit wirken und die Einsatzleitung bestärken


Figure 4: C-structure with VP-COORD under VC corresponding to (2) and f-structure projected from it Insofar, my presence is primarily supposed to have an effect in public and to encourage the operation controllers.

"die Regierung wirft ihm vor, zu dem Massaker aufgerufen und Massaker organisiert zu haber


Figure 5: C- and f-structure corresponding to (3)
The government accuses him of having called for the mass murder and of having organized massacres.

"Doch Lafontaine weiß, daß er Schröder einbinden und seine Talente nutzen muß


Figure 6: C- and f-structure corresponding to (4)
But Lafontaine knows that he must involve Schröder and capitalize on his talents.

As we have just seen, the rule disjunct that expands a VC as a coordination of VPs makes it possible to correctly analyze VP coordinations with distribution of arguments/adjuncts from the left as well as those with distribution of verbal elements from the right. What the rule disjuncts still fails to analyze, however, is the phenomenon of conjunction reduction or argument/adjunct cluster coordination, which is exemplified in (5) and (6).

Based on the observation that these coordinated argument/adjunct clusters consist of the same categories as the non-verbal material in VPs, we introduce a c-structure category called VPargs, which generates a flat sequence of constituents that can appear in VPs. To make sure that VPargs actually expands to an argument/adjunct cluster, not to a single argument/adjunct, we impose a minimal length of two such constituents, and for efficiency reasons, we impose a maximal length of three. The corresponding rule in the grammar is the following: ${ }^{3}$
(9) VPargs --> \{ DP [std]
| PP[std]
| ADVP[std]
| PREDP[std]
| XPpred[std]
\} \#2\#3.
VPargs is introduced by a special rule for argument/adjunct cluster coordinations, which takes the following form in the grammar: ${ }^{4}$

```
VPargs-COORD --> VPargs: ! $ ^ ;
    [ COMMA
        VPargs: ! $ ^; ]*
    CONJco
    VPargs: ! $ ^.
```

Finally, this special category is introduced in the VC rule, similarly to the coordination of VPs, so that the relevant part of the VC rule then looks as follows:

```
VC --> { VPargs-COORD V
    | @(VP-COORD)
    | VC Vaux
    | ... }.
```

The two special rules above in combination with the introduction of VPargs-COORD in the VC rule make it possible to build up the c-structures shown in Figures 7 and 8 for (5) and (6) respectively. From these, the f-annotations in the rules project linguistically sound f-structures, as can be verified in Figures 7 and 9. Note that the latter c-structure exhibits all three VP coordination phenomena we have addressed, i.e. distribution of an argument/adjunct (es) from the left, distribution of verbal elements (geben wird) from the right, and argument/adjunct cluster coordination (or conjunction reduction).

[^106]
"In den ersten vier Monaten stiegen die Exporte um 8,4 und die Importe um sechs Prozent


Figure 7: C- and f-structure corresponding to (5)
In the first four months, the exports rose by 8.4 percent, and the imports, by six.


Figure 8: C-structure corresponding to (6)
that there will be no compromises in the struggle against the PKK and no concessions for the Kurdish civil rights activists
"daß es im Kampf gegen die PKK keine Kompromisse und für die kurdischen Bürgerrechtler keine Zugeständnisse geben wi


Figure 9: F-structure corresponding to (6), projected from c-structure in Figure 8
that there will be no compromises in the struggle against the PKK and no concessions for the Kurdish civil rights activists

### 3.4 Remaining problems

One issue that our special coordination rules cannot solve is the violation of subject-verb agreement between a distributed verb form in the plural and singular subjects in the (partial) VP conjuncts. Examples (12) and (13) exhibit this issue.
(12) Dafür heuerten zunächst der baden-württembergische Verband, später auch andere For that signed at first the of Baden-Württemberg federation, later also other Profis an. professionals on.
'At first, the Baden-Württemberg federation signed on for that; later, other professionals did so, too.' (TIGER sentence \# 9682)
(13) Dies kündigten [...] Ursula Engelen-Kefer im "Mitteldeutschen Express" (Halle) This announced [...] Ursula Engelen-Kefer in the "Mitteldeutscher Express" (Halle) und der Zweite IG-Metall-Vorsitzende Klaus Zwickel im Sender Rias an. and the Second IG Metall President Klaus Zwickel in the radio station Rias . '[...] Ursula Engelen-Kefer announced this in the "Mitteldeutscher Express" (Halle) and the Second President of the IG Metall, Klaus Zwickel, in the Rias radio station.' (TIGER sentence \# 1034)

In (12), the verb features plural agreement, which is satisfied by the subject of the second VP conjunct, but conflicts with the subject of the first VP conjunct. In (13), the verb conflicts in number with the
subjects of both VP conjuncts.
It is well-known that subject-verb agreement in number and person, while generally being a very stable wellformedness constraint in German, is often violated in the context of coordinations. It is therefore not surprising to find violations of subject-verb agreement also in the context of (partial) VP coordinations. Furthermore, we would like to point out that all formal accounts of (partial) VP coordinations that we are aware of, including Maxwell \& Manning (1996), fail to account for number and/or person mismatches between verbs and their subjects in these constructions.

Another class of coordinations that our special rules cannot deal with are the ones that Maxwell \& Manning (1996) analyze using a stack. An example is given in (14).
(14) Nach [...] stimmten 127799 Dresdner für und 58778 gegen die

According to [...] voted 127,799 Dresdeners for and 58,778 against the
stadtnahe Variante der A 17.
city-adjacent variant of the A 17.
'According to [...], 127,799 Dresdeners voted for and 58,778, against the variant of the A 17 in close proximity to the city.' (TIGER sentence \# 6102)

An approximate TIGERSearch query that we have run indicates that non-constituent coordinations whose analyses would require a stack are very rare: We have found fewer than 30 sentences containing these in the entire non-evaluation part of the TIGER Corpus. Now it may well be that we miss some because the secondary edges that characterize these constructions were not annotated reliably outside of S and VP coordinations. Nevertheless, we claim that right node raising from coordinated PPs, as we find it in (14), is very infrequent in comparison with right node raising from coordinated VPs.

Finally, an issue which would be easy to solve if the solution did not affect efficiency so adversely is the fact that VPs can be coordinated using just commas, i.e. without an explicit conjunction. (15) illustrates this use of a comma instead of a conjunction.
(15) Nach einer Hochrechnung aus der Nacht kommt Kwasniewski auf 34,8 According to a projection from the night comes Kwasniewski up to 34.8
Prozent, Walesa auf 33,3 Prozent percent, Walesa up to 33.3 percent.
'According to a projection of last night, Kwasniwski achieves 34.8 percent, Walesa, 33.3 percent.' (TIGER sentence \# 6251)

While this is a general problem with VP coordinations that lack an explicit conjunction, this phenomenon seems to be particularly frequent in VP coordinations involving conjunction reduction. As a consequence, we still cannot capture a large portion of these, despite the special rules we have introduced for them.

## 4 Further Benefit of Our Analysis: a Flat VP

By introducing special rules for (partial) VP coordinations, we obliterate the need for a recursive rightbranching VP (or VPx) rule. Instead, we can now formulate a flat VP rule that attaches all arguments and adjuncts of a (non-coordinated) VP as sisters at the same level. This is highly desirable from a grammar developer's point of view, since (i) it allows for the formulation of more general rules, (ii) it makes it possible to express hard linear precedence constraints on VP arguments/adjuncts, and (iii) it facilitates the design of learning features that can act as soft constraints on the constituent order in VPs in parse or realization ranking models.

### 4.1 More general rules

In previous versions of the German ParGram LFG, clauses with the main verb in second position were captured by a very different set of rules than clauses with the main verb in clause-final position. This was motivated by the observation that the "headless" VPs in clauses with the main verb in second position could be analyzed much more efficiently by a flat rule than by the recursive right-branching VP (or VPx) rule used for the analysis of VPs with a verbal head. Now that both "headless" and headed VPs are analyzed by means of a flat rule, a lot of the rule code can be shared between the two constructions. This is desirable both conceptually, as it is a more general description of the phrase structure of German, and from an engineering point of view, as code sharing leads to better maintainability.

### 4.2 Hard constraints on constituent order in the VP (especially for generation)

Although the order of arguments and adjuncts in German VPs is very free, there are positional constraints on certain types of arguments. For example, PREDPs (i.e. predicative arguments) have to occur next to the VP-final verb, and the expletive subject pronoun $e s$ has to occur at the left edge of the VP. In previous versions of the German ParGram LFG, the verb-adjacent position of PREDPs was enforced by attaching a PREDP within the VC rather than the VP (or VPx) rule; no restriction on the placement of expletive pronouns was expressed.

With the flat VP rule that we can use now, it is far easier to formulate constraints on the placement of certain types of constituents within VPs. This is particularly important in the context of generation, where the order of constituents needs to be controlled relatively tightly to avoid extremely marked or even ungrammatical string realizations. The German ParGram LFG therefore now states linear precedence constraints on expletive pronouns, the sentential negation adverb nicht and PREDPs within VPs and thereby prevents extremely marked or ungrammatical string realizations from being passed to the probabilistic realization ranker associated with the grammar (Cahill et al. 2007a,b).

### 4.3 Capturing soft constraints in the form of learning features for a statistical model

Just like hard constraints on the constituent order in VPs can be expressed more easily with a flat VP rule, learning features (or properties) that can potentially capture soft constraints on the constituent order in VPs are easier to formulate in this case. Learning features for statistical models used for parse or realization ranking (Forst 2007a,b, Cahill et al. 2007a,b) are typically based on templates that consider certain local c- or f-structure configurations. For example, the feature template cs_sub_rule implemented in XLE counts the number of times the context free rule that it takes as an argument appears in the analyses or generated trees that the statistical model has to rank. With a flat VP rule, the order of constituents in VPs can be captured by local features such as cs_sub_rule VP DP PP ADVP PP VC or cs_sub_rule VP DP PP PP ADVP VC, whereas features would have to be non-local at considerable depth to capture the same orders expressed in terms of VP (or VPx) recursions.

## 5 Conclusions and Future Directions

We have proposed a solution for problematic cases of VP coordination in German which, to our knowledge, have not been implemented in any computational LFG so far. Our approach is inspired by the proposal of Maxwell \& Manning (1996), and just like that approach, captures argument cluster coordination (or conjunction reduction) and many instances of right node raising. Unlike the approach of Maxwell \& Manning (1996), which is a general theory of non-constituent coordination, our solution is targeted at problematic cases of VP coordination. As a result of this, it is less general and admittedly
fails to capture right node raising from categories other than VP, but it is also far more efficient computationally and hence suitable for implementation. By modifying the grammar in the way described above, we have improved both coverage and parse quality ( $83.45 \%$ F-score with the new rules as opposed to $82.98 \%$ previously), without adversely affecting efficiency. As a welcome side effect, the replacement of the recursive VPx rule by a flat VP rule makes it possible to formulate more general rules and easier to express both hard and soft constraints on constituent order in the VP domain.

In future work, we will examine in greater detail examples of coordination which Maxwell \& Manning (1996) analyze using a stack, like, e.g., (14). Preliminary experiments indicate that it is possible to obtain the semantically intended f-structures for these coordinations by distributing the DP in the PP at the right edge using f-annotations. We will try to find ways to make such rules more general without sacrificing efficiency.

Finally, another research topic we want to pursue is the type of right node raising illustrated by (16).
(16) Die Regierung begrüßte und die Opposition kritisierte gestern nach der Sitzung The government welcomed and the opposition criticized yesterday after the meeting die Entscheidung der Regierung.
the decision of the government.
'Yesterday after the meeting, the government welcomed and the opposition criticized the decision of the government.'

Our preliminary solution covers most cases of right node raising where the raised constituent needs to be distributed into a CProot conjunct. However, it tends to overgenerate because the boundary between the end of the second CProot conjunct and the beginning of the raised constituent(s) is difficult to determine. We intend to collect more data in order to find an empirical basis for the restrictions to be introduced.

## References

Berman, Judith. 2003. Topics in the Clausal Syntax of German. Studies in Constraint-Based Lexicalism. CSLI Publications.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2003. TIGER: Linguistic Interpretation of a German Corpus. Journal of Language and Computation Special Issue.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER Treebank. In Proceedings of the Workshop on Treebanks and Linguistic Theories, Sozopol, Bulgaria.

Brants, Sabine, and Silvia Hansen. 2002. Developments in the TIGER Annotation Scheme and Their Realization in the Corpus. In Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Spain.

Cahill, Aoife, Martin Forst, and Christian Rohrer. 2007a. Designing Features for Parse Disambiguation and Realisation Ranking. In Proceedings of LFG07 Conference, pp. 128-147, Stanford, California. CSLI Online Publications.

Cahill, Aoife, Martin Forst, and Christian Rohrer. 2007b. Stochastic Realisation Ranking for a Free Word Order Language. In Proceedings of the Eleventh European Workshop on Natural Language Generation, pp. 17-24, Saarbrücken, Germany. DFKI GmbH. Document D-07-01.

Crouch, Dick, Mary Dalrymple, Ron Kaplan, Tracy King, John Maxwell, and Paula Newman. 2009. XLE documentation. Technical report, Palo Alto Research Center, Palo Alto, CA.

Dipper, Stefanie. 2003. Implementing and Documenting Large-scale Grammars - German LFG. PhD thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 9, Number 1.

Forst, Martin. 2007a. Disambiguation for a Linguistically Precise German Parser. PhD thesis, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 13, Number 3.

Forst, Martin. 2007b. Filling Statistics with Linguistics - Property Design for the Disambiguation of German LFG Parses. In Proceedings of the ACL Workshop on Deep Linguistic Processing, Prague, Czech Republic.

Haider, Hubert. 2003. Deutsche Grammatik generativ: Vorstudien zur Theorie einer projektiven Grammatik. Tübingen, Germany: Günther Narr Verlag.

König, Esther, Wolfgang Lezius, and Holger Voormann. 2003. TIGERSearch User's Manual. IMS, University of Stuttgart, Stuttgart, Germany.

Lezius, Wolfgang. 2002. Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. PhD thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 8, Number 4.

Maxwell, John T., and Christopher Manning. 1996. A Theory of Non-constituent Coordination based on Finite State Rules. In Proceedings of the First LFG Conference, Grenoble, France.

Müller, Stefan. 2002. Complex Predicates: Verbal Complexes, Resultative Constructions, and Particle Verbs in German. Studies in Constraint-Based Lexicalism. Stanford, CA: CSLI Publications.

| Martin Forst | Christian Rohrer |
| :---: | :---: |
| martin.forst@microsoft.com | rohrer@ims.uni-stuttgart.de |
| Natural Language Engineering | IMS |
| Powerset/Microsoft | University of Stuttgart |
| 475 Brannan St | Azenbergstr.12 |
| San Francisco, CA 94110 | 70174 Stuttgart |
| USA | Germany |

# F-STRUCTURE TRANSFER-BASED STATISTICAL MACHINE TRANSLATION 

Yvette Graham Josef van Genabith Anton Bryl Dublin City University

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper, we describe a statistical deep syntactic transfer decoder that is trained fully automatically on parsed bilingual corpora. Deep syntactic transfer rules are induced automatically from the f-structures of a LFG parsed bitext corpus by automatically aligning local f-structures, and inducing all rules consistent with the node alignment. The transfer decoder outputs the n -best TL f-structures given a SL f-structure as input by applying large numbers of transfer rules and searching for the best output using a log-linear model to combine feature scores. The decoder includes a fully integrated dependency-based tri-gram language model. We include an experimental evaluation of the decoder using different parsing disambiguation resources for the German data to provide a comparison of how the system performs with different German training and test parses.


## 1 Introduction

In this paper, we describe a statistical deep syntactic transfer decoder used as the transfer component of a Transfer-Based Machine Translation (TBMT) system to transfer source language (SL) deep structures to the target language (TL). Deep syntactic transfer rules are induced automatically from the functional structures of a Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Bresnan, 2001; Dalrymple, 2001) parsed bitext corpus. Firstly, local f-structures are automatically aligned, before all rules consistent with the node alignment are induced automatically. The transfer decoder applies large numbers of transfer rules to the input SL f-structure and searches for the best TL output $f$-structure using a log-linear model to combine feature scores.
The paper is structured as follows: in Section 1, we give our motivation for using deep syntax in MT, Section 2 describes the architecture of deep syntactic Transfer-Based MT, Section 3 describes the main focus of this paper, statistical transfer between source and target deep syntactic structures, in Section 4, we give an experimental evaluation of the transfer decoder in the context of a hybrid system that uses LFG functional structures (f-structures) as the intermediate representation for transfer, training and testing the system using two different disambiguation models for the German data for German to English translation, and Section 5 gives our plans for future work.

## 2 Motivation

In TBMT, among the different types of intermediate structures used for transfer are deep syntactic structures. For example, Bojar and Hajič (2008) use the Functional Generative Description (FGD) (Sgall et al., 1986) Tectogrammatical Layer (T-layer), labeled ordered dependency trees, while Riezler and Maxwell (2006) use the LFG f-structure, an attribute-value structure encoding of bilexical labeled dependencies.

Deep syntactic structures are more language independent than other representations used for MT such as surface form strings and phrase-structure trees, and therefore should provide a better means of forming generalizations about how to translate from one language to another. For example, automatic translation between very distant language pairs can require complex re-ordering of words between source and target. For many languages incorrect word order in TL output results in one of two problems; the output is either (i) ungrammatical or (ii) grammatical with incorrect meaning. Since the permitted word order of a sentence in many languages is strongly influenced by the dependency relations between the words of the sentence, explicitely including these relations in the translation model should help produce correct TL word order, especially when translating between very distant language pairs. In addition, using a language specific generator designed to generate from structures in which these relations between words are explicitely represented could also help to produce better quality output with respect to word order.

As well as dependency relations, many theories of deep syntax also include morphological analysis, so that words in the surface form are represented in the deep syntactic structure in lemma form with a set of features encoding grammatical information, like case, person, number, tense, etc. Explicitly representing this grammatical information may be important for translation from morphologically poor languages into morphologically richer ones. For example, when translating from English into German the red wine has at least three possible translations: der rote Wein, den roten Wein and dem roten Wein. In this example, the value of the feature case in the TL needs to be known in order to choose the correct morphological inflection of the determiner der and adjective rot. If the case of the noun in the English phrase is established this information should help select the best phrase in German. Including this grammatical information present in the source and target deep syntactic structure should therefore help produce the correct morphology in the TL.

## 3 Deep Syntactic Transfer-Based MT

Deep Syntactic Transfer-Based MT is composed of three parts; (i) parsing to deep syntactic structure, (ii) transfer from SL deep structure to TL deep structure and (iii) generation of TL sentence (Figure 1). Each stage in the three stage pipeline architecture could be carried out using fully automatically learned (statistical) resources, hand-crafted resources or a hybrid of statistical and handcrafted resources. For example, for parsing Riezler and Maxwell (2006) use hand-crafted grammars in addition to automatically learned disambiguation models. The parsing step in their system is therefore a hybrid of hand-crafted and statistical methods. For transfer, they use mostly automatically induced transfer rules as well as some hand-crafted rules. In addition, they carry out hand-


Figure 1: Deep Syntax Transfer-Based MT Pipeline Architecture
selected corrections of the word alignment prior to rule induction. ${ }^{1}$ They also use a statistical search and statistical model to transfer SL structures to TL structures. The transfer component of their system therefore is also a hybrid. Finally, for generation, they used a hand-crafted generation grammar and a statistical model, including a TL model, for example, to select the best output. Thus the generation step in their system is also a hybrid of hand-crafted and automatically learned resources.

The focus of our work is to investigate methods of automatically learning how to translate from training data. The transfer step in our system is trained fully automatically without any hand-crafted rules or human-selected corrections to any part of the rules or word-alignment. ${ }^{2}$ Our system uses hand-crafted resources for parsing and generation (Kaplan et al., 2004; Riezler et al., 2002). The bitext training data is automatically parsed (Kaplan et al., 2002) and the same type of grammar is used for generation. The transfer stage of our system is fully statistical, but the experimental evaluation in this paper is evaluating the decoder in the context of a hybrid system, using hand-crafted resources for parsing and generation. ${ }^{3}$ Figure 1 shows the Transfer-Based MT system pipeline with each stage labeled either statistical or hybrid for our system.

## 4 Statistical Transfer

### 4.1 Transfer Rule Induction

To induce transfer rules automatically from the parsed corpus, we use the RIA rule induction tool (Graham and van Genabith, 2009). Figure 2 shows some

[^107]

Figure 2: Example Transfer Rules
example transfer rules produced by the tool. The transfer rule induction algorithm takes as input (i) a dependency structure pair and (ii) a one-to-one set of alignments between nodes of the dependency structure pair.

### 4.1.1 Local F-Structure Alignment

Prior to rule induction a set of one-to-one correspondences between the local f-structures of each pair of parsed sentences in the bilingual corpus must be established. For automatic alignment of local f-structures we take the parsed bilingual corpus and extract the predicate values from each pair of f-structures to reconstruct a lemmatized version of the bitext. Figure 3 shows an example of a bitext corpus that is first parsed, then reconstructed from the f-structure representation. The order of the predicates in the reconstructed version of the bitext (Figure 3(c)) is determined by the location of the local f-structure within the overall f-structure. The predicate values are ordered via a depth-first traversal of the underlying dependency graph encoded in the f-structure. For example, the order of the predicates in the reconstructed corpus (Figure 3(c)) of the German f-structure in Figure 3(b) is ähneln und bill bob since ähneln is the predicate of the main f -structure with daughter und that in turn has daughters bill and bob. In order for the depth-first traversal not to loop if the f-structure contains instances of reentrancy or argument sharing we temporarily ignore these dependencies when reconstructing the corpus from the f-structures. The reconstructed bitext is then input to Giza++ (Och et al., 1999) and automatic word alignment is run
in both language directions. The output is then input to Moses to compute the symmetrization of the bidirectional alignment. We currently use the intersection in order to get a reliable set of one-to-one correspondences between words.

The aligned parsed bitext is used as input to the rule induction step. We use the RIA open source rule induction tool (Graham and van Genabith, 2009) to induce transfer rules. For each input f-structure pair and its node alignment, RIA induces all transfer rules consistent with the node alignment. The following section provides the definition for consistent transfer rules.

### 4.1.2 Consistent Transfer Rules

As in Phrase-Based Statistical Machine Translation (PB-SMT), where a word alignment for each example sentence pair is first established before phrases consistent with that word alignment are extracted (Och et al., 1999; Koehn et al., 2003), we induce transfer rules that are consistent with the node alignment. We define a consistent transfer rule using a simplification of the actual training dependency structures and temporarily consider them as acyclic graph structures by ignoring edges that cause cycles in the graph or edges that share an end node with another edge. Definition 1 applied to a (simplified) dependency structure pair yields a set of rules containing no variables by constraining rule induction using both the alignments between nodes and the position of the nodes within the two structures:

## Definition 1.

Given a one-to-one set of alignments $A$ between nodes in dependency pair $(F, E),(\bar{f}, \bar{e})$ is a rule consisting of nodes $\left(N_{f}, N_{e}\right)$, rooted at $\left(r_{f}, r_{e}\right)$, with descendents $\left(D_{f}, D_{e}\right)$ of $r_{f}$ and $r_{e}$ in $F$ and $E$ respectively, if

$$
\begin{array}{ll} 
& N_{f}=r_{f} \cup D_{f} \\
\Lambda & N_{e}=r_{e} \cup D_{e} \\
\Lambda & \forall f_{i} \in N_{f}:\left(f_{i}, e_{j}\right) \in A \rightarrow e_{j} \in N_{e} \\
\Lambda & \forall e_{j} \in N_{e}:\left(f_{i}, e_{j}\right) \in A \rightarrow f_{i} \in N_{f} \\
\Lambda & \exists e_{j} \in N_{e}:\left(r_{f}, e_{j}\right) \in A \\
\Lambda & \exists f_{i} \in N_{f}:\left(f_{i}, r_{e}\right) \in A
\end{array}
$$

## Definition 2.

For any rule $(\bar{f}, \bar{e})$ in dependency pair $(F, E)$ rooted at $\left(r_{f}, r_{e}\right)$ consisting of nodes $N_{f}$ and $N_{e}$, where $(\bar{s}, \bar{t})$ is also a rule in $(F, E)$ rooted at $\left(r_{s}, r_{t}\right)$ consisting of nodes $N_{s}$ and $N_{t}$ where $r_{s} \neq r_{f}, r_{t} \neq r_{e}$, iff $r_{s} \in N_{f}$ and $r_{t} \in N_{e}$, there is a rule $(\bar{a}, \bar{b})$ rooted at $\left(r_{f}, r_{e}\right)$ with nodes $r_{s}$ and $r_{t}$ replaced by variable $x_{k}$, where k is an index unique to the transfer rule, consisting of nodes:

$$
\begin{aligned}
& N_{a}: N_{f} \backslash N_{s} \cup x_{k} \\
& N_{b}: N_{e} \backslash N_{t} \cup x_{k}
\end{aligned}
$$

a) Maria ass

Der Bucheneinband ist schwarz.
Der elektrische Traktorschalter ist rot.
Der Ölkannenhersteller hat seine Schlüssel verloren
Bill und Bob ähneln sich.
…

Mary ate the cake
The book cover is black.
The tractor MM electrical switch is red
The oil can manufacturer lost his keys. Bill and Bob resemble each other

## PARSE

b)



Figure 3: Alignment of Local F-structures


Figure 4: Consistent Transfer Rules

Definition 2 allows the introduction of variables into transfer rules. Any rule that contains another rule nested within it can be used to form a new rule by replacing the nested rule with a single variable in its LHS and RHS. To help visualize what is considered a consistent transfer rule, Figure 4(b) shows the example dependency structure in Figure 4(a) divided into parts by a number of boxes with corresponding parts of the dependency structure pair labeled with the numbers 1-6. Each consistent transfer rule can be realised by assigning a binary value to each pair of boxes, so that boxes assigned 1 are included in the rule and boxes assigned 0 are left out. Combinations of binary values for nodes are constrained and this can be visualized by only allowing adjoining boxes in Figure 4(b) to be labeled 1 for any rule. Figures 4(c), 4(d) and 4(e) show example consistent rules with the binary value combinations that encode them.

### 4.2 Translation Model

As in PB-SMT, a Transfer-Based SMT translation model can be defined as a combination of several feature functions combined using a log-linear model:

$$
p(e \mid f)=\exp \sum_{i=1}^{n} \lambda_{i} h_{i}(e, f)
$$

### 4.2.1 Transfer Rule Probabilities

In PB-SMT the translation of an input sentence into an output sentence is modeled by breaking down the translation of the sentence into the translation of a set of phrases. Similarly, for Transfer-Based SMT, the transfer of the SL structure $\mathbf{f}$ into a TL structure $\mathbf{e}$ can be broken down into the transfer of a set of rules $\{\bar{f}, \bar{e}\}$ :

$$
p\left(\bar{f}_{1}^{I} \mid \bar{e}_{1}^{I}\right)=\prod_{i=1}^{I} \phi\left(\bar{f}_{i} \mid \bar{e}_{i}\right)
$$

We compute all rules from the training corpus and estimate the translation probability distribution by relative frequency of the rules:

$$
\phi(\bar{f}, \bar{e})=\frac{\operatorname{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_{i}} \operatorname{count}\left(\bar{e}, \bar{f}_{i}\right)}
$$

This is carried out in both the source-to-target and target-to-source direction and each model is used as a feature.

### 4.2.2 Lexical Weighting

We adapt a standard lexical-weighting method used in PB-SMT to hierarchical deep syntactic structure. In PB-SMT, lexical weighting is used as a back-off since it provides richer statistics and more reliable probability estimates. Adapting this feature to deep syntax is straightforward. In PB-SMT the lexical translation probability of a phrase pair is calculated based on the alignment between the words in the phrase pair. For deep syntax, we simply calculate the same probability via the alignment of lexical items in the LHS and RHS of a transfer rule. The lexical translation probability of a RHS, $\bar{e}$, given the LHS, $\bar{f}$, is estimated as follows:

$$
\operatorname{lex}(\bar{e} \mid \bar{f}, a)=\prod_{i=1}^{\text {length }(\bar{e})} \frac{1}{|\{j \mid(i, j) \in a\}|} \sum_{\forall(i, j) \in a} w\left(e_{i} \mid f_{j}\right)
$$

We use lexical weighting in both language directions.

### 4.2.3 A Dependency-Based Language Model

The overall system employs a language model at two different stages; a trigram dependency-based language model is used as a feature in the log-linear model by the transfer decoder and a standard trigram language model is used after generation to select the single best TL output. Riezler and Maxwell (2006) used a dependency-based language model in their system, but this was only done after decoding by calculating dependency-based language model scores on the n-best output of the decoder. ${ }^{4}$ We take an approach that is more in keeping with SMT and use language modeling during decoding. This section describes how

[^108]a)

Figure 5: Dependency-Based Language Model Example for F-structure of The cat likes to sleep
we have fully integrated a dependency-based language model into the transfer decoder.

Since our statistical search produces dependency structures where words are organized in a graph as opposed to a standard language model that deals with linear sequences of words, we estimate the probability of a dependency structure using the preceding context of each word within the dependency graph. In a standard trigram language model, the probability of the $i^{\text {th }}$ word in the context of its preceding i-1 words is approximated by the probability of observing it preceeded by its two preceding words:

$$
P\left(w_{1}, \ldots, w_{m}\right) \approx \prod_{i=1}^{m} P\left(w_{i} \mid w_{i-2}, w_{i-1}\right)
$$

The dependency-based language model approximates the probability of each word in the structure as the probability of observing it preceeded by its parent and grandparent words:

$$
P\left(w_{1}\left(\ldots, w_{m}\right)\right) \approx \prod_{i=1}^{m} P\left(w_{i} \mid \operatorname{parent}\left(\operatorname{parent}\left(w_{i}\right)\right), \operatorname{parent}\left(w_{i}\right)\right)
$$

If all dependency relations between local f-structures that cause either argument sharing or reentrancy are ignored, the underlying pred-only structure is an acyclic graph. We ignore such dependency relations when extracting the dependency-based language model so that each node in the structure can be assumed to have at most a single parent node. Figure 5(a) shows an example f-structure for the English sentence The cat likes to sleep. Figure 5(b) shows the simplified graph that used for language modeling where the reentrancy involving sleep and cat is ignored. As in standard language modeling, where the start of a sentence is represented by the special symbol $\langle s\rangle$, we add a root node to the structure with this symbol. We also add the end symbol to the leaf nodes $\langle/ s\rangle$. Figure 5(c) shows the probability approximation of the f-structure shown in Figures 5(a) and (b).

### 4.2.4 Other Features

Other features included in the log-linear model for ranking TL hypothesis structures include:

- Word Penalty
- Phrase Penalty
- Fragmented Structure Penalty
- Fragmented Rule Penalty
- Grammatical Mismatch Penalty

The word penalty and phrase penalty are taken almost directly from PB-SMT. The word penalty is used to counterbalance the dependency-based language model's bias for shorter TL structures and the phrase penalty is used to counterbalance the bias of transfer rule probabilities toward smaller rules. All other things being equal, it is better to transfer the structure using large transfer rules, as the chunk of structure that forms the RHS was already observed together in the corpus and therefore can be assumed to cause no problems with regard to creating unusual TL word combinations, which can happen when combining smaller rules. In addition, as the system can produce structures that are missing dependency relations between two nodes in the TL structure, the fragmented structure penalty is used to allow the model to bias towards more complete structures. A fragmented rule penalty is also used to disprefer rules that were induced from training data that had received a fragment parse from the parser. These rules tend to lead to bad TL structures that cause problems for the generator. It would be possible to completely filter out such rules to ensure they were never used, but in theory it is better to leave them in and allow the system to bias against their use as it is still possible in some cases that a fragmented rule leads to the best solution for a given input, for example when no non-fragmented rule is available to translate the word. Finally, the grammatical mismatch penalty is used to penalize rules by the amount of mismatching grammatical information in the LHS of the rule and the SL structure. All else being equal, rules that have a small amount of LHS grammatical information matching that of the SL structure are dispreferred.

### 4.3 Decoding

### 4.3.1 Top-down Transfer Rule Application

Decoding takes a single SL structure as input and involves a statistical search for the n -best TL structures. TL solutions are created via a top-down application of transfer rules to the SL structure beginning at the root (or main) f-structure. When the LHS of a rule unifies with the SL structure, the RHS produces a


Figure 6: Example top-down application of transfer rules
portion of the TL structure. Figure 6 shows an example application of three rules to the dependency structure for the German sentence Die Katze schläft gern 'The cat likes to sleep' shown in Figure 6(a). Figure 6(b) shows the first transfer rule to be applied to the root node of the SL structure to produce the TL structure portion shown in Figure 6(c). Transfer rule variables map arguments in the SL structure to the desired position when creating a TL solution. For example, variable $X_{0}$ in Figure 6(b) maps the subject of schlafen to the subject of like in the TL structure labeled with id number 1 shown in Figure 6(c). Next Katze in the SL structure is translated (Figures 6(d) and 6(e)), before finally die is translated (Figures 6(f) and 6(g)).

### 4.3.2 Beam Search

As with all SMT systems, the number of possible output translations given a single SL input is too large to exhaustively rank each possible output. We therefore employ a standard search algorithm, beam search, to produce the nbest TL solutions.
Partial translations (or translation hypotheses) are constructed by applying transfer rules to the SL structure. While TL translations are constructed, beam search manages the large search space by ranking translation hypotheses and pruning the search by dropping lower scoring hypotheses. A number of stacks are used to organize translation hypotheses into groups of comparable hypothe-
ses, according to the portion of SL structure that has already been translated to produce each hypothesis, i.e. hypothesis stack $N$ stores TL translation hypotheses with $N$ nodes covered in the SL structure. For example, Figure 7(a) shows the hypothesis stacks for decoding the f-structure of Die Katze schläft gern containing 4 nodes and therefore requiring stacks 1-4 for decoding, each stack storing translation hypotheses for solutions covering one to four nodes, respectively.

Transfer rules are indexed by root node so that they can be retrieved quickly to translate SL structure nodes. For example, in Figure 7(a) the rules rooted at node Katze are stored together. Since rules are applied top-down to the SL structure (see Section 4.3.1) rules beginning at the root node of the SL structure (or main SL f-structure) are first used to construct hypotheses. For example, in Figure 7(b) the rule that translates the root node of the SL structure schlafen as doze is first used to construct a hypothesis and since it covers one SL node it is stored in hypothesis stack 1. Figure 7(c) shows the next three hypotheses that are constructed: snooze, sleep and like sleep. Hypotheses are ordered within each stack according to their score, high-to-low from bottom-to-top. We currently use histogram pruning. When a stack becomes full, lower scoring solutions are pruned by being popped off the top of the stack.

For efficiency, each partial translation is only stored once in memory even though it may be part of several different future hypotheses. For example, hypothesis stack 2 in Figure 7(d) contains four translations constructed by expanding hypothesis doze by four different rules, each translating the word Katze into a different TL word. These new hypotheses are represented by a reference to the most recently applied transfer rule (rules translating Katze) and a reference back to the previous hypothesis. Figure 6 shows an example of decoding. Figure 7(e) shows an example of how per single completed translation, the structure for the lion likes to doze, is represented in the hypothesis stacks and Figure 7(f) shows all hypotheses are represented when the decoder has completed translating a single SL input structure. The n-best translated structures can be retrieved from the final stack.

### 4.3.3 Efficient Dependency-Based Language Modeling

An important feature in an SMT decoder is the language model and integrating one can be a more challenging task than other features since the language model score of a translation hypothesis cannot be calculated by simply combining the language model scores of the phrases (or rules) that it is composed of.

Although the search space is limited by beam search, during decoding large numbers of TL hypothesis structures need to be ranked. At each expansion of a translation hypothesis (via joining of an existing hypothesis with a new rule) a language model score for the newly created hypothesis needs to be calculated. Since this is carried out very many times per single decoding run, it is vital that


Figure 7: Beam Search Decoding
the method of calculating this score is highly efficient.
In our system, we pre-compute a dependency-based language model score for each transfer rule prior to beam search. This score is calculated only once for each rule even though a single rule may be part of several translation hypotheses. Then during decoding, when a translation hypothesis is expanded by adding a new rule, the new hypothesis score can be calculated quickly by combining the score of the old hypothesis, the rule score and a score calculated based on the probabilities of trigrams where the old hypothesis and rule join together. The probability of a TL hypothesis, $h_{n}$, that was produced by combining hypothesis $h_{n-1}$ and rule $r$ can be calculated as follows:

$$
\text { hyp_score }\left(h_{n}\right)=\text { hyp_score }\left(h_{n-1}\right) * \text { join_score }\left(h_{n-1}, r\right) * \text { rule_score }(r)
$$

Since hyp_score ( $h_{n-1}$ ) and rule_score $(r)$ are already computed, only join_score ( $h_{n-1}, r$ ) needs to be computed when hyp_score $\left(h_{n}\right)$ is computed.
Figure 8 shows how the language model scores are efficiently calculated when decoding the f-structure for the German sentence Die Werbung spiegelt die Vielfalt der britischen Universität wider 'The advertisement reflects the diversity of the British university'. We begin with the German f-structure graph shown in Figure 8(a) with nodes labeled by id numbers. Figure 8(b) shows the initial empty translation hypothesis that has probability 1 .
Figures 8(c), 8(f) and 8(i) show example transfer rules that can be applied to the German f-structure. Dependency-based language model scores are precomputed for each rule by identifying all trigrams within the RHS structure and calculating the product of their individual probability estimations retrieved from the language model; we will call this the rule_score (see Figure 8(d) for Rule A, Figure $8(\mathrm{~g})$ for Rule $B$ and Figure $8(\mathrm{j})$ for Rule $C$ ). In addition, for each rule, n-grams located at the RHS root node and frontier nodes are recorded. For example, Rule B in Figure 8(g) has a single root node bigram advertisement the located at node 2 while Rule $A$ in Figure 8(d) has two frontier bigrams $\langle s\rangle$, reflect and diversity, of located at nodes 2 and 6 , respectively. This information is used to calculate the language model score of joining a rule and a hypothesis.
Figure 8(e) shows the translation hypothesis established by applying Rule $A$ to the German structure. The language model score for the structure is established by combining the score of the previous hypothesis (since this is the first rule for this hypothesis, the previous hypothesis is the empty hypothesis and is therefore 1 ), the join score (since we are joining the rule with the empty hypothesis this score is also 1) and the rule score (see Figure 8(d)).
Figure 8(h) shows the translation hypothesis created by expanding Hypothes$i s_{1}$ by Rule $_{B}$. Since this expansion involved adding a rule at node 2 in the TL structure, the joining trigrams are derived by creating lists of words via all possible combinations of the frontier bigrams belonging to Hypothesis ${ }_{1}$ labeled 2 and the root bigrams of Rule $_{B}$, also labeled 2 (see root n-grams in Figure 8(g)). For this example, this results in a single word sequence $<s>r e-$


Figure 8: Efficient Dependency-based Language Modeling
flect advertisement the which forms two trigrams $<s>$-reflect-advertisement and reflect-advertisement-the. The score for Hypothesis $s_{2}$ is then calculated by combining the hypothesis score for Hypothesis ${ }_{1}$, the join score and the precomputed rule score for Rule B.

## 5 Experimental Evaluation

In our experimental evaluation of the system, we investigate the effects of the disambiguation model used to select the best parse. Riezler and Maxwell (2006) used an English disambiguation model for parsing both the German and English data when translating from German to English. If a single disambiguation model is used for both languages, the f-structures of a given pair of training sentences are likely to be quite similar, and this may help the rule induction process. However, another approach is to use language-specific disambiguation models for parsing. In this case, it is more likely that the actual best f-structure for each sentence of the training data is selected. Although a more authentic German parse may help the overall MT system, at the same time this is likely to increase the dissimilarity between the parses of the German-English sentences pairs, which may increase the difficulty of transfer.

We conduct an empirical investigation into which approach achieves better machine translation output for our system, by training and testing the system using (i) an English disambiguation model (Kaplan et al., 2004; Riezler et al., 2002) to select the best parse for both German and English sentences, and compare with results when (ii) a German disambiguation model (Forst, 2007) is used for selecting the best German parse and an English disambiguation model (Kaplan et al., 2004; Riezler et al., 2002) is used to select the best parse for the English sentences.

### 5.1 Training

The system was trained separately for each configuration. Training data for both configurations used data restricted by sentence length of 5-15 words from the Europarl (Koehn et al., 2005) and Newswire parallel corpora, which resulted in approximately 360,000 German-English sentence pairs, and a held-out development set of 500 sentences pairs. Both sides of the training corpus were parsed with the XLE parse engine (Kaplan et al., 2002). For Configuration 1, an English disambiguation model (Kaplan et al., 2004; Riezler et al., 2002) was used when parsing both the German and English data. For Configuration 2, a German disambiguation model (Forst, 2007) was used when parsing the German data and the English disambiguation model (Kaplan et al., 2004; Riezler et al., 2002) for the English data. The single best parse for each sentence, according to the appropriate disambiguation model, was used for training for both configurations.

For node alignment, Giza++ (Och et al., 1999) was run in both language

| Config. | BLEU | NIST | Coverage | Connected TL structure |
| :---: | :---: | :---: | :---: | :---: |
| 1 | 0.1121 | 3.5685 | $92.2 \%$ | $33.4 \%$ |
| 2 | 0.0730 | 2.6643 | $91.8 \%$ | $47.2 \%$ |

Table 1: Machine Translation System Results for Configuration 1: English disambiguation model for both German and English data, and Configuration 2: German disambiguation model for German data and English disambiguation model for English data
directions and the intersection was obtained using Moses (Koehn et al., 2007). We used both a dependency-based language model from the parsed TL side of the Europarl corpus and a conventional language model using lower-cased TL sentences, both trained on approximately $1,250,000$ sentences. The SRILM toolkit (Stolcke, 2002) was used for both language models. Minimum Error Rate Training (Och, 2003) was carried out using ZMERT (Zaidan, 2009) to train weights for each configuration on 500 randomly selected held-out development set sentences optimizing for Bleu.

### 5.2 Testing

The system was tested in a single language direction, German to English on 500 randomly selected held-out test set German sentences and the single best TL translation produced by the system was evaluated using automatic metrics with a single reference translation.

For each configuration, the German sentences were parsed with the same parsing engine and grammar as was used for training, and the single best fstructure according to the disambiguation model was selected as input to the decoder.

TL decoder output structures can be fragmented, and we automatically repair them if necessary. Automatic repair involves adding edges (in the form of FIRST/REST equations with nodes ordered via the position of their translations in the SL structure) to any TL structure that does not already form a single connected graph. For each test sentence, the 100-best TL decoder output structures were repaired automatically, before being input to the generator and a maximum of 50,000 sentences were generated per test sentence. A standard language model was used to select the final TL output.

### 5.3 Results

The Bleu (Papineni et al., 2002) and NIST (Doddington, 2002) scores for both system configurations on the test set are shown in Table 1. According to the automatic metrics, Configuration 1 achieves a Bleu score of 0.1121 outperforming Configuration 2, which achieves 0.073 , almost 4 Bleu points lower than Configuration 1. Configuration 1 also has higher system coverage, i.e. it was able to produce at least some output for $92.2 \%$ of the test set, while Configuration 2
achieves $91.8 \%$ coverage. The number of TL structures output from the decoder that already formed a single connected graph and therefore did not require any repair was, however, higher for Configuration 2 (47.2\%) than Configuration 1 (33.4\%).

### 5.4 Discussion

The results obtained in the experimental evaluation are contrary to our initial expectations. With the current system translating from English to German, using the English disambiguation model for both languages outperforms automatic evaluation results when the system is run on parse data disambiguated by language specific models. We had expected that the more authentic parses for the German data should lead to an overall increase in translation results, even if the difficulty of transfer is increased slightly by the slight increase in nonisomorphism across the f-structure representations for the parsed sentence pairs in the training data. The transfer rule induction algorithm is designed to induce rules that capture non-isomorphism, and therefore increasing non-isomorphism should not effect the system to this degree. One suspected cause of the problems for Configuration 2 may lie in the grammar used with this disambiguation model. The number of features in the grammar is higher than that of the German grammar used with the English disambiguation model of Configuration 1. When the data is parsed this leads to the German f-structures of Configuration 2 containing far more atomic features than those of Configuration 1. In fact, for the German development set parses, the ratio of number features in the fstructures for Configuration 1 compared to Configuration 2 is approximately 1:4. We suspect that due to the higher number of features of Configuration 2, transfer rules do not generalize as well to unseen data. The SL atomic features are used in our system to guide the selection of transfer rules. The smaller set of features of Configuration 1 may be a better guide for transfer than the larger set of Configuration 2.

## 6 Future Work

The size of the training corpus used in the evaluation is small compared to corpora usually used for training SMT systems. We would like to perform further extrinsic evaluation of the two disambiguation models when the system is trained on a larger corpus not restricted by sentence length. This would provide each configuration with richer statistical estimates and higher coverage of transfer rules on unseen SL structures.

## 7 Conclusion

We presented a SMT transfer decoder that uses deep syntactic structures, as the intermediate representation for transfer that applies state-of-the-art methods of PB-SMT to deep syntactic transfer. In the experimental evaluation the decoder
achieves better results using an English disambiguation model for parsing German data, than when a German disambiguation model is used.

## Acknowledgements

This work was partly funded by a Science Foundation Ireland Ph.D. scholarship P07077-60101.

## References

Ondřej Bojar and Jan Hajič. 2008. Phrase-Based and Deep Syntactic English-to-Czech Statistical Machine Translation. In Proceedings of the third Workshop on Statistical Machine Translation, Columbus, Ohio, June 2008.
Joan Bresnan. 2001. Lexical-Functional Syntax., Blackwell Oxford, 2001.
Miriam Butt, Helge Dyvik, Tracy H. King, Hiroshi Masuichi and Christian Rohrer. 2002. The Parallel Grammar Project. In Proceedings of the 19th International Conference on Computational Linguistics (COLING'02), Workshop on Grammar Engineering and Evaluation, pages 1-7. Tapei, ROC.
Aoife Cahill, Michael Burke, Ruth O'Donovan, Josef van Genabith and Andy Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired WideCoverage PCFG-based LFG approximations. In Proceedings of the 42nd ACL.
Aoife Cahill and Josef van Genabith. 2006. Robust PCFG-Based Generation using Automatically Acquired LFG Approximations. In Proceedings of COLING-ACL 2006, pages 1033-1040, Sydney, Australia.
Mary Dalrymple. Lexical Functional Grammar, Academic Press, San Diego, CA; London. 2001.
George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In Proceedings of HLT 2002.
Martin Forst. 2007. PhD thesis. Disambiguation for a Linguistically Precise German Parser. University of Stuttgart.
Yvette Graham, Deirdre Hogan and Josef van Genabith. 2007. Automatic Evaluation of Generation and Parsing for Machine Translation with Automatically Acquired Transfer Rules. In Proceedings of the 2007 Workshop on Using Corpora for NLG: Language Generation and Machine Translation, at MT Summit XI, Copenhagen, September 2007.
Yvette Graham and Josef van Genabith. 2008. Packed Rules for Automatic Transfer Rule Induction. In Proceedings of the European Association of Machine Translation Conference 2008, Hamburg, Germany.
Yvette Graham and Josef van Genabith. 2009. An Open Source Rule Induction Tool for Transfer-Based SMT. To appear in The Prague Bulletin of Mathematical Linguistics.
Ronald M. Kaplan, Stefan Riezler, Tracy H. King, John T. Maxwell, and Alexander Vasserman. 2004 Speed and Accuracy in Shallow and Deep Stochastic Parsing. In Proceedings of Human Language Technology Conference/North American Chapter of the Association for Computational Linguistics Meeting, Boston, MA, May 2-7 2004.

Ronald M. Kaplan, Tracy H. King and John T. Maxwell. 2002. Adapting Existing Grammars: the XLE Experience. In Proceedings of COLING 2002, Taipei, Taiwan.

Ronald Kaplan and Joan Bresnan. 1982. Lexical Functional Grammar, a Formal System for Grammatical Represenation. In J. Bresnan, editor, The Mental Representation of Grammatical Relations, pages 173-281, MIT Press, Cambridge, MA.
Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical Phrase-based Translation. In Proceedings of the HLT-NAACL 2003, pages 48-54, Edmonton, May/June 2003.
Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In Proceedings of MT Summit 2005.
Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison Burch, Richard Zens, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Evan Herbst, Brooke Cowen, Wade Shen, Christine Moran and Ondřej Bojar. 2007. Moses: Open Source Toolkit for Statistical Machine Translation In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics
John T. Maxwell III and Ronald M. Kaplan. 1991. A Method for Disjunctive Constraint Satisfaction. In Current Issues in Parsing Technology, Masaru Tomita editor, pages 173-190, Kluwer Aca demic Publishers.
Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), Sapporo, Japan, pages 160-167.
Franz Josef Och, Christoph Tillmann Hermann and Franz Josef Ney. 2000. Improved Alignment Models for Statistical Machine Translation. In Proceedings of the 1999 Conference on Empirical Methods in Natural Language Processing (EMNLP'99). College Park, MD, pages 20-28.
Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. A Method for Automatic Evaluation of Machine Translation. In Proceedings of ACL 2002, Philadelphia, pages 311-318.
Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell, and Mark Johnson. 2002. Parsing the Wall Street Journal using Lexical Functional Grammar and Discriminitive Estimation Techniques . (grammar version 2005) In Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL), Philadelphia, July 2002.
Stefan Riezler and John T. Maxwell III. 2006. Grammatical Machine Translation. In Proceedings of HLT-ACL, pages 248-255, New York.
Petr Sgall, Eva Hajičova and Jarmilla Panevova. 1986. The Meaning of the Sentence and its Semantic and Pragmatic Aspects. Dordrecht: Reidel and Prague: Academia 1986.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In Proceedings of the International Conference on Spoken Language Processing, Denver, Colorado, September 2002.
Omar Zaidan. 2009. Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems. In The Prague Bulletin of Mathematical Linguistics, No. 91:79:88.

# 'med' - THE SYNTAX AND SEMANTICS OF CONCOMITANCE IN NORWEGIAN 

Dag Haug<br>University of Oslo

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)
2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

The paper discusses the syntax and semantics of the Norwegian preposition med, which denotes a variety of concomitance relations. The c-structural, f -structural and semantic properties of the preposition are examined. Special emphasis is put on the constructions which involve syntax-semantics mismatches, such as bare noun phrases denoting sets of states instead of sets of individuals and it is shown how this can be dealt with in Glue semantics.


## 1 Introduction

Norwegian has a preposition med 'with' which has a variety of meanings, more or less closely related to 'concomitance' in a wide sense - just like its English counterpart:
(1) Gutten kom med faren.
boy.ART came with father.ART
The boy came with the father.
(2) Morderen forsvant med våpenet.
murder.ART disappeared with weapon.ART
The murderer disappeared with the weapon.
(3) Badekulturen forsvant med Romerrikets fall. bathing culture.ART disappeared with Roman empire.GEN fall
The bathing culture disappeared with the fall of the Roman empire.
(4) Mordet ble utført med en pistol.
murder.ART was committed with a gun
In example (1) we have the core meaning of concomitance - the father accompanies the boy and participates in the coming event. In example (2) there is also concomitance, but it is less symmetrical: the subject 'controls' the concomitance and is responsible for implicating the object of med in the event. In example (3), on the other hand, we have the opposite assymetry: the natural reading is that object of med, the fall of the Roman empire, somehow causes the event of the bathing culture disappearing. And in example (4) we have an instrumental reading, which can however also be described as a kind of specialized concomitance - the gun somehow participates in the matrix event, or to put it in other terms, there is a contextually inferrable relation between the gun and the murder event.

Even though these examples differ semantically, the syntax remains the same: we always have a prepositional phrase consisting of $\mathrm{P}+\mathrm{DP}$. But med frequently

[^109]occurs with a 'bare' (i.e. determinerless) NP. ${ }^{1}$ There are often important semantic effects:
(5) Skredderen satt der med istykkerrevet skjorte tailor.ART sat there with torn shirt The tailor sat there wearing a torn shirt.
(6) Skredderen satt der med ei istykkerrevet skjorte.
tailor.ART sat there with a torn shirt
The tailor sat there with a torn shirt.
In example (5) it is clear that the tailor wears the torn shirt: for the articleless NP to be felicitous, there must be a relation of inalienable possession. In (6) the preferred interpretation is that the tailor does not wear the shirt, but only works on it.

Finally, there are cases where med seems to embed a 'small clause' ${ }^{2}$ consisting of a DP and a predicate:
(7) Han kom med lua på hodet.
he came with cap.ART on head.ART
He came with the cap on his head.
(8) Hev deigen itre timer med et håndkle over (seg). let rise dough.ART for three hours with a cloth over (REFL)
Let the dough rise for three hours with a cloth over (it).
(9) Tjeneren kom inn med Johannes' hode på et fat. servant.ART came in with John's head on a plate The servant entered with John's head on a plate.
(10) Det er ikke lett $\stackrel{\circ}{a}$ få bilder av bygninger med blader på trcerne.
it is not easy to get pictures of buildings with leaves on trees.ART It is not easy to get pictures of buildings when there are leaves on the trees.
(11) Fødselen foregår med ski på beina.
birth.ART takes place with skis on legs.ART
The birth takes place with (the mother or the baby) wearing skis. ${ }^{3}$
So there is much syntactic and semantic variation between the different constructions of med, and yet they all seem to be interrelated as instances of a meaning of 'concomitance' between the object of med and some element in the matrix clause.

[^110]The very general meaning of concomitance which med expresses is reminiscent of the possession relation expressed by genitives, and from a discourse functional perspective, the two have similar, but 'inverse' functions: med, and English with, can be used to anchor the reference of a possessor via a possessum (the boy with the knife), just like a genitive will anchor the reference of a possessum (the boy's knife). In both cases, the semantic relation between the boy and the knife is the same.

In this paper, I will first examine the syntax of med at c- and f-structure, then develop the idea that the meaning(s) of med is essentially the mirror image of the meaning(s) of genitives, and show how this can be this can be matched with the syntax in a Glue semantics approach.

## 2 c-structure

When med embeds a DP, the c-structure is straightforward:
(12) med faren
with father.ART
'with the father'
(13)

with father.ART
Like in other Norwegian prepositional phrases, the P can be stranded by its complement:
(14) mannen som gutten kom med man.ART REL boy.ART came with the man whom the boy came with


When the complement of med is bare, it is reasonable to assume the same cstructure modulo the category of the complement:


In constructions where med embeds a 'small clause', the question arises whether the DP and the predicate makes up a constituent or whether we have a ternary branching structure. For the latter speaks the fact that it is not possible to front the predicate and the DP together, whereas it is possible to front the whole prepositional phrase. ${ }^{4}$ Thus example (10) can be turned into (17), but not (18).
(17) Med blader på trcerne er det vanskelig å ta bilder av with leaves on trees.ART is it difficult to take pictures of bygningene.
buildings.ART
With leaves on the trees, it is difficult to take pictures of the buildings.
(18) *Blader på trarne er det vanskelig å ta bilder avbygningene med. leaves on trees is it difficult to take pictures of buildings.ART with
In fact (18) invites the reading where på trcerne is restrictive and forms a constituent with blader, but in that case we get an instrumental reading which does not make sense. It therefore seems right to analyze the PP in (10) and (17) as in (19).


## 3 f-structure

### 3.1 Grammatical functions

In the 'normal' case where med embeds a DP or an NP, it is clear that that phrase bears the OBJ-function. This is clearly a thematic argument: in example (1), for example, the preposition med relates its object faren to gutten in the matrix clause.

[^111]When med embeds a 'small clause', on the other hand, it takes an OBJ and an ховJ, and the object is non-thematic:

## (20) Jeg hadde noen timer til overs med kona bortreist

I had some hours free with wife.ART away from home
I had some hours free with my wife away from home.
Here, the with-relation holds between the subject of the matrix clause and the state of the wife being away, not between the subject and the wife. This is very different in the following example:
(21) Jeg hadde noen timer til overs med ei bortreist kone. I had some hours free with a away from home wife I had some hours free with a wife away from home.

Semantically, med is always a two-place relation which denotes a concomitance relation between its object and an element in the matrix clause. This does not necessarily mean that med is bivalent in the syntax as well, ${ }^{5}$ but there are good arguments from binding facts that it does take a subject whenever it embeds a 'small clause':
(22) Hev deigen itre timer med et håndkle over (seg). let rise dough.ART for three hours with a cloth over (REFL)
Let the dough rise for three hours with a cloth over (it).
deigen is the object of the matrix clause, but it can bind an (optional) reflexive in the complement of med. Since Norwegian does not generally allow inanimate object binders, there must be a local subject present. In other words, the semantic form is here
(23) 'med $\langle\mathrm{SUBJ}, ~ X C O M P\rangle, ~$ OBJ'

Now what about the cases where med embeds a DP or an NP? Is it still the case that there is a subject present? Again, the binding facts suggest yes: As Lødrup (1999) showed, null possessors in inalienable possession constructions are generally bound in the same way as (simple) reflexives in Norwegian. Since possessors in the complement of med can be bound by inanimate objects in the main clause, we need a subject position in med here too:

## (24) Han leverte bilen med full tank

he returned car.ART with full tank
He returned the car with the tank full.
In other words, we should conclude that the semantic form of med is here
(25) med ' $\langle\mathrm{SUBJ}, \mathrm{OBJ}\rangle$ '

[^112]But does this apply to cases where the object of med is a DP and not an NP? med + NP patterns with med + 'small clause' rather than with med + DP in several respects, and at first sight the binding of inalienables seems to work differently in DPs and NPs. In example (24) the tank definitely is a part of the car which is returned, whereas (26) suggests that the subject returned the car together with some external tank:
(26) Han leverte bilen med en full tank

He returned car.ART with a full tank
He returned the car (along) with a full tank.
Consider the following three examples:
(27) Kirurgen arbeidet med nesa brekt (small clause)

Surgeon.ART worked with nose.ART broken
The surgeon worked with his nose broken.
(28) Kirurgen arbeidet med brekt nese (bare NP)

Surgeon.ART worked with broken nose
The surgeon worked with his nose broken.
(29) Kirurgen arbeidet med ei brekt nese (indef DP)

Surgeon.ART worked with a broken nose
The surgeon worked on a broken nose.
(27) and (28) are close paraphrases, both saying that nose of the surgeon was broken as he was working, whereas (29) is very different and says that the surgeon is working on a broken nose. However, the difference disappears if we look at non-unique inalienables, e.g. by substituting tå 'toe' for nese 'nose'.
(30) Kirurgen arbeidet med tåa brekt (small clause)

Surgeon.ART worked with toe.ART broken
The surgeon worked with his toe broken.
(31) Kirurgen arbeidet med brekt tå (bare NP)

Surgeon.ART worked with broken toe
The surgeon worked with his toe broken.
(32) Kirurgen arbeidet med ei brekt tå (indef DP)

Surgeon.ART worked with a broken toe.
The surgeon worked on a broken toe/with his toe broken.
In (32), unlike (29), the toe can belong to the surgeon. In other words, the difference in binding an inalienable in an indefinite DP (29) versus a bare NP (28) does not have to do with the category but rather with the fact that the indefinite article strongly suggests non-uniqueness, which does not make sense for noses belonging to a certain person, but is ok for toes belonging to a certain person. Finally, notice that an object can bind a non-unique inalienable in an indefinite NP:

He parked the car with a wheel on the sidewalk.
All in all, then, we can conclude that there is always a subject present in med, so the semantic form is either as in (23) or as in (25). These will fit one of the two following phrase structure rules for Norwegian prepositions: ${ }^{6}$
PP $\quad \rightarrow$

$$
\begin{array}{cc}
\text { P } & \text { NP }  \tag{34}\\
\uparrow=\downarrow & (\uparrow \text { OBJ })=\downarrow \\
((\uparrow \text { SUBJ PRED })=\text { 'pro' }) &
\end{array}
$$

| $\mathrm{PP} \rightarrow$P NP $\{\mathrm{AP} \mid \mathrm{PP}\}$ <br>  $\uparrow=\downarrow$ $(\uparrow$ OBJ $)=\downarrow$ <br>  $((\uparrow$ SUBJ PRED $)=$ 'pro' $)$ $(\uparrow$ XCOMP SUBJ $)=\downarrow$ |  |
| :---: | :---: | :---: | :---: |
|  | $(\uparrow$ XCOMP $)=\downarrow$ |

### 3.2 Control

The subject of med is never overtly realized, but is always a PRO which must be anaphorically bound. The binder is normally an argument of the matrix verb, in most cases the SUBJ or OBJ, as we saw above, but it can also be the implied agent in a passive, and even a participant implied by a verbal noun in the matrix clause as fødsel 'birth' in example (11).

Even more interestingly, the subject of med can also be anaphorically bound by the matrix event itself. This has been noted for other constructions as well Kortmann (1991):
(36) For three weeks the city had sweltered in heat and humudity, producing tension all around.

In more rigid approaches to the syntax-semantics interface such examples inevitably pose problems because the definite event description does not really correspond to any particular syntactic item (the verb denotes a set of events), but in Glue semantics we can capture this nicely, as we will see.

## 4 Semantics

### 4.1 Introduction

As we have already noted, the semantics of med are similar to that of genitives. So what exactly do genitives mean? Several answers have been forthcoming, but we will follow Partee (1983/1997). The first thing to notice is that they are ambiguous:

[^113]John's book
John's neighbour
In (37) we have a very general relation of 'possession': this could be a book that John owns, or a book he has written, or a book that is about him. In (38) we have a different situation, though, where John fills a slot in the valency of neighbour, so that the relationship is defined by the head noun as it were. We find the same ambiguity in med:
(39) Mannen med boka
man.ART with book.ART
(40) En mann med døde foreldre
a man with dead parents
In example (39), there is a general possessive relation between the man and the book, whereas in (40), the relation is defined by the lexical semantics of the word parent.

Genitives have been analysed by Partee (1983/1997) as being ambiguous between denoting a contextually supplied relation $R_{c}$ and linking a noun to the argument structure of the head of the genitive: ${ }^{7}$
a. John's ${ }_{2}: \lambda P . \iota x . P(x) \wedge R_{c}(\mathbf{j}, x)$
b. book: $\lambda$ P.book $(x)$
c. John's book $\lambda x \cdot \operatorname{book}(x) \wedge R_{c}(\mathbf{j}, x)$
a. John's ${ }_{1}: \lambda R . \iota x . R(\mathbf{j}, x)$
b. neighbour : $\lambda x \cdot \lambda y \cdot P(x, y)$
c. John's neighbour $\iota x$.neighbour $(\mathbf{j}, x)$

In the following we develop a similar semantics for med as essentially ambiguous between a linking function as in example (42) and denoting a general possessive/comitative relation as in (41). But in both uses, med essentially serves to relate two entities and so we will first have a look at what kinds of entities it can relate.

### 4.2 The semantic type of med's complements

Ignoring the NP-internal restrictive cases, the subject of med will always have the type of individuals. This individual can be either a participant in the matrix event (as in e.g. examples 1 and 2), or the matrix event itself (as in example 3 and 4). In a Glue-based approach we do not need to distinguish these types in the lambda calculus: they are both simply individuals, and the meaning constructors will tell us how to combine them with other elements in the clause.

What about the semantic type of the complement of med? In the 'prototypical' case, this is also an individual: a 'normal' individual in examples (1), (2) and (4),

[^114]and an event, the fall of the Roman empire, in example (3). Again, we do not need to differentiate the semantic types of (non-quantificational) DPs.

However, when med embeds a 'small clause' (as in examples 7-11), the complement is not an individual state, but a set of states: for example John's head on a plate does not refer to a particular state, but to an infinity of states characterized by John's head being on a plate. To denote a particular state, it needs to be hooked up to a definite time (which should then be related to the time of the matrix event).

Finally, we have the cases like example (5). A bare NP would normally denote a set of individuals: we would expect istykkerrevet skjorte 'torn shirt' to have the following meaning:
(43) $\quad \lambda x \cdot \operatorname{shirt}(x) \wedge \operatorname{torn}(x)$

But this is clearly not what we have. What example (5) means is that the tailor's shirt is torn at the moment of sitting. In other words, the semantics is in fact similar to the cases where med embeds a 'small clause' and (5) can be paraphrased as
(44) Skredderen satt der med skjorta istykkerrevet tailor.ART sat there with shirt.ART torn
The tailor sat there with the shirt torn.
This means that the 'bare NP' in example (5), despite initial appearances, must denote a set of states (and not an individual state, for the same reasons as above). But this is exactly the kind of syntax-semantics mismatch that Glue is designed to capture, and we will see in section 4.5 how it can be done.

There is another thing to be noticed about the cases where med embeds a small clause or an NP: as was observed by Sæbø (2009), it is very often the case that the complement contains a variable which is bound by the subject of med. The variable is typically provided by a relational noun, as in example (24), but it can also come from a preposition without an object (as in the version of (8) without the reflexive), or from an inalienable possession, as in (44). This means that in many cases the semantic type of the complement is not a set of states, but rather a function from individuals to sets of states.

To sum up, then, the subject argument of med (disregarding the restrictive case) always has the type of an individual whereas the object can be either an individual (whenever med embeds a full DP), or a set of states (whenever it embeds a 'small clause' or a 'bare' NP), or a function from individuals to sets of states whenever there is an unfilled argument slot in the embedded state description.

### 4.3 Semantics of $m e d+\mathbf{D P}$

This is the most straightforward case: med takes two individual type arguments and says that there is a contextually definable relation $R_{c}$ between them. More technically, it constructs a set of states of relations holding between the two individuals:

$$
\begin{equation*}
\text { med }: \lambda x \cdot \lambda y \cdot \lambda s \cdot R_{c}(x, y, s):(\uparrow \mathrm{SUBJ})_{\sigma} \multimap(\uparrow \mathrm{OBJ})_{\sigma} \multimap \uparrow \sigma \tag{45}
\end{equation*}
$$

Again we need the set of states to hook the state up temporally to the matrix event. In other words we need some kind of constructional meaning to bind the $s$ variable and introduce a temporal relation $\supseteq$ between $s$ and the matrix event $e$.

This constructional meaning constructor is independently required by so-called depictives like raw in He ate the meat raw, where we need to map the set of states of the meat begin raw onto the set of those events of him eating the meat which are surrounded by a state of the meat being raw. Pylkkänen (2002, p. 28) proposes such a depictive operator, which also takes care of linking raw to both the secondary predication and the matrix event, but in an LFG analysis with anaphoric control, we only need to link the two eventualities. This gives us an @ depictive template with the following meaning constructor:

$$
\begin{align*}
& \lambda P . \lambda Q . \lambda e . \exists s . P(s) \wedge Q(e) \wedge s \supseteq e:  \tag{46}\\
& \left(m_{\sigma}\right) \multimap\left(\left((\mathrm{ADJ} \in m)_{\sigma} \mathrm{EV}\right) \multimap(\mathrm{ADJ} \in m)_{\sigma}\right) \multimap \\
& \quad\left(\left((\mathrm{ADJ} \in m)_{\sigma} \mathrm{EV}\right) \multimap(\mathrm{ADJ} \in m)_{\sigma}\right)
\end{align*}
$$

Where $m$ refers to the f-structure of med.
For example, whenever med has an instrumental reading, it typically relates the matrix event to an object via the relation $R_{c}$ interpreted as an instrumental thematic role. For example, the predicate kill with the knife should have the following meaning:

$$
\begin{equation*}
\lambda e . \exists s . R_{c}(e, \mathbf{k}, s) \wedge \operatorname{kill}(e) \wedge s \supseteq e: f_{\sigma} \mathrm{EV} \multimap f_{\sigma} \tag{47}
\end{equation*}
$$

This meaning can be derived as in figure 1 in the appendix. Although the event variable does not have a direct representation in the syntax, it is present in the semantic structure as ( $f_{\sigma} \mathrm{EV}$ ) and can therefore be accessed as an antecedent by the pronominal subject of med. We start by hypothesizing an event e1 and let this event serve as the antecedent of the PRO subject of med. When we combine this with with the knife, we get a pair of the event constant $\mathbf{e} \mathbf{1}$ and a set of states of a relation $R_{c}$ holding between $\mathbf{e 1}$ and the knife $\mathbf{k}$. The depictive template turns the set of states into an event modifier restricting sets of events to those which are included in the time of $R_{c}$, so that we get a pair of the event constant $\mathbf{e} 1$ and a function from events to truth values. Applying this function to $\mathbf{e} \mathbf{1}$ yields the proposition that there is a state $s$ of there being a relationship $R_{c}$ between the killing event $\mathbf{e} \mathbf{1}$ and the knife $\mathbf{k}$ and this relationship holds throughout the run time of the killing event. Finally we discharge the hypothetical event $\mathbf{e} \mathbf{1}$ to get a set of events. ${ }^{8}$

There are even cases where we want med to relate two events, as in example (3). The object of med has the following meaning:

$$
\begin{equation*}
\text { Romerrikets fall }=\iota e . f a l l(e) \wedge t h e m e(e, \mathbf{r e}):(\text { ADJ OBJ } \uparrow)_{\sigma} \tag{48}
\end{equation*}
$$

med relates the two events and says that there is a relation between them; the secondary predication rule says this state holds at least througout the runtime of the

[^115]matrix event. The derivation straightforwardly follows the same lines as in the previous example, and the contextual relation $R_{c}$ is in this case interpreted as one of causation.

## 4.4 med + 'small clause'

As noted above, it is remarkable that in uses of med with 'small clauses', the 'small clause' typically contains a relational noun which has an unsatured argument slot. It is generally the case that this slot is controlled by the subject of med, as in examples (7-8). This corresponds to the cases where the genitive links an argument to a head as in example (38).

For these cases, we need the following meaning constructor: ${ }^{9}$

$$
\begin{equation*}
\lambda x . \lambda P . P(x):(\uparrow \text { SUBJ })_{\sigma} \multimap \forall H\left(H \multimap(\uparrow\{\mathrm{XCOMP} \mid \mathrm{OBJ}\})_{\sigma}\right) \multimap \uparrow \sigma \tag{49}
\end{equation*}
$$

med here combines with its subject and with its XCOMP, which 'lacks' a semantic resource $H$, and is therefore a function from such a resource to a set of states (simplified in the meaning constructor as $(\uparrow \text { XCOMP })_{\sigma}$ since we do not need to go into the internal structure of the set), and then constructs a set of states such that the subject of med fills the missing slot in the XCOMP. One example would be the version of example (8) without the reflexive pronoun seg, whose derivation is shown in figure 2 of the appendix.

For simplicity, we ignore the quantificational reading of the indefinite et hånd$k l e$ 'cloth' and just represent it with the constant $\mathbf{h}$. This combines with over to give a function from entities $y$ to states $s$ of the cloth being over $y .{ }^{10}$ med then combines with a hypothetical subject, later to be discharged. The result is a function which takes a function from anything to a meaning for the XCOMP of med, to produce a meaning for the whole med-phrase. et håndkle over provides exactly this, since there is an empty slot corresponding to the object of over. The result is a set of states of the cloth being over the hypothetical subject of med. When the hypothetical subject is discharged, we get a function from entitites to such states.

Figure 3 in the appendix shows how to combine the meaning of med et håndkle over with (a simplified version of) the matrix clause hev deigen. First, the PROsubject of med creates a copy of its antecedent. Next, this copy fills the empty subject slot in med et håndkle over while the antecedent resource, the object of the main verb, is still available. Then the @ depictive template is applied, turning med into a modifier of events. This modifier, and the object resource, can now be applied to a (simplified) ${ }^{11}$ version of the matrix verb to yield a set of events

[^116]of letting the dough rise such that there is a state of a cloth being over the dough which holds throughout the run time of the event.

In the above case, the small clause under med contained an unbound variable. There are also cases where there is no unbound variable in the 'small clause'. The semantic type of the complement is then just a set of states and med says that there is a state which falls under the description provided by its complement and which stands in a contextually defined relation to the subject of med:

$$
\begin{align*}
& \lambda x . \lambda P \cdot \lambda s . \exists t \cdot P(t) \wedge R_{c}(x, t, s):  \tag{50}\\
& (\uparrow \mathrm{SUBJ})_{\sigma} \multimap(\uparrow\{\mathrm{xCOMP} \mid \mathrm{OBJ}\})_{\sigma} \multimap \uparrow \sigma
\end{align*}
$$

We do not provide a full derivation here, but consider briefly example (10). The subject of med is the matrix event, which is the state of it being difficult to take pictures of the building. The complement of med is the state of there begin leaves on the trees. med blader på trarne denotes a set of states of some contextually definable relation, say causation, holding between it being difficult to take pictures of the buildings and there being leaves on the trees.

## 4.5 med + bare NP

The most interesting case is provided by the examples where med embeds a bare NP. These involve some interesting syntax-semantics mismatches:
(51) Hvor mye veier denne ATVen med full tank? How much weighs this ATV.ART with full tank
How much does this ATV weigh with its tank full?
(52) \#Hvor mye veier denne ATVen med en full tank? How much weighs this ATV.ART with a full tank
(53) Hvor mye veier denne ATVen med tanken full?

How much weighs this ATV.ART with tank.TANK full
As the examples show, the bare NP construction patterns with the 'small clause' construction, not with the case where med embeds an indefinite DP. Both (51) and (53) introduce a predication over the tank, which is in both cases not just any tank, but the tank of the ATV - in other words, we have a case of inalienable possession. On the other hand, example (52), if it can be made sense of at all, must refer to how much the ATV weighs together with some full tank; there is no predication, only restrictive modification; and there is no inalienable possession.

Notice that Norwegian actually allows bare NPs to a much higher degree than other European languages do. But as shown by Borthen (2003), bare NPs in possessive context (widely defined, and including intensional and negated possession),
@ TRANSITIVE. For a discussion of how these work in conjunction with secondary predication, see Haug (2008), although the constructions discussed there involve functional rather than anaphoric control.
such as those introduced by e.g. ha 'have', trenge 'need', dele ut 'hand out', $f a ̈$ 'get' etc. and med, differ from other bare NPs in several respects.
'Normal' bare NPs typically refer to conventional situation types:
De ser på TV
They watch TV
They cannot be freely modified, probably because of the restriction to conventional situation types:
(55) *De ser på gammel TV

They watch old TV
But in possessive contexts there is no such restriction:
De har gammel TV
They have old TV
They have an old TV (i.e. their TV is old).
Also, it is clear that the adjective in a bare NP occurring in a possessive construction often has a predicative reading, as in (24) above or in the following example with the verb trenge 'to need'.

## (57)

For å kjøre herfra til Bergen trenger du full tank
To drive from here to Bergen need you full tank
To drive from here to Bergen you need a full tank.
Despite appearances, it does not make sense to paraphrase this as 'You need an X such that X is a full tank' - the meaning is rather 'You need that your tank be full'. In other words, the bare NP denotes a state.

In an event-based semantics, we can model this as a constructional meaning. Recall first that according to the standard view all stage-level predicates (whether introduced by adjectives, verbs or prepositions) must have a state (or event) argument. The distinctive feature of these adjectives is that they are hooked up to times, and to achieve that we need the state/event argument. In other words, an adjective like 'happy' will need to have a lexical entry as the following:
(58) happy

$$
\begin{aligned}
& \lambda P \cdot \lambda x \cdot \lambda s \cdot P(x, s) \wedge \operatorname{happy}(x, s): \\
& \left((\mathrm{ADJ} \in \uparrow)_{\sigma} \operatorname{VAR} \multimap(\operatorname{ADJ} \in \uparrow)_{\sigma} \operatorname{RESTR}\right) \multimap \\
& \left((\mathrm{ADJ} \in \uparrow)_{\sigma} \mathrm{VAR} \multimap(\operatorname{ADJ} \in \uparrow)_{\sigma} \operatorname{RESTR}\right)
\end{aligned}
$$

In normal restrictive contexts, the state argument does not really play a role, but is closed off by the determiner, for example the quantifier $a$ :
(59) a
$\lambda P . \lambda Q . \lambda x . \exists s . P(x, s) \wedge Q(x):$
$(\operatorname{SPEC} \uparrow)_{\sigma} \mathrm{VAR} \longrightarrow(\operatorname{SPEC} \uparrow)_{\sigma} \operatorname{RESTR} \multimap$ $\forall H\left[(\operatorname{SPEC} \uparrow)_{\sigma} \multimap H\right] \multimap H$

But when there is no determiner the state argument is available to certain typeshifting operations such as the possessive bare NP-construction, which we can model through a template @ pOSS-NP:
(60) @ POSs-NP

$$
\lambda P . \lambda s . \iota x . P(x, s):(\uparrow \mathrm{OBJ})_{\sigma} \mathrm{VAR} \multimap(\uparrow \mathrm{OBJ})_{\sigma} \mathrm{RESTR} \multimap(\uparrow \mathrm{OBJ})_{\sigma}
$$

To see how this works in conjunction with med, consider the derivation in figure 4. First, full combines with the (in this context) relational noun tank, which denotes the set of individuals such that they are the tank of $\mathbf{x} \mathbf{2}$. The result is a function from individuals that are the tank of $\mathbf{x} 2$ to states of that tank being full. @ POSS-NP then turns this into the set of states of the contextually uniquely identifiable tank of $\mathbf{x} \mathbf{2}$ being full. We now discharge $\mathbf{x} \mathbf{2}$ so that we get a function from individuals to states of their tank being full. This is what med, with its hypothetical subject $\mathbf{x} \mathbf{1}$ looks for, so we get a set of states of $\mathbf{x 1}$ 's tank being full. When we discharge $\mathbf{x} \mathbf{1}$ we again get a function from individuals to states of their tank being full, and this time the meaning constructor tells us to combine this with the subject of med. The derivation then proceeds in a similar way to that in figure 3 .

## 5 Conclusion

We have seen that med can be constructed with bare NPs, full DPs and 'small clauses'. Syntactically, the NP and DP constructions are similar, since they both involve med heading a binary branching PP and taking two f-structure functions, a subject and an object. In the small clause construction, on the other hand, we have a ternary branching structure, and med takes a subject, an XCOMP and a (nonthematic) object which is the subject of the XCOMP. But semantically, the NP construction patterns with the 'small clause' construction.

Glue semantics lets us deal with this syntax-semantics mismatch in an elegant way. We can treat bare NPs in the complement of med as other bare NPs occurring in possessive contexts. Apart from that we only need one meaning constructor (45) for med in the cases where it takes an individual-type object (i.e. a DP), one meaning constructor (50) for the cases where the complement is a set of states, and one (49) for the cases where the complement is a function from individuals to states, i.e. the state description has an empty slot. Finally, we have seen how Glue semantics lets us deal with other syntax-semantics mismathces, where a semantic argument which is not present in the syntax, either the implicit participants of a verbal noun such as fødsel or the event argument of a finite verb, can bind the subject of med.
pro

Figure 1: Semantic derivation of kill with the knife

## et håndkle

over
(m
$\lambda y . \lambda s . \operatorname{over}(\mathbf{h}, y, s)$
$(m \text { XCOMP OBJ })_{\sigma}$
$(m \text { XCOMP OBJ })_{\sigma} \multimap(m \text { XCOMP })_{\sigma}$
h:
$(m \text { XCOMP SUBJ })_{\sigma}$
$\lambda x \cdot \lambda y \cdot \lambda$ s.over $(x, y, s)$ ( $o$ XCOMP OBJ $) \multimap(m \text { XCOMP })_{\sigma}$

$\lambda x . \lambda s . \operatorname{over}(\mathbf{h}, x, s)$
$(m \text { SUBJ })_{\sigma} \multimap m_{\sigma}$

Figure 2: Semantic derivation of med et håndkle over 'with a cloth over (it)',

## pro

$<x, x>$ :
med et håndkle over $\quad \frac{(f \text { OBJ })_{\sigma} \quad\left(p_{\sigma} \text { ANT }\right) \multimap\left(p_{\sigma} \text { ANT }\right) \otimes p_{\sigma}}{<\iota x \cdot \operatorname{deig}(x), \iota x \cdot \operatorname{deig}(x)>:}$
$\begin{array}{ll}\lambda x . \lambda s . \operatorname{sever}(\mathbf{h}, x, s): & <\iota x . \operatorname{deig}(x), \iota x . \operatorname{deig}(x)> \\ & (f \mathrm{OBJ})_{\sigma} \otimes(m \mathrm{SUBJ})_{\sigma}\end{array}$
$\begin{aligned} &<\lambda \operatorname{s.over}(\mathbf{h}, \iota x \cdot \operatorname{deig}(x), s), \iota x \cdot \operatorname{deig}(x)>: \\ & m_{\sigma} \otimes(f \text { ОВJ })_{\sigma}\end{aligned}$
@ DEPICTIVE
$<\lambda Q . \lambda e . \exists s . Q(e) \wedge \operatorname{over}(\mathbf{h}, \iota x \cdot \operatorname{deig}(x), s) \wedge s \supseteq e, \iota x . \operatorname{deig}(x)>: \quad m_{\sigma} \multimap\left(\left(\left(f_{\sigma} \mathrm{EV}\right) \multimap f_{\sigma}\right) \multimap\left(\left(f_{\sigma} \mathrm{EV}\right) \multimap f_{\sigma}\right)\right)$
$\left(\left(\left(f_{\sigma} \mathrm{EV}\right) \multimap f_{\sigma}\right) \multimap\left(\left(f_{\sigma} \mathrm{EV}\right) \multimap f_{\sigma}\right)\right) \otimes(f \mathrm{OBJ})_{\sigma}$
$\lambda e . \exists \operatorname{s.over}(\mathbf{h}, \iota x . \operatorname{deig}(x), s) \wedge \operatorname{heve}(\iota x . \operatorname{deig}(x), e) \wedge s \supseteq e:$
$\left(f_{\sigma} \mathrm{EV}\right) \multimap f_{\sigma}$

full
$\lambda P . \lambda x . \lambda s . P(x) \wedge \operatorname{full}(x, s)$
tank
$\lambda x . \operatorname{tank}(x, \mathbf{x} 2)$
$\left(t_{\sigma} \mathrm{VAR}\right) \multimap\left(t_{\sigma} \mathrm{RESTR}\right)$
onsenP
@ poss-NP
$\lambda P . \lambda s . \iota x . P(x, s):$ $\lambda P . P(\mathbf{x} \mathbf{1}):$
$\forall H\left(H \multimap t_{\sigma}\right) \multimap m_{\sigma}$

[^117]. Sall tank
Figure 4: Semantic derivation of med full tank

## References

Aa, Leiv Inge. 2006. Med som innleiande funksjonsord til småsetningar med og utan relativisering. Norsk Lingvistisk Tidsskrift pages 193-228.

Asudeh, Ash, Dalrymple, Mary and Toivonen, Ida. 2008. Constructions with Lexical Integrity: Templates as the Lexicon-Syntax Interface. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG08 Conference, CSLI.

Borthen, Kaja. 2003. Norwegian Bare Singulars. Ph. D.thesis, NTNU, Trondheim.
Dalrymple, Mary, Dyvik, Helge and King, Tracy Holloway. 2004. Copular Complements - Closed or Open? In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG04 Conference, CSLI Publications.

Haug, Dag. 2008. Tense and Aspect for Glue Semantics: The Case of Participial XADJs. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG08 Conference, CSLI.

Kortmann, Bernd. 1991. Free adjuncts and absolutes in English: problems of control and interpretation. London: Routledge.

Lødrup, Helge. 1999. Inalienables in Norwegian and Binding Theory. Linguistics 37(3), 365-388.

Partee, Barbara H. 1983/1997. Genitives - a Case Study. In Johan van Benthem and Alice ter Meulen (eds.), Handbook of Logic and Language, pages 464-470, Elsevier.

Pylkkänen, Liina. 2002. Introducing Arguments. Ph. D.thesis, Massachussets Institute of Technology, Cambridge, MA.

Sæbø, Kjell Johan. 2009. Possession and Pertinence: The Meaning of have. Natural Language Semantics 17(4), 369-397.

Sæbø, Kjell Johan, Fabricius-Hansen, Cathrine and Haug, Dag Trygve Truslew. Forthcoming. Closed Adjuncts. In Cathrine Fabricius-Hansen and Dag Trygve Truslew Haug (eds.), Big Events and Small Clauses.

# ADAPTING STOCHASTIC LFG INPUT FOR SEMANTICS 

Annette Hautli and Tracy Holloway King<br>Universität Konstanz<br>Microsoft

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

LFG c(onstituent)-structure and f(unctional)-structure analyses provide the detailed syntactic structures necessary for subsequent semantic analysis. The f-structure encodes grammatical functions as well as semantically relevant features like tense and number. The c-structure, in conjunction with the $\phi$-mapping, provides the information on linear precedence necessary for semantic scope and anaphora resolution. In this paper, we present a system in which a stochastic LFG-like grammar of English provides the input to the semantic processing. The LFG-like grammar uses stochastic methods to create a c-structure and a proto f-structure. A set of ordered rewrite rules augments and reconfigures the proto f-structure to add more information to the stochastic output, thereby creating true LFG f-structures with all of the features that the semantics requires. Evaluation of the resulting derived f-structures and of the semantic representations based on them indicates that the stochastic LFGlike grammar can be used to produce input to the semantics. These grammars combine the advantages of LFG structures, e.g. the explicit encoding of grammatical functions, with the advantages of stochastic systems, e.g. providing connected parses in the face of less-than-ideal input.


## 1 Introduction

LFG c(onstituent)-structure and f(unctional)-structure analyses provide the detailed syntactic structures necessary for subsequent semantic analysis (Dalrymple, 1999, 2001). The f-structure encodes grammatical functions as well as semantically relevant features like tense and number. The c-structure, in conjunction with the $\phi$-mapping, provides the information on linear precedence necessary for semantic scope and anaphora resolution.

LFG has also proven an excellent theory for use in computational linguistics due to its computational and mathematical tractability (Maxwell and Kaplan, 1989, 1993, 1996). Cross-linguistic theoretical and implementational work has resulted in large-scale LFG grammars for typologically varied languages (Butt et al., 1999, 2002). These LFG grammars face two challenges. First, some constructions may be outside the scope of the grammar. This can arise when the input is ungrammatical, e.g. contains typos, or when a construction is not covered by the grammar, e.g. the construction is too computationally costly or too rare to warrant inclusion in the grammar. Second, even highly efficient LFG implementations can be significantly slower than state-of-the-art stochastic parsers.

In this paper, we present a project where the output of a stochastic LFG-like grammar of English (Cahill et al., 2008) serves as input to the XFR semantic representation (Crouch and King, 2006), mapping f-structures into semantic representations. The XFR semantics is used for meaning-sensitive applications such as question answering (Bobrow et al., 2007) and search, expecting as input well-formed

[^118]LFG c- and f-structures as created by the English ParGram grammar (referred to here as the XLE grammar) which runs on the XLE LFG parser (Crouch et al., 2009). The stochastic LFG-like grammar, created at Dublin City University and referred to here as the DCU grammar, uses stochastic methods to create a c-structure and a proto f-structure (Cahill et al., 2002). These proto f-structures do not necessarily obey LFG's completeness and coherence conditions, especially when long-distance dependencies are involved, and do not have all of the f-structure features that the XLE grammar provides.

Therefore, we augmented and reconfigured the output of the DCU grammar and created a set of rewrite rules that add the information needed to obtain true LFG fstructures with all of the features that the semantics requires (Hautli, 2009). For most open class items, general rules could be used; other lexical items, such as pronouns and determiners, required more specific, lexicalized rules to create the appropriate $f$-structure facts.

The project results suggest that the proto LFG structures of stochastic grammars such as the DCU grammar can be used for meaning-sensitive applications. They combine the advantages of LFG structures, e.g. the explicit encoding of grammatical functions in f-structure, in conjunction with the advantages of stochastic systems, e.g. providing connected parses in the face of less-than-ideal input. The initial results also suggest that stochastic grammars producing the proto LFG structures can be used when no XLE LFG grammar is available but a treebank of the language is: in such situations, it can be faster to create a stochastic grammar instead of a rule-based one (Cahill et al., 2005). When an XLE LFG grammar does exist, the DCU grammar can be used in conjunction with the XLE LFG grammar to replace it in out-of-coverage sentences. The XLE LFG grammars produce distinctive fragment parses when sentences are out of coverage (Riezler et al., 2002); these can be replaced by the DCU proto f-structures to provide spanning c - and f -structures.

Section 2 provides an overview of the rule-based XLE grammar and the DCU stochastic LFG-like grammar. The rewrite rules that apply to the DCU structures to create XLE-style f-structures are described in section 3, with its evaluation following in section 4 . Section 5 discusses the approach and points to future work.

## 2 The Grammars

### 2.1 The English XLE Grammar

XLE is an efficient rule-based grammar development platform, developed by the Palo Alto Research Center (PARC). It consists of cutting-edge algorithms for parsing and generating Lexical-Functional Grammars, along with a user interface for writing and debugging such grammars (Crouch et al., 2009). The platform is also used in the ParGram project (Butt et al., 1999, 2002) for the development of parsers for several languages including Arabic, Chinese, German, French, Norwegian, Turkish, Urdu, and Welsh.

The English XLE LFG grammar is designed to handle well-edited English text (e.g. newspaper text, technical manuals) and is part of a larger system that maps text to an Abstract Knowledge Representation (AKR) (Bobrow et al., 2007) via the XLE parser and the XFR ordered rewrite system. The output of the system is used for applications such as search, question-answering (Bobrow et al., 2007), and redaction (Bier et al., 2009). The basic system pipeline is shown in Figure 1.


Figure 1: Pipeline architecture of the XLE-based system
The text is broken into sentences and words, using finite-state transducers (FST). The morphology analyzes each word and passes the information on to the broadcoverage XLE LFG grammar, which outputs c- and f-structure. These are then processed by the semantic and AKR rewrite rules.

Like all LFG grammars, the output of the syntax is a tree (c-structure) encoding linear order and constituency and an attribute-value matrix (f-structure) encoding predicate argument structure and semantically important features such as number and tense. These structures are more articulated than those usually found in LFG textbooks and papers because they contain all the features needed by subsequent processing and applications. Sample XLE c- and f-structures for The boys hopped. are shown in Figure 2. ${ }^{1}$

XLE outputs a packed representation of all possible solutions which allows subsequent processing to choose between different analyses for ambiguous sentences. In order for the grammar to be robust, XLE uses Optimality Theory marks (OT marks) in the syntax rules to indicate which analyses are dispreferred (Frank et al., 1998). In addition, the grammar can produce well-formed fragments if there is no analysis that spans the entire input (Riezler et al., 2002). The combination of these capabilities makes XLE robust in the face of ill-formed input and shortfalls in the coverage of the grammar.

[^119]

Figure 2: XLE output for The boys hopped.

On top of the syntactic c- and f-structure output, a semantic representation is derived using the XFR rewriting system. The semantic XFR system consists of rules that rewrite the syntactic structure to a semantic one, using external resources to replace words with concepts and grammatical functions with semantic roles (Crouch and King, 2006). In this paper, the XFR rewrite system is further used to rewrite DCU f-strctures to XLE-like f-structures (§3).

### 2.2 The DCU LFG-like parser

Extensive efforts at Dublin City University have resulted in the development of an automatic treebank annotation algorithm which annotates Penn-Treebank style trees (Marcus et al., 1994) with LFG f-structure information (Cahill, 2004). The annotated treebank can be used as a training resource for stochastic versions of unification and constraint-based grammars and for the automatic extraction of such resources (Cahill and McCarthy, 2002). The treebank is annotated such that solving the annotated functional equations produces LFG-like f-structures. The annotations describe what are called "proto-f-structures", which

- encode basic predicate-argument-modifier structures;
- may be partial or unconnected (i.e. in rare cases a sentence may be associated with two or more unconnected f-structure fragments rather than a single fstructure);
- may not encode some re-entrancies, e.g. in the case of wh- or other movement or distribution phenomena (e.g. of subjects into VP coordinate structures) (Cahill and McCarthy, 2002).

The basis of the annotation algorithm are treebank trees. These can either be Penn-II Treebank trees or, for novel text where there is no treebanked analysis, the output of a statistical parser such as that of Charniak (Charniak, 2000) or Bikel
(Bikel, 2002). After obtaining the tree, the nodes in the tree are annotated with f-structure equations. An example for Boys hopped. with the annotations is shown in Figure 3.

## Unannotated tree:

(S (NP (NNS Boys) ) (VP (VBD hopped)) (. .)

```
Annotated DCU tree:
(S
    (NP [up-subj=down]
            (NNS boys [up-pred='boy', up-num=pl,up-pers=3]))
        (VP [down-stmt_type=decl]
        (VBD hopped [up-pred='hop',up-tense=past])
    (. .))
```

Figure 3: Annotated DCU tree representation for Boys hopped.

After the annotation, all equations are percolated up the tree and unified at the topmost node. This process results in the f-structure in Figure 4.


Figure 4: DCU f-structure for Boys hopped.

Evaluating the DCU annotation algorithm against existing gold standards shows that it can outperform hand-crafted, wide-coverage constraint grammars. The current DCU system achieves an f-score of 82.73 against the PARC 700 Dependency Bank (King et al., 2003), compared to $80.55 \%$ for the hand-crafted XLE LFG parsing system (Cahill et al., 2008). However, there are two issues with the f-structures produced by the DCU gramamrs. First, the PARC 700 Dependency Bank has a reduced feature set and contains only a subset of the features that are found in the very detailed ParGram f-structures. Therefore, the XFR semantics would fail due to missing f-structure features. Second, many of the features are present in the DCU f-structures in a different form than those of the XLE ones, and so they must be reformatted in order for the semantics to process them. Both of these problems will be illustrated in the next section.

To summarize, with the DCU LFG-like grammar and the XLE grammar, we have two different approaches to obtaining LFG analyses. On the one hand, the rule-based XLE grammar has very detailed feature structures, but faces coverage issues. On the other hand, the stochastic DCU grammar has the drawback of a less detailed f-structure, but with more connected parses. In this project we aimed to combine the advantages of both approaches.

## 3 Hybridization with XFR Augmentation Rules

The reasons for hybridizing the XLE-based system (Figure 1) are two-fold. In the case of English, the language used in this project, the stochastic grammar can be used in place of the rule-based grammar for out-of-coverage sentences, thereby supplying more connected input to the semantics. In the case of other languages, if no rule-based grammar is available but a treebank of the language is, it can be used to create a stochastic grammar for that language (Cahill et al., 2005).

### 3.1 The Overall Architecture

To produce the full, detailed f-structures needed by the semantics, we apply XFR rules to map DCU proto f-structures to XLE-style f-structures. The XFR ordered rewrite rules consume a set of input facts and replace it with another set of facts (§3.2). These rules can create a link between the stochastic DCU grammar and the rule-based XLE grammar output. The system using the DCU output as input for the XLE semantics is shown in the pipeline in Figure 5.


Figure 5: Hybridized pipeline

First, a sentence breaker splits running text into sentences, which are then processed by a probabilistic treebank based parser (PTBP) and annotated by the DCU annotation algorithm (§2.2). The DCU proto f-structure output is then reformatted by a script in order to be compatible with the input format expected by the XFR system. After that, the XFR ordered rewrite rules are applied to create XLE-style fstructures. In the final step, the rewritten f-structures are fed into the XFR semantic rules.

### 3.2 The XFR Rules

Input to the system is a set of facts representing the f -structures obtained by the DCU parser and the output is a set of rewritten facts representing the full f-structures that are fed into the XFR semantic system. The XFR system operates on a source f -structure and transforms it incrementally into the target structure. The order of the rules is important because each rule has the potential to change the set of input
facts that the subsequent rules will encounter: rules can prevent following rules from applying by removing facts that they would otherwise have applied to; they can also enable the application of later rules by introducing facts that these rules require. See the XLE documentation (Crouch et al., 2009) for details of the XFR system rule notation.

The rewriting works as follows: if a set of f-structure features (or part of an f -structure) matches the left-hand side of a rule, then the rule applies to produce the features on the right-hand side of the rule. A XFR rule which rewrites the DCU proto f-structure for the subject boys is shown in Figure 6.

$$
\text { input: }\left[\text { subj }\left[\begin{array}{ll}
\text { pred } & \text { boy } \\
\text { num } & \text { pl } \\
\text { pers } & 3
\end{array}\right]\right]
$$

## XFR rule

```
subj(%X,%Subj), pred(%Subj,%Pred), num(%Subj,%Num), pers(%Subj,3)
==>
SUBJ(%X,%Subj), PRED(%Subj,%Pred),
NUM(%Subj,%Num), PERS(%Subj,3),
NTYPE(%Subj,%Ntype), NSYN(%Ntype,common),
NSEM(%Ntype,%Nsem), COMMON(%Nsem,count).
```

output: $\left.\left[\begin{array}{lll}\text { SUBJ }\left[\begin{array}{ll}\text { PRED } & \text { 'boy' }^{\prime} \\ \text { NTYPE } & {\left[\begin{array}{ll}\text { NSEM }[\text { COMMON } & \text { count }\end{array}\right]} \\ \text { NSYN common }\end{array}\right] \\ \text { NUM } & \mathrm{pl} \\ \text { PERS } & 3\end{array}\right]\right]$

Figure 6: Rewriting of the noun boys

The XFR rule in Figure 6 works as follows. The material before the arrow (==>) contains the input f-structure facts which must be matched for the rule to apply. The material after the arrow contains the f-structure facts created by the application of the rule. Forms beginning with a percent sign (\%) are variables. For example, in Figure 6, the variable $\% \mathrm{X}$ is the f-structure which contains a sub $j$; that subj is then referred to by the variable $\%$ Subj. This $\%$ Subj f-structure must have pred attribute with value $\%$ Pred and a num attribute with value $\%$ Num in order for the XFR rule to match.

Given the input f-structure in Figure 6, the left-hand side of the rule goes through the list of XFR facts and matches with the subj fact, whose pred argument has the value boy and also matches the subject's num and pers attributes with their values. The rule rewrites these facts to those on the right-hand side of
the rule, resulting in the output f-structure shown in Figure 6. This is a very simple example of an ordered rewrite rule but the principle remains the same for more complicated constructions.

In total, the XFR system mapping from DCU to XLE f-structures consists of 162 rewrite rules.

In the remainder of this section we discuss several classes of issues that arose: the correction of core predicate-argument structures which did not conform to the XLE analysis or which were simply incorrect; the addition of default values which are necessary for the semantics; and the lexicalization of rules to provide features for particular predicates.

### 3.2.1 Core Predicate-Argument Structure

Of primary importance was correcting syntactic constructions where the core predi-cate-argument structure provided in the DCU f-structure differed from that in the XLE, and often in the theoretical LFG, analysis.

There were a few places where the original DCU analysis did not capture functional control as a re-entrant f-structure. For example, in the DCU analysis of sentences like This seems to be a post-1990 problem., what traditional LFG analysis would consider a functionally controlled subject was represented only once within the f -structure, as the subject of the matrix verb. In order to correctly represent functional control, the identity relation between controller and infinitival subject is done by creating a re-entrant f-structure for the controlled SUBJ under the XCOMP. The creation of functional control structures is shown in Figure 7.



Figure 7: Predicate-Argument Structure: Creation of functional control for This seems to be a post-1990 problem.

A second set of phenomena where the core predicate argument structure had to be changed was imperatives and certain participial constructions. These structures lack subjects in the original DCU f-structures. This occurs because the f-structures produced by the DCU parser are not subject to the LFG completeness requirement whereby all arguments of a predicate must be present in the f-structure, even if they are not realized in the c-structure. These constructions were identified and the appropriate subject information was provided. An example is provided in Figure 8.

$$
\begin{aligned}
& -1\left[\begin{array}{l}
\text { adjunct }[1 \text { pred appear, stmt_type declarative, tense pres } \\
\text { pred }]
\end{array}\right.
\end{aligned}
$$

Figure 8: Predicate-Argument Structure: Insertion of a subject for Do not appear

### 3.2.2 Adding Default Values: Verbs

The addition of default features for f -structures was one of the most important tasks of the XFR mapping. This was particularly the case for verbs which govern many semantically-relevant features such as tense, mood, and aspect (TNS-ASP) and subcategorization frame information. All of these are used by the semantics rules and so must be present in the input f -structure.

The features such as TENSE, VTYPE, MOOD and PASS IVE posed a challenge, because of feature sparseness in the DCU structures. For example, the DCU grammars provide a passive + feature when the verb is passive, but no passive feature when the verb is active. In the augmentation rules, these missing features and values are provided by judicious use of rules inserting default values.

The system also adds subcategorization features. The XLE grammar's verb lexicon has almost 9,800 verb stems with an average of 2.8 subcategorization frames each. ${ }^{2}$ Subcategorization features are essential for the semantic lexical look-up that aids in mapping the verb's arguments to thematic roles in the semantics. The features dealing with subcategorization include the core argument structure of the PRED and a feature encoding the subcategorization frame name, which is not part of the DCU f-structures.

[^120]An example for the sentence He pushes it. is shown in Figure 9. Notice the insertion of negative values for PASSIVE, PERF, and PROG as well as insertion of the _SUBCAT-FRAME feature.

Figure 9: Default Values: Augmentation of verb-related features for He pushes it.

### 3.2.3 Lexicalization: Nouns and Pronouns

In some cases, more specific lexically-based information had to be provided in the f -structure. This arose when the semantics depends on f-structure features which are present in the XLE structures but not the DCU ones and which are not predictable from the general syntactic configuration.

For example, in the DCU grammar, proper nouns are correctly identified as such, but are not categorized by type. The morphology used in the XLE parser types many proper nouns (e.g. locations (Detroit, France), organizations (IBM, Congress), people (Mary, Smith), and gender for first names (Mary vs. John)). Such information is valuable for the semantic interpretation, especially for anaphora resolution and more accurate concept look-up. For this project, we extracted this information from the morphology and incorporated it into the XFR rules.

Similarly, many time-related nouns, such as months, days, and seasons, which the semantics expects to have identified with special f-structure date/time features, were lexicalized in the XFR rules, as the DCU output did not distinguish them from other nouns. However, all other nouns are accounted for by a general rule for modifying and inserting common noun features and rewritten accordingly.

An example for the rewriting of proper nouns (i.e. Masha) and time expressions (i.e. fall) is shown in Figure 10.

As a final example of lexicalization, personal, possessive, demonstrative, interrogative, and relative pronouns have to be mapped individually based on the lexical item due to the lack of relevant features on the DCU side. Examples of rewritten



Figure 10: Lexicalization: Rewriting of proper nouns and time expressions for Masha is going there next fall.
pronouns were shown in Figure 9 for the personal pronouns he and it. In particular, note the addition of the PRON-TYPE, PERS, GEND-SEM, and HUMAN features. ${ }^{3}$

## 4 Development and Evaluation

To develop the XFR augmentation rules that map from the DCU proto f-structures to XLE-style f-structures, we created a testsuite of 430 sentences which covered core and some peripheral syntactic phenomena in English and which was based on the testsuites used by the XLE grammar developers to test syntactic coverage. For example, the testsuite covers syntactic phenomena such as extraposition, extraction, gerunds, sentential subjects and different clause types (declarative, interrogative and imperative sentences).

A schematic overview of the system development and evaluation is provided in Figure 11. The development and evaluation sentences were parsed by the XLE parser to obtain full f-structures. The same sentences were parsed by the DCU parser, creating proto f -structures. In the next step, augmentation ordered rewriting

[^121]

Figure 11: Overview of development (and evaluation)
rules reformatted and rewrote the DCU f-structures. The resulting f-structures were compared to those produced by the xle parser using XLE's triples match (Crouch et al., 2009). The same process was done at the level of the semantic representations.

### 4.1 Evaluation Measures

To compare the f-structures and semantic representations, we use the standard evaluation measures $f$-score, precision and recall, which compare the features of the representations to be evaluated in relation to the features of a (often gold standard) reference set. ${ }^{4}$ In this project, precision measures how many features in the transferred DCU f -structures are correct, whereas recall focuses on the completeness of the transferred DCU structures. Here precision is usually higher than recall; this is true for f-structures and for semantic representations. However, recall is the more important measure here as it shows how complete the transferred DCU structures are in relation to the original XLE structures.

### 4.2 Syntactic F-structure Evaluation

The f-structures in the transferred DCU f-structures were compared to those in the xLE f-structures. The results are shown below for declaratives, interrogatives, and imperatives.

[^122]F-Structure Matching of Declaratives The matching of declaratives varies quite a bit, depending on how many proper nouns are included in the sentence. ${ }^{5}$ Expressions like Eiffel Tower and President George W. Bush initially get a proper noun feature from the DCU parser, but no information as to what kind of proper noun they are (e.g. location, person, etc.). In order to insert this information, one would have to list every item in the rewrite rules, something which cannot be done in the general case for unknown text. This is why sentences with proper nouns usually have a lower matching score.

Another difference between the DCU and XLE grammars is that the DCU parser always treats hyphenated forms as single units. For example, in the noun phrase a high-interest loan, the head noun loan has a single modifer high-interest in the DCU analysis. This loses certain semantic relationships which are needed for the semantic matching.

| precision | recall | f-score |
| :---: | :---: | :---: |
| 72.63 | 65.61 | $\mathbf{6 8 . 9 4}$ |

Table 1: Matching results for declaratives with proper nouns
If we consider sentences that do not contain proper nouns (e.g. I'll go. or He laughed every third year.), the results are as in Table 2. Note the significantly higher scores.

| precision | recall | f-score |
| :---: | :---: | :---: |
| 87.13 | 82.67 | $\mathbf{8 4 . 8 4}$ |

Table 2: Matching results for indicatives without proper nouns
Many clauses have a perfect f-score of 100 , but this is countered by issues concerning coordination and the correct assignment of adjuncts in other sentences.

F-Structure Matching of Interrogatives A major issue is the DCU parsing of interrogative clauses. The training data for the DCU grammar is a corpus from the Wall Street Journal, which does not contain many matrix interrogatives. Due to the lack of training data, interrogatives are often analyzed incorrectly, e.g. the subject of the sentence is often analyzed as an object. Any mismatch in grammatical functions is a serious issue for the semantic processing. Judge et al. (2006) propose a method to add more interrogatives to the training data to alleviate this problems by building a QuestionBank. This bank consists of a corpus of 4,000 annotated questions used to train parsers in question answering technology and the evaluation of question parsing. Unfortunately, this DCU parser option was not available for this project.

[^123]Thirty-one sentences of the development data were interrogative sentences. Matching the rewritten interrogatives against the original XLE interrogatives gives the results in Table 3. The reason for the relatively low matching figures is that these clauses get incorrect analyses due to the lack of interrogative sentences in the DCU training data.

| precision | recall | f-score |
| :---: | :---: | :---: |
| 45.17 | 43.15 | $\mathbf{4 4 . 1 3}$ |

Table 3: Matching results for interrogatives

F-Structure Matching of Imperatives As with interrogatives, imperatives are relatively rare in the Wall Street Journal corpus used to train the DCU parser. Our development set contains 25 imperatives. The figures for the matching of imperatives are higher than those for interrogatives, as more features could be added by the XFR rules and the analyses in general are closer to the XLE f-structures.

| precision | recall | f-score |
| :---: | :---: | :---: |
| 61.83 | 49.10 | $\mathbf{5 4 . 7 4}$ |

Table 4: Matching results for imperatives

### 4.3 Semantic Representation Evaluation

In order to determine whether enough information is included in the f-structures for them to be input to the semantic representations used by meaning-sensitive applications, we prepared an evaluation of semantic representations of 66 queries and answers, chosen for their coverage of phenomena of interest to the semantics (e.g. negation, anaphora). ${ }^{6}$ These were parsed by both the DCU and XLE parser, then the DCU f-structures were processed by the XFR rules. Both XLE and DCU f -structures were used as input to the semantic system, and the resulting semantic representations were compared against each other. An example for a passage and query (and answer) is the following:
(1) P: Although Mary likes vegetables she eats them raw.

Q; Does Mary like vegetables?
A: Yes

The matching figures for the semantic representations of the passage and the query are shown in Table 5.

[^124]| precision | recall | f-score |
| :---: | :---: | :---: |
| 64.04 | 60.27 | $\mathbf{6 2 . 1 0}$ |

Table 5: Query-passage pair match results for the semantic representation

The overall figures for the semantic matching are lower than those in $\S 4.2$ for the syntactic matching due to the large number of interrogatives. However, on the positive side, some of the passage sentences are significantly longer than those in the development set, which had been chosen as representative of isolated syntactic phenomena. The ability to correctly process these data provides evidence that the hybrid system works on data which combines simple syntactic constructions into the complex sentences found in naturally occurring text.

## 5 Future Work and Discussion

Given the initial positive results of the project, the next step is to build a fully integrated hybrid DCU-XLE system (Figure 5) that can be run over large corpora and compare the results with those of the standard XLE system (Figure 1). Of particular importance is the behavior of the hybrid DCU-XLE system in application contexts. Having such a system raises some issues that were unimportant in the initial project. We address two of these here. We first discuss the issues arising from the different treatment of ambiguity in the two systems. We then discuss efficiency: back-of-the-envelope calculations show that the two systems should be roughly similar in efficiency, but this remains to be tested empirically.

### 5.1 Ambiguity

The XLE LFG grammar can efficiently produce multiple analyses for a given sentence (Maxwell and Kaplan, 1991). A maximum entropy model is applied to the output of the grammar to rank the parses (Riezler et al., 2002) and an n-best subset of the parses is then passed to the semantics. The more parses that are passed forward, the more processing that the semantics and AKR rules must perform, although the impact of this is mitigated by the ability of the XFR system to operate on the packed structures produced by the XLE grammar (Crouch, 2005). In fact, the XFR system uses the same packing mechanism and code that the XLE parser does. For meaning sensitive applications, the n-best, instead of the single best, parses are used in order to increase the chances that the correct parse is available.

On the DCU side, our system used the single parse produced by the DCU grammar. In theory, it would be possible to obtain ranked output from the DCU parser, e.g. by taking the n-best trees produced by the PTBG. In order for the semantics to operate on them efficiently, these parses would have to be packed. However, packing unpacked input can be difficult and inefficient. As such, the hybrid DCU-XLE approach seems best suited for applications and situations where a single parse
provides sufficient information. Search, as opposed to question-answering, is one possible application of this type.

### 5.2 Efficiency

The efficiency of the hybrid DCU-XLE approach was not systematically explored. The XLE system can process sentences in documents with an average of $\sim 20$ words per sentence (e.g. Penn Treebank WSJ sentences) at less than a second from text to semantic output. Half of the time is spent on the syntax (i.e. creating the fstructure). ${ }^{7}$ The exact percentage of time spent on the parsing step depends on how many parses are passed forward to the semantics rules: when more parses are passed forward, the processing by the XFR rules slows.

XLE has a number of performance variables that can be set to trade speed for accuracy (Crouch et al., 2009). The one-second-a-sentence results use relative aggressive settings with the result that $\sim 1.1 \%$ of the sentences time out or run out of memory.

This version of the XLE grammar uses c-structure chart pruning to trim the context-free c-structure forest before unification (Crouch et al., 2009). C-structure pruning eliminates a subtree if there is another subtree analysis available and if the subtree is significantly less probable than the most-probable subtree. The chart pruner uses a simple stochastic CFG model where the probability of a tree is the product of the probabilities of each of the rules used to form the tree. The probability of a rule is basically the number of times that that form of the rule occurs in the training data divided by the number of times the rule's category occurs in the training data, plus a smoothing term. If a subtree's probability is lower than the best probability by a given factor, then the subtree is pruned. This approach ensures that there is always at least one tree and that only highly improbable subtrees are eliminated. The resulting c-structure forest is often still very large, but it is often significantly smaller than the original one. Using c-structure pruning speeds the XLE parser by $\sim 40 \%$ for English, while maintaining accuracy.

The DCU parser runs with a similar level of efficiency and hence should not significantly change the speed of the overall system. In parsing the British National Corpus (BNC) (Wagner et al., 2007), which has an average sentence length of 18 words, the PTBG, annotation, and unification took an average of 1.48 seconds per sentence. ${ }^{8}$ This longer per-sentence parse time is somewhat misleading because the parser in the DCU project in Wagner et al. (2007) was configured to provide analyses for all sentences, no matter how long, complex, or grammatical; if the occasional missed analysis is acceptable for a given application, more efficient processing settings can be used.

[^125]The XFR rules used to map from the DCU output to the semantics input are relatively few in number and add a negligible amount of time to the processing.

### 5.3 Conclusion

This paper reported on a project to use a stochastic parser (the probabilistic DCU parser) that produces proto f-structures as the input to a semantic parser, in the place of a rule-based LFG parser (the XLE parser). The f-structures were augmented using a set of ordered rewrite XFR rules similar to the rules that create semantic structures. When evaluating the DCU-based system against XLE output on the f-structure level, the results are promising in that the DCU-based f-structures can be used by the semantics to produce well-formed semantic structures. This provides the opportunity to build hybrid systems using different grammar versions depending on their ability to parse the input data. The disadvantages of the DCU parser, which assigns fewer features to the proto f-structures, can be overcome by the XFR rules providing full LFG structures with detailed syntactic and semantic features identical to those produced by XLE LFG grammar.

As more researchers wish to build meaning-sensitive applications on top of ParGram-style XLE grammars, our work suggests that hybrid systems can be built using DCU grammars for the syntactic processing step (e.g. for Spanish for which there is a DCU ParGram grammar but no XLE one).

## References

Bier, Eric, Chow, Richard, Gollé, Philippe, King, Tracy Holloway and Staddon, Jessica. 2009. The Rules of Redaction: Identify, Protect, Review (and Repeat). In IEEE Volume on Access Control, pages 46-53, IEEE Computer and Reliability Societies.

Bikel, Daniel M. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In The Proceedings of HLT 2002.

Bobrow, Daniel G., Cheslow, Bob, Condoravdi, Cleo, Karttunen, Lauri, King, Tracy Holloway, Nairn, Rowan, de Paiva, Valeria, Price, Charlotte and Zaenen, Annie. 2007. PARC's Bridge and Question Answering System. In Grammar Engineering Across Frameworks, pages 46-66, CSLI Publications.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In Proceedings of COLING2002, Workshop on Grammar Engineering and Evaluation, pages 1-7.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. A Grammar Writer's Cookbook. CSLI Publications.

Cahill, Aoife. 2004. Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations. Ph. D.thesis, School of Computing, Dublin City University.

Cahill, Aoife, Burke, Michael, O’Donovan, Ruth, Riezler, Stefan, van Genabith, Josef and Way, Andy. 2008. Wide-Coverage Deep Statistical Parsing Using Automatic Dependency Structure Annotation. Computational Linguistics 34(1), 81-124.

Cahill, Aoife, Forst, Martin, Burke, Michael, McCarthy, Mairéad, O'Donovan, Ruth, Rohrer, Christian, van Genabith, Josef and Way, Andy. 2005. TreebankBased Acquisition of Multilingual Unification Grammar Resources. Journal of Research on Language and Computation; Special Issue on "Shared Representations in Multilingual Grammar Engineering" pages 247-279.

Cahill, Aoife and McCarthy, Mairéad. 2002. Automatic Annotation of the Penn Treebank with LFG F-Structure Information. In Proceedings of the LREC Workshop on Linguistic Knowledge Acquistion and Representation: Bootstrapping Annotated Language Data, LREC Workshop on Linguistic Knowledge Acquisiton and Representation.

Cahill, Aoife, McCarthy, Mairéad, van Genabith, Josef and Way, Andy. 2002. Parsing with PCFGs and Automatic F-Structure Annotation. In LFG02 Proceedings, pages 76-95, CSLI On-line Publications.

Charniak, Eugene. 2000. A Maximum Entropy-inspired Parser. In Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics, pages 132-139.

Crouch, Dick. 2005. Packed Rewriting for Mapping Semantics to KR. In Proceedings of the International Workshop on Computational Semantics.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ron, King, Tracy, Maxwell, John and Newman, Paula. 2009. XLE Documentation, http://www2.parc.com/isl/groups/nltt/xle/doc/.

Crouch, Richard and King, Tracy Holloway. 2006. Semantics via F-structure Rewriting. In LFG06 Proceedings, CSLI On-line Publications.

Dalrymple, Mary (ed.). 1999. Semantics and Syntax in Lexical Functional Grammar. The MIT Press.

Dalrymple, Mary. 2001. Lexical Functional Grammar (Syntax and Semantics). Academic Press.
de Paiva, Valeria and King, Tracy Holloway. 2008. Designing Testsuites for Grammar-based Systems in Applications. In Proceedings of the Coling 2008 Workshop on Grammar Engineering Across Frameworks, pages 49-56.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas and III, John T. Maxwell. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In LFG98 Proceedings, CSLI On-line Publications.

Hautli, Annette. 2009. Adapting Stochastic Output for Rule-Based Semantics. Masters Thesis, University of Konstanz.

Judge, John, Cahill, Aoife and van Genabith, Josef. 2006. QuestionBank: Creating a Corpus of Parse-Annotated Questions. In Proceedings of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics, pages 497-504.

King, Tracy Holloway, Crouch, Richard, Riezler, Stefan, Dalrymple, Mary and Kaplan, Ron. 2003. The PARC700 Dependency Bank. In Proceedings of the EACL03: 4th International Workshop on Linguistically Interpreted Corpora (LINC-03).

Marcus, Mitchell, Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen and Schasberger, Britta. 1994. The Penn Treebank: Annotative Predicate Argument Structure. In ARPA Human Language Technology Workshop.

Maxwell, John T. and Kaplan, Ronald M. 1989. An Overview of Disjunctive Constraint Satisfaction. In Proceedings of the International Workshop on Parsing Technologies, pages 18-27.

Maxwell, John T. and Kaplan, Ronald M. 1991. A Method for Disjunctive Constraint Satisfaction. Current Issues in Parsing Technologies .

Maxwell, John T. and Kaplan, Ronald M. 1993. The Interface between Phrasal and Functional Constraints. Computational Linguistics 19, 571-590.

Maxwell, John T. and Kaplan, Ronald M. 1996. Unification-based Parsers that Automatically Take Advantage of Context Freeness. In LFG96 Conference.

Melamed, I. Dan, Green, Ryan and Turian, Joseph P. 2003. Precision and Recall of Machine Translation. In Proceedings of HLT/NAACL.

O’Donovan, Ruth, Burke, Michael, Cahill, Aoife, van Genabith, Josef and Way, Andy. 2005. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II and Penn-III Treebanks. Computational Linguistics 31, 329-366.

Riezler, Stefan, King, Tracy Holloway, Kaplan, Ron, Crouch, Dick, III, John T. Maxwell and Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 271-278.

Wagner, Joachim, Seddah, Djamé, Foster, Jennifer and van Genabith, Josef. 2007. C-Structures and F-Structures for the British National Corpus. In LFG07 Proceedings, CSLI On-line Publications.

# INTERMEDIARY AGENTS AND UNEXPRESSED PRONOUNS 

Anna Kibort<br>University of Cambridge and Surrey Morphology Group

Proceedings of the LFG09 Conference Miriam Butt and Tracy Holloway King (Editors) 2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

The aim of this paper is to show the full range of possible participant-function mappings available for the classes of verbs in Polish which denote predicates entailing an 'intermediary agent'. An intermediary agent is a semantic participant that can be conceptualised as an instrument or means with which the event is accomplished, or alternatively as the causer or instigator of this event. The particular verb classes involved include verbs of emission of smell, sound, or light, verbs expressing expansion of an aggregate or a mass/abstract entity (corresponding roughly to the English SWARM verbs), and verbs expressing physical or psychological states due to a stimulus which can be interpreted as an intermediary agent. I discuss how to model the identified alternations with LMT and offer argument structure models of all the variants. I argue that a certain type of clause which is often regarded as impersonal (due to the lack of a lexically expressed nominative subject, as well as the defocusing of the instigator) can be analysed as having a 'pro-drop' subject (an unexpressed/incorporated pronoun or pronominal inflection) which may co-refer with an overtly expressed instrument or other oblique argument.


## 1 The set of constructions under consideration

I begin with a discussion of the class of verbs in Polish which includes verbs of emission of smell, sound, or light - examples of which are given in (1)-(3), respectively; and verbs expressing expansion of an aggregate or a mass/abstract entity - examples of which are given in (4). ${ }^{1}$ Some of the verbs are reflexiva tantum, and others are reflexive variants of non-reflexive verbs which are reflexive when used inchoatively:
(1) a. pachnieć
b. śmierdzieć c. cuchnqć
(2)
a. grzmieć
b. szumieć
c. huczeć
(3) a. mienić się
b. bielić się
c. migotać 'glitter, shimmer'
(4) a. roić się 'swarm, teem'
b. kipieć 'seethe, effervesce'
c. pęcznieć 'swell, bulge'
d. mrowić się 'teem, swarm'
e. wrzeć 'seethe, throb'
f. przelewać się 'overflow'

These predicates can be thought of as denoting events that typically involve two entities as participants. One is the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept. The other entity is the location in which the event takes place, where the event is present and/or propagated.

It appears that in Polish the events in question can be conceptualised in three different ways, resulting in three different syntactic constructions forming a set of so-called 'alternations'. Argument alternations have been extensively discussed in syntactic literature since the beginning of generativism, and the work of Rappaport and Levin (1988), Pinker (1989), and Jackendoff (1990), has been particularly influential in formalising the differences between the semantic contents of the alternants. Dowty's (1991) theory of proto-roles attributes the different argument configurations to the different entailments produced by the related predicates, and Dowty (2000) offers an extensive discussion of the differences in the meanings between the English alternants involving swarm and spray/load verbs. The work presented in this paper follows from this tradition and assumes that the different syntactic frames correlate with different meanings, not only of the verbs themselves (resulting, for example, in the holistic vs partitive effect of the alternation), but also of the participants in the events denoted by the verbs. Hence, while the entities referred to by the arguments may be the same between the alternants, the semantic roles a particular entity fulfils in the different alternants may be different. This last distinction corresponds to Jackendoff's

[^126]functional representation of arguments at 'action tier' and the representation of their conceptual roles at 'thematic tier' (1990:126ff).

Since it would be difficult to talk about 'agentivity' of any participants in the events discussed here, while referring to the semantic roles entailed by the predicates I follow Siewierska (2008:121) in using the term 'instigator' for the causal participant of an event the most broadly.

Apart from sharing some semantics, the verbs listed above can be identified as belonging to one class due to their participation in a particular set of syntactic alternations, which results in their use in the following three constructions.

### 1.1 The oblique place + oblique emitter construction

First, they are commonly used in a syntactic frame where the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept is expressed through an instrumental nominal or other oblique (a prepositional phrase). There is no overt lexical element realising a nominative subject, and the verb bears the 3 SG.N inflection. This type of clause commonly includes an optional locative which is often topicalised. However, in this syntactic frame the instrumental or prepositional phrase expressing the emitter is also optional.

I refer to this syntactic frame as the 'oblique place + oblique emitter construction', even though I reserve judgement on the question of whether they are both arguments of the predicate, or whether the location might be an adjunct: ${ }^{2}$

b. Śmierdziało moczem $w$ catym korytarzu. smelt.3SG.N urine(M).INS in whole corridor 'There was a smell of urine in the whole corridor.'
a. Na forach grzmiało od głosów niezadowolenia.
on forums roared.3SG.N from voices(NONVIR) ${ }^{3}$.GEN discontent(N).GEN
'[Internet] forums were roaring with voices of discontent.'
b. W głowie szumiało od muzyki.
in head throbbed.3SG.N from music(F).GEN
'The [my/his/her] head was throbbing with music.'
a. Na ulicach mienito sie od światecznych dekoracji.
on streets glistened.3SG.N REFL from festive.PL.GEN decorations(NONVIR).GEN
'The streets glittered with festive decorations.'
b. W ogrodzie bieli
in garden appear-white.3SG.(N) REFL from hoarfrost(M).GEN
'The garden is glistening with hoarfrost.'
a. W ogrodzie roiło się od pszczót. in garden swarmed.3SG.N REFL from bees(NONVIR).GEN 'The garden was swarming with bees.'
b. W głowach kipiało nam od pomysłów. in heads seethed.3SG.N us.DAT from ideas(NONVIR).GEN
'Our heads were seething with ideas.'
c. W sercu pęczniało od gniewu.
in heart swelled.3SG.N from anger(M).GEN
'The [my/his/her] heart was swelling with anger.'

[^127]This is a common construction in Polish and the naturally occurring clauses may display different word orders from the ones illustrated above, different collocations, and include additional lexical material. However, the reason why I selected the particular examples above for illustration is that they allow me to demonstrate the alternations available to these predicates with the minimum number of lexical elements and minimal pragmatic adjustments to improve their felicitousness.

### 1.2 The subject place + oblique emitter construction

The second syntactic frame in which these predicates can be found involves the location expressed via a nominative subject. The predicate agrees with the subject, while the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept is expressed through an instrumental nominal or other oblique (a prepositional phrase) as in (5)-(8). I will refer to this syntactic frame as the 'subject place + oblique emitter construction': ${ }^{4}$
(9) a .

Dom pachnie kawq.
house(M).NOM emit-fragrance.3SG.(M) coffee(F).INS
'The house smells of coffee.'
b. Caly korytarz śmierdziat moczem.
whole.M.NOM corridor(M).NOM smelt.3SG.M urine(M).INS
'The whole corridor smelt of urine.'
(10) a. Fora
grzmiaty od głosów niezadowolenia.
forums(NONVIR).NOM roared.3PL.NONVIR from voices(NONVIR).GEN discontent(N).GEN
'[Internet] forums were roaring with voices of discontent.'
b. Glowa szumiata od muzyki.
head(F).NOM throbbed.3SG.F from music(F).GEN
'The [my/his/her] head was throbbing with music.'
(11)

Ulice mienity się od światecznych dekoracji. streets(NONVIR).NOM glistened.3PL.NONVIR REFL from festive.PL.GEN decorations(NONVIR).GEN 'The streets glittered with festive decorations.'
b. Ogród bieli się od szronu. garden(M).NOM appear-white.3SG.(M) REFL from hoarfrost(M).GEN 'The garden is glistening with hoarfrost.'
(12) a. Ogród roit się od pszczót. garden(M).NOM swarmed.3SG.M REFL from bees(NONVIR).GEN 'The garden was swarming with bees.'
b. Glowy kipialy nam od pomystów. heads(NONVIR).NOM seethed.3SG.NONVIR us.DAT from ideas(NONVIR).GEN 'Our heads were seething with ideas.'
c. Serce peczniało od gniewu.
heart(N).NOM swelled.3SG.N from anger(M).GEN
'The [my/his/her] heart was swelling with anger.'

### 1.3 The subject emitter + oblique place construction

Finally, one more alternation available to these predicates, resulting in a third type of syntactic frame, has the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept expressed through a nominative subject. The predicate agrees with the subject, and - if felicitous - the location can be expressed as an optional locative: ${ }^{5}$

[^128](13) a. Ta kawa pięknie pachnie w calym domu.
this coffee(F).NOM beautifully emit-fragrance.3SG.(F) in whole house
'This coffee smells beautifully in the whole house.'
b. Mocz śmierdziat $w$ catym korytarzu. urine(M).NOM smelt.3SG.M in whole corridor 'The urine smelt in the whole corridor.'
(14) a. Na forach grzmialy głosy niezadowolenia. on forums roared.3PL.NONVIR voices(NONVIR).NOM discontent(N).GEN
'On [internet] forums were roaring voices of discontent.'
b. Muzyka szumiata w głowie. music(F).NOM throbbed.3SG.F in head
'The music was throbbing in the [my/his/her] head.'
(15) a. Na ulicach mienity się świateczne dekoracje. on streets glistened.3PL.NONVIR REFL festive.NONVIR.NOM decorations(NONVIR).NOM 'On the streets glittered festive decorations.'
b. Szron bieli się w ogrodzie.
hoarfrost(M).NOM appear-white.3SG.(M) REFL in garden
'Hoarfrost is glistening in the garden.'
(16) a. W ogrodzie roity sie pszczoly.
in garden swarmed.3PL.NONVIR REFL bees(NONVIR).NOM
'In the garden were swarming bees.'
b. W głowach kipiaty nam pomysty. in heads seethed.3PL.NONVIR us.DAT ideas(NONVIR).NOM
'In our heads were seething [new] ideas.'
c. Gniew pecznial w sercu. anger(M).NOM swelled.3SG.M in heart 'Anger was swelling in the [my/his/her] heart.'

## 2 Modelling alternations at argument structure

I assume that the three constructions are related, that is, that they share the base verbal lexeme, and that the relations between the three variants of the lexeme are best captured at the level of argument structure. In the remainder of the paper, I provide argument structure models for all three of them.

A follow-up question pertinent to the first construction, the oblique place + oblique emitter one, is whether it is indeed impersonal as is often assumed. It evidently lacks a lexically expressed nominative subject, by which it fulfils a structural criterion of impersonality; and it defocuses the instigator, by which it fulfils the key functional criterion of impersonality (Siewierska 2008:116, 121-122). However, Polish is a pro-drop language, and applying these criteria to pro-drop languages can be tricky, as we would obviously not want to analyse all basic pro-drop clauses with omitted lexical (pronominal) subjects as impersonal.

In section 3 below I argue for a pro-drop analysis of the oblique place + oblique emitter construction. However, the sections immediately below prepare the ground by discussing the mechanism of variable participant-function mappings and by applying it to the class of verbs in question. In the process, I account first for the remaining two constructions: the subject place + oblique emitter one, and the subject emitter + oblique place one.

### 2.1 Participants competing for the same argument status

The subject place + oblique emitter construction (illustrated in 1.2 ) and the subject emitter + oblique place construction (illustrated in 1.3) as a pair bear close resemblance to many well documented pairs of clauses that exhibit alternative mappings of semantic participants to grammatical functions.

Many different types of such alternations have been identified where, holding constant both the (base of the) predicate and the participants selected for expression, there are two (and sometimes
more) ways of matching the same set of grammatical functions with the participants which are available for mapping. A common type of alternation involves two arguments within a verb phrase, either of which can be specified as an object (OBJ) or an oblique ( $\mathrm{OBL}_{\theta}$ ). An example is the locative alternation (see Levin 1993:49-55 for extensive references, and particularly Rappaport and Levin 1988, Pinker 1989, Jackendoff 1990, and Dowty 2000 for discussion of the different semantic contents of the English locative alternation variants). The example below is from Ackerman (1991; 1992) and Ackerman and Moore (2001) who used the locative alternation in their construction of a theory of mapping between semantic arguments and grammatical functions:
(17) a. The peasant loaded (the) hay onto the wagon.

OBJ OBL $_{\theta}$
b. The peasant loaded the wagon with (the) hay.

OBJ $\quad$ OBL $_{\theta}$
Another type of alternation which results from the different possibilities of matching up the same sets of grammatical functions and participants is the so-called material-product alternation in English (Levin 1993:57). This alternation is even more relevant to the constructions discussed in this paper in that the set of grammatical functions in both types of alternation includes the subject, not just the arguments within the VP. Specifically, when the material-product alternation involves the intransitive variants of verbs such as grow, develop, evolve, hatch, and mature, both the raw material and product arguments may be expressed either as the subject or as the object of a preposition:

| (18) a. | That a corn |  |
| :---: | :---: | :---: |
| SUBJ | will grow | into an oak tree. |
| b. An oak tree | will grow | OBL $_{\theta}$ |
| SUBJ |  | from that acorn. $^{\text {OBL }_{\theta}}$ |

By analogy, pairs of clauses made up of the Polish subject place + oblique emitter construction and subject emitter + oblique place construction may be represented in the following way:


Differences in the interpretation of the variants (such as the holistic vs partitive effect of the locative alternation) is regarded as evidence that the variants do not actually involve 'the same' predicates, but that they share the base lexeme and that the predicates are related to each other by some sort of lexical mechanism. However, it has not been clear whether it is possible to establish which variant is more basic, at least in English ${ }^{6}$ - in this respect, they seem to have equal status, and

[^129]formalisations of the alternations may be able to reflect this (as does the account of Markantonatou and Sadler 1996, who propose underspecified verb entries to account for some argument alternations).

### 2.2 Modelling the locative alternation with LMT

Modelling this type of alternation with textbook Lexical Mapping Theory (LMT) (e.g. Bresnan 2001: Chapter 14) is problematic. Taking the locative alternation in (17a,b) as an example, the most widely used versions of LMT would produce the following representations for the two variants, respectively:
$\left.\begin{array}{cccc}\text { (21) a. load } & \begin{array}{cc}\text { ag } \\ {[-\mathrm{o}]} & \text { th } \\ {[-\mathrm{r}]} & {[-\mathrm{o}]}\end{array} & \text { b. load } & \langle a g \\ & \text { th } & \text { loc }\end{array}\right\rangle$

Kordoni (2003:259-260) discusses the difficulty which the alternation poses for the assignment of the syntactic pre-specifications $[+/-\mathrm{r} / \mathrm{o}$ ] to the arguments, and states the problem succintly: 'the attempt to account for two different linkings to the respective grammatical functions from the same array of thematic roles clearly fails'.

Solutions to extending the capability of LMT that have been offered in the literature are twofold. First, it has been argued that the role of the hierarchy of thematic roles has to be reconsidered. The most widely used versions of LMT have a fixed hierarchy of thematic roles which determines the ordering of argument positions, as in (21a,b). However, there are many different hierarchies on offer (Newmeyer 2002 cites 18) and none of them appear to capture correctly all generalisations involving the realisation of arguments in terms of their semantic roles (Levin and Rappaport Hovav 2005: Chapter 6). Furthermore, Ackerman and Moore (2001:27) cite Gawron (1983) as a good general critique of the shortcomings associated with delimiting classes of verbs and identifying finite lists of discrete semantic roles.

Second, many authors have argued for the dissociation in the argument structure of the tier of semantic participants from the tier of syntactic argument positions, specifically to be able to account for morphosemantic operations on the predicate (e.g. Grimshaw 1988:1, T. Mohanan 1990/1994:15ff, Ackerman 1991:12; 1992:57ff, Joshi 1993, Alsina 1996:37, Ackerman and Moore 2001:40ff, Falk 2001:105).

Following these insights, I propose that the tier of semantic participants is distinct from the tier of valency slots. I follow Ackerman and Moore (after Dowty 1991) in assuming that an argument of a predicate is a set of predicate entailments that is specific to a participant in the event denoted by the predicate, that sets of proto-properties can be ordered from most proto-agentive to most proto-patientive, and that the linking of entailment sets to valency slots can be regulated by a well-formedness condition (2001:44-45). ${ }^{7}$

Furthermore, following Zaenen (1993:151) and Ackerman \& Moore (2001:44ff), I argue that the point of reference which should remain constant in modelling argument structure is the syntactic representation of the predicate's valency rather than the semantic representation of the participants with which argument positions are linked. I assume that the following valency template is available to a base predicate: ${ }^{8}$

$$
\begin{array}{cccccc}
<\arg _{1} & \arg _{2} & \arg _{3} & \arg _{4} & \ldots & \arg _{\mathrm{n}}> \\
{[-\mathrm{o} /-\mathrm{r}]} & {[-\mathrm{r}]} & {[+\mathrm{o}]} & {[-\mathrm{o}]} & & {[-\mathrm{o}]}
\end{array}
$$

Note that the pre-specification of the ordered valency slots corresponds to LFG's hierarchy of syntactic functions, but it is based on LMT's atomic values instead of final grammatical functions. As in all widely used models of LMT, the syntactic pre-specification of the arguments determines their availability for the mapping of particular grammatical functions. In order to retain the principle of monotonicity for the tractability of syntactic information (e.g. Bresnan 2001:45-46), I assume that the only mechanism that can intervene at the level of argument-to-function mapping is

[^130]a mechanism of increasing markedness, but the primitives $[+/-\mathrm{r} / \mathrm{o}$ ] cannot be either changed or deleted.

Since argument positions are linked with particular types of predicate entailments corresponding to semantic participants, if the predicate does not have a particular set of entailments, the slot corresponding to that set of entailments is not invoked. Thus, for a particular predicate, the angled brackets contain all and only the selected valency slots for the arguments associated with that predicate, both core and non-core $\left(\arg _{n}[-\mathrm{o}]\right.$ indicates the availability of multiple non-core arguments), and there are no 'empty slots' in any particular predicate's argument structure.

Finally, I retain the widely accepted LMT 'Markedness Hierarchy of Syntactic Functions':

$$
[-\mathrm{o}] /[-\mathrm{r}] \text { SUBJ }>[-\mathrm{r}] /[+\mathrm{o}] \text { OBJ, }[-\mathrm{o}] /[+\mathrm{r}] \mathrm{OBL}_{\theta}>[+\mathrm{o}] /[+\mathrm{r}] \mathrm{OBJ}_{\theta}
$$

and use the following, revised, Mapping Principle for mapping argument structures to surface grammatical functions, based on the Markedness Hierarchy: 'The ordered arguments are mappped onto the highest (i.e. least marked) compatible function on the markedness hierarchy' (see Zaenen 1993:151 for a similar approach).

With the proposed model of LMT, ${ }^{9}$ I arrive at the following representations for (17a,b). The referent of the semantic participant $y$ is the 'hay', and the referent of the semantic participant $z$ is the 'wagon'. In variant (22a), the role of 'hay' is more patient-like (we may call it a 'theme', for example) and therefore it maps onto the second argument position which is normally the syntactic slot for objects. In variant (22b), the role of 'wagon' is more patient-like (i.e. 'wagon' is conceptualised as the affected participant), and so it is 'wagon' which maps onto the second argument position where it is assigned the grammatical function of the object (unless the predicate undergoes passivisation, for example):
(22) a. The peasant loaded (the) hay onto the wagon.

|  | OBJ |  |  |
| :---: | :---: | :---: | :---: |
|  |  |  |  |
|  | $x$ | $\boldsymbol{y}$ | $\boldsymbol{z}$ |
| load | $\mid$ | $\mid$ | $\mid$ |
|  | (arg | $\arg$ | $\arg \rangle$ |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ |
|  | SUBJ | OBJ | OBL $_{\theta}$ |

b. The peasant loaded the wagon with (the) hay. OBJ OBL $_{\theta}$


In other words, in (22a), the predicate (load something on/into something) entails an agent, theme, and location: peasant fulfills the semantic role of the agent, hay fulfills the semantic role of the theme, and wagon - of the location. In (22b), the predicate (load something with something) entails an agent, patient/affected entity, and instrument/theme/means (or whatever this last role is best called): peasant fulfills the semantic role of the agent, wagon fulfills the semantic role of the patient, and hay - of the instrument/theme/means. I have indicated that the two predicates are related by coding the participants with the same letters for both predicates. The variants result from the fact that two of the participants are capable of fitting slightly different roles in the two predicates.

Thus, as a result of the shift of perspective on semantic participants - from classifying them into discrete roles to seeing them as sets of semantic entailments of the predicate - it is now expected that the same semantic participants may align with the available argument positions in two

[^131](or more) different ways, as is exemplified by locative and other alternations. It is also expected that the semantic participants may 'change order' and re-associate with different argument positions for derived, morphosemantically altered, predicates.

Thus, even though there may be a default ordering of semantic participants, evidently it can be altered, the alteration being driven by the change in the interpretation of the predicate together with its sets of entailments. As was illustrated in (22), the most straightforward way to model this with LMT is to allow the same semantic participants to 'realign' and link to different argument positions for different types of clauses which may or may not differ in valency. Note, however, that although the ordering of the semantic participants is relaxed in this version of LMT (it is no longer fixed, because it does not depend on any fixed thematic hierarchy), the ordering of syntactic argument positions (i.e. the argument positions together with their syntactic pre-specifications) has to remain fixed, representing a presumably universal valency template available to a predicate.

Despite its more limited application, Kordoni (2003:262-263), following the LMT version of Zaenen (1993), offers a compatible solution for her analysis of the locative alternation in German: instead of following a hierarchy of thematic roles, she fixes the ordering of the argument positions and refers to them by 'conventional labels in the spirit of Zaenen (1993), such as agent, patient and nonpatient, [which] are used in order to indicate that the verb supports three arguments, each of which is associated with some general lexico-semantic entailments'. Consequently, the arguments receive their correct syntactic pre-specifications, but crucially they express different participants in the two variants: patient=locatum and nonpatient=location in the locative variant ('Peter filled water in the tank'), while patient=location and nonpatient=locatum/means in the 'with'-variant ('Peter filled the tank with water'):
(23) a. Peterfüllte Wasser in den Tank.



### 2.3 Modelling the emitter-place alternations with LMT

The Polish verbs in (1)-(4) denote events that typically involve two entities as participants. One is the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept. The other entity is the location in which the event takes place - that is, the location where the smell, sound, light, or aggregate/abstract concept is present and/or propagated.

For example, the verb pachnieć 'emit fragrance' typically involves two participants: the emitter of the fragrance $(x)$, and a location ( $y$ ). Two simple mapping options involving these two participants are:
(24) a.


[^132]b. pachnie $\quad \begin{aligned} & \text { emit-fragrance.3SG.(M) }\end{aligned}$
dom kawa
kawa
coffee(F).INS
OBL $_{\text {INS }}$

|  | $y$ | $x$ |
| :---: | :---: | :---: |
| pachnieć | $\langle\arg$ | $\mid$ |
|  | $[-\mathrm{r}]$ | $\arg \quad\rangle$ |
|  | SUBJ | $[-\mathrm{o}]$ |
|  | $\left(\right.$ OBL $\left._{\text {INS }}\right)$ |  |

In (24a), the predicate entails (in the sense of Dowty 1991, Ackerman and Moore 2001, Grimm 2007; see also Donohue and Donohue 2004 regarding instruments) an 'instigator/causer' participant which emits the fragrance, and an optional location. In (24b), the predicate entails an 'instigator/causer' participant which propagates the fragrance, and an optional oblique participant (a kind of 'instrument', or 'means' - this latter term is due to Kordoni 2003:262) with which the propagation is achieved. Kawa 'coffee' and dom 'house' can map in two different ways, because they can each fulfil two different semantic roles. One of the roles that both of them are capable of fulfilling is that of an 'instigator'.

Similarly, the verb roic sie 'swarm' typically involves two participants: the entity which swarms $(x)$, and a location $(y)$, with the following two simple mapping options:

| roity sie | pszczoty | w ogrodzie |
| :--- | :--- | :--- |
| swarmed.3PL.NONVIR REFL | bees(NONVIR).NOM | in garden(M).LOC |
|  | SUBJ | OBL $_{\text {LOC }}$ |


|  | $x$ <br> $\mid$ <br> roićsię | $\langle\arg$ |
| :---: | :---: | :---: |
|  | $[-\mathrm{r}]$ | $\mid$ |
|  | SUBJ | $\arg \quad\rangle$ |
|  |  | $[-\mathrm{O}]$ |
|  | $\left(\right.$ OBL $\left._{\text {LOC }}\right)$ |  |


| b. roit sié <br> swarmed.3PL.M REFL | ogród <br> garden(M).NOM <br> SUBJ | od pszczól <br> from bees(NONVIR).GEN <br> OBL $_{\theta}$ | cf. (20a) |
| :--- | :--- | :--- | :--- |

In (25a), the predicate entails an 'instigator' ('agentive', 'causal') participant, and an optional location. In (25b), the predicate entails an 'instigator' or 'causal' participant projecting the activity of swarming, and an optional oblique participant (a kind of 'instrument', or 'means') with which the activity is achieved. Pszczoty 'bees' and ogród 'garden' can map in two different ways, because they can each fulfil two different semantic roles. One of the roles that both of them are capable of fulfilling is that of an 'instigator'.

## 3 Identifying a 'dummy' instigator

In order to analyse the oblique place + oblique emitter construction in Polish (the one illustrated in section 1.1), I need to bring up more tools. I have already demonstrated that both the place and the emitter participants of the predicates under discussion can be conceptualised as having semantic roles which fit oblique argument functions - the functions of locative and instrumental/prepositional obliques, respectively. In the two constructions discussed in section 2 , that was the end of the story. The resulting clauses are active and uncontroversially personal, with nominative subjects.

The oblique place + oblique emitter construction presents an additional problem of having no overt subject, with the verb bearing what looks like the default non-agreeing inflection (3SG.N). In the following subsections I offer an analysis which involves identifying three rather than two
semantic participants for this construction, finding a pro (unexpressed/incorporated pronominal) syntactic subject, and establishing the identity of the subject by co-referring it with the argument expressing the emitter or the location.

### 3.1 Distinguishing between semantic participants and referents

All predicates discussed in section 2 had different referents associated with each of the predicate's arguments. But, obviously, this may not always be the case. I use standard coindexing at the level of semantic participants to indicate their coreference. An example is Piotr robi sobie zastrzyk 'Peter is giving himself an injection', where the agent (subject) and the patient (object) co-refer:
(26) a.
Peter(M).NOM make.3SG.(M) self.DAT injection(M).ACC
'Peter gives/is giving himself an injection.'
b.
$\left.\begin{array}{cccc} & \text { Piotr } & \text { zastrzyk } & \text { sobie } \\ \text { robić } & \mid & \mid & \mid \\ & \text { arg } & \arg & \arg \\ & {[-\mathrm{o}]} & {[-\mathrm{r}]} & {[+\mathrm{o}]}\end{array}\right\rangle$
c.
$x_{i}$
$<$
$<\arg$
$[-\mathrm{o}]$
SUBJ
$y$
$\underset{\sim}{\arg }$
$[-r]$
OBJ


The LMT representation in (26) corresponds to diagrams found in traditional descriptive linguistic work on diathesis, such as Geniušienė's (1987), cited in Klaiman (1991:66):
(27)
(i) 'Ordinary transitive diathesis'

| Person1 | Person2 |
| :---: | :---: |
| Agent | Patient |
| Subject | Object |

(ii) 'Diathetical semantic reflexive'

| Person1 |  |
| :---: | :---: |
| Agent |  |
| Subject |  |

(iii) 'Nondiathetical semantic reflexive'

| Person1 |  |
| :---: | :---: |
| Agent | Patient |
| Subject | Object |

Specifically, the 'Agent' and 'Patient' in Geniušienè's diagrams correspond to semantic participants in LMT, such as the $x$ and $y$ in example (26). When they co-refer, Geniušienè represents the referent as one 'Person'; otherwise they are represented as two distinct referents, 'Person1' and 'Person2'. ${ }^{11}$

An example of an instrument co-referring with an agent is found in sentences such as Piotr zastonit soba slońce 'Peter blocked/shaded the sun with himself':
(28) a.

Piotr zasłonit soba stońce.
Peter(M).NOM blocked.3SG.M self.INS sun(N).ACC
'Peter blocked/shaded the sun with himself.'
b.


If three independent referents were associated with this predicate, as in Piotr zasłonit slońce

[^133]parawanem 'Piotr blocked/shaded the sun with a screen', the representation would simply be:
(29) a. Piotr zastonit stońce parawanem. Peter(M).NOM blocked.3SG.M sun(N).ACC screen(M).INS
'Peter blocked/shaded the sun with a screen.'
b.

|  | $x$ | $y$ | $z$ |
| :---: | :---: | :---: | :---: |
|  | $\mid$ | $\mid$ | $\mid$ |
| zastonić | $\langle\arg$ | $\arg$ | $\arg \rangle$ |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ |
|  | SUBJ | OBJ | OBL $_{\text {INS }}$ |

The sentence Piotr zasłonit stońce 'Peter blocked/shaded the sun' is obviously ambiguous with regard to whether the action is accomplished with Peter as the instrument causer, or with a distinct entity as an instrument used by Peter. If one wished to be explicit about the ambiguity, one could represent it as:
(30) a. Piotr zasłonit stońce.
Peter(M).NOM blocked.3SG.M sun(N).ACC
'Peter blocked/shaded the sun.'
b.


However, oblique participants/arguments are optional in Polish, and when they are not there, it is not due to any operation on argument structure that removes them, but they are simply not selected for expression. Therefore, we can also represent Piotr zasłonit stońce 'Peter blocked/shaded the sun' simply as:

```
(31) a. Piotr zasłonit stońce.
    Peter(M).NOM blocked.3SG.M sun(N).ACC
    'Peter blocked/shaded the sun.'
```

b.


To sum up, zastonić 'block/shade/cover' entails three semantic participants: an agent, patient, and instrument/means, but it involves only two (rather than three) referents when the agent and the instrument co-refer.

## 3.2 pro-drop constructions in Polish

One more building block of analysis, necessary to account for the oblique place + oblique emitter construction in Polish (the one illustrated in section 1.1), involves a discussion of Polish pro-drop constructions.

The most familiar instances of pro-drop in Polish are clauses formed from a personal predicate with a dropped personal pronoun, such as sentence (32) 'He saw that the door was open and went in' occurring in the following context: 'Peter didn't waste his time: [he saw that the door was open and went in]'.
(32) Zobaczyl, $\dot{z} e ~ d r z w i ~ s q ~ o t w a r t e ~ i ~ w s z e d t . ~$ saw.3SG.M that doors are open and went-in.3SG.M
'[He/Someone/They] saw that the door was open and went in.'
Other familiar instances are clauses formed from personal predicates with a dropped indefinite pronoun, both the pronoun referring to humans, such as sentence (32) 'Someone/They saw that the
door was open and went in' occurring in the following context: 'Someone may have not had an intention to burgle, but [they saw that the door was open and went in]', and the pronoun referring to non-humans, as in the so-called 'weather constructions' exemplified in (33) and 'adversity impersonals' exemplified in (34):
(33) Wiało, jakby chciato powyrywać drzewa z korzeniami. blew.3SG.N as-if wanted.3SG.N pull-out.INF trees with roots
'[It/Something] was blowing as if it wanted to pull out trees with their roots.'
Rzucito go w bok.
threw.3SG.N him to side
'[It/Something] threw him to the side.'
Contrary to tradition, predicates expressing weather phenomena and natural forces are now beginning to be recognised more widely as syntactically and/or morphologically personal in many languages in which weather verbs do not preclude the use of a lexical subject such as 'rain', 'wind', 'sky', 'universe/world/time' etc., and are capable of carrying corresponding inflection (e.g. all East Caucasian languages except Nakh - Daniel, Khalilova and Molochieva 2008; several Oceanic languages - Moyse-Faurie 2008; various Afroasiatic - Tosco and Mettouchi 2008; see also a 2008 discussion thread in lingtyp).

Even when they occur without a lexical subject, Polish weather clauses, adversity impersonals, and other apparently subjectless clauses involving verbs of physical or psychological states do not lack a syntactic subject. They can be analysed as a construction with an optionally unexpressed pronominal subject, where the understood subject is the indefinite pronoun referring to non-humans ( pro $_{\mathrm{INDEF}}$ ). ${ }^{12}$ As expected, this subject can be found to participate in syntactic control and raising - see, for example, (33) above; or Wiało rzucajac gałęziami '[It/Something] was blowing, throwing branches'; Zdawało się padać '[It/Something] seemed to rain'; etc. A construction with a dropped indefinite pronoun subject does not present problems for LFG, as it falls under the standard analysis of unexpressed pronouns (e.g. Bresnan 2001:144-177).

The same line of argumentation, and the same LFG analysis, can be applied to the apparently subjectless Polish construction in 1.1 which uses the class of verbs including verbs of emission of smell, sound, or light, and verbs expressing expansion of an aggregate or a mass/abstract entity. The verbs themselves are obviously not impersonal, since they easily admit and commonly appear with an overt nominative subject, and fully agree with an overt subject's inflectional properties. This was demonstrated in sections 1.2 and 1.3 which showed alternative syntactic frames available for those verbs. There are also no morphosyntactic restrictions that would prevent these verbs from agreeing with a subject in a person other than third - that is, the predicates in question have a complete inflectional paradigm of personal verbs. Furthermore, any Polish verb that can express an event whose causer/instigator is non-human may occur with an overt indefinite pronoun (cos' 'something') expressing the subject, for example: coś pachnie/s'mierdzi/szumi/huczy/mieni się/bieli się/roi się/kipi etc. 'something emits fragrance/smells/hums/rumbles/glitters/glistens/teems/seethes' etc., also coś mnie mdli/dusi/skręca etc. 'something nauseates/chokes/convulses me' etc. When they occur without a lexical subject, the unexpressed pro $_{\text {INDEF }}$ subject is capable of syntactic control and raising - e.g. Pachnie, jakby chciato cię omamić '[It/Something] smells as if it wanted to charm you'; W ogrodzie bieli się od szronu, przypominajac o nadchodzacym Nowym Roku 'In the garden [it/something] is glistening with hoarfrost, reminding about the up-coming New Year'; Zdawało się roić od pszczót '[It/Something] seemed to swarm with bees', etc.

I argue, therefore, that the oblique place + oblique emitter construction in Polish is only functionally impersonal, but it is not subjectless. It has a fairly ordinary syntactic subject which is the pro $_{\text {INDEF }}$, which behaves syntactically like any other pro subject, and which can be given a standard syntactic analysis of an 'unexpressed/incorporated' pronoun or pronominal inflection.

### 3.3 The instigator in the oblique place + oblique emitter construction

The section finally explains the relevance of the distinction between semantic participants and referents for the analysis of the oblique place + oblique emitter construction in Polish, by bringing

[^134]together the discussion from section 3.1 with the account of the syntactic subject of this construction in section 3.2.

In terms of grammatical functions, I have argued that the oblique place + oblique emitter construction in Polish (as exemplified in 1.1) has a syntactic pro ${ }_{\text {INDEF }}$ subject, and that the subject expresses an unspecified or undisclosed non-human instigator/causer of the denoted event. Apart from the subject, the construction may also have up to two oblique arguments, one expressing the entity which emits the smell, sound, or light, or the entity which is the expanding aggregate or mass/abstract concept, and the other expressing the location where the event takes place/is present/is propagated.

Therefore, in terms of semantic participants, the functionally impersonal variant of the predicate entails an unspecified instigator/causer, an optional instrument/means with which the activity of the instigator/causer is achieved, and an optional location. However, clauses are wellformed even if no arguments are lexically expressed (provided that the context ensures that they are felicitous), as is illustrated below in (35a-d) and (36a-d).

The following are proposed representations of the oblique place + oblique emitter construction. The unspecified instigator/causer participant ( $z$ ), which does not have an independent referent, may co-refer with either the emitter (for which I have retained the label $x$ ) or the location (for which I have retained the label $y$ ). By coding the semantic participants in this construction with the same letters as in the other two constructions, I capture the way in which the predicates in all three constructions are related.

## a. W domи pachnie kawq.

in house emit-fragrance.3SG.(N) coffee(F).INS
'There is a smell of coffee in the house.' [lit. '(It) smells of coffee in the house.']
b. Ale pachnie.
how emit-fragrance.3SG.(N)
'What a fragrance!' [lit. 'How (it) emits fragrance.']
c. Pachnie w tym domu.
emit-fragrance.3SG.(N) in this house
'There is a fragrance in this house.' [lit. '(It) emits fragrance in this house.']
d. Pachnie kawq.
emit-fragrance.3SG.(N) coffee(F).INS
'There is a smell of coffee.' [lit. '(It) smells of coffee.']
e.

f.

|  |  | $z_{(i / j)}$ | $y_{(i)}$ | $x_{(j)}$ |
| :---: | :---: | :---: | :---: | :---: |
| pachniećc | $\langle\quad\|$ | $\arg$ | $\arg$ | $\mid$ |
|  |  | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ | $[-\mathrm{o}]$ |
|  | SUBJ | $\left(\right.$ OBL $\left._{\text {LOC }}\right)$ | $\left(\mathrm{OBL}_{\text {INS }}\right)$ |  |
|  |  |  |  |  |

a. W ogrodzie bieli się od szronu.
in garder appear-white.3SG.(N) REFL from hoarfrost(M).GEN
'The garden is glistening with hoarfrost.' [lit. '(It) glistens with hoarfrost in the garden.']
b. Ale sie bieli.
how REFL appear-white.3SG.(N)
'How it is glistening!' [lit. 'How (it) glistens.']
c. Bieli się wogrodzie.
appear-white. 3 SG .(N) REFL in garden
'It is glistening in the garden.' [lit. '(It) glistens in the garden.']
d. Bieli się od szronu. appear-white.3SG.(N) REFL from hoarfrost(M).GEN 'It is glistening with hoarfrost.' [lit. '(It) glistens with hoarfrost.']
e.



f.

$z_{(i / j)}$
$\mid$
$\arg$
$[-\mathrm{r}]$
SUBJ

arg

( $\mathrm{OBL}_{\mathrm{LOC}}$ )

$\arg >$

( $\mathrm{OBL}_{\theta}$ )

This completes the proposal of how to capture the linking between semantic participants and grammatical functions in the oblique place + oblique emitter construction, taking into consideration the interpretation of the participant roles and the morphosyntactic behaviour of the arguments present in this construction - in particular, the obligatory and syntactically active unexpressed pronominal subject, which contrasts with the implied and optionally lexicalised reflexive pronoun in (28) (cf. (30)). However, extending beyond LMT, there remains the technical question of how exactly to handle the coreference between a 'PRO' argument (the inflectionally expressed pro INDEF ) and another argument within a simple predicate despite their possible different featural specifications. I leave this issue up for further research, but just note that a sample solution for the coference of a nominal and reflexive elements bearing different featural specifications has been offered in HPSG by Trawiński (2007).

Note also that the $\mathrm{pro}_{\text {INDEF }}$ could be expressed overtly as cos' 'something' in all sentences in (35) and (36). Its overt expression itself does not, however, resolve the ambiguity of its reference. Hence, sentences with the overt pro $_{\text {INDEF }}$ would also have the representations in (35f) and (36f). It seems that the ambiguity of the reference of pro $_{\text {INDEF }}$ can only be resolved with the help of additional linguistic material or extralinguistic context.

## 4 Intermediary agents in other pro $_{\text {INDEF }}$-drop constructions

The following is a summary view of the three different syntactic frames available for the class of Polish predicates discussed in the sections above, that is verbs of emission of smell, sound, or light, and verbs expressing expansion of an aggregate or a mass/abstract entity. I use pachnieć 'emit fragrance' as the example, and retain the coding of the semantic participants throughout as: $x=$ emitter; $y=$ location; and $z=$ the unspecified instigator/causer. Note that this class of verbs is intransitive:
(37) a.
pachnieć $\left.\begin{array}{ccc}x & \mid & \begin{array}{c}y \\ \arg \\ {[-\mathrm{r}]} \\ \text { SUBJ }\end{array} \\ & \text { arg } & {[-\mathrm{O}]} \\ \left(\mathrm{OBL}_{\mathrm{LOC}}\right)\end{array}\right\rangle$
b.
pachniec

c.
$\left.\begin{array}{ll}y_{(i)} & x_{(j)} \\ \mid & \mid \\ \arg & \arg \quad \\ {[-\mathrm{O}]} & {[-\mathrm{o}]} \\ \left(\mathrm{OBL}_{\mathrm{LOC}}\right) & \left(\mathrm{OBL}_{\mathrm{INS}}\right)\end{array}\right\rangle$

An analogous LMT analysis can be applied to two more classes of predicates in Polish, most of which are typically used transitively: predicates denoting some physical or psychological states, and predicates used in the so-called 'adversity impersonal' construction. I discuss them briefly in the following subsections.

### 4.1 Intermediary agents co-occurring with experiencers

Polish verbs denoting various physical or psychological states typically entail an experiencer participant and an (optional) stimulus participant:

Mdli/Dusi/Skręca/Ciagnie/Boli/Swędzi/Kłuje mnie. nauseate/choke/convulse/pull/ache/itch/stab.3SG.(N) me.ACC
'[Something] makes me nauseous/choke/convulse/contract my muscles/ache/itch/gives me shooting pains.'
All of these verbs typically appear with an experiencer marked for accusative case. However, they also frequently collocate with a particular oblique expression of the stimulus, for example:
(39) a. Mdli/Dusi/Skręca mnie od tego zapachu. nauseate/choke/convulse.3SG.(N) me.ACC from this smell 'This smell makes me nauseous/choke/convulse.' [lit. '(It) makes me nauseous/choke/convulse from this smell.'
b. Mdli/Dusi/Skręca mnie z bólu/zazdrości. nauseate/choke/convulse.3SG.(N) me.ACC from pain/envy 'The pain/envy makes me nauseous/choke/convulse.' [lit. '(It) makes me nauseous/choke/convulse from pain/envy.'
This construction in Slavonic has frequently been regarded as impersonal (e.g. Franks 1995:70ff; Babby 1998:6ff; Nagórko 1998:266; Saloni and Świdziński 1998:150; Śpiewak 2000:169). However, contrary to the common assumption that these predicates do not accept a nominative subject, in modern Polish their morphosyntax does not disallow it. Furthermore, the verbs have a full personal paradigm. Consider the following examples:
(40) a. Wszystkie zapachy mnie mdlity. all.NONVIR.NOM smells(NONVIR).NOM me.ACC nauseated.3PL.NONVIR
Nawet zapach kawy mnie mdlit.
even smell(M).NOM coffee(F).GEN me.ACC nauseated.3SG.M
'All smells made me nauseous. Even the smell of coffee made me nauseous.'
b. Ból skręcat mnie niemiłosiernie.
pain(M).NOM convulsed.3SG.M me.ACC mercilessly
'The pain convulsed me mercilessly.'
c. Bolała/Swędziała mnie głowa.
ached/itched.3SG.F me.ACC head(F).NOM
'My head ached/itched.'
d. Coś mnie dusi. something(N).NOM me.ACC choke.3SG.(N)
'Something makes me choke.'
e. Dusity mnie te zapachy.
choked.3PL.NONVIR me.ACC these.NONVIR.NOM smells(NONVIR).NOM
'Those smells made me choke.'
I offer the following LMT representations for the two syntactic frames available to these predicates. I use the verb mdlic' 'nauseate' as an illustration and code its semantic participants throughout as: $x=$ stimulus; $v=$ experiencer; $z=$ the unspecified instigator/causer. The first syntactic frame, in (41), models the second clause in example (40a):
(41) a.
$\left.\begin{array}{ccl} & \text { zapach } & \text { mnie } \\ \text { mdlić } & \langle\arg & \mid \\ & \arg { }^{\arg } & {[-\mathrm{r}]} \\ & \text { SUBJ } & \text { OBJ }\end{array}\right\rangle$
b.
mdlić $\left.\begin{array}{ccc}x & v \\ \mid & \mid \\ \arg & \arg \\ {[-\mathrm{r}]} & {[-\mathrm{r}]} \\ \text { SUBJ } & \text { OBJ }\end{array}\right\rangle$

And the following syntactic frame models examples (38) and (39a):
(42) $a$.

b.

|  | $z_{(i)}$ <br> $\mid$ <br> mdlić$\left\langle\begin{array}{ccc} & v & x_{(i)} \\ & \text { arg } & \arg \end{array}\right.$ | $\arg$ |  |
| :---: | :---: | :---: | :---: |
|  | $[-\mathrm{r}]$ | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ |
|  | SUBJ | OBJ | $\left(\mathrm{OBL}_{\theta}\right)$ |

### 4.2 Intermediary agents co-occurring with patients

Finally, the so-called 'adversity impersonals' can be exemplified in Polish by the following sentences:
(43) a. Zasypało droge.
covered.3SG.N road(F).ACC
'The road got covered (with snow or sand).' [lit. '(It) covered the road.']
b. Zasnuto las. enveiled.3SG.N forest(M).ACC
'The forest got enveiled (with fog or smoke).' [lit. '(It) enveiled the forest.']
c. Bito człowieka w twarz.
beat.3SG.N man(M).ACC into face
'One was beaten in the face (by rain/sleet/hail).' [lit. '(It) beat one in the face.']
Apart from typically appearing with a patient which is expressed through a direct object, adversity impersonals may also include an instrumental argument which is commonly interpreted as denoting the 'cause' (Wierzbicka 1966, Doros 1975, Siewierska 1988):
(44) a. Zasypało droge śniegiem. covered.3SG.N road(F).ACC snow(M).INS
'The road got covered with snow.' [lit. '(It) covered the road with snow.']
b. Las zasnuło mgła.
forest(M).ACC enveiled.3SG.N fog(F).INS
'The forest got enveiled with fog.' [lit. '(It) enveiled the forest with fog.']
c. Biło deszczem w twarz.
beat.3SG.N rain(M).INS into face
'The rain beat one/you in the face.' [lit. '(It) beat in the face with rain.']

[^135]This construction has also frequently been regarded as impersonal (e.g. Wierzbicka 1966 and Wlodarczyk 1993 for Polish; or Mel'čuk 1979 for a cognate Russian construction). However, like the other constructions discussed in this paper, this one also happily accepts a nominative causer/instigator. First, we find clauses corresponding to the ones in (44) where the same 'cause' participant is expressed through a nominative subject rather than an instrumental argument:
(45) a.
zasypat
snow(M).NOM covered.3SG.M road(F).ACC
'Snow covered the road.'
b. Mgła zasnuta las. fog(F).NOM enveiled.3SG.F forest(M).ACC
'Fog enveiled the forest.'
c. Deszcz bit w twarz.
rain(M).NOM beat.3SG.M into face
'The rain beat one/you in the face.'
And second, we find clauses corresponding to the ones in (44) where the 'cause' participant remains expressed through an instrumental argument, but additionally there is a nominative subject denoting a natural force (or exceptionally an agent). Its referent is different from the instrumental nominal; it is interpreted as the actual (rather than unspecified or unidentified) instigator of the event which uses the participant expressed through the instrumental as its instrument or means:
(46) a. Huragan zasypal droge śniegiem.
storm(M).NOM covered.3SG.M road(F).ACC snow(M).INS
'The storm covered the road with snow.'
b. Niewidzialna ręka zasnuła las mgła. invisible.F.NOM hand(F).NOM enveiled.3SG.F forest(M).ACC fog(F).INS
'An invisible hand enveiled the forest with fog.'
c. Wichura biła deszczem w twarz.
strong-wind.(F).NOM beat.3SG.F rain(M).INS into face
'The strong wind beat one/you with rain in the face.'
Siewierska (1988:276) remarks that the construction in (44), which contains both an accusative argument and an instrumental one, bears a striking resemblance to the passive. Ss it could be seen to be derived from the construction in (46), it has been classified by some linguists as passive. However, both (45) and (46) have their legitimate and morphologically regular passives, as in (47) and (48), respectively:

## (47) a. Droga została zasypana przez śnieg.

road(F).NOM became.3SG.F covered.PART.SG.F by snow
'The road got covered with snow.'
b. Las zostat zasnuty przez mgłę.
forest(M).NOM became.3SG.M enveiled.PART.SG.M by fog 'Fog enveiled the forest.'
c. Człowiek byt bity w twarz przez deszcz. man(M).NOM was.3SG.M beat.PART.SG.M into face by rain 'One was beaten in the face by the rain.'
(48) a. Droga została calkowicie zasypana śniegiem przez huragan. $\operatorname{road}(F)$.NOM became.3SG.F completely covered.PART.SG.F snow(M).INS by storm 'The road got completely covered with snow by the storm.'
b. Las zostat zasnuty mgła jakby przez niewidzialna rękę. forest(M).NOM became.3SG.M enveiled.PART.SG.M fog(F).INS as-if by invisible hand 'The forest got enveiled with fog as if by an invisible hand.'
c. Człowiek byt dosłownie bity deszczem w twarz przez tę wichure. man(M).NOM was.3SG.M literally beat.PART.SG.M rain(M).INS into face by this strong-wind 'One was literally beaten in the face with the rain by this strong wind.'

Instead of a passive analysis of the construction in (43) and (44), I suggest that it should instead be analysed in the way analogous to the other constructions discussed in this paper. Namely, I suggest that the predicates in (43)-(44) involve an unspecified instigator/causer, a patient, and an instrument.

I offer the following LMT representations for the three syntactic frames available to the predicates which are found in 'adversity impersonals'. I use the verb zasypać 'cover [by spilling/pouring a grainy substance]' as an illustration and code its semantic participants throughout as: $z=$ instigator/causer/agent, $v=$ patient, $x=$ instrument/means/theme. The first syntactic frame, in (49), models the examples in (45):
(49) a.

$\begin{array}{cccc}\text { b. } & z_{i} & v & x_{i} \\ & & \mid & \mid \\ & & & \\ & \text { zasypać } & \langle\arg & \arg \\ & {[-\mathrm{o}]} & {[-\mathrm{r}]} & \\ & & \text { SUBJ } & \text { OBJ }\end{array}$
The following syntactic frame models the examples in (46):
(50) a.


And, finally, the following syntactic frame models the examples in (43)-(44):
(50) a.

$\begin{array}{lllll}\text { b. } & & z_{i / j} & v & x_{j} \\ & & \mid & \mid & \mid \\ & & \\ \text { zasypać } & \langle & \arg & \arg & \arg \\ & & {[-\mathrm{o}]} & {[-\mathrm{r}]} & {[-\mathrm{o}]} \\ & \text { SUBJ } & \text { OBJ } & \left(\mathrm{OBL}_{\mathrm{INS}}\right) & \end{array}$

## 5 Summary

In the sections above I propose an analysis of the morphosyntax of Polish clauses with verbs of emission, SWARM verbs, verbs expressing physical or psychological states due to a stimulus which can be interpreted as an intermediary agent, and verbs which are used in the so-called 'adversity impersonal' construction. I show the full range or possible participant-function mappings available for these classes of verbs and offer argument structure analyses for all the patterns of mapping. In order to do this, I have to extend the existing accounts of both the variable syntactic expression of semantic participants (in particular, the oblique-subject alternation), and pro-drop.

In particular, in order to model the fact that the same base predicate may have two (or more) options of matching its participants with grammatical functions without undergoing any morphosyntactic operations such as passivisation, I use a representation of argument structure in which the tier of semantic participants is dissociated from the tier of argument positions. Furthermore, in order to model the argument structure of nominativeless Polish clauses involving the predicates in question, I demonstrate that they are not impersonal and identify their subject as expressing a 'dummy' (unidentified) instigator/causer which may co-refer with an instrument or other oblique semantic participant. In this way, the paper brings together two phenomena which are usually treated independently: the oblique-subject alternation and pro-drop.

## References

Ackerman, Farrell. 1991. Locative alternation vs. locative inversion. In: Halpern, Aaron L. (ed.) The Proceedings of the Ninth West Coast Conference on Formal Linguistics 1990. Stanford, CA: CSLI Publications. 1-13.
Ackerman, Farrell. 1992. Complex predicates and morpholexical relatedness: locative alternation in Hungarian. In: Sag, Ivan A. and Anna Szabolcsi (eds) Lexical Matters. Stanford, CA: CSLI Publications. 55-83.
Ackerman, Farrell and John Moore. 2001. Proto-Properties and Grammatical Encoding. A Correspondence Theory of Argument Selection. Stanford, CA: CSLI Publications.
Alsina, Alex. 1996. The Role of Argument Structure in Grammar: Evidence from Romance. Stanford, CA: CSLI Publications.
Alsina, Alex \& Sam A. Mchombo. 1988. Lexical mapping in the Chichewa applicative construction. Paper presented to the Summer Working Group on Argument Structure and Syntax, CSLI, Stanford University. Revised and expanded from a paper presented at the 19th Annual African Linguistics Conference, Boston University, April 14-17, 1988.
Babby, Leonard H. 1998. Voice and Diathesis in Slavic. Position paper at the 'Comparative Slavic Morphosyntax' conference in Bloomington, Indiana, 5-7 June 1998.
Bresnan, Joan. 2001. Lexical-Functional Syntax. Oxford: Blackwell.
Butt, Miriam. 2006. Theories of Case. Cambridge University Press.
Daniel, Michael, Zaira Khalilova and Zarina Molochieva. 2008. Impersonal domain in East Caucasian. Paper given at SLE 41, Workshop on 'Impersonal constructions: a cross-linguistic perspective', University of Bologna at Forlì.
Donohue, Cathryn and Mark Donohue. 2004. On the special status of instrumentals. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG04 Conference, University of Canterbury. Stanford, CA: CSLI Publications. 209-225.
Doros, Aleksander. 1975. Werbalne konstrukcje bezosobowe w jezykach rosyjskim i polskim na tle innych jezzyków słowiańskich. Wrocław: Ossolineum.
Dowty, David. 1991. Thematic Roles and Argument Selection. Language 67:547-619.
Dowty, David. 2000. 'The garden swarms with bees' and the fallacy of 'argument alternation'. In: Ravin, Yael and Claudia Leacock (eds) Polysemy. Theoretical and Computational Approaches. Oxford: Oxford University Press. 111-128.
Falk, Yehuda. 2001. Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax. Stanford, CA: CSLI Publications.
Franks, Steven. 1995. Parameters of Slavic Morphosyntax. Oxford: Oxford University Press.
Gawron, Jean Mark. 1983. Lexical Representations and the Semantics of Complementation. PhD thesis, University of California, Berkeley.
Geniušienė, Emma. 1987. The Typology of Reflexives. Berlin: Mouton de Gruyter.
Grimm, Scott. 2007. Instrumental Subjects. PhD thesis, Stanford University.
Grimshaw, Jane. 1988. Adjuncts and argument structure. Occasional Paper 36. Center for Cognitive Science, MIT, Cambridge, MA.
Jackendoff, Ray. 1990. Semantic Structures. Cambridge, MA: The MIT Press.
Joshi, Smita. 1993. Selection of Grammatical and Logical Functions in Marathi. PhD thesis, Stanford University.
Kibort, Anna. 2001. The Polish passive and impersonal in Lexical Mapping Theory. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG01 Conference, University of Hong Kong, Hong Kong. Stanford, CA: CSLI Publications. 163-183.

Kibort, Anna. 2004. Passive and Passive-Like Constructions in English and Polish. PhD thesis, University of Cambridge. Available on-line.
Kibort, Anna. 2006. On three different types of subjectlessness and how to model them in LFG. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG06 Conference, University of Konstanz, Germany. Stanford, CA: CSLI Publications. 289-309.
Kibort, Anna. 2007. Extending the applicability of Lexical Mapping Theory. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG07 Conference, Stanford University. Stanford, CA: CSLI Publications. 250-270.
Kibort, Anna. 2008a. Impersonals in Polish: an LFG perspective. In: Siewierska, Anna (ed.) Impersonal Constructions in Grammatical Theory. Oxford: Wiley-Blackwell. 246-289.
Kibort, Anna. 2008b. On the syntax of ditransitive constructions. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG08 Conference, University of Sydney. Stanford, CA: CSLI Publications. 312-332.
Klaiman, Miriam H. 1991. Grammatical Voice. Cambridge: Cambridge University Press.
Kordoni, Valia. 2003. Valence alternations in German: an LMT analysis. In: Butt, Miriam and Tracy Holloway King (eds) Proceedings of the LFG03 Conference, University at Albany, State University of New York. Stanford, CA: CSLI Publications. 250-268.
Levin, Beth. 1993. English Verb Classes and Alternations. Chicago, IL: University of Chicago Press.
Levin, Beth and Malka Rappaport Hovav. 2005. Argument Realization. Cambridge: Cambridge University Press.
Markantonatou, Stella and Louisa Sadler. 1996. Linking indirect arguments. Essex Research Reports in Linguistics 9:24-63.
Mel'čuk, Igor A. 1979. Syntactic or lexical zero in natural language. Berkeley Linguistic Society 5:224260.

Mohanan, Tara. 1990/1994. Argument Structure in Hindi. Stanford, CA: CSLI Publications.
Moyse-Faurie, Claire. 2008. Impersonal constructions in a few Oceanic languages. Paper given at SLE 41, Workshop on 'Impersonal constructions: a cross-linguistic perspective', University of Bologna at Forlì.
Nagórko, Alicja. 1998. Zarys gramatyki polskiej (ze slowotwórstwem). Warszawa: PWN.
Newmeyer, Frederick J. 2002. Optimality and functionality: a critique of functionally-based optimalitytheoretic syntax. Natural Language and Linguistic Theory 20:43-80.
Pinker, Steven. 1989. Learnability and Cognition: The Acquisition of Argument Structure. Cambridge, MA: The MIT Press.
Rappaport, Malka and Beth Levin. 1988. What to do with $\theta$-roles. In: Wilkins, Wendy (ed.) Syntax and Semantics 21: Thematic Relations. New York, NY: Academic Press. 7-36.
Saloni, Zygmunt and Marek Świdziński. 1998. Składnia wspótczesnego języka polskiego. Warszawa: PWN.
Schaefer, Ronald P. and Francis O. Egbokhare. 2009. Toward a notion of possible verb in Emai. In: Austin, Peter K., Oliver Bond, Monik Charette, David Nathan and Peter Sells (eds) Proceedings of Conference on Language Documentation and Linguistic Theory 2, 13-14 November 2009, SOAS, London. School of Oriental and African Studies, University of London. 263-273.
Siewierska, A.1988. The passive in Slavic. In: Shibatani, Masayoshi (ed.) Passive and Voice. Amsterdam: John Benjamins. 243-289.
Siewierska, Anna. 2008. Impersonalization from a subject-centred vs. agent-centred perspective. In: Siewierska, Anna (ed.) Impersonal Constructions in Grammatical Theory. Oxford: Wiley-Blackwell. 115-137.
Śpiewak, Grzegorz. 2000. The Lexical-Conceptual Structure of Nominativeless Constructions in Polish. Towards a Unified Account. Ph.D. thesis, Uniwersytet M. Curie-Skłodowskiej, Lublin.
Talmy, Leonard. 2000. Toward a Cognitive Semantics. Cambridge, MA: The MIT Press.
Tosco, Mauro and Amina Mettouchi. 2008. Speakers, entities and situations: impersonal configurations in Afroasiatic. Paper given at SLE 41, Workshop on 'Impersonal constructions: a cross-linguistic perspective', University of Bologna at Forlì.
Trawiński, Beata. 2007. Referential relations in Polish. Paper given at FDSL-7, University of Leipzig.
Wierzbicka, Anna. 1966. Czy istnieją zdania bezpodmiotowe. Język polski 46(3):177-196.
Wlodarczyk, Hélène. 1993. Sentences without a nominative NP in Polish. In: Hentschel, Gerd and Roman Laskowski (eds) Studies in Polish Morphology and Syntax. (Specimina Philologiae Slavicae; Band 99). München: Verlag Otto Sagner. 209-227.
Zaenen, Annie. 1993. Unaccusativity in Dutch: integrating syntax and lexical semantics. In: Pustejovsky, James (ed.) Semantics and the Lexicon. Dordrecht: Kluwer. 129-161.

# RELATIONAL NOUNS AND ARGUMENT STRUCTURE EVIDENCE FROM HUNGARIAN 

Tibor Laczkó

University of Debrecen

# Proceedings of the LFG09 Conference <br> Miriam Butt and Tracy Holloway King (Editors) 

2009

## CSLI Publications

http://csli-publications.stanford.edu/


#### Abstract

The fundamental goal of this paper is to argue for postulating that a clearly identifiable group of (underived) relational nouns has an argument structure in addition to what is often called a lexical conceptual structure - at least in languages like Hungarian. First, I place relational nouns in a general typological setting, pointing out that it is necessary to attribute at least lexical conceptual structures (LCS) to them. Then I claim that Hungarian body part nouns like kéz 'hand' have both inalienable and non-inalienable uses, which has become grammaticalized. In the former, they must be associated with an argument structure (AS), and in the latter, their LCS complement has been eliminated (hence, they have no argument structure, either). Next, I discuss the treatment of relational nouns of the szomszéd 'neighbour' type, which behave partially differently. The essence of my analysis is that they also have ASs, and I account for the optionality of their possessor argument by invoking the standard suppression operation. I also show that there is independent motivation for assuming that Norwegian body part nouns also have Ass, although their behaviour is different from that of their Hungarian counterparts. Finally, I demonstrate that there are several (LFG) analyses of different phenomena in a variety of languages whose crucial assumption, shared by my approach in this paper, is that possessors are true syntactic arguments of relational nouns.


## 1. Introduction

On the grounds of Bresnan's (2001) view of the architecture of LexicalFunctional Grammar (LFG), in this paper I make the following assumption without any justification. The lexical form of a predicate contains two interrelated dimensions:

- argument structure (AS), cf. Bresnan's a-structure;
- lexical conceptual structure (LCS), cf. Bresnan's lexical semantics.

This is a relatively wide-spread view across generative approaches. For a classical example in the Chomskyan paradigm, see Grimshaw (1990), which is also relevant from the perspective of the present paper inasmuch as it extensively discusses argument taking nominal predicates as well.

In my representations I will indicate LCS complements between \{ and \} and AS arguments between $<$ and $>$. Consider the following simplified example:

$$
\begin{array}{cccc}
\text { kick, } \mathrm{V} & \left\{\begin{array}{c}
\mathrm{x} \\
\text { agent, } \\
{[-\mathrm{o}]}
\end{array}\right. & \left.\begin{array}{c}
\mathrm{y} \\
\text { theme } \\
{[-\mathrm{r}]}
\end{array}\right\} \tag{1}
\end{array}
$$

In this paper I argue that there is a group of Hungarian underived relational nouns whose morpho-syntactic behaviour calls for an analysis that postulates that they have both LCSs (which is a trivial assumption) and ASs.

As regards the analysis of English underived relational nouns like hand and neighbour, just like in the case of the analysis of derived nominals like destruction and assassination, there is a whole range of radically different approaches. One extreme view is that underived relational nouns do not have ASs at all: they only have LCSs, cf. Uriagereka (1995), for instance. Other authors attribute strict ASs, typical of verbal predicates, to a set of relational nouns, and they account for the predictable optionality of (some of) their arguments by assuming lexical processes (e.g., suppression) trivially taken to apply to genuine verbal predicates, cf. Barker (1995). Some others do not assume such processes, so for them fewer noun types have ASs, cf. Castillo (2001); yet others consider the arguments of relational nouns to be optional to begin with, cf. Asudeh (2005).

Hungarian relational nouns have been much less studied so far in general, and from a generative linguistic perspective in particular. In his special, nongenerative, functional-semantic framework, Hadrovics (1969) offers a comprehensive analysis of Hungarian possessive constructions, based on a remarkable empirical investigation. It also deals with noun phrases headed by words I consider underived relational nouns; however, there is no explicit, coherent and consistent treatment of these nouns in this work. For detailed criticism, see Laczkó (2008). On the generative side, Szabolcsi (1994), in a Government and Binding Theory (GB) framework, and Laczkó (1995), in an LFG framework, assume without any justification or argumentation that certain classes of underived relational nouns assign theta roles or have argument structures, respectively. Alberti (1995) argues against Szabolcsi's (1994) theory of theta-role assignment in Hungarian possessive constructions and endorses a strongly LCS-based-style account on which even underived nouns standardly not considered relational are assumed to assign a theta-role. Laczkó (2008) appears to be the first attempt at a comprehensive generative account of Hungarian underived relational nouns.

The paper has the following structure. In section 2, I posit relational nouns in a general typological setting, pointing out that it is necessary to attribute at least LCSs to them. In section 3, I claim that Hungarian body part nouns like kéz 'hand' have both inalienable and non-inalienable uses, which have become grammaticalized. In the former, they must be associated with an argument structure, and in the latter, their LCS complement has been eliminated (hence, they have no argument structure, either). In section 4, I discuss the treatment of relational nouns of the szomszéd 'neighbour' type, which behave partially differently. The essence of my analysis is that they also have argument structures, and I account for the optionality of their possessor argument by invoking the standard suppression operation. In section 5, I show that there is independent motivation for assuming that

Norwegian body part nouns also have argument structures, although their behaviour is different from that of their Hungarian counterparts. In section 6, I demonstrate that there are several (LFG) analyses of different phenomena in a variety of languages whose crucial assumption, shared by my approach in this paper, is that possessors are true syntactic arguments of relational nouns. This is followed by some concluding remarks in section 7 .

## 2. Relational nouns, typology, and LCS

The behaviour of relational nouns across languages has been extensively discussed in the typological literature, fundamentally in the context of inalienability and the semantics of possessive constructions. The most salient types of these nouns include

- kinship terms (e.g., father, child);
- body parts (e.g., hand, eye);
- part-whole relationships (e.g., middle, edge).

In a great number of languages these inalienable nouns typically exhibit distinct formal (morphological or syntactic) properties. For a comprehensive overview, see Nichols (1988). Here is a randomly selected example. In Nanai if a relational noun, which by default has an inalienable interpretation, is used in a non-inalienable/alienated sense then it has to be used with a special possessive paradigm including a morpheme encoding non-inalienable possession (NIP).
a. naj-dili-ni
person-head-3SG
'(the) person's (own) head'
b. naj-dili-ŋo-ni
person-head- NIP-3SG
'(the) head in the possession of the person'
These well-attested cross-linguistic phenomena minimally call for a clearly definable LCS approach to the nouns in question - otherwise it would hardly be possible to formulate the relevant rules for the relevant processes.

The following Hungarian facts point in the same direction.
(A) There are nouns in the relevant categories which practically can only be used in possessive constructions; that is, in possessive inflectional forms (cf. head-marking). This especially holds for part-whole nouns, cf.
b. a kör közep-e
the circle.NOM middle-3SG
'the middle of the circle'
(B) There are a few nouns that have two different stem variants when they are inflected in the possessive paradigm, depending on whether they are used in an inalienable or in an alienable sense. For instance, gyapjú 'wool' has different stem versions when it is taken to belong to a sheep inalienably and when it is interpreted as alienably belonging to a shepherd. The noun tető 'roof' also exhibits stem alternation depending on its inalienable vs. alienable use, for instance, whether it is an inalienable part of a house or it is alienably related to the owner of the house.
a. a juh gyapj-a the sheep.NOM wool-3SG 'the sheep's wool'
b. a juhász gyapjú-ja
the shepherd.NOM wool-3SG
'the shepherd's wool'
(5)
a. a ház tete-je
the house.NOM roof-3SG
'the roof of the house'
b. János tető-je

John.NOM roof-3SG
'John's roof'
(C) In Hungarian, too, it is only a part-whole inalienable noun that can be used as an "extra" argument. First consider the following English examples.
(6) a. John kicked Peter's leg.
b. John kicked Peter on the leg.

[^136]In Hungarian this extra argument can be expressed either as a fully-fledged possessive noun phrase or as a reduced constituent, a bare noun.

| a. János | megrúg-ta | Péter | láb-á-t. |
| :--- | :--- | :--- | :--- |
| John.NOM | kick-PAST.3SG.DEF | Peter.NOM | leg-3SG-ACC |
| 'John kicked Peter's leg.' |  |  |  |

b. János megrúg-ta Péter-t a láb-á-n.

John.NOM kick-PAST.3SG.DEF Peter-ACC the leg-3SG-SUP
'John kicked Peter on his leg.'
c. János láb-on rúg-ta Péter-t.

John.NOM leg-SUP kick-PAST.3SG.DEF Peter-ACC
'John kicked Peter on the leg.' ${ }^{2}$
The relevant empirical generalization is that part-whole inalienability is a necessary but not sufficient condition on the existence of these constructions, many of which are set phrases and not infrequently are used figuratively:
(8) a. fül-ön fog valaki-t
ear-SUP grab somebody-ACC
literally: 'grab somebody on the ear'
figuratively: 'catch somebody (to make them do something)'
b. szív-en üt valaki-t
heart-SUP hit somebody-ACC
literally: 'hit somebody on the heart'
figuratively: 'affect somebody emotionally’
The detailed formal analysis of such constructions is left to future research.
The important point from our present perspective is that the appropriate encoding of inalienability at least in the LCS of Hungarian nouns with the properties discussed above is necessary for the pertinent generalizations to be made.

## 3. Arguments for argument structures in Hungarian

In this section I present three considerations which can be taken to provide arguments for assuming that certain groups of Hungarian relational nouns also have ASs in addition to their LCSs.

[^137](A) Intuitively, the generalization that quite a few such nouns practically can only be used in possessive constructions (with the possessive inflectional paradigm), can be captured in the most principled manner by assuming that they have an AS which contains a possessor argument, and if this argument is not realized, the principle of completeness is violated. Consider (3), repeated here for convenience's sake.
(3) a. a ceruza hegy-e the pencil.NOM edge-3SG 'the edge of the pencil'
b. a kör közep-e the circle.NOM middle-3SG 'the middle of the circle'

The Hungarian word hegy is interesting because it means both 'edge' (of a long-shaped object) and 'mountain'. Although the etymological relationship is clear, speakers sense homophony here. In its relational meaning the word is almost exclusively used in a possessive DP. The word közép 'middle', again, almost always occurs in a possessive construction. Compare (9a) $)^{3}$ and (9b).
(9) Context: Lödd le azt a korongot! 'Shoot that disk.'
a. A közep-é-t céloz-d!
the middle-3SG-ACC aim-IMPER.2SG
'Aim at its middle.'
b. *A közep-et céloz-d!
the middle-ACC
aim-IMPER.2SG
'Aim at the middle.'

As this pair of examples illustrates, in this context, which represents a typical context for such strongly relational words, the word cannot be used in a nonpossessive construction (also see point (C) below). In actual fact, it seems that közép 'middle' is generally restricted to possessive constructions when used in fully-fledged DPs/NPs, with the possibly rare exception of its abstract notional use:

[^138](10) A számtani közép különbözik
the arithmetic middle.NOM differ.PRES.3SG
a mértani közép-től.
the geometric mean-ABL
'The arithmetical mean is different from the geometrical mean.'
The presence of the possessor always triggers the use of possessive inflection on the noun head in Hungarian, given that in this language possessive constructions are obligatorily head-marking. Thus, a corresponding morphological generalization is that these strongly relational nouns almost exclusively occur in the possessive paradigm.

Naturally, there can be other ways of capturing the fact that these nouns require the possessive context, but by far the most straightforward solution is to attribute strict ASs to them. On the treatment of their rare occurrence in non-possessive constructions, see section 4.
(B) A closely related property of relational nouns is that when they occur within possessive noun phrases, without a strongly influential linguistic or situational context, the possessor constituent is, as a rule, interpreted as the salient element of the inalienable relationship.

| Péter | fel-emel-te | a | kez-é-t. |
| :--- | :--- | :--- | :--- |
| Peter.NOM | up-raise-PAST.3SG.DEF | the | hand-3SG-ACC |

'Peter raised his hand.'

We can think of several situations in which Peter did not raise his own inalienable hand. For instance, Peter could work in a factory producing artificial limbs, or he could be a hand surgeon, or he could be a sculptor working on a hand etc. In all these cases he could have raised a limb he was working or operating on. It is worth pointing out in this connection that in theory (11) could also be interpreted as Peter raising someone else's inalienable hand. What is significant is that none of these interpretations is available without a strongly influential context. It is not the case then that there is a continuum of more likely and less likely interpretations in no particular context or in a neutral context, and the primary inalienable interpretation happens to be at the top. It is the only interpretation. Again, although there can be other plausible solutions, this fact can be most naturally captured by assuming that the possessor is a genuine argument of the noun bearing this inalienable relation to the noun. On the use of nouns like kéz 'hand' in non-inalienable meanings, see point (C) below and section 4.1.
(C) Most importantly, certain types of Hungarian relational nouns in clearly definable constructions must not be used in non-possessive noun phrases on the relevant inalienable reading. Compare (12a) and (12b).

| a. Péter fel-emel-te | a | kez-é-t. |
| :--- | :--- | :--- | :--- |
| Peter.NOM up-raise- PAST.3SG.DEF | the hand-3SG-ACC |  |
| 'Peter $_{(i)}$ raised his $_{(\text {(i) }}$ hand.' |  |  |

b. Péter fel-emel-te a kez-et.

Peter.NOM up-raise-PAST.3SG.DEF the hand-ACC
'Peter raised the hand. / * Peter $_{(\mathrm{i})}$ raised his $_{(\mathrm{i})}$ hand.'
Without any context or in a neutral context, the interpretation of (12a), in which the noun kéz 'hand' occurs in a possessive DP (with possessive prodrop), is that Peter raised his own inalienable hand (also see the discussion in point (B) above). What is of crucial significance here is that this inalienable interpretation of the non-possessive DP containing the noun kéz 'hand' is strictly unavailable in (12b). In other words, the relationship between Peter and the hand in (12b) is strictly non-inalienable/alienated. This sentence only admits readings on which Peter raised someone else's hand or an "alienated" limb, e.g., an artificial hand in a hospital. The rule then is that for the inalienable interpretation to be available the noun must be used in a possessive construction, as in (12a).

I propose to capture these facts by making the following key assumption: kéz 'hand' in its inalienable use has an AS and its possessor argument must be realized; hence the ungrammaticality of (12b) in the relevant sense. I give the simplified lexical form of this word in this use in (13).

$$
\begin{array}{ll}
\text { kéz, N: ‘HAND } \quad & \{\text { body, body-part }\} ’  \tag{13}\\
& <\text { whole }> \\
{[-\mathrm{r}]} \\
(\text { POSS })
\end{array}
$$

Note that I use relatively informal but informative labels in this representation. The terms "body" and "body part" in the LCS are rather straightforward. The semantic role label "whole" is meant to encode the generalization that body-part-body relationships are a subset of part-whole relationships. These terms are also comparable with Barker and Dowty's (1993) non-verbal proto-roles. In addition, I assume that the "whole" role has the $[-r]$ intrinsic specification; that is, it is a theme-like argument, and it is mapped onto the (POSS) grammatical function. On the nature of (POSS) see section 6; on the mapping principles I assume, see Laczkó (2004).

There are two important points to be made at the end of this section. On the one hand, it has to be explored how this approach can be augmented in such a way that it should accommodate an appropriate and consistent analysis of kéz 'hand' as used in (12b). I tackle this task in section 4. On the other hand, in languages like Norwegian nouns corresponding to the Hungarian kéz 'hand' type behave differently inasmuch as they admit the inalienable
interpretation even in sentences corresponding to (12b). I discuss these issues in section 5 .

## 4. Extending the analysis

In this section I intend to extend the analysis in two directions. On the one hand, I show how I analyze kéz 'hand' type relational nouns when they are used in a non-inalienable sense. On the other hand, I discuss how I envisage the analysis of the szomszéd 'neighbour' type, which differs from the previous type in a significant respect.

### 4.1. The non-inalienable use of kéz 'hand'

Consider the example in (12b), repeated here for convenience.
(12) b. Péter fel-emel-te a kez-et.

Peter.NOM up-raise-PAST.3SG.DEF the hand-ACC 'Peter raised the hand. / * Peter $_{(\mathrm{i})}$ raised $^{\text {his }}(\mathrm{ij})$ hand.'

The challenge here is twofold. On the one hand, we have to account for the omissibility of the possessor argument of kéz 'hand' we postulated in section 3. On the other hand, we have to ensure in a principled manner that the sentence in (12b) must not have a reading on which Peter raised his own hand (see the starred translation in (12b)). The first solution that jumps to mind is a kind of a suppression analysis developed by Barker (1995) for English relational nouns. The essence of his approach is as follows. ASs are attributed to several groups of relational nouns in their relevant uses. In order to account for cases in which no argument is present in the construction, Barker assumes that the argument in question is suppressed. In the spirit of the standardly applied general suppression analysis (in the treatment of passivization, for instance), the basic idea is that the suppressed argument (which does not appear in the constituent structure) is existentially bound in the AS of the nominal predicate. Barker's analysis can be translated into our framework and analysis so far in the following way.


Obviously, this solution easily captures the optionality of the possessor argument. However, it fails to exclude the inalienable interpretation of kéz 'hand' in (12b). The reason for this is as follows. If the existentially bound argument is present in the AS then it is a mystery why its trivial, most natural interpretation is strictly prohibited.

On the basis of the above considerations, I assume that a deeper, more radical process takes place in the form of a lexical redundancy rule: semantically it creates a non-inalienable/alienated noun from an inalienable one. This means that it affects even the LCS of the noun by removing its relational complement. As a consequence, no complement from the LCS is projected to AS, thus, there is no AS at all. Consider:

b. kéz $z_{2}, \mathrm{~N}$ : 'HAND ${ }_{2}$ \{ body-part \}'

The important point is that this radical change creates a noun that is used in a non-inalienable, non-relational sense from the relevant perspective. For instance, (12b) can have a reading on which Peter raised John's inalienable hand, but in this situation the hand was non-inalienable from Peter's perspective.

The issue of the inalienable vs. alienable uses of body part nouns has also been addressed in the typological literature. It is quite widely assumed that an inalienable $\rightarrow$ alienable process takes place in these cases, cf. Nichols (1988). However, the opposite view can also be found. For example, rather interestingly and surprisingly Herslund and Baron (2001: 13) claim that body part nouns like nose are non-relational to begin with and their relational use or interpretation is the result of a special process.

My stance is this: in the case of this type of relational nouns (body part nouns) the formal treatment of the inalienable vs. non-inalienable contrast is unquestionably necessary. It is crucial that this contrast is not absolute: it is relative in the sense that it is perspective-based. (For example, if I raise Mary's inalienable hand then that limb is viewed, and in my analysis, it is treated, as being non-inalienable.) As regards the appropriate linking of the members of this contrast: I assume that the process goes from inalienable to non-inalienable.

### 4.2. The szomszéd 'neighbour' type

Relational nouns like szomszéd 'neighbour', apa 'father', nagymama 'grandmother', etc., which I call "social relation" nouns, behave rather differently in a significant respect. Consider the following example.
(16) Péter bosszant-ja a szomszéd-ok-at. Peter.NOM annoy-PRES.3SG.DEF the neighbour-PL-ACC 'Peter is annoying the neighbours.'

This sentence, as opposed to (12b), can be felicitously interpreted in such a way that Peter annoys his own neighbours (despite the fact that the relational noun occurs in a non-possessive construction). Interestingly, (16) can also mean that Peter annoys the speaker's (and/or the listener's) neighbours. These nouns appear to behave in a more uniform manner across languages. I claim that the explanation for this is that they cannot be interpreted inalienably. Let us take a look at Asudeh's (2005) example.
(17) Context - Hermit A to Hermit B:

John saw a neighbour downtown. \#It wasn't his neighbour, though.
The special context ensures that the noun neighbour in the first sentence cannot be interpreted as related either to the speaker (Hermit A) or to the listener (Hermit B), given that hermits, by definition, do not have neighbours. (Compare this with the possible interpretations of (16) discussed above.) Thus, the only appropriate scenario for the first sentence is that John saw one of his own neighbours. It is for this reason that the second sentence is semantically deviant in this context. A person can only be the neighbour of someone (another neighbour): no non-inalienable/alienated interpretation is available in the case of these social relation nouns. Compare (17) with (18).
(18) John raised a hand. It wasn't his hand, though.

As I argued above, body part nouns do also admit the noninalienable/alienated interpretation; hence the contrast between (17) and (18).

Given the absolutely inalienable character and the consequential behaviour of relational nouns of the szomszéd 'neighbour' type, that is, social relational nouns, the following alternative analyses suggest themselves.
(A) It can be assumed that these relational nouns have no ASs: they only have LCSs, in the spirit of Uriagereka (1995), for instance.
(B) It can be postulated that they have LCSs and optional ASs, that is, the AS argument is optional, in which case the corresponding LCS complement is existentially bound, cf. Asudeh (2005).
(C) In the spirit of Barker (1995), we can develop a suppression account by assuming that there are two lexical entries for words like szomszéd 'neighbour'. The basic lexical form is szomszéd ${ }_{l}$ 'neighbour ${ }_{1}$ ', which is a relational noun with an LCS and an obligatory AS, and we derive szomszéd $_{2}$ 'neighbour ${ }_{2}$ ' from it by suppressing, that is existentially binding, the argument in AS.

Notice that all the three solutions enable us to capture the relevant empirical generalizations about the behaviour of szomszéd 'neighbour' nouns: the optionality of the syntactic expression of the complement and its obligatory presence in interpretation (cf. strict inalienability). In the representational system of lexical forms I have introduced for the purposes of the present paper, below I show the crucial aspects of these solutions. ${ }^{4}$

The (A) solution would require the following simple lexical form.

```
szomszéd, N: 'NEIGHBOUR { sr/m
```

This entry only contains the LCS dimension, no AS. The relevant complement is always present in the LCS and its optional syntactic reflex has the status of adjunct-like modification; therefore, it does not appear in this lexical representation.

The (B) version would make the entire AS, containing the possessor argument, optional.


This representation is my "translation" of Asudeh's (2005) treatment of the English word neighbour into my framework for the analysis of szomszéd 'neighbour' in Hungarian. The significant parts of his lexical entry for my present purposes are as follows.
(21) neighbour: N
$(\uparrow$ PRED $)=$ 'neighbour'
$\left((\uparrow \text { OBL })_{\sigma}=\left(\uparrow_{\sigma} \mathrm{ARG}\right)\right)$

[^139]I take the equations $(\uparrow \operatorname{PRED})=$ 'neighbour' and $\left((\uparrow \mathrm{OBL})_{\sigma}=\left(\uparrow_{\sigma} \mathrm{ARG}\right)\right)$ to indicate that in his approach AS is optional. According to him, when there is no (OBL) argument realized, the corresponding LCS complement is existentially bound. (It is an additional complication and contrast that in the case of the Hungarian noun here and in all my previous work I attribute the (POSS) function to the relevant complement/argument.)

The (C) option is the classical suppression analysis formulated in my current framework. Consider:
a. szomszéd ${ }_{1}$, N: ‘NEIGHBOUR ${ }_{1}$

b. szomszéd ${ }_{2}$, $\mathrm{N}^{\text {: 'NEIGHBOUR }} 2$, $\quad$ sr $\left./ \mathrm{m}_{1}, \mathrm{sr} / \mathrm{m}_{2}\right\}$ '
$<\mathrm{sr} / \mathrm{m}_{1}>$
Ø
日x

The standard suppression idea, adopted here, too, is that szomszéd ${ }_{1}$ 'neighbour ${ }_{1}$ ' does have a strict AS, cf. (22a), and when the suppression lexical redundancy rule is applied to this lexical entry, the argument in the AS will be associated with the $\varnothing$ grammatical function, and at the same time, it will be existentially bound, cf. (22b). Thus, the fundamental difference between (B) and $(\mathrm{C})$ is that in the former the entire AS is optional and existential binding takes place in LCS, while in the latter AS is obligatory and suppression and existential binding takes place at this level.

Remember that I rejected this Barker (1995) style suppression in AS analysis in the case of kéz 'hand' nouns for reasons discussed above. However, given the significantly different behaviour of szomszéd 'neighbour' nouns, I find this solution absolutely tenable and feasible here. Moreover, of the three theoretically possible analyses of szomszéd 'neighbour' presented above, I opt for this version, that is, alternative (C). My main motivation for this is that it is this kind of account that is closest in spirit to my analysis of kéz 'hand' nouns. Consider the following comparison.

- Similarities:
- in the analysis of both types the base form of the noun has obligatory LCS and AS;
- in both cases some kind of a reduction (deletion or suppression) takes place at some level.
- Differences:
- in the case of the kéz 'hand' type, suppression and existential binding affects the AS;
- in the case of the szomszéd 'neighbour' type, the deletion of a complement affects the LCS.


## 5. On relational nouns in Norwegian

Norwegian presents a significant contrast to Hungarian inasmuch as in the former even body part relational nouns exhibit the same behaviour as social relation nouns. That is, the Norwegian counterpart of (12b) is grammatical even on the inalienable reading. Consider:
(23) Peter løftet hånden

Peter.NOM raised hand.DEF
'Peter raised the/his hand.'
Thus, the following question needs to be addressed. In the previous section I pointed out that the behaviour of Hungarian social relation nouns like szomszéd 'neighbour' does not necessarily call for a strict AS and subsequent suppression analysis, and I opted for this version for the sake of at least partial similarity to the analysis of body part nouns like kéz 'hand'. Now the lack of the behavioural contrast between these two major relational noun types in Norwegian may suggest that in this language there is no strong motivation for the strict AS and subsequent LCS complement deletion analysis, and it should be restricted to languages like Hungarian.

My claim is that there is indirect, independent, LFG-analysis-based motivation for the assumption that even Norwegian body part nouns should be analyzed as having an AS with a possessor argument. Lødrup (2009) discusses possessor raising in Norwegian. Consider his examples.
(24) Han kysset henne på føttene. (transitive)
he kissed her on feet.DEF
lit.: 'He kissed her on the feet.'
(25) Han tråkket henne på føttene. (unergative)
he stepped her on feet.DEF
lit.: 'He stepped her on the feet.'
The essence of Lødrup's account is illustrated in the simplified f-structure representations of these sentences in (26) and (27).


The most important point from our present perspective is that in this analysis the body part noun is assumed to have an AS containing a possessor argument, which is functionally controlled by a thematic object of the matrix predicate in (26) and by a non-thematic object of the matrix predicate in (27). Thus, if this approach to these phenomena is tenable then it provides independent motivation for us to assume that Norwegian body part nouns do have Ass.

In the next section I also discuss that, following from the general control principles of LFG, the (POSS) function in this analysis must be taken to be semantically unrestricted.

## 6. A note on possessor arguments

In this section I discuss independent, LFG-analysis-specific considerations supporting the view that it is reasonable to assume that (at least) body part type relational nouns have ASs and they contain a possessor argument. There are several linguists who tend to query the nature of this assumption. For instance, one of my anonymous reviewers has written this. "It is dubious, in general, that "possessor arguments" are listed in the argument structure of any nominal head." Ash Asudeh (p.c., 28.07.2009.) has expressed a similar concern.

From a Hungarian perspective, I have argued for this assumption in numerous papers, especially in the context of analyzing complex event nominalization, cf. Laczkó (1995), (2000), (2004), for example. (In Laczkó (2007) I develop a radically novel analysis of Hungarian possessive DPs but my proposal here is compatible with the relevant aspects of that analysis.) In addition, I have consistently claimed that the (POSS) function, at least in languages like Hungarian, is semantically unrestricted.

In the previous section I showed that Lødrup (2009), in his LFG analysis of possessor raising constructions in Norwegian, postulates that body part nouns have a possessor argument. Moreover, on his account these possessors are functionally controlled. Given the standard assumptions about functional control in LFG, this automatically means that the (POSS) function of this argument must be taken to be semantically unrestricted, cf. Bresnan (1982).

According to Bresnan's (2001) analysis even originally non-relational nouns in English possessive constructions have ASs: a possessive predicative template augments their lexical forms:
a. hat $_{1}, \mathrm{~N}$ 'HAT $<>$ '
b. hat ${ }_{2}, \mathrm{~N}$ 'HAT-OF $<(\uparrow$ POSS $)>$ '

In other words, the template creates a relational noun from an ordinary noun, and the argument in this new argument structure is mapped onto the (POSS) function. Although she does not overtly discuss this, Bresnan independently needs to postulate the subcategorizable and semantically unrestricted nature of (POSS) for her functional control treatment of verbal gerunds in English.

Sells (2009) discusses examples of the type shown in (29) and (30) ${ }^{5}$ in relation to the treatment of adnominal clauses with genitive subjects in Altaic and East Asian languages. What is important for our present purposes is that in this generalized schema an ordinary, originally non-relational noun is assumed to have a relational counterpart with an AS containing a possessor

[^140]argument when it occurs in this construction type: "informally, my bought book means that I stand in some relation $\mathbf{R}$ to the book and in this case the relation is that 'I bought it'" (Sells 2009, page 5 of the handout). This view is straightforwardly comparable to Bresnan's (2001) approach outlined above. Furthermore, in this case, too, for theory-internal reasons the (POSS) function must be taken to be semantically unrestricted, cf. functional control.

(the $\mathrm{NP}_{\text {poss }}$ position is empty)


We can conclude that there are several LFG analyses of several phenomena in a variety of languages that assume that relational nouns do have ASs, which contain an argument bearing the (POSS) function, and this function is semantically unrestricted. Here by relational nouns I mean both inherently relational nouns and those that are converted from ordinary, nonrelational ones.

## 7. Concluding remarks

In this paper I have argued for postulating that in Hungarian certain groups of (underived) relational nouns have argument structures in addition to their lexical-conceptual structures.

My claim is that these nouns come in the following two major varieties.
(A) Body part nouns like kéz 'hand' have both inalienable and noninalienable/alienated uses, and this is grammaticalized in the language.
(B) Social relation nouns like szomszéd 'neighbour' are always inalienable.

I postulate that Type (A), in its inalienable use, has obligatory argument structure, and I capture its alienated use by assuming that its complement is deleted from its lexical-conceptual structure (naturally, this also results in the elimination of its argument structure).

For the sake of some partial uniformity in the analysis, I assume that Type (B) also has obligatory argument structure, and I capture the optionality of the argument by invoking suppression, and corresponding existential binding.

Several part-whole relational nouns appear to reject suppression. Compare, in this respect, (9b), repeated here for convenience, with (31).
(9) Context: Lödd le azt a korongot! 'Shoot that disk.'
b. *A közep-et céloz-d!
the middle-ACC aim-IMPER.2SG
'Aim at the middle.'

Although in this particular context the interpretation of (9b) would be straightforward, and it could be easily accommodated by a suppression analysis, the relational noun közép 'middle' does not admit this.
(31) Context: Mosd le az ablakpárkányokat! 'Wash the window sills.'

Figyel-j a perem-ek-re!
pay.attention-IMPER.2SG the edge-PL-SUBL
'Pay attention to the edges.'
Here we have a minimal pair contrast. The contextual setting is exactly of the same type in (9) as in (31), and in the latter example the suppression-based interpretation is readily available.

## References

Alberti, Gábor (1995) Role assignment in Hungarian possessive constructions. In Kenesei, István (ed.) Approaches to Hungarian, Vol. 5. Levels and Structures. Szeged, JATE, 11-28.
Asudeh, Ash (2005) Relational nouns, pronouns, and resumption. Linguistics and Philosophy 28(4): 375-446.
Barker, Chris and David Dowty (1993) Non-verbal thematic proto-roles. In Schafer, Amy (ed.), Proceedings of NELS 23, vol. 1, GSLA, Amherst, 4962.

Barker, Chris (1995) Possessive Descriptions. Stanford, CA: CSLI Publications.
Baron, Irene, Michael Herslund and Finn Soerensen (eds.) (2001) Dimensions of Possession. Typological Studies in Language 47. Amsterdam: John Benjamins.
Bartos, Huba (2000) Az inflexiós jelenségek szintaktikai háttere [The syntactic background of inflectional phenomena]. In Kiefer, Ferenc (ed.) Strukturális magyar nyelvtan 3. Morfológia [Structural Hungarian Grammar 3. Morphology]. Budapest: Akadémiai Kiadó, 653-762.
Bresnan, Joan (1982) Control and complementation. In Bresnan, Joan (ed.) The Mental Representation of Grammatical Relations. Cambridge, MA: MIT Press, 282-390.
Bresnan, Joan (2001) Lexical-Functional Syntax. Oxford: Basil Blackwell.
Castillo, Juan Carlos (2001) Thematic Relations between Nouns. Ph.D. Dissertation. University of Maryland at College Park.
Chisarik, Erika and John Payne. 2003. Modelling possessor constructions in LFG: English and Hungarian. In Butt, Miriam and Tracy H. King (eds.) Nominals: Inside and Out. Stanford: CSLI Publications, 181-199.
É. Kiss, Katalin (2002) The Syntax of Hungarian. Cambridge: Cambridge University Press.
Grimshaw, Jane (1990) Argument Structure. Cambridge, Mass.: The MIT Press.
Hadrovics László (1969) A funkcionális magyar mondattan alapjai [Fundamentals of Hungarian Functional Syntax]. Budapest: Akadémiai Kiadó.
Herslund, Michael and Irene Baron (2001) Introduction: Dimensions of possession. In Baron, Herslund and Soerensen (2001:1-25).
Laczkó, Tibor (1995) The Syntax of Hungarian Noun Phrases: A LexicalFunctional Approach. Frankfurt am Main: Peter Lang.
Laczkó, Tibor (2000) Derived nominals, possessors and lexical mapping theory in Hungarian DPs. In Butt, Miriam and Tracy H. King (eds.) Argument Realization. Stanford: CSLI Publications, 189-227.
Laczkó, Tibor (2004) Grammatical Functions, LMT, and Control in the Hungarian DP Revisited. In Butt, Miriam and Tracy H. King (eds.) Proceedings of the LFG '04 Conference. On-line publication: 2004 CSLI Publications, ISSN 1098-6782, http://wwwcsli.stanford.edu/publications/LFG2/lfg04.html.
Laczkó, Tibor (2007) Revisiting Possessors in Hungarian DPs: A New Perspective. On-line publication: 2007 CSLI Publications, ISSN 10986782, http://www-csli.stanford.edu/publications/LFG2/lfg04.html, 343362.

Laczkó, Tibor (2008) A relációs főnevek [Relational nouns]. In Kiefer, Ferenc (ed.) Strukturális magyar nyelvtan 4. A szótár szerkezete [Structural Hungarian Grammar. Volume 4. The Structure of the Lexicon]. Budapest: Akadémiai Kiadó. 2008, 323-503.
Lødrup, Helge (2009) Looking possessor raising in the mouth: Norwegian possessor raising with unergatives. This volume.
Nichols, Johanna (1988) On alienable and inalienable possession. In Shipley, William (ed.) In Honor of Mary Haas: From the Haas Festival Conference on Native American Linguistics. Berlin: Mouton de Gruyter, 557-609.
Sells, Peter (2009) Adnominal clauses with genitive subjects across Asia. Paper presented at LFG'09, Cambridge University.
Szabolcsi, Anna (1994) The noun phrase. In Kiefer, Ferenc and Katalin É. Kiss (eds.),The Syntactic Structure of Hungarian. Syntax and Semantics 27, New York: Academic Press, 179-274.
Uriagereka, Juan (1995) Warps: Some thoughts on categorization. In Echepare, Ricardo and Viola Miglio (eds.) University of Maryland Working Papers in Linguistics 3, 256-308. College Park: University of Linguistics.

## Acknowledgements

I am grateful for very useful comments from the participants of LFG '09, especially from Ash Asudeh. My thanks are due to Helge Lødrup for help with the Norwegian data. The research reported here was supported, in part, by an OTKA research grant (No. K72983) and the Research Group for Theoretical Linguistics of the Hungarian Academy of Sciences at the Universities of Debrecen, Pécs and Szeged.
email: laczkot "at" delfin "dot" unideb "dot" hu

# LOOKING POSSESSOR RAISING IN THE MOUTH: NORWEGIAN POSSESSOR RAISING WITH UNERGATIVES 

Helge Lødrup<br>University of Oslo<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>\section*{CSLI Publications}<br>http://csli-publications.stanford.edu/


#### Abstract

It has always been a standard assumption in the literature that possessor raising (as in She kissed him on the cheek) is limited to transitive verbs, and the theory of possessor raising has been designed to capture this restriction. This paper shows that Norwegian possessor raising is productive not only with transitive verbs, but also with unergative verbs (e.g. Han tråkket henne på føttene 'he stepped her on the feet'). The 'raised' possessor is argued to be non-thematic with unergative verbs. Apart from this difference, possessor raising with transitive and unergative verbs is very similar syntactically. It is proposed that possessor raising is a unitary phenomenon in Norwegian, and it is shown how Lexical Functional Grammar can give a unified analysis in a natural way. The similarity of possessor raising to equi ("control") and raising in the verbal domain is also discussed.


## 1 Introduction ${ }^{1}$

The term possessor raising (or possessor ascension) is used for sentences such as (1). The intuition behind this term is that the object is 'raised' from the possessor position of the body part noun phrase. No raising is involved in the analysis to be proposed here, but the term is kept as a descriptive term.

## (1) She kissed him on the cheek.

Possessor raising is a traditional topic in syntactic research. It has been claimed to belong to "the core of the grammatical function changing processes that are allowed by universal grammar" (Baker 1988:11), even if it has never received the same attention as for example the passive. However, it was often discussed in the theoretical literature some time ago, both in Relational Grammar (Frantz 1981:30-31, Perlmutter and Postal 1983, Blake 1990:99-103), and Principles and Parameters Theory (Baker 1988:268-277). There are also descriptions of possessor raising in various languages (e.g. Blake 1984, O'Connor 1996). The literature on the grammar of possession also contains some discussion of possessor raising, including criticism of its

[^141]traditional conception (see e.g. Chappell and McGregor 1996, Heine 1997, Payne and Barshi 1999).
A complication is that the term possessor raising has been used for two phenomena, which are in principle completely different. The term will be used here about the grammatical construction found in sentences such as (1) above. The relevant sentences have a transitive verb, which gives its internal role to an object. This object is understood as the possessor of a body part noun, which is the object of a locative preposition (see Levin 1993:71-72 on possessor raising in English).
It should be mentioned that possessor raising to subject is also often assumed (e.g. Blake 1990:99-102, Baker 1988:274). An example is (2). Possessor raising to subject is difficult to delimit, and needs further study. It plays a minor role in this paper.
(2) Han verker i leddene.
he aches in joints.DEF
His joints ache.
Possessor raising seems to be a common phenomenon in the languages of the world. In this respect, it differs from the other process sometimes called possessor raising. This construction, which will be called the dative external possessor construction, seems to be common in Europe only (Haspelmath 1999). An example is the French (3) (from Vergnaud and Zubizarreta 1992:597).
(3) Le médecin leur a examiné la gorge.
the doctor them.DAT has examined the throat
The doctor examined their throats.
In (3), the body part noun is the direct object, while the possessor is realized as a dative. In (1), on the other hand, the body part noun is the object of a preposition, while the possessor is the direct object.
The dative external possessor construction is less restricted than possessor raising. The dative possessor is not included in the verb's regular valency, differing from the direct object possessor. There are fewer restrictions on the syntax and semantics of the verb, and there are connections to other uses of the dative (see e.g. Guéron 1985, Neumann 1996, Lee-Schoenfeld 2006).
The distinction between possessor raising and the dative external possessor construction is in principle sharp. One difference that is important to this paper is that the dative external possessor construction has no transitivity requirement, as shown in the French sentences (4)-(5).
(4) Une pierre lui tombe sur la tête. a stone him.DAT falls on the head A stone falls on his head.
(5) Je lui ai marché sur les pieds. I him.DAT have stepped on the feet I stepped on his feet.

Possessor raising, on the other hand, has always been assumed to require a transitive verb. This is not only a descriptive fact of English (as stated e.g. in Levin 1993:71-72) and other languages, it has also been important to theories of possessor raising.
As possessor raising was originally conceived in Relational Grammar, the initial object of the verb was the body part noun phrase with the possessor. This possessor raised out of this object to become a new object. The requirement for an initial object was crucial, because it made the rule conform to "The Relational Succession Law" of Relational Grammar (Perlmutter and Postal 1983), which says that a raised element must take the syntactic function of the element that it raises out of.
The transitivity requirement was also important to Baker 1988. He proposed an analysis of possessor raising in which the body part noun incorporates 'abstractly' into the verb (Baker 1988:273). In his framework, this causes the (underlying) possessor to be governed by the verb, just like an object (Baker 1988:274). Baker points out that this analysis is only possible when the raised noun phrase is the possessor of a transitive verb's direct object, or of an unaccusative verb's surface subject. He claims that "this prediction is correct across languages" (Baker 1988:274).
It has been noticed that English has a couple of fixed expressions which could be taken to represent possessor raising with an intransitive verb. An example is (6).
(6) Don't look a gift horse in the mouth.

## 2 The basic Norwegian facts

Norwegian is like English in not having a productive dative external possessor construction, ${ }^{2}$ as shown by (7)-(8), which are word-by-word translations of the French examples (3)-(4).

[^142](7) *Legen har undersøkt dem halsen. doctor.DEF has examined them throat.DEF The doctor has examined their throats. [intended]
(8) *En stein faller ham på hodet. a stone falls him on head.DEF A stone falls on his head. [intended]

Norwegian also lacks morphological case, apart from a nominative - oblique distinction in some pronouns. It also does not have any direct equivalents to the "free" datives of case languages such as German, except in some fixed expressions and archaisms (Western 1921:142-44, Faarlund et al. 1997:723). An example is (9) (from the author Bjørnstjerne Bjørnson, taken from Western 1921:142-43).
(9) Han gadd ikke staa nogen til regnskap. he bothered not stand anybody to account He did not bother to account to anybody.

Faarlund et al. 1997:719-20 say that possessor raising in Norwegian is primarily used when the verb is transitive. However, they say that a couple of intransitive verbs can be used in more or less fixed expressions. (A similar claim can be found in Lødrup 1999:385 note 10.) Their examples are (10)(11).
(10) Dei lo sjefen opp i ansiktet. they laughed boss.DEF up in face.DEF They laughed in the boss' face.
(11) Alt maset gjekk meg på nervane. all nagging.DEF went me on nerves.DEF All the nagging got on my nerves.

Western 1921:142-44 also gives examples with intransitive verbs, such as (12) (from the author Vilhelm Krag).

The closest parallel to a dative external possessor construction in Modern Norwegian is a construction with a PP possessor, as in example (ii). This possessor could be taken to be external or internal to the noun phrase, see Lødrup 2009 for discussion.
(ii) Legen har undersøkt halsen på dem. doctor.DEF has examined throat.DEF on them The doctor has examined their throats.
(12) Rædselen fra 1801 sad endnu Hvermand i Blodet.
fear.DEF from 1801 sat still everybody in blood.DEF
The fear from 1801 was still in everybody's blood.
Western 1921:143 says that the intransitive construction is primarily possible in the literary language. Today, most of his examples must be considered unacceptable and/or fixed expressions.
It is striking that most of Western's examples have non-agentive verbs and inanimate subjects. It has never been noticed that Norwegian has a productive option for sentences such as (13) with agentive intransitive verbs, i.e. unergative verbs.
(13) Han tråkket henne på føttene.
he stepped her on feet.DEF
He stepped on her feet.
Example (13) looks similar to sentences with the dative external possessor construction in e.g. French and German (compare example (5) above). However, it is impossible to assume that this construction exists in Norwegian since it is not productive with transitive and unaccusative verbs (see (7)-(8) above). A better alternative would be to take example (13) as an instance of possessor raising.
Regular possessor raising with transitive verbs requires that the verb denote some form of physical contact (see Levin 1993:71-73). Looking at Norwegian unergative verbs that fit this description, it is difficult to avoid the conclusion that (what could be seen as) possessor raising is completely productive. Examples such as (14)-(18) are easily found by searching the internet. (Sentences found on the internet are marked "auth".)
(14) Da bokset bestandig mormor ham i magen. (auth)
then boxed always grandma him in stomach.DEF
Grandma would then always give him a punch in his stomach.
(15) Frøken ... pirket meg i ryggen med pekestokken. (auth)

Miss poked me in back.DEF with pointer.DEF
The teacher poked me in my back with the pointer.
(16) Mor smilte og rusket ham i håret. (auth)

Mother smiled and rumpled him in hair.DEF
Mother smiled and rumpled his hair.
(17) det er [ikke] noe lurt å ... klå henne på rompa. (auth)
it is not any smart to paw her on behind.DEF
It isn't very smart to paw her behind.
(18) Eminem spyr ham i ansiktet. (auth)

Eminem vomits him in face.DEF
Eminem vomits in his face.

There are also sentences in which there is physical contact in an extended sense only, such (10) above and (19).
(19) Legen bør da lyse deg i halsen. (auth)
physician.DEF should then light you in throat.DEF
The physician should then shine a light in your throat.
Without a body part noun as an object of the preposition, the construction is not possible, consider examples (20)-(21).
(20) *Hun tråkket ham på mobilen.
she stepped him on cell-phone.DEF
She stepped on his cell phone. [intended]
(21) *Hun pirket ham i stolen. she poked him in chair.DEF She poked on his chair. [intended]

Nouns denoting garments can be used in some cases; an example is (22). These nouns can also be used in regular possessor raising with transitive verbs, as in example (23).
(22) Antonsen river han i skjorten. (auth)

Antonsen tears him in shirt.DEF
Antonsen tears his shirt.
(23) jeg holdt arrestanten i kraven. (auth)
$I$ held prisoner.DEF in collar.DEF
I held the prisoner by his/her collar.

## 3 Grammatical properties of unergative possessor raising

Examples (14)-(19) with unergative verbs look like possessor raising sentences with transitive verbs. They also share important grammatical properties. For example, a transitive and an intransitive verb can coordinate, both in the active and the passive, as shown in (24)-(25).
(24) Han både kysset og tråkket henne på føttene.
he both kissed and stepped her on feet.DEF
He both kissed (her feet) and stepped on her feet.
(25) Hun ble både kysset og tråkket på føttene.
she was both kissed and stepped on feet.DEF
Her feet were both kissed and stepped on.

An important point to be made is that the objects of the intransitive verbs should be considered non-thematic, in the sense that they do not get a thematic role from their governing verb. (They get a possessor role from the body part noun, see section 5.1.) This kind of possessor raising will be called unergative possessor raising, to distinguish it from regular possessor raising with transitive verbs. Evidence that the objects are non-thematic will be given in the discussion to follow.
The option of non-thematic objects is explained by the unergativity of the verbs in question. Unergative verbs are known to take a non-thematic object in certain constructions (consider 'Burzio's generalization'). A clear parallel is resultatives. It is well known that unergatives can also take a non-thematic object in resultative sentences (see e.g. Levin and Rappaport Hovav 1995). An example is (26).
(26) Han tråkket føttene hennes gule og blå.
he stepped feet.DEF her yellow and blue
He stepped her feet black and blue.
An alternative analysis in which the object is thematic would have to assume that the verbs in question are transitive in possessor raising sentences only. ${ }^{3}$ This kind of analysis could not account for the data given below, and in section $4 .{ }^{4}$

[^143]An argument that the object is non-thematic with unergative possessor raising concerns the adjectival passive. Unergative possessor raising differs from regular possessor raising with transitive verbs in not having adjectival passives, as shown in (27)-(28). This is expected when the object is nonthematic, because an adjectival passive can only be derived from a passive participle with a thematic subject (Carrier and Randall 1992, Levin and Rappaport Hovav 1995:43-46).
(27) Nykysset på munnen kom hun løpende. newly-kissed on mouth.DEF came she running Recently kissed on her mouth, she came running.
(28) *Nyklasket på baken kom hun løpende. newly-slapped on behind.DEF came she running Recently slapped on her behind, she came running. [intended]

The analysis of the object as non-thematic implies that the possessive object with unergative possessor raising is a regular, or direct, object, and not an oblique, or indirect, object. The reason is that a non-thematic object can only be a direct object; non-thematic indirect objects do not exist. This fact strengthens the parallel to regular possessor raising with transitive verbs, because it is uncontroversial that the object with regular possessor raising is a direct object. ${ }^{5}$
The syntactic restrictions on unergative possessor raising follow automatically when it is assumed that the possessor is a non-thematic direct object. The verb must be unergative to license the non-thematic object (when some fixed expressions and archaisms are put aside). Sentences (29)-(32) are therefore impossible.
(29) *Vannet rant henne i håret.
water.DEF ran her in hair.DEF
The water ran in her hair. [intended]
(30) *Han helte henne vann i håret.
he poured her water.DEF in hair.DEF
He pored water in her hair. [intended]
(31) *En tann knakk ham i underkjeven.
a tooth broke him in lower-jaw.DEF
A tooth broke in his lower jaw. [intended]

[^144](32) *Ola knekte ham en tann i underkjeven.

Ola broke him a tooth in lower-jaw.DEF
Ola broke a tooth in his lower jaw. [intended]
The non-thematicity of the object with unergative possessor raising makes it different from the dative possessor in the dative external possessor construction. The dative possessor is often assumed to have a semantic relation to the verb, and it could be considered a benefactive or a malefactive, or an affectee (see e.g. Guéron 1985, Lee-Schoenfeld 2006).
The non-thematicity of the object with unergative possessor raising also makes it different from the (equivalents of) "free" datives that can be found in some fixed expressions and archaisms in Norwegian (Western 1921:14244, Faarlund et al. 1997:723), such as (8) above, reproduced here as (33).
(33) Han gadd ikke staa nogen til regnskap.
he bothered not stand anybody to account
He did not bother to account to anybody.
The "dative" nogen 'anybody' in (33) must be assumed to get a thematic role in its (derived) object function. There is no alternative source for a thematic role. (For example, the object of the preposition cannot be a source.) With unergative possessor raising, on the other hand, the object gets its only thematic role in its function as a possessor of the body part noun. A related difference concerns the option of alternating with a PP: The "dative" alternates with a PP with the preposition for 'for', like many regular indirect objects. With unergative possessor raising, on the other hand, the object cannot alternate this way.

## 4 Unergative - transitive alternations

4.1 The case of nappe 'remove, pull'

Investigating the range of unergative possessor raising, it is an important fact that the line between unergative verbs and transitive verbs is thin. Consider for example the verb nappe 'remove, pull'. This verb can take an object that denotes a thing being suddenly removed from its owner, as in (34).
(34) Han nappet lua hennes.
he removed cap.DEF her
He removed her cap.
When the verb takes an oblique (and no object), the verb has a slightly different meaning. There is not necessarily anything that is removed, and the verb just denotes a "pulling" movement. An example is (35).
(35) Jeg napper i barten hans.

I pull in moustache.DEF his I pull his moustache.

The meaning of (35) is exactly the meaning found with possessor raising, as in (36).
(36) Jeg ... napper ham i barten. (auth) I pull him in moustache.DEF I pull his moustache.

It is clear, then, that the basis for possessor raising must be the intransitive verb in (35), and not the transitive verb in (34).
A group of verbs show the same pattern as nappe 'remove, pull' (e.g. hugge 'cut', trykke 'press', sprute 'splash', rekke 'come up to'). These verbs can take a thematic object, but this object cannot normally denote a person. Even so, they can take an object denoting a person in possessor raising. This raising must be unergative possessor raising, based on an intransitivized version of the verb. ${ }^{6}$
The resultative construction gives many parallel examples of verbs being intransitivized to add a "new" non-thematic object (Levin and Rappaport Hovav 1995:37-39). For example, the verb spise 'eat' can intransitivize and get a non-thematic object in a resultative sentence. An example is (37).
(37) De spiste kjøleskapet tomt.
they ate fridge.DEF empty
They ate the fridge empty.

### 4.2 The case of $d r a$ 'pull'

A group of verbs that behave somewhat different from nappe 'remove, pull' can be represented by the verb dra 'pull'. This verb can take a thematic object
${ }^{6}$ The verb $t a$ is interesting in this resepct. Its basic meaning is 'take', but intransitive $t a$ with an oblique means 'touch', as in (i). This meaning is not available for transitive ta. The verb also means 'touch' in sentences with possessor raising, such as (ii), which must be unergative possessor raising.
(i) Han tok på skulderen hennes.
he took on shoulder.DEF her
He touched her shoulder.
(ii) Mamma påstår at Elton John tok henne på skulderen en gang. (auth) mama claims that Elton John took her on shoulder.DEF one time Mama claims that Elton John once touched her on the shoulder.
that denotes the thing or person being pulled along, as in (38). If the thing or person does not move, it is not realized as an object of the verb, but rather as the object of a preposition, as in (39).
(38) Han dro henne til hulen.
he pulled her to cave.DEF He pulled her to the cave.
(39) Han dro i henne.
he pulled in her He pulled her.

With possessor raising, the object of $d r a$ 'pull' can get a thematic or a nonthematic interpretation, as shown in (40)-(41).
(40) Han dro henne etter håret.
he pulled her after hair.DEF
He pulled her by her hair.
(41) Han dro henne i håret.
he pulled her in hair.DEF
He pulled her hair.
Sentence (40) has a thematic object. Its meaning is that 'he' pulled 'her' along by the hair, for example to his cave. This is regular possessor raising. Sentence (41) has a non-thematic object. Its meaning is that 'he' pulls 'her' hair without making 'her' move, maybe violently to cause 'her' pain, or softly to caress 'her'. This is unergative possessor raising.
The behavior of verbs such as nappe 'remove, pull' and dra 'pull' gives strong evidence for the non-thematicity of the object with unergative possessor raising.

### 4.3 Interaction with the conative alternation

Some transitive verbs that allow regular possessor raising participate in the conative alternation (Levin 1993:41-42). They then take an oblique and no object; compare (42) and (43).
(42) Hunden slikket hånden hans.
dog.DEF licked hand.DEF his
The dog licked his hand.
(43) Hunden slikket på hånden hans.
dog.DEF licked on hand.DEF his
The dog licked on his hand.

This pattern can be found with several verbs (e.g. klore 'scratch', gni 'rub', operere 'operate', male 'paint'). When these verbs take an oblique, they are syntactically identical to the intransitive verbs that take unergative possessor raising. This means that possessor raising sentences with these verbs, such as (44), are syntactically ambiguous.
(44) Hunden slikket ham på hånden.
dog.DEF licked him on hand.DEF
The dog licked his hand.
Example (44) has one analysis as regular possessor raising with the transitive verb slikke 'lick', and one analysis as unergative possessor raising with the corresponding intransitive verb. One might expect this syntactic ambiguity to be correlated with semantic ambiguity, but it does not seem to be.

## 5 Grammatical treatment

5.1 Regular possessor raising with transitive verbs

The treatment of regular possessor raising with transitive verbs is rather straight forward in Lexical Functional Grammar (LFG). Possessor raising is structure sharing between the object of the verb and the possessor of the oblique object. The shared argument realizes both the object role of the verb and the possessor role of the body part noun. The simplified f-structure is as in (45) for She kissed him on the cheek.


The verbs in question are equipped with the equation in (46). ${ }^{7}$

$$
\begin{equation*}
(\uparrow \mathrm{OBJ})=\left(\uparrow \mathrm{OBL}_{\theta} \mathrm{OBJ} \text { POSS }\right) \tag{46}
\end{equation*}
$$

[^145]This equation can be simplified if obliques are treated as in Bresnan 1982. She sees the preposition of an oblique as a kind of case marker. The preposition and its object are co-heads, which gives a "flat" functional structure for the oblique with the preposition represented as a feature. The equation in (46) could then be replaced by the one in (47).
$(47)(\uparrow \mathrm{OBJ})=\left(\uparrow \mathrm{OBL}_{\theta} \mathrm{POSS}\right)$

The analysis given is not affected by the criticism that has been raised against the traditional conception of possessor raising, in which the possessor (in some sense) was moved out of the body part noun phrase.
Payne and Barshi 1999:7 say that "there is no 'raising' of anything" in sentences such as She kissed him on the cheek. What they have in mind is that the corresponding sentence without the oblique (She kissed him) is a complete sentence with a regular object. This intuition is captured here by treating 'him' as a regular object that realizes the verb's patient role. At the same time, the analysis given accounts for the relation between the object and the understood possessor of the body part noun. Body part nouns usually require a possessor to be syntactically realized (when they denote actual parts of a body, and not for example objects for anatomical study, see e.g. Vergnaud and Zubizarreta 1992), and it can be assumed that they assign a thematic role to their possessor.
The analysis given is also not affected by the criticism of possessor raising in e.g. Blake 1990:102-3, Chappell and McGregor 1996:6-7, Heine 1997:163-64. They claim that sentences with and without possessor raising are not synonymous, so the possessor raising rule is not "meaning preserving". This kind of criticism is without force here. There is no "underlying structure" in which the body part noun is an object, and there is no concept of "meaning preservation" involved.

### 5.2 Unergative possessor raising

Sentences with regular and unergative possessor raising are very similar grammatically, except for the differences that follow from the thematicity or non-thematicity of the object (for example concerning the adjectival passive, see section 3). The natural analysis is, then, that possessor raising in Norwegian does not distinguish between intransitive and transitive verbs. To say it another way, possessor raising should be one rule that unifies a thematic POSS with an object that is thematic or non-thematic. The rule in (46) is all that is needed to do exactly this - it only needs to apply to intransitive as well as transitive verbs.
To be more exact, the intransitive verbs in question must be unergative, as shown above. This follows automatically from the way syntactic features are
assigned to thematic roles in Lexical Mapping Theory (Bresnan and Zaenen 1990:49, Bresnan 2001:309). A patient gets the feature [-r(estricted)], while an agent gets the feature [-o(bject)]. A non-thematic argument can only get the feature $[-\mathrm{r}]$; this follows from the understanding of "restricted" as "restricted to a particular thematic role". A verb can only take one [-r] argument in Norwegian (as in many other languages, see Bresnan and Moshi 1990, Alsina and Mchombo 1993). This is the reason a non-thematic object can only occur with unergative verbs, and not with unaccusative and transitive verbs. They already have a [-r] argument, the subject and the object, respectively. (Note that a locative does not stand in the way of a nonthematic object, because a locative gets the feature [-o].)
A reviewer asks where the unergative possessor raising structures are created. This question can be split in two, because there are two requirements for an unergative verb to undergo possessor raising: It needs an object and the equation in (46). Taking a non-thematic object is a general option for unergative verbs. The simplest assumption is that these objects are 'inserted freely', in the sense that an unergative verb can always have a lexical entry with a non-thematic object. Whether a resulting sentence is well formed depends upon other factors, including the availability of a thematic role from another source. The equation in (46) is inserted by a lexical rule, which is semantically conditioned. As mentioned above, possessor raising is only possible with verbs that denote physical contact (see Levin 1993:71-73), sometimes in an extended sense.
The account given of possessor raising to object could be generalized to include possessor raising to subject with unaccusative verbs. This task must be left to future research, because too little is known about possessor raising to subject. It is difficult to delimit, and it is not clear what the semantic conditions are.

### 5.3 The problem of possessives

The analysis given predicts that there can be no realized POSS with the body part noun in sentences with possessor raising. The reason is that this would create a conflict between the PRED of the POSS and the PRED of the object (and possibly other conflicts). Consider (48) and (49) with realized possessives. These sentences have a redundancy of expression that makes them less than perfect. Even so, they are not really unacceptable. ${ }^{8}$

[^146](48) Han kysset henne på føttene hennes.
he kissed her on feet.DEF her
He kissed her on her feet.
(49) Han tråkket henne på føttene hennes.
he stepped her on feet.DEF her
He stepped on her feet.

Examples (48)-(49) have a transitive and an unergative verb, respectively. An important difference between them is that example (48), with the transitive, could be an ordinary transitive sentence without possessor raising. With this analysis, it would be difficult to find anything wrong with it from a formal point of view. Only example (49), with the unergative, has no well formed analysis. It is difficult to avoid the unsatisfying conclusion that even if (48) and (49) are rather similar with respect to well-formedness, (48) is generated by the grammar, while (49) is not.

### 5.4 Possessor raising compared to equi and raising

Possessor raising shares interesting properties with equi (i.e. "control") and raising in the verbal domain. Standard examples of equi and raising are (50) and (51), in which the object is thematic and non-thematic, respectively.
(50) We persuaded him to come.
(51) We expected him to come.

Treating equi and raising the same way was proposed in Bresnan 1982 (see also Bresnan 2001:267-301). Her theory of control and complementation allows structure sharing between positions with or without thematic roles, as long as the shared argument gets a thematic role from at least one predicate (see the discussion of the Coherence Condition in Bresnan 2001:63). This is called functional control. Both persuade and expect have the equation in (52).
(52) $(\uparrow \mathrm{OBJ})=(\uparrow$ XCOMP SUBJ $)$

One similarity between possessor raising on the one hand, and raising and equi in the verbal domain on the other hand, concerns what functions share an argument. The function on the left-hand side of the equations in (46) and (52) is OBJ. Another option is SUBJ, both with raising and equi verbs (such
(ii) Han tråkket henne på den skadede tåen hennes. he stepped her on the hurt toe.DEF her He stepped on her hurt toe.
as seem and try), and with verbs that take possessor raising to subject. ${ }^{9}$ (The latter case is illustrated in example (2) above.)
Both with possessor raising and raising and equi in the verbal domain, the choice between SUBJ and OBJ for each verb follows from the lexical rule of functional control (Bresnan 1982:322), which picks out the core function that is lowest on the relational hierarchy
Both with possessor raising and raising and equi in the verbal domain, an argument is shared between a core function and an embedded 'subject' function. This embedded 'subject' function is SUBJ with raising and equi, and POSS with possessor raising. SUBJ and POSS are functions that are closely related, as has been discussed in different frameworks over the years. From an LFG point of view, it is important that they are both unrestricted (Laczkó 1997), and highest in their domain on the relational hierarchy.

Another similarity between possessor raising on the one hand, and raising and equi in the verbal domain on the other hand, concerns syntactic ambiguity. It was shown in section 4 that sentences can be syntactically ambiguous between unergative and regular possessor raising. In the same way, sentences can be ambiguous between raising and equi. (This has been discussed several times, see e.g. Ruwet 1991.) One example is the Norwegian accusative with infinitive construction. In the analysis of Lødrup 2008, some verbs always take raising in this construction (e.g. føle 'feel' and kjenne 'feel'), some always take equi (e.g. huske 'remember', oppdage 'discover'), and some can take both (e.g. se 'see' and høre 'hear'). Apart from the differences that follow from the raising-equi distinction, accusative with infinitive sentences share syntactic and semantic properties. The syntactically ambiguous sentences are not necessarily semantically ambiguous, similar to the ambiguity seen in sentences with possessor raising (section 4.3).

### 5.5 Possessor raising in transformational syntax

It was shown that the classical LFG theory of control and complementation in Bresnan 1982 can give a unified analysis of Norwegian possessor raising. It is interesting to see how the same data can be accounted for in transformational syntax.
Possessor raising has been treated in different ways through the years. Baker 1988:273 noted that a raising analysis was incompatible with then current Principles and Parameters theory. Creating a "new" object to a verb was prohibited by the Projection Principle. There was also a problem with the object role of the verb, because role assignment could not take place after
${ }^{9}$ In addition, $\mathrm{OBJ}_{\theta}$ is an option with some equi verbs in some languages.
It should be mentioned that it is not clear that possessor raising to subject is possible with a non-thematic subject.
movement. (It should be noted that this problem would not arise with unergative possessor raising.)
Current (versions of) Minimalism, on the other hand, would not prohibit a raising analysis of possessor raising with transitive verbs. However, there would have to be one difference from the classical conception of possessor raising. The PP must be present in underlying structure, because it cannot be "built" in the derivation. The possessor would have to raise from the object of this PP. In (versions of) Minimalism, it would not only be possible to raise the possessor to be a "new" object of the verb, it could also get a new thematic role after movement (see e.g. Hornstein 1999, Davies and Dubinsky 2004, Lee-Schoenfeld 2006). A Minimalist movement analysis along these lines is sketched in Ussery and Moonan 2004.
Unergative possessor raising would not in itself create new problems for Minimalism. Given the assumptions mentioned, it would be possible to give a unified analysis, in which regular and unergative possessor raising are treated the same way - just like in the classical theory of control and complementation in LFG (Bresnan 1982).

## 6 Conclusion

In Norwegian possessor raising, the raised possessor can be thematic or nonthematic. Apart from the differences that follow from this, possessor raising sentences share syntactic properties. LFG gives a framework that can account for this situation in a simple and enlightening way, by allowing structure sharing between positions with or without thematic roles.

## REFERENCES

Åfarli, Tor A. and Kristin M. Eide 2003 Norsk generativ syntaks. Oslo: Novus.
Alsina, Alex and Sam A. Mchombo 1993 Object asymmetries and the Chichewa applicative construction. In Sam A. Mchombo (ed.) Theoretical Aspects of Bantu Grammar. Stanford, CA: CSLI Publications. 17-45.
Baker, Mark C. 1988 Incorporation: A Theory of Grammatical Function Changing. Chicago: University of Chicago Press.
Blake, Barry J. 1990 Relational Grammar. London: Routledge.
Blake, Barry J. 1984 Problems for possessor ascension: Some Australian examples. Linguistics, 22, 4, 437-453.

Bresnan, Joan 1982 Control and complementation. In Joan Bresnan (ed.) The Mental Representation of Grammatical Relations. Cambridge, MA: MIT Press. 282-390.
Bresnan, Joan 2001 Lexical-Functional Syntax. Oxford: Blackwell.
Bresnan, Joan and Lioba Moshi 1990 Object asymmetries in comparative Bantu syntax. Linguistic Inquiry 21, 147-85.
Bresnan, Joan and Annie Zaenen 1990 Deep unaccusativity in LFG. In Katarzyna Dziwirek, Patrick Farrell and Errapel Mejías-Bikandi (eds.) Grammatical Relations: A Cross-Theoretical Perspective. Stanford, CA: CSLI Publications. 45-57.
Carrier, Jill and Janet H. Randall 1992 The argument structure and syntactic structure of resultatives. Linguistic Inquiry 23, 173-234.
Chappell, Hilary and William McGregor 1996 Prolegomena to a theory of inalienability. In Hilary Chappell and William McGregor (eds.) The Grammar of Inalienability: A Typological Perspective on Body Part Terms and the Part-Whole Relation. Berlin: Mouton de Gruyter. 3-30.
Davies, William D. and Stanley Dubinsky 2004 The Grammar of Raising and Control: A Course in Syntactic Argumentation. Malden, MA: Blackwell.
Faarlund, Jan Terje 2004 The Syntax of Old Norse. Oxford: Oxford University Press.
Faarlund, Jan Terje, Svein Lie and Kjell Ivar Vannebo 1997 Norsk referansegrammatikk. Oslo: Universitetsforlaget.
Falk, Hjalmar and Alf Torp 1900 Dansk-norskens syntax $i$ historisk fremstilling. Kristiania: Aschehoug.
Frantz, Donald G. 1981 Grammatical Relations in Universal Grammar. Bloomington: Indiana University Linguistics Club.
Guéron, Jacqueline 1985 Inalienable possession, PRO-inclusion and lexical chains. In Jacqueline Guéron, Hans-Georg Obenauer and Jean-Yves Pollock (eds.) Grammatical Representation. Dordrecht: Foris. 43-86.
Haspelmath, Martin 1999 External possession in a European areal perspective. In Doris L. Payne and Immanuel Barshi (eds.) External Possession. Amsterdam: Benjamins. 109-135.
Heine, Bernd 1997 Possession: Cognitive Sources, Forces, and Grammaticalization. Cambridge: Cambridge University Press.
Hellan, Lars 1988 Anaphora in Norwegian and the Theory of Grammar, Dordrecht: Foris.
Hornstein, Norbert 1999 Movement and control. Linguistic Inquiry 30, 1, 6996.

Kaplan, Ronald M. and Joan Bresnan 1982 Lexical-Functional Grammar: A formal system for grammatical representation. In Joan Bresnan (ed.) The Mental Representation of Grammatical Relations. The MIT Press, Cambridge, MA. 173-281.

König, Ekkehard and Martin Haspelmath 1998 Les constructions à possesseur externe dans les langues de l'Europe. In Jack Feuillet (ed.) Actance et Valence dans les Langues de l'Europe. Berlin: Mouton de Gruyter. 525-606.
Laczkó, Tibor 1997 Action nominalization and the possessor function within Hungarian and English noun phrases. Acta Linguistica Hungarica 44, 413-75.
Lee-Schoenfeld, Vera 2006 German possessor datives: Raised and affected. Journal of Comperative Germanic Linguistics 9, 2, 101-142.
Levin, Beth 1993 English Verb Classes and Alternations: A Preliminary Investigation. Chicago: University of Chicago Press.
Levin, Beth and Malka Rappaport Hovav 1995 Unaccusativity: At the Syntax - Lexical Semantics Interface. Cambridge MA: MIT Press.
Lødrup, Helge 1999 Inalienables in Norwegian and binding theory. Linguistics 37, 3, 365-388.
Lødrup, Helge 2008 The diversity and unity of the accusative with infinitive construction: A Norwegian perspective. Linguistics 46, 5, 891-916.
Lødrup, Helge 2009 External and internal possessors with body part nouns: The case of Norwegian. Manuscript. http://folk.uio.no/helgelo/
Neumann, Dorothea 1996 The dative and the grammar of body parts in German. In Hilary Chappell and William McGregor (eds.) The Grammar of Inalienability: A Typological Perspective on Body Part Terms and the Part-Whole Relation. Berlin: Mouton de Gruyter. 745779.

O’Connor, Mary C. 1996 The situated interpretation of possessor-raising. In Masayoshi Shibatani and Sandra A. Thompson (eds.) Grammatical Constructions: Their Form and Meaning. Oxford: Clarendon Press. 125-156.
Payne, Doris L. and Immanuel Barshi 1999 External possession: What, where, how and why. In, Doris L. Payne and Immanuel Barshi (eds.) External Possession. Amsterdam: Benjamins. 3-29.
Perlmutter, David M. and Paul M. Postal 1983 The Relational Succession Law. In David M. Perlmutter (ed.) Studies in Relational Grammar 1. Chicago: University of Chicago Press. 30-80.
Ruwet, Nicolas 1991 Raising and control revisited. In John Goldsmith (ed.) Syntax and Human Experience. Chicago: University of Chicago Press. 56-81.
Sæbø, Kjell Johan 2009 Self intensification and focus interpretation. In Bergljot Behrens and Cathrine Fabricius-Hansen (eds.) Structuring Information in Discourse: The Explicit/Implicit Dimension, Oslo Studies in Language 1, 1, 109-29. http://www.journals.uio.no/osla
Ussery, Cherlon and Robert Moonan 2004 Possessor Raising in English [Abstract]
http://www.cas.sc.edu/ling/activities/Englishpossessorraising.pdf

Vergnaud, Jean-Roger and Maria Luisa Zubizarreta 1992 The definite determiner and the inalienable construction in French and English. Linguistic Inquiry 23, 595-652.
Western, August 1921 Norsk riksmåls-grammatikk. Kristiania: Aschehoug.

Helge Lødrup
University of Oslo
Department of Linguistics and Scandinavian Studies
Pb 1102, Blindern,
NO-0317 Oslo, Norway
helge.lodrup@ilf.uio.no
http://folk.uio.no/helgelo/index.html

# ON THE STATUS OF CLITIC REFLEXIVES AND RECIPROCALS IN ITALIAN AND SERBIAN 

Maja Miličević<br>University of Belgrade<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

This paper discusses the status of Italian and Serbian clitic reflexives and reciprocals with respect to the phenomenon of split intransitivity. Assuming an analyses that treats clitics as non-argumental markers of the verb's derived intransitivity, numerous proposals have been put forward as to whether clitic reflexives and reciprocals are unaccusative or unergative. The issue still appears as problematic in the literature, due to the fact that compelling empirical evidence is available for both views.

As a solution to this problem, we adopt the approach of Alsina (1996), according to whom both verbal arguments remain implicitly present in the clitic forms, making reflexives and reciprocals behave as unaccusative in some contexts, and as unergatives in others. In addition, we look at patterns of reflexive and reciprocal marking of intransitive verbs in Italian and Serbian, and we show that reflexives are more closely related to non-derived unaccusatives, while reciprocals have a closer relation to non-derived unergatives. This is formally captured in the framework of Correspondence Theory (Ackerman and Moore 2001), in an analysis indicating that in reflexive-marked forms there is a progressive loss of agentivity, while the reciprocal-marked forms are characterised by a gradual decrease in patienthood.


## 1 Introduction

Italian and Serbian belong to the group of languages that use clitics as predominant reflexive and reciprocal markers. Both the Italian si and the Serbian se have caused much dispute, as they have been analysed both as short forms of argumental reflexive pronouns, and as morphological spell-outs of the verbs' derived intransitivity. ${ }^{1}$ Most current approaches assume them to be non-argumental morphological markers (see De Alencar and Kelling 2005 for a recent view to the contrary); however, assuming that clitic-marked reflexives and reciprocals are indeed intransitive, another important problem arises: are these forms unaccusative or unergative? ${ }^{2}$

Both views have been advocated in the literature, and proponents of both have offered compelling evidence in favour of their positions, pointing to somewhat of a paradox: clitic reflexives and reciprocals sometimes display unaccusative, and sometimes unergative behaviour. Consequently, any approach based on the assumption that the process of intransitive reflexive and reciprocal formation requires

[^147]a reduction of one of the verbal arguments (the external one in the case of unaccusative analyses, the internal one in unergative approaches) has to make additional stipulations to account for the full range of empirical facts.

The present paper builds on a somewhat different approach, formulated by Alsina (1996). This particular approach attempts to resolve the above paradox by arguing that clitic reflexives and reciprocals are neither exclusively unaccusative or exclusively unergative. Specifically, Alsina claims that the contradictory behaviour of clitic reflexives and reciprocals is best explained if it is assumed that no argument reduction takes place in the course of their formation, and that both their external and internal arguments are implicitly kept, even though they are mapped onto a single syntactic function.

Our paper brings additional evidence for this view, based on the patterns of reflexive and reciprocal marking with a wider range of intransitive verbs. We argue that another important problem for the unaccusative and unergative analyses (which we label as 'reductionist') lies in the fact that reflexives and reciprocals behave alike with respect to the unaccusativity diagnostics, which indicates that they are formed through the same argument structure alternation, ${ }^{3}$ but reflexives seem to have a closer relation to non-derived unaccusatives, whereas reciprocals appear to be closer to non-derived unergatives. This 'closer relation' manifests itself in the fact that intransitive verbs close to reflexives (in that they involve a single participant) are typically unaccusative (e.g. sedersi 'sit down' in Italian), while those close to reciprocals (in that they obligatorily involve at least two participants) are typically unergative (e.g. rukovati se 'shake hands' in Serbian). We formalise this using the Correspondence Theory of Ackerman and Moore (2001), an approach based on standard Lexical Mapping Theory on the one hand, and Dowty's (1991) Proto-Role approach to thematic roles on the other.

The paper is structured as follows. Section 2 gives an overview of the reductionist approaches and the evidence they use when arguing in favour of an unaccusative or unergative analysis. Section 3 introduces the non-reductionist analysis proposed by Alsina (1996), explaining how this approach deals with the unaccusative/unergative dilemma. Section 4 presents additional evidence for the nonreductionist approach, capturing it in a formal LMT-style representation. Lastly, section 5 gives some concluding remarks.

## 2 Reductionist approaches

The approaches which assume that Italian and Serbian clitic-marked reflexives and reciprocals are unaccusative or unergative can quite straightforwardly be described as reductionist approaches, as they necessarily assume a reduction of one of the

[^148]verb's arguments. To be precise, if the clitic forms are unaccusative, their external argument was suppressed in the process of reflexive derivation, and their internal argument was consequently promoted to a subject, similarly to what happens in passivisation (see Bouchard 1984; Grimshaw 1990; Van Valin 1990); if clitic reflexives and reciprocals are unergative, their internal argument was reduced and the subject simply remained the only argument present in the syntax (see Grimshaw 1982; Wehrli 1986; Chierchia 2004; Reinhart and Siloni 2004). Some of the pieces of evidence quoted in favour of each of these approaches are presented in the following two sections.

### 2.1 The unaccusative view

In Italian, clitic reflexives and reciprocals pattern with unaccusatives with respect to auxiliary selection, as the verbs associated with reflexive and reciprocal clitics invariably select the auxiliary essere 'be' (1a), typically used with unaccusatives (1b), and not avere 'have', reserved for transitive and unergative verbs (1c).
(1) a. Silvia e Sandra si sono vestite.

Silvia and Sandra REFL be.PRES.3pl dress.PAST.PART
'Silvia and Sandra dressed.'
b. Sandra è uscita.

Sandra be.PREs.3sG go.out.PAST.PART
'Sandra went out.'
c. Silvia ha pianto.

Silvia have.PRES.3sG cry.PAST.PART
'Silvia cried.'
Another oft-cited proof for the unaccusative view concerns the fact that absolute participles in Italian can be formed only from unaccusative verbs, as in (2a). Reflexives permit them too, as demonstrated by (2b).
(2) a. Arrivato Gianni, dovevamo partire.
arrive.PASt.part Gianni must.Imp.1pl leave.Inf
'Once Gianni had arrived, we had to leave.'
b. Vestitisi i bambini, potevamo uscire.
dress.PAST.PART-REFL the children can.IMP.1pl go.out.INF
'Once the children got dressed, we could go out.'
Moreover, reflexives pattern with unaccusatives in being incompatible with constructions having derived subjects, such as passives, raising predicates, predicative 'be' and frighten-type psychological predicates (see Bouchard 1984; Rizzi 1986; Wehrli 1986; Grimshaw 1990); the latter two are illustrated by the Italian examples in (3).
a. *Gianni si è affezionato.

Gianni Refl be.pres.3sg affectionate
‘Gianni is affectionate to himself.' (from Rizzi 1986, 83)
b. *Gianni si preoccupa.

Gianni REFL worry.PRES.3sG
'Gianni worries himself. ${ }^{4}$ (from Belletti and Rizzi 1988, 296)
Several other diagnostics are mentioned by scholars adhering to the unaccusative view, but they will not be discussed here. It should only be added that the reason for which only examples with reflexives are given is their higher frequency in the literature; reciprocals display the same behaviour in these contexts.

### 2.2 The unergative view

One of the facts most typically quoted in favour of the unergative approach is the impossibility to have partitive ne cliticisation in Italian with reflexives, alongside unergative verbs (Alsina 1996; Reinhart and Siloni 2004):
(4) a. Ne sono rimasti tre.
of-them be.PREs.3pl remain.PAST.PART three
'Three of them remained.' (from Rosen 1988, 64)
b. ${ }^{*} \mathrm{Ne}$ hanno telefonato tre.
of-them have.PREs.3pL telephone.PAST.PART three
'Three of them telephoned.' (ibid: 63)
c. *Se ne sono difesi parecchi.

REFL of-them be.PREs.3PL defend.PAST.PART several
‘Several defended themselves.' (ibid: 94)
Reinhart and Siloni (2004) argue that reduced relatives also indicate a patterning with unergative verbs: while unaccusatives allow reduced relatives (5a), unergatives do not (5b), and neither do reflexives and reciprocals ( $5 \mathrm{c}-5 \mathrm{~d}$ ).


[^149]d. *I ragazzi baciatisi ieri sono miei alunni. the children kiss.PAST.PART-REC yesterday be.PRES.3pl my pupils 'The children who kissed yesterday are my pupils.' (from Siloni 2008, 461)

Lastly, Marelj (2004) notes that the test of Left-Branch Extraction can be used to diagnose unaccusative verbs in Serbian, which displays very few syntactic reflexes of split intransitivity. Namely, in Serbian it is possible to extract possessives, demonstratives or interrogative elements from postverbal objects, including the surface subjects of unaccusatives (6a), but not from postverbal unergative subjects (6b), meaning that reflexives should not allow it if they are unergative forms. This is precisely what happens in (6c).
(6)

> a. Moj je stigao brat.
> my be.PREs.3SG arrive.PAST.PART brother 'My brother arrived.'
> b. *Moj je plakao brat. my be.PRES.3sG cry.PAST.PART brother 'My brother cried.'
> c. *Moj se obukao brat. my REFL dress.PAST.PART brother 'My brother dressed.'

In sum, it is clear that the evidence is inconclusive, as clitic reflexives and reciprocals pattern with unaccusatives in some contexts, and with unergatives in others. Most authors account for the mixed pattern by questioning the reliability of some of the unaccusativity tests. However, even though such questioning might be justified, there does not appear to exist a principled solution that would offer a unified explanation for different tests.

A problem closely related to this one is that the tests are usually taken as straightforward diagnostics based only on the surface behaviour of different forms, without taking into account what exactly it is that makes unaccusatives and unergatives behave differently in each of these contexts. It is commonly argued that unaccusative verbs are inadmissible in constructions that require the presence of an external argument, whereas unergatives are banned from those asking for an internal one. However, the real situation seems to be more complex than this and the mere presence or absence of external and internal arguments does not account for all the manifestations of unaccusativity and unergativity. Therefore, looking at these accounts only, it remains unclear why and how intransitive reflexives and reciprocals should display both unaccusative and unergative properties.

[^150]
## 3 The non-reductionist approach

This state of affairs led to a formulation, by Alsina (1996), of an account that does not assume a reduction of either argument, but suggests instead that both the external and the internal argument are implicitly retained, providing the verb with potential for both unaccusative and unergative behaviours. Before describing Alsina's account in more detail, it should be mentioned that there are at least two other approaches that assume both semantic arguments to be present in the subject of intransitive reflexives and reciprocals. Reinhart and Siloni (2005) propose that a 'bundling' of thematic roles takes place, but that only unergative derivations are possible. Similarly, Rákosi (2008) argues for a 'unification' of theta-roles, allowing again only for unergative derivations in reciprocals, and for default unergative and some lexicalised unaccusative derivations in reflexives; however, Rákosi deals primarily with data from Hungarian, which has a closed class of intransitive reflexive and reciprocal predicates, so his analysis cannot be straightforwardly extended to Italian and Serbian. Even more importantly, neither of these approaches proposes that both arguments (in the sense of valence slots) are retained in reflexive and reciprocal formation.

It should be highlighted at this point that the account we adopt assumes a threelevel representation of verbal argument structure, consisting of thematic structure ( $\theta$-structure, the level of semantic roles), argument structure (a-structure, the level of valence slots), and surface syntax (f-structure, the level of grammatical functions); see Alsina (1996), Ackerman and Moore (2001), a.o. The middle level, the a-structure, is central for our discussion of clitic reflexive and reciprocal formation.

The mapping principles we assume are those proposed by Bresnan and Kanerva (1989) and Bresnan and Zaenen (1990), based on the $[+/-r]$ (thematically restricted) and $[+/-o]$ (objective) features. The key elements of the mappings between different levels of argument structure are given in (7) and (8); specifically, (7) shows the features that can be mapped onto each grammatical function, while (8), taken from Kelling (2001), explains how the semantic properties of arguments (i.e. their thematic roles) affect their mapping onto syntactic functions (filtered by the a-structure level).
(7) Featural specifications of grammatical functions

SUBJ $\quad[-r],[-o]$
OBJ $[-r],[+o]$
OBL $_{\theta} \quad[+r],[-o]$
$\mathrm{OBJ}_{\theta} \quad[+r],[+o]$
(8) Intrinsic features of thematic roles

| Thematic roles: | Features: | Possible mappings: |
| :--- | :--- | :--- |
| Agent | $[-\mathrm{o}]$ | SUBJ/OBL |
| Theme/Patient | $[-\mathrm{r}]$ | SUBJ/OBJ |
| Locative | $[-\mathrm{o}]$ | SUBJ/OBL |

The $[-r]$ feature indicates a mapping onto syntactic functions that can be occupied by any argument regardless of its original thematic role; such functions are subject (SUBJ) and object (OBJ), while all others (obliques and restricted objects) are $[+r]$. The $[-o]$ feature defines functions other than object, i.e. SUBJ and $\mathrm{OBL}_{\theta}$, whereas $\mathrm{ObJ}^{2}$ and $\mathrm{OBJ}_{\theta}$ (restricted object) are $[+o]$.

Returning now to reflexive and reciprocal formation, Alsina (1996) proposes that, when clitic reflexives and reciprocals are formed, the two arguments of the transitive input verb are bound in the a-structure and jointly mapped onto the SUBJ function, in an instance of many-to-one argument linking. This type of binding relation is obtained when a reflexive or reciprocal marker such as the clitic si/se is added to a transitive verb. The reflexivising morpheme is assumed to have the structure si/se $\left\langle[\ldots]_{1}[\ldots]_{1}\right\rangle$, which makes the two arguments of the transitive verb map onto a single syntactic function (SUBJ), as indicated by the numerical indices. Alsina uses a somewhat different annotation and a different set of features, but his account can easily be captured in an LMT representation of argument structure (exemplified on the Italian verb difendere 'defend'):
(9) Formation of clitic reflexives and reciprocals

| $\theta$-structure: |  | Agent | Patient |
| :--- | ---: | :---: | :---: |
| a-structure: | difendersi $i_{\text {REFL } / R E C}$ | $\left\langle\operatorname{Arg} 1_{1}\right.$ | $\left.\operatorname{Arg} 2_{1}\right\rangle$ |
|  |  | $[-o]$ | $[-r]$ |

f -structure:
As can be seen from (9), both arguments of the predicate and both its thematic roles remain implicitly present in clitic reflexives and reciprocals. Clearly, the proposed analysis rests on the assumption that the mapping between different levels of representation can be many-to-one. This is contrary to what is assumed in most theories (cf. the Theta Criterion of Chomsky 1981, the Uniformity of Theta Assignment Hypothesis (UTAH) of Baker 1988, and the Function-Argument Biuniqueness constraint of Bresnan 1982), but in the specific case of reflexives and reciprocals there appears to be good reason to allow it. ${ }^{6}$

When it comes to explaining the unaccusative/unergative paradox, Alsina suggests that with phenomena sensitive to the presence/absence of the internal argument, and insensitive to the involvement of an external argument, reflexives and reciprocals behave like unaccusatives; by contrast, when the phenomenon is dependent upon whether the argument is an external or an internal one, their unergative properties manifest themselves. For instance, the principal requirement of absolute participle formation in Italian (see example (2) above) is the presence of an internal argument, which can be either the object of a transitive verb or the subject of an unaccusative verb, as long as it triggers participial agreement. Since there is nothing in this rule that refers to the external argument, reflexivised and reciprocated verbs satisfy the condition and can therefore be used in absolute participle constructions.

[^151]Similarly, the unaccusative-like auxiliary selection with reflexives and reciprocals in Italian can be explained if it is assumed that essere 'be' has to be used whenever there is an internal argument mapped onto the subject position, while avere 'have' is the correct choice only if this is not the case. In contrast, ne-cliticisation, which also targets only internal arguments, additionally imposes a ban on the external ones, and as a consequence of this reflexive and reciprocal forms are ruled out in constructions with ne. ${ }^{7}$ Moreover, additional evidence for this view comes from the data on reflexive and reciprocal marking with intransitive verbs that are derived through somewhat different lexical operations, or are not derived at all.

## 4 Additional evidence for the non-reductionist account

In section 2 above we pointed out that reflexives and reciprocals pattern together with respect to the main unaccusative diagnostics, and are commonly analysed as products of the same mapping process. ${ }^{8}$ However, the patterns of reflexive marking with non-derived intransitive verbs indicate that while reflexives are more closely related to unaccusatives, reciprocals have a closer connection to unergatives. Clearly, this is highly problematic for both reductionist views, as they assume a reduction of the same argument in both cases.

Specifically, reflexive-marked verbs that take single semantic participants, and are thus similar to reflexives, are typically unaccusative (e.g. sedersi 'sit down' or inginocchiarsi 'kneel down' in Italian), while reflexive-marked verbs that, similarly to reciprocals, must have two semantic participants (an Agent and a Comitative), are unergative (e.g. takmičiti se 'compete' or svadjati se 'argue' in Serbian). A few lexicalised forms exist that constitute exceptions to this tendency (e.g. the unergatives smejati se 'laugh' or igrati se 'play' in Serbian), but a systematic reflexivemarking pattern pointing in this direction does not seem to be attested. Reductionist approaches can only explain such a distribution of reflexive marking by assuming that a different argument is eliminated in reflexive and reciprocal forms (external vs. internal); this, however, would be contrary to the empirical facts, which demonstrate their parallel behaviours on unaccusativity tests. In the next two sections, we present more detailed data on reflexive-marked intransitive verbs and we elaborate Alsina's theoretical account to capture these facts.

### 4.1 Descriptives

Both Italian and Serbian use an extremely wide range of reflexive-marked forms. ${ }^{9}$ The theoretical account presented above deals with what we can call 'proper' reflexives and reciprocals, i.e. those clitic reflexives and reciprocals whose semantics

[^152]remains the same as that of their transitive alternant, and whose clitics can thus be replaced by argumental reflexive/reciprocal pronouns. For instance, the (a) and (b) version of the Serbian sentence in (10), containing a reflexive clitic and an argumental reflexive pronoun respectively, have the same meaning, and differ only in emphasis.
a. Marko se dobro brani. Marko refl well defend.pres.3sG
b. Marko sebe dobro brani.

Marko refl.pro well defend.pres.3sG
'Marko defends himself well.'
In contrast, there are numerous cases in which clitic reflexives and reciprocals have a somewhat changed semantics compared to their transitive alternants. The sentences in (11) illustrate the point, as the act of throwing does not entail the same actions in the two sentences (see Kayne 1975 for similar examples in French).
a. Ivan se bacio kroz prozor.
Ivan REFL throw.PAST.PART out window
'Ivan threw himself out the window.'
b. Ivan je bacio Jovana kroz prozor.

Ivan be.PREs.3SG throw.PAST.PART Jovan.ACC out window
'Ivan threw Jovan out the window.'
We will refer to cases similar to the one in (11) as 'extended' reflexives and reciprocals, given that they are formed through a sort of reflexive (or reciprocal) derivation, but one that is not based solely on the mapping of two arguments onto one syntactic function. Consequently, what we wish to add to Alsina's account is that, in addition to proper clitic reflexives and reciprocals, there also exist extended reflexives and reciprocals, whose semantics is changed (jointly with their morphosyntax), with respect to the semantics of their transitive version, or whose transitive version is not instantiated at all. In particular, we argue that there are two distinct continua that relate proper clitic reflexives and reciprocals to non-derived intransitive verbs. As has already been pointed out, the reflexive continuum is related to unaccusative verbs, and the reciprocal continuum to unergative verbs. The key elements in the continua are the semantic shift that happens in the formation of the verbs along the continuum, and the morphosyntactic consequences of this shift.

Starting from reflexives, some verbs that are commonly treated as proper reflexives are actually characterised by a difference in meaning between the clitic form and the transitive use (see (11) above, also in Serbian/Italian sakriti se/nascondersi 'hide', preobraziti se/trasformarsi 'transform'). A related group of verbs are the verbs of nontranslational motion and verbs of change of body posture (Kemmer 1993). Among these verbs we find predicates that have a transitive alternant, but for which the alternant has a changed meaning, and is normally used with inanimate objects, as shown in (12) for the Italian verb alzarsi 'get up'.
a. Matteo si è alzato.

Matteo Refl be.PRES.3sG get.up.PAST.PART
'Matteo got up.'
b. Matteo ha alzato la mano. Matteo have.PRES.3SG raise.PAST.PART the hand 'Matteo raised his hand.'

Moving further down the continuum, the next group is given by those verbs of nontranslational motion and body posture that do not have a transitive alternant at all (Italian sedersi 'sit down'). Lastly, we reach verbs of translational motion, such as Italian arrivare 'arrive' or partire 'leave'; most of them do not receive the reflexive marking, and they roughly correspond to the change of location class postulated in Sorace's (2000) Auxiliary Selection Hierarchy, which represents the class of core unaccusative verbs.

Moving on to reciprocal forms, some verbs often treated as proper reciprocals actually undergo a change in meaning with respect to the canonic one; examples are the Serbian forms videti se 'see each other', čuti se 'hear from each other' and naći se 'meet up' (literally 'find each other') - these verbs' semantics typically equals 'meet', 'talk on the phone' or 'keep in touch', and they are only rarely used with their literal meanings. The next class comprises reciprocal-marked verbs which can be used transitively, but only with objects of the type $X$ and $Y$, or with a comitative complement in addition to the object, conforming to Levin's (1993: 5859) 'transitive simple reciprocal alternation' (e.g. Serbian pomiriti se 'reconcile', Italian unirsi 'unite'). This is illustrated by the Serbian sentences in (13).
a. Sanja i Mita su se pomirili. Sanja and Mita be.Pres.3PL REC reconcile.PAST.PART 'Sanja and Mita reconciled.'
b. Pomirili smo Sanju i Mitu. reconcile.PAST.PART be.Pres.1PL Sanja.acc and Mita.ACC 'We reconciled Sanja and Mita.'
c. Pomirili smo Sanju sa Mitom. reconcile.PAST.PART be.pres.1pl Sanja.acc with Mita.Ins 'We reconciled Sanja with Mita.'

What follows are the reciprocal-marked verbs participating in Levin's 'intransitive simple reciprocal alternation' (1993: 62-63), which do not have a transitive alternant and can either take an $X$ and $Y$ subject or a comitative argument (Serbian svadjati se 'argue', takmičiti se 'compete'). An example is given in (14).
(14) a. Iva i Aca se stalno svadjaju. Iva and Aca REC always argue.PRES.3PL 'Iva and Aca always argue.'
b. Iva se stalno svadja sa Acom.

Iva REC always argue.Pres.3sG with Aca.INS
'Iva always argues with Aca.'
They are followed by similar non-reciprocal-marked verbs, which belong to Sorace's non-motional controlled process verbs, which are the core group of unergative verbs on her Hierarchy; examples in Italian are collaborare 'collaborate', negoziare 'negotiate', etc.

It is clear from the above that the area between proper clitic reflexives and reciprocals and non-derived intransitive verbs is quite blurred, and that several related criteria determine the groups on the extended continua: the presence of reflexive or reciprocal marking, the existence of a transitive alternant, and the difference in meaning or use between the reflexive or reciprocal-marked predicate and its transitive variant. Specifically, the members of the first class have the marking and they have a transitive alternant whose meaning is only minimally different, and whose syntactic behaviour is essentially the same. The members of the second group also have the marking and an alternant, but they differ from the alternant in the syntactic conditions of use (e.g. type of object, requirement for a comitative argument). The members of the third group preserve the marking despite not having a transitive alternant, because they encode types of actions similar to the ones denoted by the verbs from the previous group. The final (non-derived) group contains verbs with no reflexive or reciprocal marking and with no transitive alternant. These defining properties indicate that at least some of these forms are products of regular operations on the verbs' argument structure, but operations that do not create proper reflexive and reciprocal predicates.

Crucially, the above data show that the reflexive and reciprocal continua take different directions in their passage into non-derived intransitive verbs. The fact that the continuum of reflexive use gradually passes into the domain of unaccusative verbs and that of reciprocal use into the domain of unergative verbs confirms the claim that both arguments must be kept in the formation of intransitive reflexives and reciprocals.

### 4.2 Theoretical approach

While the theoretical account presented in section 3 treats reflexivisation and reciprocation as morphosyntactic processes, it is clear from the data in 4.1 that some reflexive-marked and reciprocal-marked forms also involve a change in verbal semantics. In order to explain the patterns of semantic shift, it is necessary to elaborate Alsina's proposal further.

This can be achieved by relying on the Correspondence Theory of Ackerman and Moore (2001). The Correspondence Theory is an approach to argument linking based on standard Lexical Mapping Theory on the one hand, and the Proto-Role approach of Dowty (1991) on the other. Specifically, Ackerman and Moore propose two complementary principles of argument selection, the Syn-
tagmatic Selection Principle and the Paradigmatic Selection Principle; the former regulates morphosyntactic operations, that is, changes in the way arguments are mapped onto syntactic functions (e.g. in the active/passive alternation or the locative inversion), while the latter deals with morphosemantic changes such as the causative/inchoative alternation. Standard LMT is the basis for the Syntagmatic Selection Principle, i.e. for morphosyntactic operations, which yield realignments of grammatical functions without any changes in the semantics of predicates. The Paradigmatic Selection Principle, on the other hand, is crucial in explaining the operations motivated by morphosemantic changes, which cannot be captured by LMT's mapping principles.

The analysis of clitic-marked reflexives and reciprocals presented in section 3 treats them as being derived via a morphosyntactic operation, i.e. through a change in the way the verb's arguments are mapped onto syntactic functions, and without any modifications in the verb's meaning. This analysis accounts for proper reflexives and reciprocals, but it cannot explain the forms in the extended continua, which do undergo a semantic change in addition to the morphosyntactic one.

In order to capture the gradual increase in the level of semantic change, our theoretical approach has to capture the extended reflexive and reciprocal continua described in the previous section. Given that verbal semantics influences syntax through thematic roles, a continuum can only be enabled by a non-categorical view of thematic roles, which is the central idea of Dowty's Proto-Role approach.

Dowty (1991) argues that two proto-thematic-roles - Proto-Agent and ProtoPatient - are sufficient to describe all thematic relations a predicate can express. This is possible because proto-roles are determined by predicate entailments, and a single entailment can suffice for a specific proto-role to be assigned to an argument; however, the prototypicality becomes stronger as the number of properties increases. The properties that contribute to each of the roles are listed in (15) and (16), taken from Dowty (1991, 572).
(15) Contributing properties for the Agent Proto-Role:
a. volitional involvement in the event or state
b. sent[i]ence (and/or perception)
c. causing an event or change of state in another participant
d. movement (relative to the position of another participant)
e. (exists independently of the event named by the verb)
(16) Contributing properties for the Patient Proto-Role:
a. undergoes change of state
b. incremental theme
c. causally affected by another participant
d. stationary relative to movement of another participant
e. (does not exist independently of the event, or not at all)

These properties influence the syntactic realisation of verbal arguments through the mapping principles introduced in section 3, and they also underlie the principles of argument selection postulated by Ackerman and Moore. Rather than citing the (Syntagmatic) Argument Selection Principle (Dowty 1991, 576), which assumes that the assignment of grammatical functions is determined directly from proto-role entailments, we rely on the principles proposed in Zaenen (1993, 150), according to which proto-role entailments determine intrinsic argument classification, which in turn enables the mapping onto surface grammatical functions:
(17) If a participant has more patient properties than agent properties, it is marked $[-r]$.
If a participant has more agent properties than patient properties, it is marked $[-o]$.
An equal number of properties leads to the assignment of $[-r]$.
When the sole participant of a verb has neither agent nor patient properties it is marked $[-o]$.

These principles are a 'graded' elaboration of the ones discussed in section 3, as they allow for different degrees of agentivity and patienthood. As the central part of the Syntagmatic Argument Selection Principle, they can also be related to the Paradigmatic Argument Selection Principle, given in (18), from Ackerman and Moore (2001, 67).

Let $\mathrm{P}\left(\ldots, \arg _{i}, \ldots\right)$ and $\mathrm{P}^{\prime}\left(\ldots, \arg ^{\prime}{ }_{i}, \ldots\right)$ be related predicates, where $\arg _{i}$ and $\arg ^{\prime}{ }_{i}$ are corresponding arguments. If $\arg _{i}$ and $\arg _{i}{ }_{i}$ exhibit different grammatical encodings and $\arg _{i}$ is more prototypical with respect to a particular proto-role than $\arg ^{\prime}{ }_{i}$, then $\arg _{i}$ 's encoding will be less oblique than $\arg ^{\prime}{ }_{i}$ 's encoding.

This principle is intended to regulate morphosemantic changes, i.e. to relate semantically non-equal realisations of the same predicate. It finds direct applications in the analysis of psychological predicates and causatives, where it can explain, for instance, alternative (more and less oblique) causee encodings of some languages by relying on different degrees of proto-agentivity entailed by the predicate; however, the authors point out that different degrees of proto-agentivity can have morphosyntactic consequences in other domains of grammar as well. One example is the selection of the suffix -age vs. -(e)ment in the derivation of deverbal nouns in French; Kelling (2001) proposes that -age is chosen when the input verb possesses many Proto-Agent properties, while -(e)ment is selected if there are fewer.

A similar application of the above principles can be found for clitic reflexives and reciprocals. Specifically, in proper clitic reflexives and reciprocals the arguments keep the degree of agentivity and the degree of patienthood that are defined by the transitive verb from which they are formed (recall the example in (10)); the only change concerns the fact that in the reflexive/reciprocal form the grammatical subject is the function characterised by both agentivity and patienthood. The approach of Alsina (1996), presented in section 3, and falling within the domain
of the Syntagmatic Argument Selection Principle, can thus fully account for this operation. ${ }^{10}$

A bigger change occurs in the extended forms. Along the extended continuum, there is a decrease in one of the proto-properties characterising the grammatical subject. In the case of reflexives, agentivity is progressively lost, and the extended reflexives gradually blend into unpaired reflexive-marked intransitives and non-reflexive-marked unaccusatives. In the case of reciprocals, the subject progressively loses its patienthood, becoming more prominently a Proto-Agent than a Proto-Patient, until an end point is reached at which it turns into reciprocal-marked unpaired unergatives, and unmarked unergatives. As a consequence of different degrees of agentivity and patienthood, verbs at different points of the continua allow different morphosyntactic behaviours: those that undergo a minor semantic shift remain transitive and can appear with the same type of object as the proper forms, with a moderate change in meaning; those whose proto-property decreases further also keep their transitivity, but can only take a specific type of objects, different from the ones allowed with their transitive alternant; when the change in the given property grows even bigger, the verb becomes intransitive, after which it also looses the reflexive/reciprocal marking.

A simplified schema of the process, without the syntactic consequences, is shown in (19) for reflexives, and in (20) for reciprocals. 'P-A' and 'P-P' denote Proto-Agent and Proto-Patient properties respectively, while the ' + ' and the '-' signs symbolise the increase/decrease in the number of proto-properties. The top two lines show the Italian and Serbian verbs that exemplify each of the different classes; they are followed by their English translations. The exact specification of the Proto-Role entailments involved in each step is left for future work.
(19) The continuum reflexive $\rightarrow$ unaccusative

| vestirsi/ <br> obući se <br> 'dress' |  | buttarsi/ <br> baciti se 'throw onself' |  | alzarsi/ <br> podići se <br> 'get up' |  | sedersi/ <br> [sesti] <br> 'sit down' | arrivare/ <br> stići <br> 'arrive' |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| P-A | P-P | $\mathrm{P}-\mathrm{A}^{-}$ | P-P ${ }^{+}$ | $\mathrm{P}-\mathrm{A}^{--}$ | P-P ${ }^{++}$ | $\mathrm{P}-\mathrm{P}^{R}$ | P-P |
| Arg $1_{1}$ | $\operatorname{Arg} 2{ }_{1}$ > | $\left\langle\operatorname{Arg} 1_{1}\right.$ | Arg2 ${ }_{1}$ > | $\left\langle\operatorname{Arg} 1_{1}\right.$ | Arg2 ${ }_{1}$ > | $\left\langle\operatorname{Arg} 1_{1}\right\rangle$ | $\left\langle\operatorname{Arg} 1_{1}\right\rangle$ |
| [-o] | [-r] | [-o] | $[-r]$ | [-o] | $[-r]$ | $[-r]$ | $[-r]$ |
|  |  | $\mathrm{SUBJ}_{1}$ |  | SUBJ ${ }_{1}$ |  | $\mathrm{SUBJ}_{1}$ | $\mathrm{SUBJ}_{1}$ |

(20) The continuum reciprocal $\rightarrow$ unergative

| baciarsi/ poljubiti se 'kiss' | vedersi/ <br> videti se <br> 'meet up' |  | riconciliarsi/ pomiriti se 'reconcile' |  | [competere]/ takmičiti se 'compete' | collaborare/ saradjivati 'collaborate' |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathrm{P}-\mathrm{A} \quad \mathrm{P}-\mathrm{P}$ | $\mathrm{P}-\mathrm{A}^{+}$ | P-P ${ }^{-}$ | P-A ${ }^{++}$ | $\mathrm{P}-\mathrm{P}^{--}$ | $\mathrm{P}-\mathrm{A}^{R}$ | P-A |
| $\left\langle\operatorname{Arg} 1_{1} \quad \operatorname{Arg} 2_{1}\right\rangle$ | $\left\langle\mathrm{Arg}_{1}\right.$ | $\operatorname{Arg} 2{ }_{1}$ > | $\left\langle\mathrm{Arg} 1_{1}\right.$ | $\operatorname{Arg} 2{ }_{1}$ > | $\left\langle\operatorname{Arg} 1_{1}\right\rangle$ | $\left\langle\operatorname{Arg} 1_{1}\right\rangle$ |
| $[-o] \quad[-r]$ | [-o] | [-r] | [-o] | [-r] | [-o] | [-o] |
| $\mathrm{SUBJ}_{1}$ | $\mathrm{SUBJ}_{1}$ |  | $\mathrm{SUBJ}_{1}$ |  | $\mathrm{SUBJ}_{1}$ | $\mathrm{SUBJ}_{1}$ |

[^153]In sum, we argue that proper reflexives and reciprocals are created through a morphosyntactic operation, where the alignment of arguments changes, but not their meaning, while with the extended reflexives and reciprocals, a morphosemantic operation is also at work, as the semantic content of the arguments is altered. In this specific case, the reflex of the Paradigmatic Selection Principle consists in the inability of these forms to freely alternate with their transitive versions. And most importantly for the central topic of this paper, any theoretical approach that aims at explaining both reflexive and reciprocal formation must be able to account for these processes, which go in opposite directions for reflexives and reciprocals.

## 5 Conclusion

The account proposed by Alsina (1996), based on the joint mapping of two arguments onto the SUBJ function, seems to provide a satisfactory explanation for some of the most problematic facts concerning the status of clitic reflexives and reciprocals in Italian and Serbian. Most importantly, it can account for the mixed unaccusative/unergative behaviour of these forms, both with respect to unaccusativity diagnostics and the divergent marking patterns of non-derived intransitive verbs. The latter is achieved by incorporating Alsina's account in a wider context of the Correspondence Theory of Ackerman and Moore (2001).

Moreover, even though in this paper we deal only with Italian and Serbian, it should be possible to apply the same approach to at least some other languages, as the unaccusative/unergative paradox is not limited to Serbian and Italian, or to the Slavic and Romance families. Clearly, before making any further claims, more crosslinguistic data needs to be examined. In addition, there are a number of related verbal forms that could be compared to the reflexives and reciprocals analysed in this paper. One such case are psychological predicates, which also show an interesting pattern (cf. the Italian pairs spaventarsi 'get scared' - spaventare se stesso 'scare oneself' and rispettarsi - rispettare se stesso 'respect oneself'). Studying similar forms is necessary if we are to formulate a comprehensive account of reflexives and reciprocals.

## References

Ackerman, F. and J. Moore (2001). Proto-Properties and Grammatical Encoding: A Correspondence Theory of Argument Selection. Stanford: CSLI Publications.

Alsina, A. (1996). The Role of Argument Structure in Grammar: Evidence from Romance. Stanford: CSLI Publications.

Baker, M. (1988). Incorporation: A Theory of Grammatical Function Changing. Chicago: Chicago University Press.

Belletti, A. and L. Rizzi (1988). Psych-verbs and $\theta$-Theory. Natural Language and Linguistic Theory 6, 291-352.

Bouchard, D. (1984). On the Content of Empty Categories. Dordrecht: Foris Publications.

Bresnan, J. (1982). Polyadicity. In J. Bresnan (Ed.), The Mental Representation of Grammatical Relations, pp. 149-172. Cambridge, MA and London: The MIT Press. Revised version of Bresnan, J. (1980). Polyadicity: Part 1 of a Theory of Lexical Rules and Representations. In T. Hoekstra, H. van der Hulst and M. Moortgat (Eds.), Lexical Grammar, pp. 97-121. Dordrecht: Foris Publications.

Bresnan, J. and J. Kanerva (1989). Locative inversion in Chichewa: A case study of factorization in grammar. Linguistic Inquiry 21, 1-50.

Bresnan, J. and A. Zaenen (1990). Deep unaccusativity in LFG. In K. Dziwirek, P. Farrell, and E. Mejías-Bikandi (Eds.), Grammatical Relations: A CrossTheoretical Perspective, pp. 45-57. Stanford: CSLI Publications.

Chierchia, G. (2004). A semantics for unaccusatives. In A. Alexiadou, E. Anagnostopoulou, and M. Everaert (Eds.), Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface, pp. 22-59. Oxford: Oxford University Press.

Chomsky, N. (1981). Lectures on Government and Binding. Dordrecht: Foris Publications.

De Alencar, L. F. and C. Kelling (2005). Are reflexive constructions transitive or intransitive? Evidence from German and Romance. In M. Butt and T. H. King (Eds.), Proceedings of the LFG05 Conference, pp. 1-20. Univeristy of Bergen.

Dowty, D. (1991). Thematic proto-roles and argument selection. Language 67, 547-619.

Grimshaw, J. (1982). On the lexical representation of Romance reflexive clitics. In J. Bresnan (Ed.), The Mental Representation of Grammatical Relations, pp. 87-148. Cambridge, MA and London: The MIT Press.

Grimshaw, J. (1990). Argument Structure. Cambridge, MA and London: The MIT Press.

Kayne, R. (1975). French Syntax: The Transformational Cycle. Cambridge, MA and London: The MIT Press.

Kelling, C. (2001). Agentivity and suffix selection. In M. Butt and T. H. King (Eds.), Proceedings of the LFG01 Conference, pp. 1-20. Univeristy of Hong Kong.

Kemmer, S. (1993). The Middle Voice. Amsterdam and Philadelphia: John Benjamins.

Levin, B. (1993). English Verb Classes and Alternations. Chicago and London: University of Chicago Press.

Marelj, M. (2004). Middles and Argument Structure across Languages. LOT Dissertation Series 88. Utrecht: LOT Publications.

Progovac, Lj. (2005). A Syntax of Serbian: Clausal Architecture. Bloomington: Slavica.

Rákosi, G. (2008). The inherently reflexive and the inherently reciprocal predicate in Hungarian: Each to their own argument structure. In E. König and V. Gast (Eds.), Reciprocals and Reflexives: Theoretical and Typological Explorations, pp. 411-450. Berlin and New York: Mouton de Gruyter.

Reinhart, T. and T. Siloni (2004). Against the unaccusative analysis of reflexives. In A. Alexiadou, E. Anagnostopoulou, and M. Everaert (Eds.), Unaccusativity Puzzle: Explorations of the Syntax-Lexicon Interface, pp. 159-180. Oxford: Oxford University Press.

Reinhart, T. and T. Siloni (2005). The Lexicon-Syntax parameter: Reflexivization and other arity operations. Linguistic Inquiry 36, 389-436.

Rizzi, L. (1986). On chain formation. In H. Borer (Ed.), Syntax and Semantics 19: The Syntax of Pronominal Clitics, pp. 65-95. Orlando: Academic Press.

Rosen, C. (1988). The Relational Structure of Reflexive Clauses: Evidence from Italian. London and New York: Garland.

Siloni, T. (2008). The syntax of reciprocal verbs: An overview. In E. König and V. Gast (Eds.), Reciprocals and Reflexives: Theoretical and Typological Explorations, pp. 451-498. Berlin and New York: Mouton de Gruyter.

Sorace, A. (2000). Gradients in auxiliary selection with intransitive verbs. Language 76, 859-890.

Van Valin, R. D. (1990). Semantic parameters of split intransitivity. Language 66, 221-260.

Wehrli, E. (1986). On some properties of French clitic se. In H. Borer (Ed.), Syntax and Semantics 19: The Syntax of Pronominal Clitics, pp. 263-283. Orlando: Academic Press.

Zaenen, A. (1993). Unaccusativity in Dutch: Integrating syntax and lexical semantics. In J. Pustejovsky (Ed.), Semantics and the Lexicon, pp. 129-161. Dordrecht: Kluwer Academic Publishers.

Please send any comments to mm510@cantab.net.

# BEYOND IDENTITY: THE CASE OF A COMPLEX HUNGARIAN REFLEXIVE 

György Rákosi

University of Debrecen

Proceedings of the LFG09 Conference
Miriam Butt and Tracy Holloway King (Editors)

2009
CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

It is a well-known typological universal that long distance reflexives are generally monomorphemic and complex reflexives tend to be licensed only locally. I argue in this paper that the Hungarian body part reflexive maga 'himself' and its more complex counterpart önmaga 'himself, his own self' represent a non-isolated pattern that adds a new dimension to this typology. Nominal modification of a highly grammaticalized body part reflexive may reactivate the dormant underlying possessive structure, thereby granting the more complex reflexive variant an increased level of referentiality and syntactic freedom. In particular, the reactivation of the possessive structure in önmaga is shown to be concomitant with the possibility of referring to representations of the self, as well as a preference for what appears to be coreferential readings and the loss or dispreference of bound-variable readings.


## 1. Introduction

According to an established typology, complex reflexives are expected to be local and relatively well-behaved from a binding theoretical perspective, whereas long distance reflexives tend to be monomorphemic (see Faltz 1985, Pica 1987 and subsequent work, as well as Dalrymple 1993 and Bresnan 2001 in the LFG literature). Polymorphemic reflexives, however, are not uniform as they may show different types of morphological complexity. In particular, body part reflexives, which owe their complexity to their historical origin as possessive structures, are often grammatical outside of the local domain in which their antecedent is located. This is a prima facie problem for the typology, since long distance reflexives are not expected to be morphologically complex.

The existence of long distance uses of body part reflexives can be explained under the assumption that these reflexives have a syntactically active possessive structure. Kornfilt (2001) argues that it is exactly such a structure that licenses the Turkish kendisi 'himself' both in local and nonlocal contexts. But this assumption is not necessary, and others have rejected the possessive analysis of non-strictly local body part reflexives (see, for example, the analysis Beavers \& Koontz-Garboden 2006 put forward for the English colloquial reflexive his ass).

In this paper, I bring evidence from Hungarian to argue for a constrained application of the possessive analysis to complex body part
reflexives. The primary reflexive strategy in Hungarian involves the use of the highly grammaticalized body part reflexive maga 'himself', which has a more complex variant önmaga 'himself, his own self'. I will argue that only the more complex önmaga can project a possessive structure in one of its two uses, in essence reactivating an underlying structure that appears to have been lost during the grammaticalization of the primary reflexive. The possessive reanalysis correlates with changes in the syntax and semantics of the complex anaphor önmaga. In particular, the reflexive becomes grammatical as a subject and it shows invariable 3 SG agreement. Bound variable readings are lost or are dispreferred, and the reflexive can refer to representations of the self, rather than encoding true identity with the referent of the antecedent.

I will use this analysis to argue that on closer inspection, complex body part reflexives which allow for long distance uses do not refute Faltz's (1985) typology. They simply fall outside of the scope of this typology and in fact add a new dimension to it.

The structure of the paper is as follows. In section 2, I give a brief summary of how morphological complexity is known to interact with the size of the reflexive binding domain, paying special attention to body part reflexives. In section 3, I describe the morphology of the two Hungarian reflexives discussed here, and briefly overview the available literature. In section 4, I show that önmaga 'himself, his own self' is less constrained syntactically than the primary reflexive maga 'himself', but this cannot be explained by simply analyzing önmaga as an emphatic reflexive element. This paves the way for a presentation of the peculiar syntactic and semantic properties of önmaga in section 5. I conclude in section 6 by showing how the possessive analysis can account for the observed properties of önmaga, and round up in section 7 with a cross-linguistic outlook on the implications of the current analysis.

## 2. Complex reflexives

In his thorough typological survey of reflexives, Faltz (1985) distinguishes between pronominal and compound (here: complex) reflexives. The third person Norwegian seg, the German sich or the Russian sebja are representatives of the first strategy. The second strategy consists of two broader morphological types. What Faltz calls adjunct reflexives are complexes of a pronoun plus an emphatic marker, like the English himself or the Norwegian seg selv 'himself'. The other major group consists of body part reflexives (or head reflexives in Faltz's terminology), which start their historical development as a possessive structure and can then become grammaticalized to differing degrees. The Basque bere burua, for example, is still ambiguous between the readings 'himself' and 'his head' (Faltz 1985: 32 ).

The fundamental typology on the correlation between the morphological form of reflexives and their domain of licensing consists of two partially independent statements (see Faltz 1985, Pica 1987, and Cole et al. eds. 2001, among others, and Dalrymple 1993 and Bresnan 2001 in LFG): ${ }^{1}$

## (1) Complex reflexive typology

a. Long distance reflexives are monomorphemic.
b. Complex reflexives need local antecedents. ${ }^{2}$

Despite occasional skepticism (cf. Büring: 2005, fn. 37), the typology does seem to be making good predictions for adjunct reflexives. The following Norwegian data serve to illustrate the point (Bresnan 2001: 284):
a. Ola overgår seg selv/*seg.

Ola surpasses
'Ola surpasses himself.'
b. Ola bad oss snakke om *seg selv/seg. Ola asked us talk.INF about 'Ola ${ }_{\mathrm{i}}$ asked us to talk about $\mathrm{him}_{\mathrm{i}}$.

Seg selv is a complex reflexives and is only licensed locally, and only seg can be used as a long distance reflexive form. ${ }^{3}$

Interestingly, reported instances of complex reflexives that do not obey (1b) since they are grammatical both with local and non-local antecedents are all body part reflexives. Let me mention here three such reflexive forms.

The first is the colloquial English his ass, discussed by Beavers \& Koontz-Garboden (2006). ${ }^{4}$ They argue that this reflexive form is a universal pronoun in the sense of Kiparsky (2002), that is, it is grammatical with local ((3a)) as well as non-local antecedents ((3b)), and may even pick up its referent deictically from discourse ((3c)):

[^154](3) a. But most people do believe OJ bought his ass out of jailtime.
b. The more he whined about it, the more they nailed his ass.
c. I mean her ass, over there .

The problematic cases for the typology are (3b) and (3c), since they involve the complex reflexive taking a non-local antecedent.

Turkish represents an even more intriguing case (Kornfilt 2001, and also Faltz 1985 and Enç 1987). The primary reflexive kendi is a body part reflexive. Its paradigm includes inflected first and second person forms, as well as a non-inflected third person form in the singular and the plural, all of which are local ((4a)). The bare third person singular form kendi contrasts with the inflected third person form kendisi (and similarly in third person plural), which can take local or long distance antecedents, and even discourse antecedents ((4b)). The examples are from Kornfilt (2001: 198).
(4) a. Fatma [Ahmet-nin kendi-i çok beğen-diğ-in]-i biliyor. Fatma Ahmet-GEN self-ACC very admire-GER.3SG-ACC knows

b. Fatma [Ahmet-nin kendi-sin-i çok beğen-diğ-in]-i biliyor. Fatma Ahmet-GEN self-3SG-ACC very admire-GER.3SG-ACC knows 'Fatma ${ }_{\mathrm{i}}$ knows that Ahmet $_{\mathrm{j}}$ admires self $\mathrm{i} j \mathrm{j} k$ very much.'

What escape the typology in (1b) are the inflected third person reflexives, which can, but need not, take local antecedents, in contrast to inflected first and second person reflexives and non-inflected third person reflexives, which are only locall licensed, as expected.

A third problem case is the Chinese (Mandarin) ziji-ta ziji 'himself' pair. Presumably, ziji might be derived from the meaning 'nose', though this etymology is debatable (Huba Bartos p.c., and see also König \& Gast 2006: 264). $T a$ is the third person singular pronoun. Whether ziji is a body part reflexive or not, it allows for long distance uses, and, interestingly, ta ziji does the same. Pan (1998: 775-76) actually reports that if the antecedent does not c-command the reflexive, he finds ta-ziji better than ziji.
(5) [Zhangsan de jiao'ao] haile ziji/ta-ziji.

Zhagsan gen pride hurt.PERF self
'Zhangsan ${ }_{\mathrm{i}}$ 's pride hurt him $_{\mathrm{i}}$.'
That ta-ziji is thus a problem for the complex reflexive typology is also mentioned in Bresnan (2001: 301).

Summing up, body part reflexives may represent a general problem for the typology in (1), but what is especially troubling is the existence of the Turkish and Chinese reflexive pairs. The reflexive typology appears to
suggest that increasing the morphological complexity of a reflexive will decrease the size of its binding domain. This does not happen in Chinese, since both ziji and ta ziji have roughly the same distribution, involving long distance uses. And in Turkish, the morphologically more complex inflected reflexive (kendisi) has a wider distribution than the local non-inflected reflexive (kendi).

As we will see, Hungarian repeats the Turkish pattern, and thus represents another challenge. But this, as I intend to show here, is only apparent once we realize that we are dealing here with a phenomenon that is simply not covered by the typology in (1).

## 3. Hungarian reflexives: the background 3.1. The morphology of the two Hungarian reflexives

The primary Hungarian reflexive, which has roughly the same distribution as the English himself, is maga. The stem is reconstructed to have been used as a word for body, but this meaning was lost long ago and in fact native speakers do not have the intuition that the reflexive is compositional. Mag in current Hungarian means 'seed'.

However, the reflexive still shows signs of its possessive origin and it bears possessive type agreement morphology. In Table 1 below, I compare the possessive paradigm of maga 'himself' and magja 'his seed'. ${ }^{5}$ The latter represents the productive morphological pattern, and boldface is used to mark the places where the productive pattern differs from the paradigm of the reflexive. There are two important points of divergence. First, the definite article is obligatory in the possessive construction if the possessor is a (prodropped) pronoun, but the reflexive maga does not co-occur with the definite article. Second, the phonological shape of the inflectional morphology is not identical in the two paradigms.

[^155]|  | maga' ${ }^{\text {'HIMSELF }}$ ' | POSSESSIVE PARADIGM |
| :---: | :---: | :---: |
| 1SG | $\begin{aligned} & \text { magam } \\ & \text { mag.1SG } \end{aligned}$ | a mag-om the seed.1SG.POSS 'my seed' |
| 2SG | $\begin{aligned} & \hline \text { magad } \\ & \text { mag. } 2 S G \end{aligned} \text { 'yourself' }$ | a mag-od 'your seed' the seed-2SG.POSS |
| 3SG | $\begin{aligned} & \text { maga } \\ & \text { mag.3SG } \end{aligned} \text { 'himself' }$ | a mag-ja the seed-3SG.POSS 'his seed' |
| 1PL | magunk 'ourselves' mag.1PL | a mag-unk <br> the seed-1PL.POSS 'our seed' |
| 2PL | magatok 'yourselves' mag.2PL | a mag-otok 'your seed' the seed-2PL.POSS |
| 3PL | $\begin{array}{ll} \hline \text { maguk } \\ \text { mag.3PL } \end{array} \quad \text { 'themselves' }$ | a mag-juk <br> the seed-3PL.POSS 'their seed' |

Table 1.
It is the possessive paradigm that has the productive morphophonology, which is a clear indication that the reflexive is highly grammaticalized. Nevertheless, it is also evident that both paradigms utilize the same type of agreement morphology.

Önmaga is the complex of the primary reflexive maga and the nominal prefix ön- 'self'. This prefix, much like its English counterpart, normally combines with deverbal nouns ((6a)) or participles ((6b)), but it can also be attached to simple, non-eventive nouns ((6c)).
(6)
a. ön-ellát-ás
self-serve-NOMINAL.SUFFIX
'self-service'
b. ön-müköd-ö
self-operate-PARTICIPIAL.SUFFIX
'self-operating'
c. ön-hiba
self-fault
'(one's) own fault'
Thus we could draw a formal analogy between the possessive form of (6c) and önmaga:
a. ön-hibá-m
self-fault-1SG.POSS
'my own fault'
b. ön-magam self-mag.1SG
'my own self'

Note that I am translating here önmagam as 'my own self' only as an attempt at illustrating how it might differ from maga 'himself' in meaning, but the claim is certainly not that önmaga and (the Hungarian for) his own self are direct grammatical and semantic equivalents of each other.

### 3.2. The previous literature on ömaga 'his own self'

As has been stated above, the primary reflexive strategy in Hungarian involves the use of maga 'himself'. Önmaga 'his own self' has received relatively little specific attention in the pertinent syntactic literature. In fact the two reflexives are generally treated as essentially equivalent without further comment, and önmaga may even be used to illustrate basic binding data in Hungarian (as happens in É. Kiss 1994: 23-26 or É. Kiss 2002: 3540).

It does appear at first sight that the two reflexives have the same distribution, roughly similar to that of the English himself:
(8) János felismerte (ön)magá-t a kép-en. John.NOM recognized himself-ACC the picture-on 'John recognized himself in the picture.'

The occasional remark that one may find (especially in the descriptivist literature) is that önmaga is more emphatic than maga, but the nature of this difference is not spelled out in any detail. The only work which goes beyond this remark is Everaert \& Szendrői (2002). They note that only maga, but not önmaga may form part of idiomatic expressions ((9a)), and that maga tends to be adjacent with the verb and bear one accent with it $((9 b))$. The brackets in (9b) are to be interpreted disjunctively.
(9) a. János nem izgatja (*ön)magá-t. John not excites himself-ACC 'John can't be bothered.'
b. János megmutatta (magá-t) Marinak ( 'magá-t). John showed himself-ACC Mary-DAT himself-ACC 'John showed himself to Mary.'

They conclude that whereas maga is a simple NP, the more complex önmaga projects an extended nominal phrase, or DP.

Though I believe the analysis that Everaert \& Szendrői (2002) offer is a step towards a better understanding of the difference between the two Hungarian reflexives, it does not account for further peculiar properties of önmaga, which I will show to exist. In the rest of the paper, I undertake a detailed investigation of the diverging grammar of maga 'himself' and
önmaga 'his own self', and offer an alternative analysis that I believe to provide an account of the observed differences and that I hope accommodates the Hungarian data within a larger cross-linguistic domain.

## 4. Maga vs önmaga: the basics

## 4.1. Önmaga is less constrained

Maga is a nuclear anaphor, and is licensed as such only in the presence of local antecedents. As I noted in the previous section, önmaga is also acceptable in the same local binding domain, so the two are often interchangeable from a purely syntactic perspective. There are, however, constructions in which only önmaga is grammatical, and maga is ruled out. I briefly survey these contexts here.

First, maga is normally not grammatical if embedded within obviously non-argument expressions like the high-level adjunct in (10a) or the passive by-phrase with the participle in (10b). Önmaga, however, is acceptable in the selfsame contexts.

$$
\begin{array}{lllll}
\text { a. Önmaga / *maga } & \text { szerint } & \text { János } & \text { okos } & \text { ember. }  \tag{10}\\
\text { himself.nOM } & \text { according.to } & \text { John } & \text { clever } & \text { man } \\
\text { 'According to himself, John is a clever man.' }
\end{array}
$$

b. az önmaga / *maga által okos-nak tart-ott ember the himself.NOM by clever-DAT consider-PART man.NOM lit. 'the man who is considered to be clever by himself'

Second, önmaga also shows apparent long distance uses, though it sounds best if it occurs adjacent to the clause in which its antecedent is embedded.
(11) János fél, hogy *(ön)magát sem választ-ják meg. John afraid.is that himself-ACC neither elect-3PL PARTICLE lit. 'John is afraid that they will not elect himself either.'

Such long distance uses generally occur in point-of-view contexts.
Third, it has been noted in the literature that önmaga, unlike maga, can function as a nominative subject if it is no more prominent thematically than its non-subject antecedent (see Everaert \& Szendrői 2002 and Rákosi 2006, as well as É. Kiss 2002, who does not mention though that maga is in fact ungrammatical as a subject). This mainly covers object and dative experiencer verbs, like the following:
(12) János-t meglepte *(ön)maga. John-ACC suprised himself.NOM lit. 'Himself surprised John.'

In section 5.2, I will argue that once the right context is set up, önmaga is licensed as a syntactic subject by any predicate. But the immediate point is that maga is never acceptable as a syntactic subject.

I should hasten to add that even if önmaga is freer syntactically than maga, it is not as obviously free as the colloquial English his ass or the Turkish kendisi 'himself'. Compare (4b), repeated as (13a), with (13b):
a. Fatma[Ahmet-nin kendi-sin-i çok beğen-diğ-in]-i biliyor. Fatma Ahmet-GEN self-3SG-ACC very admire-GER.3SG-ACC knows 'Fatma ${ }_{i}$ knows that Ahmet $_{\mathrm{j}}$ admires self $\mathrm{i}_{\mathrm{i} j \mathrm{k}}$ very much.'
b. Fatma tudja, hogy Ahmed nagyon szereti önmagá-t. Fatma knows that Ahmet very likes himself-ACC 'Fatma ${ }_{\mathrm{i}}$ knows that Ahmet $_{\mathrm{j}}$ admires self ${ }_{*_{\mathrm{i} j}{ }^{j} *_{\mathrm{k}}}$ very much.'

Though the just discussed differences do exist, it still holds that önmaga is not an all purpose reflexive. As (13b) demonstrates, önmaga does not always allow for long distance uses, and it does not normally take discourse antecedents. Or, to be more precise, it does not do so the same way as his ass or kendisi do.

Nevertheless, this situation does represent a problem for the complex reflexive typology. Maga is a complex body part reflexive, and it behaves as expected since it is a nuclear anaphor. The even more complex önmaga, however, is not necessarily nuclear, and has a wider distribution than the primary reflexive.

## 4.2. Önmaga is not an emphatic form

One potential explanation for the less constrained syntax of önmaga could be that it is a special emphatic form, and as such, it is not subject to the bounds of binding theory. Cole, Hermon \& Lee (2001: 36) offer some arguments for such an account of the Chinese reflexive ta ziji. They claim that ta ziji can in fact be analyzed as the complex of the pronoun ta 'he' and the reflexive ziji 'himself' as an intensifying element. In essence, rather than being a complex reflexive pronominal, ta ziji would then be equivalent to the English he himself (cf. John said that he himself wanted to do it).

Irrespective of whether this analysis works for Chinese or not, it clearly cannot be applied to the case of önmaga. The Hungarian intensifier is in fact maga, complying with the known fact that primary reflexives often function as intensifiers (cf. König \& Gast 2006). Önmaga cannot or only marginally can associate with noun phrases as an intensifier:

| Maga/*önmaga az elnök | beszélt | velünk. |  |
| :--- | :--- | :--- | :--- |
| himself | the president.NOM | talked | with.1PL |
| 'The president himself talked to us.' |  |  |  |

Neither can önmaga substitute for a pronoun + intensifier unit:
(15) Ez-t [neki magá-nak]/[*önmagá-nak] add oda. this-ACC DAT.3SG himself-DAThimself-DAT give.IMP.2SG PART 'Give this to him himself.'

So the relative syntactic freedom of önmaga cannot be explained by assuming that this reflexive may function as intensifier.

## 5. Maga vs önmaga: beyond identity

## 5.1. Önmaga resembles proper nouns

In 4.1, I focused on some of the usual contexts to show that the syntax of önmaga is not identical to that of maga. It is, however, more revealing than the previous data set that önmaga, unlike maga, pattern with proper nouns in certain constructions. I discuss here two such constructions.

First, both proper nouns and önmaga can be used predicatively in identity statements. Besides, önmaga can also be interpreted as what König \& Gast (2006) call an adverbial-exclusive intensifier ( $\approx$ 'alone’). ${ }^{6}$ This is the only reading for maga, which cannot be the predicate of an identity statement. Compare:
a. Újra Péter vagyok.
again Peter am
'I am (the good old) Peter again.'
b. Újra önmagam vagyok.
again myself am
(i) 'I am myself again.'
(ii) 'I am alone again.'
c. Újra magam vagyok. again myself am
(i) *'I am myself again.'
(ii) 'I am alone again.'

[^156]Second, proper nouns can be restrictively premodified when the same realworld individual is conceptualized as corresponding to two partially nonidentical selfs. Önmaga can also be premodified this way, but maga cannot:

| a. a Kádár-kor-i | Péter |
| :--- | :--- |
| the Kádar-era-ADJECTIVAL.SUFFIX Peter |  |
| 'the Peter of the Kádár-era' |  |

Notice that the grammaticality contrast between maga and önmaga is very sharp both in (16) and in (17).

What these data suggest is that önmaga, unlike a regular reflexive pronominal, shows an increased level of referentiality. It cannot be a simple accident that it patterns with proper nouns in the contexts just discussed.

### 5.2. Representations of the self

I claimed above that maga can normally be substituted for önmaga. But now that we have reasons to suspect that the two are not equivalent to each other semantically, it is easier to realize that in certain contexts önmaga will be the better or the only option even if the antecedent is locally available.

In general, önmaga is felt to be more natural when the context is such that it facilitates a reading in which complete semantic identity does not hold between the antecedent and the reflexive. Consider these two sentences:
> a. A történelem ismétli ?magá-t/önmagá-t.
> the history.NOM repeats itself-ACC
> 'History repeats itself.'
b. János ellentmond ?magá-nak / önmagá-nak.

John.NOM contradicts himself-DAT
'John contradicts himself.'

The reflexive relation that the predicates repeat and contradict encode is a non-trivial one, for one may only repeat or contradict temporally different states of the self. In other words, (18a) asserts that the current state of history is in some sense equivalent to one of its previous states. The semantic
relation between the antecedent and the reflexive is not strict identity, and in such cases, maga sounds degraded but önmaga is perfectly natural.

The difference is stronger in the so-called 'Mme Tussaud' contexts of Jackendoff 1992 (see also Culicover \& Jackendoff 2005). (19) is meant to describe an accident upon Ringo's visit to the wax museum.
(19) Ringo fell on himself.
(i) 'The actual Ringo fell on the statue of Ringo.'
(ii) *'The statue of Ringo fell on the actual Ringo.'

Jackendoff points out that (19) can only have the reading in which the actual Ringo falls on the statue Ringo, but not vice versa. What is important for us now is that the English reflexive can apparently be used to refer to representations of the self.

The following variety of the 'Mme Tussaud' context is based on Reuland (2001: 483) and serves to illustrate the Hungarian facts:
a. Ringo megpillantotta magá-t a tükör-ben. Ringo caught.sight.of himself-ACC the mirror-in
(i) 'The actual Ringo saw his own image.'
(ii) ?"'The actual Ringo saw the image of his statue.'
b. Ringo megpillantotta önmagá-t a tükör-ben. Ringo caught.sight.of himself-ACC the mirror-in
(i) 'The actual Ringo saw his own image.'
(ii) 'The actual Ringo saw the image of his statue.'

Though there is some variation in judgments, the statue-reading is only licensed with önmaga for most speakers, whereas maga may only very marginally allow for this reading.

The clearest cases are those when an ontologically independent and fully functioning copy of the self is created. Such contexts are mostly imaginary, but we do have means of talking about strongly intensional worlds. Imagine, for example, that Peter was cloned or he traveled back in time, and walking on the corridor, he met his own copy. To describe this situation, önmaga must be used.

> Önmaga jött Péter-rel szembe a folyosó-n. himself came.3SG lit. 'Himself was coming towards Peter in the corridor.'

There are two noteworthy aspects of (21). First, just like in the English example (19), the reflexive must refer to the copy and the proper name refers to the real (i.e., the original) Peter. Second, in these 'representations of the self' contexts önmaga is grammatical as a subject by any predicate. I noted in
subsection 4.1 that it is known in the literature that önmaga can mostly be the subject of experiencer predicates. In the light of (21), we can now interpret this as derivative of the fact that experiencer predicates facilitate at least weak 'representations of the self' readings. If, after all, John is surprised by himself is true (cf. 12), then it must be the case that what surprises John is an aspect of his personality that he was not aware of. It seems that this level of conceptual differentiation is enough to license önmaga as a subject. With non-experiencer predicates, stronger contextual support is required to achieve the same affect.

There is a further peculiar property of the subject uses of önmaga. Irrespective of which form of the paradigm is used, these reflexive subjects will always trigger third person singular agreement on the verb. In (22), the subject is the first person singular reflexive, but the verb is still in its third person singular form.
(22) Önmagam jött velem szembe a folyosó-n. myself came.3SG with.1SG against the corridor-on lit. 'Myself was coming towards me in the corridor.'

Given that otherwise agreement is applied across the board in Hungarian, it is strange that now we seemingly face its absence. Notice also that my real self is referred to by the pronominal velem 'with me', rather than by an anaphor. That is also unexpected. If önmagam 'myself' was a first person singular form, then the coreferring pronominal would have to be ungrammatical in the same clause.

### 5.3. Two entries for önmaga

I concluded the last subsection with an apparent puzzle. Önmaga shows the full agreement paradigm (cf. Table 1); still it always triggers third person singular agreement if it is used as a subject. What I want to suggest now is that in fact we have two separate lexical entries for önmaga (all through the paradigm). Önmaga $a_{1}$ is a more or less regular reflexive, except for the fact that it is not strictly nuclear (4.2). It agrees with its antecedent in person and number, and it cannot be used as a subject. Önmaga $a_{2}$ is a special type of reflexive: this is the one that is used in 'representations of the self' contexts. It can be used as a subject, and it shows constant third person singular agreement with the verb.

One intuitively appealing motivation for this move is that this way we can clearly separate relations of true identity (önmaga ${ }_{1}$ ) from relations of referential differentiation (önmaga ${ }_{2}$ ). Note that in a sentence like (11), repeated here as (23), the antecedent and the reflexive clearly refer to one single conceptualization of the same individual:

We can just simply describe this fact by assuming that the sentence contains önmaga ${ }_{1}$.

An argument with more substantial weight is based on patterns of licensing bound variable and coreference readings (see Evans 1980, Reinhart 1983, 2006, Bresnan 2001 and Büring 2005, among others, for this difference). The less complex reflexive maga only seems to allow for bound variable readings, but not for coreference readings, as the following VPellipsis context testifies:
(24) János látja magá-t, de Kati nem.

John see.3SG himself-ACC but Kate.NOM not
(i) 'John sees himself, but Kate (does) not (see herself).'
(ii) *'John sees John, but Kate (does) not (see John).'
(i) is the sloppy, bound variable reading, under which the elided anaphor is understood to be locally bound by the subject of the clause in which the VP is missing. Under the strict, coreference reading (ii), what Kate does not see is John, not herself. If, however, we replace maga with önmaga, then the coreference reading becomes fully grammatical for many speakers, and marginally available for others.
(25) János látja önmagá-t, de Kati nem.

John see.3SG himself-ACC but Kate.NOM not
(i) 'John sees himself, but Kate (does) not (see herself).'
(ii) ${ }^{\sqrt{ } / \text { ?? }}$ ‘John sees John, but Kate (does) not (see John).'

Notice that the bound variable reading is still available.
And now let us consider (26), where the reflexive is the subject and the antecedent is the object.

Engem megijeszt önmagam, de téged nem.
I.ACC scares myself.NOM but you.ACC not
(i) ${ }^{* / ? ?}$ 'Myself scares me, but (yourself does) not (scare) you.'
(ii) 'Myself scares me, but (myself does) not (scare) you.'

Interestingly, now the bound variable reading (i) becomes unavailable for many speakers, or at least very marginal for others, but the coreference reading (ii) is grammatical. This is in clear contrast with (25).

One convenient way of explaining this contrast is to assume that the entry that we have in (26) is our önmaga ${ }_{2}$, which does not license bound
variable readings. Under this account, önmaga $a_{1}$ can be considered to be a regular reflexive that favours bound variable readings. This is the entry we have in (25).

Two remarks need to be added to this. First, one could object that the bound variable reading is unavailable in (26) because the construction is an instance of weak crossover. But note that weak crossover effects are not attested in Hungarian as long as the object binder linearly precedes the bound variable in the subject. (27) is identical to (26) in every respect, except for the fact that it has a possessive noun phrase subject:

$$
\begin{array}{llll}
\text { Engem } & \text { szeret az anyá-m, } & \text { de téged } & \text { nem. }  \tag{27}\\
\text { I.ACC } & \text { loves the mother-1SG.POSS but you.ACC } & \text { not }
\end{array}
$$

(i) 'My mother loves me, but (your mother does) not (love) you.'
(ii) 'My mother loves me, but (my mother does) not (love) you.'

Second, it needs to be admitted that the constraint against bound variable readings of önmaga is valid for instances of VP ellipsis, but not necessarily for cases of binding by a universal quantifier. (28) is somewhat marked, but it is acceptable nevertheless on what appears to be a bound variable reading. Notice that it is the only possible reading anyway.

$$
\begin{array}{lll}
\text { Mindenki-t } & \text { megijeszt önmaga. }  \tag{28}\\
\text { I-ACC } & \text { scares } & \text { himself.NOM } \\
\text { 'Everybody is scared by himself.' }
\end{array}
$$

Nevertheless, the contrast between (25) and (26) is real. I conclude by maintaining that önmaga has two lexical entries. But we need to weaken the claim made above: önmaga generally disallows bound variable readings if the coreference reading is otherwise available. This is clearly not the way a proper reflexive anaphor is expected to behave.

## 6. The possessive analysis of önmaga

For the sake of comparison, let me start with the proposed LFG-style entry for the reflexive maga 'himself'. I assume a standard LFG binding account in defining the Minimal Complete Nucleus as the local binding domain (see footnote 2), and in modelling the semantic relation between the antecedent and the anaphor as identity (see Dalrymple 1993, 2001). The representative entry in (29) is for the first person singular form magam 'myself'.

$$
\begin{align*}
\text { magam: } & (\uparrow \mathrm{PRED})=\text { 'PRO' }  \tag{29}\\
& (\uparrow \mathrm{PERS})=1 \\
& (\uparrow \mathrm{NUM})=\mathrm{SG} \\
& (\uparrow \mathrm{CASE})=\mathrm{NOM} \\
& (\uparrow \mathrm{PRON}-\mathrm{TYPE})=\text { REFL } \\
& (\uparrow \mathrm{NUCL})=+ \\
& \sim(\mathrm{SUBJ} \uparrow)
\end{align*}
$$

The NUCL+ feature will require the reflexive to bind to a local antecedent (which cannot be a syntactic subject).

The entry önmaga is a more or less run-of-the-mill reflexive. It agrees with its antecedent, it cannot occur as a subject, and it prefers bound variable readings.

```
önmagam \({ }_{1}:(\uparrow \mathrm{PRED})={ }^{\prime} \mathrm{PRO}{ }^{\prime}\)
    \((\) 个PERS \()=1\)
    \((\uparrow\) NUM \()=\) SG
    \((\uparrow\) CASE \()=\mathrm{NOM}\)
    ( \(\uparrow\) PRON-TYPE) \(=\) REFL
    \(\sim(\) SUBJ \(\uparrow)\)
```

The only important difference between (30) and (29) is that (30) lacks (or is underspecified for) the nuclear feature. This is to capture the fact that önmaga $a_{1}$ is not necessarily subject to the Minimal Complete Nucleus binding condition (see 4.1).

What I have dubbed önmaga $a_{2}$ is a special reflexive. It does not agree with its antecedent (or, rather, it is always third person singular), it can occur as a subject, and it prefers what looks like prima facie coreferential readings. It is this entry that I analyze here as a possessive reflexive.

What happens is that the extra nominal morphology (i.e., the prefix ön- 'self') reactivates the dormant possessive structure, which was lost during grammaticalization. This kind of special reanalysis is possible because, as we saw in 3.1., the reflexive stem has still retained the possessive morphology. The claim is that, in essence, önmaga $a_{2}$ is analogous with the possessive expression one's self-representation. The possessor is identified via the possessive agreement morphology, and the stem, mag, acts as some sort of a semantically bleached nominal.

The proposed lexical entry is as follows, once again for the first person singular form:

[^157](31) önmagam ${ }_{2}:(\uparrow P R E D)=$ 'SELF-REPRESENTATION ${ }_{i}$ '
$($ 个PERS $)=3$
$(\uparrow N \mathrm{NM})=\mathrm{SG}$
$(\uparrow$ CASE $)=$ NOM
$(\uparrow$ POSS PRED $)={ }^{\prime} \mathrm{PRO}_{\mathrm{k}}{ }^{\prime}$
$(\uparrow$ POSS PERS $)=1$
$(\uparrow$ POSS NUM $)=$ SG
This analysis gives us an immediate account of the basic facts we observed. Since önmagam ${ }_{2}$ is not a true anaphor, but a possessive structure, we expect it to be grammatical as a subject. What is more, we expect it to trigger constant third person singular agreement on the verb. (31) also describes the fact that this entry is not used in cases of semantic identity with the antecedent, but in 'representations of the self' contexts. Notice that the 'antecedent' now is referentially identified with the possessor buried inside the complex possessive structure of the reflexive. Thus, strictly speaking, what I described here somewhat sloppily as coreference between the reflexive and the antecedent is not direct coreference, but only a referential link between an individual and its representation via the underlying abstract possessive relation.

What the analysis does not capture is why the bound variable reading (between the antecedent and the possessor inside the structure of the reflexive) does not seem to be allowed in cases of VP-ellipsis. It may turn out that this really is just dispreference, contingent on the fact that this possessive structure arguably bears a level of idiomaticity.

Finally, let me add a note about to what extent the possessive analysis is motivated. I mentioned in the introduction that Beavers \& KoontzGarboden (2006) reject the possessive analysis of the English colloquial his ass, and treat it instead as a pronominal. They in fact entertain the idea of an analysis which would be analogous with (31) above, but then they reject it on the basis of the following minimal pair:
(32) a. Mary had her office painted, and Jane had hers remodeled.
b. *John got his ass a pedicure, and Pat got his a manicure.

They argue that (32b) is ungrammatical because his ass is not a possessive structure. But the conclusion is not necessary, compare now (33a) and (33b):
a. My car is faster than John's.
b. *London's fair city is nicer than Dublin's.

What I believe makes (33b) unacceptable is the general drive to avoid breaking up the internal structure of idiomatic units. Dublin's fair city clearly does not encode a true possessive relation, which makes this noun phrase
somewhat idiomatic. But it still is a possessive construction formally. I assume that similar considerations apply to the proposed entry for önmaga ${ }_{2}$.

## 7. Summary and outlook

I started this paper by pointing out that Faltz's (1985) typology does not seem to cover reflexives that are the more complex versions of highly grammaticalized body part reflexives functioning as primary reflexive strategies in their respective languages. Whereas the prime reflexive is nuclear in accordance with the typology, its more complex version need not necessarily be nuclear.

On the basis of the analysis of the Hungarian body part reflexive maga and its more complex counterpart önmaga, I argued that what happens is that the extra nominal morphology (the prefix ön- 'self') reactivates the underlying possessive structure, and creates a special reflexive form. A similar analysis is proposed in Kornfilt (2001) for the Turkish kendisi, and possibly this analysis can also be extended to the Chinese ta ziji.

| REFLEXIVES | PRONOMINAL | COMPLEX | POSSESSIVE |
| :--- | :--- | :--- | :--- |
| NORWEGIAN | seg | seg selv |  |
| ENGLISH |  | himself |  |
| HUNGARIAN |  | maga | önmaga |
| TURKISH |  | kendi | kendisi |
| CHINESE |  | ziji | ta-ziji |

Table 2.
Table 2 gives an overview of the results. In essence, reflexives in the possessive column fall outside of Faltz's (1985) typology, and they add a new dimension to it.

I argued furthermore that when the possessive structure is triggered on önmaga, then we trigger at the same time a reading which targets representations of the self, rather than asserting identity with the self. It remains to be seen to what extent this property carries over to other reflexives with an active possessive structure. It is interesting to note nevertheless that languages that do not have possessive reflexives employ primary complex reflexives in 'representations of the self' contexts, as has been shown, among others, for the English himself by Jackendoff (1992), for the Dutch zichzelf by Reuland (2001), and for the Norwegian seg selv by Lødrup (2007). In contrast, the primary Hungarian complex reflexive, maga, does not allow for such readings. This suggests that one driving force behind the maintained interest in employing possessive reflexives is the need to have a form
specialized for encoding dependencies which do not involve complete semantic identity between the antecedent and the reflexive.

## Acknowledgements

I am grateful to Huba Bartos, Martin Everaert, Volker Gast, Ekkehard König, Tibor Laczkó and the participants of LFG09 in Cambridge for comments on various aspects of this proposal. Special thanks are due to the editors of this volume for their comments on the manuscript. Any remaining erros are mine. This research was supported by the 72983 OTKA grant on the LexicalFunctional Grammar approach to the Hungarian language.

## Bibliography

Beavers, John \& Koontz-Garboden, Andrew. 2006. A universal pronoun in English? Linguistic Inquiry 37 (3). 503-513

Bresnan, Joan. 2001. Lexical-Functional Syntax. Oxford: Blackwell.
Büring, Daniel. 2005. Binding Theory. Cambridge: CUP.
Cole, Peter; Hermon, Gabriella \& Huang, C.-T. James (eds.) 2001. LongDistance Reflexives. Syntax and Semantics 33. San Diego: Academic Press.

Cole, Peter; Hermon, Gabriella \& Lee, Cher Leng. 2001. Grammatical and discourse conditions on long distance reflexives in two Chinese dialects. In P. Cole et al. (eds.) 2001. 1-46.

Culicover, Peter W. \& Jackendoff, Ray. 2005. Simpler Syntax. Oxford: OUP. Dalrymple, Mary. 1993. The Syntax of Anaphoric Binding. Stanford: CSLI.

Dalrymple, Mary. 2001. Lexical Functional Grammar. San Diego: Academic Press.

Faltz, Leonard. 1985. Reflexivisation: A study in Universal Syntax. New York: Garland.

É. Kiss, Katalin. 1994. Sentence structure and word order. In F. Kiefer \& K. É. Kiss (eds.) The syntactic Structure of Hungarian. Syntax and Semantics 27. San Diego: Academic Press. 1-90.

É. Kiss, Katalin 2002. The Syntax of Hungarian. Cambridge: CUP.
Enç, Mürvet. 1989. Pronouns, licensing, and binding. Natural Language and Linguistic Theory 7 (1). 51-92.

Evans, Gareth. 1980. Pronouns. Linguistic Inquiry 11(2). 337-362.

Everaert, Martin \& Szendrői, Kriszta. 2002. Hungarian reflexive anaphors. Talk at the Linguistic Institute of the Hungarian Academy of Sciences.
Jackendoff, Ray. 1992. Mme. Tussaud meets the Binding Theory. Natural Language and Linguistic Theory 10 (1). 1-31.

Kiparsky, Paul. 2002. Disjoint reference and the typology of pronouns. In I. Kaufmann \& B. Stiebels (eds.) More than Words. Berlin: Akademie Verlag. 179-226.

Kornfilt, Jaklin. 2001. Local and Long-Distance Reflexives in Turkish. In P. Cole et al. (eds.) 197-226.
König, Ekkehard \& Gast, Volker. 2006. Focused assertion of identity: a typology of intensifiers. Linguistic Typology 10 (2). 223-276.
Laczkó, Tibor. 2007. Revisiting possessors in Hungarian DPs: A new perspective. In: M. Butt and T.H. King (eds.) Proceedings of the LFG07 Conference. Stanford: CSLI. 343-362. http://csli-publications.stanford.edu/LFG/12/lfg07.html

Lødrup, Helge. 2007. A new account of simple and complex reflexives in Norwegian. The Journal of Comparative Germanic Linguistics 10 (3). 183-201.

Pan, Haihua. 1998. Closeness, prominence, and binding theory. Natural Language and Linguistic Theory 16 (4). 771-815.
Pica, Pierre. 1987. On the nature of the reflexivization cycle. In J. McDonough \& B. Plunkett (eds.) Proceedings of the Seventeenth Annual Meeting of the North Eastern Linguistic Society. Volume 17. 483-500.

Rákosi, György. 2006. Dative Experiencer Predicates in Hungarian. PhD. Dissertation. Utrecht. Published as volume 146 of the LOT Dissertation Series. http://www.lotpublications.nl/publish/issues/Rakosi/index.html

Reinhart, Tanya. 1983. Anaphora and Semantic Interpretation. London \& Sydney: Croom Helm.
Reinhart, Tanya. 2006. Interface Strategies. Optimal and Costly Computations. Linguistic Inquiry Monographs 45. Cambridge, MA: The MIT Press.

Reuland, Eric. 2001. Primitives of Binding. Linguistic Inquiry 32 (3). 439492.

# AUTOMATIC ACQUISITION OF LFG RESOURCES FOR GERMAN - AS GOOD AS IT GETS 

Ines Rehbein and Josef van Genabith<br>Universität des Saarlandes<br>Dublin City University

Proceedings of the LFG09 Conference
Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

We present data-driven methods for the acquisition of LFG resources from two German treebanks. We discuss problems specific to semi-free word order languages as well as problems arising from the data structures determined by the design of the different treebanks. We compare two ways of encoding semi-free word order, as done in the two German treebanks, and argue that the design of the TiGer treebank is more adequate for the acquisition of LFG resources. Furthermore, we describe an architecture for LFG grammar acquisition for German, based on the two German treebanks, and compare our results with a hand-crafted German LFG grammar.


## 1 Introduction

Traditionally, deep, wide-coverage linguistic resources are hand-crafted and their creation is time-consuming and costly. Much effort has been made to overcome this problem by automatically inducing linguistic resources like rich, deep grammars, lexicons and subcategorisation frames from corpora. Most work so far has concentrated on English, like that of Hockenmaier and Steedman [2002], Nakanishi et al. [2004] and Cahill et al. [2002, 2004]. They present successful approaches for the acquisition of deep linguistic resources from the Penn-II treebank, using different grammar frameworks like CCG, HPSG and LFG. English, however, is a configurational language, where strict word-order constraints help to disambiguate predicate-argument structure. Porting these approaches to a semi-free word order language, we have to ask: How good can it get? Can we expect similar results when dealing with (semi-) free word order? Can data-driven methods cope when dealing with ambiguous data structures and sparse data, caused by a rich(er) morphology in combination with case syncretism? And, furthermore, what impact does treebank design have on the automatic acquisition of linguistic resources like deep grammars?

This paper describes approaches to treebank-based acquisition of LFG resources for a semi-free word order language, based on the method of Cahill et al. [2002, 2004, 2008], Burke et al. [2004] and O’Donovan et al. [2005], who presented the large-scale acquisition of LFG grammars and lexical resources from the English Penn-II and Penn-III treebanks. They also presented work on data-driven multilingual unification grammar development for Spanish, Chinese and German. While results point to treebank-based grammar acquisition being a universal method, results for other languages are by far lower than the ones achieved for English and the English Penn treebank.

There are different possible reasons for this: first of all, the size of the English Penn-II treebank, which is much larger than most treebanks for other languages, might be responsible for the good results on English. Another reason might be the configurational English word order, where strict constraints determine the grammatical function of a lexical unit in a certain surface position. Finally, the good results for English might be due to the data structures employed in the Penn-II
treebank, which might be optimised for the task at hand and thus improve performance on the English data.

In this paper we develop different f-structure Annotation Algorithms for German, based on two German treebanks with crucially different annotation schemes, adapted to feature sets of varying granularity as represented in three different gold standards. We discuss problems specific to the annotation schemes of the two treebanks as well as to language-specific properties of German, where the variability in word order and the richer morphology (compared to English) often result in data sparseness, causing severe problems for data-driven methods. Finally, we compare the performance of our data-driven grammar acquisition architectures with the hand-crafted German ParGram LFG of Dipper [2003], Rohrer and Forst [2006], and Forst [2007].

The paper is structured as follows: Section 2 gives an overview of typological properties of German and their representation in two different German treebanks. Section 3 describes the LFG grammar acquisition architecture for German, focusing on the differences to the work of Cahill et al. [2003, 2005] and Cahill [2004]. Section 4 reports on the automatic generation of LFG f-structures and discusses problems specific to semi-free word order and to the design of the German treebanks. Section 5 presents a comparison of our best automatically acquired LFG grammar with related work, namely the hand-crafted ParGram LFG for German. The last section concludes.

## 2 Typological Properties of German and their Representation in Two German Treebanks

German, like English, belongs to the Germanic language family. Despite being closely related, there are crucial differences between the two languages. One of them is the semi-free word order in German, which contrasts with the more configurational English; another, but related difference concerns the richer morphology in German, compared to the rather impoverished English morphology. Both properties are reflected in the treebank data structures used to represent syntactic analyses of the particular languages.

### 2.1 TiGer and TüBa-D/Z: Two German Treebanks

The TiGer treebank [Brants et al., 2002] and the TüBa-D/Z [Telljohann et al., 2005] are two German treebanks with text from the same domain, namely newspaper text. Both treebanks are annotated with phrase structure trees, dependency (grammatical relation) information and POS tags, using the Stuttgart Tübingen Tag Set (STTS) [Schiller et al., 1995]. Differences regard the set of categorial node labels used for syntactic annotation and the set of grammatical function labels. TiGer annotates 25 different syntactic categories and distinguishes between 44 different grammatical functions, while the TüBa-D/Z uses 26 different syntactic categories and 40

"But without the Tigers there will be no peace"
Figure 1: TiGer treebank tree
grammatical function labels. The main differences between the two treebanks are: (1) the flatter annotation in TiGer compared to the more hierarchical annotation in TüBa-D/Z, (2) the annotation of unary nodes in the $\mathrm{TüBa}-\mathrm{D} / \mathrm{Z}$ and no unary nodes in TiGer, (3) TüBa-D/Z uses topological fields to annotate the semi-free German word order, which allows for three possible sentence configurations (verbfirst, verb-second and verb-final), and (4) TiGer annotates Long Distance Dependencies through crossing branches, while TüBa-D/Z encodes LDDs with the help of grammatical function labels (see Figures 1 and 2).

## 3 Automatic Annotation of LFG F-Structures

Cahill et al. [2003, 2004, 2005, 2008] presented a modular architecture for automatically annotating the English Penn-II treebank with LFG f-structures (Figure 3 ), which enables them to automatically extract deep, wide-coverage grammars which yield results in the same range as the best hand-crafted grammars for English [Briscoe and Carroll, 2002, Kaplan et al., 2004]. The f-structure Annotation Algorithm (AA) exploits lexical head information, and categorial, configurational and functional information as well as traces and co-indexation annotated in the Penn-II treebank. After determining the head of each constituent, the main module of the AA uses left-right context annotation principles to assign the most probable f -structure equation to each node in the tree (Figure 3). These principles express annotation generalisations and have been hand-crafted by looking at the most frequent grammar rules for each node in the Penn-II treebank and are also applied to unseen low-frequency rules. A sample partial left-right context annotation rule for NPs is given in Table 1. The left-context rule states that all adjectives or adjectival phrases to the left of the head of an NP should be annotated as an adjunct, while the right-context rule specifies that an NP to the right of the head of an NP is an

"However, there won't be considerable reinforcements for the next playing season."
Figure 2: TüBa-D/Z treebank tree


Figure 3: Architecture of the English f-structure Annotation Algorithm (AA)
apposition. The creation of these left-right-context rules needs linguistic expertise and crucially depends on configurational properties of English.

| left-context | head | right-context |
| :---: | :---: | :---: |
| JJ, ADJP: $\downarrow=\in \uparrow$ ADJUNCT | NN, NNS, <br> $\uparrow=\downarrow$ | $\mathrm{NP}: \downarrow=\in \uparrow \mathrm{APP}$ |

Table 1: Left-right context annotation rule used in the English AA
Coordinations are treated seperately. After adding f-structure equations to all nodes in the tree, the Catch-All and Clean-Up module deals with overgeneralisations. Finally, traces are resolved.

The German LFG AA, like the English one, is highly modularised and proceeds as follows (Figure 4). First it reads in the treebank trees encoded in the NEGRA export format and converts each tree into a tree object. Then it applies head-finding rules which we developed in the style of Magerman [1995], in order to determine the head of each local node. ${ }^{1}$ The head-finding rules specify a set of candidate heads, depending on the syntactic category of the node, and also the

[^158]

Figure 4: Architecture of the German f-structure Annotation Algorithm
direction (left/right) in which the search should proceed. For prepositional phrases, for example, we start from the left and look at all child nodes of the PP. If the leftmost child node of the PP has the label KOKOM (comparative particle), we assign it as the head of the PP. If not, we check if it is a preposition (APPR), a preposition merged with a determiner (APPRART), an apposition (APPO), and so on. If the left-most child node does not carry one of the candidate labels, we take a look at the next child node, working our way from left to right.

For some of the nodes these head-finding rules work quite well, while for others we have to accept a certain amount of noise. This is especially true for the flat NPs in the TiGer treebank. A Special Cases module checks these nodes at a later stage in the annotation process and corrects possible errors made in the annotation.

After determining the heads, the tree is handed over to the Macros module which assigns f-structure equations to each node. This is done with the help of macros. Sometimes these macros overgeneralise and assign an incorrect grammatical function. In order to deal with this, the Special Cases module corrects inappropriate annotations made by the Macros module. Finally the Validation module takes a final look at the annotated trees and makes sure that every node has been assigned a head and that there is no node with two child nodes carrying the same governable grammatical function.

The most important difference in the design of the English and the German AAs concerns the application of left-right context annotation rules described above. For English, these rules successfully specify the correct annotation for the majority of local nodes in a given tree. For German, however, these rules do not work as well as for English. Table 2 illustrates this point by showing different possibilities for the surface realisation of a (rather short) German sentence. Some of the examples are highly marked, but all of them are possible surface realisations of (1).
(1) Die Anklage legt ihm deshalb Betrug zur Last.
the prosecution lies him therefore fraud to the burden.
The prosecution therefore charges him with fraud.
The f-structure-annotated grammar rule for the sentence in (1) (Figure 5) tells us that the first NP Die Anklage (the prosecution) is the subject of the sentence,
running the AA on parser output trees with erroneous or no GF labels in the trees, we also make use of head-finding rules for other syntactic categories.

In TüBa-D/Z, heads are marked for most categorial nodes. However, there are some open issues, like the one concerning the head of the middle field or of proper name nodes, or the annotation of appositions, which are considered to be referentially identical and therefore bear no head marking in the TüBa-D/Z.

```
S NP VVFIN PPER PROAV NN PP
    \uparrowSUBJ=\downarrow \=\downarrow \ \D= }\downarrow\quad\downarrow\in\uparrow\mathrm{ MO }\quad\uparrow\textrm{OA}=\downarrow\quad\uparrow\textrm{OP}=
```

Figure 5: Grammar rule and f-structure equations for the sentence in (1)

| Die Anklage | legt | ihm | deshalb | Betrug | zur Last. |
| :--- | :--- | :--- | :--- | :--- | :--- |
| Betrug | legt | ihm | deshalb | die Anklage | zur Last. |
| Ihm | zur Last | legt | die Anklage | deshalb | Betrug. |
| Zur Last | legt | ihm | die Anklage | deshalb | Betrug. |
| Deshalb | legt | ihm | die Anklage | Betrug | zur Last. |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |

Table 2: Variable word order in German (sentence (1))
while the noun Betrug (fraud) should be annotated as an accusative object, and the pronominal adverb deshalb (therefore) is an element of the modifier set. Table 2, however, illustrates that these constituents can occur in very different positions to the left or right of the head of the sentence. This shows that, unlike for a strongly configurational language such as English, the specification of left-rightcontext rules for German is not very helpful.

Instead of developing horizontal and strongly configurational context rules, the AA for German makes extended use of macros, using different combinations of information such as part-of-speech (POS) tags, node labels, edge labels and parent node labels (as encoded in the TiGer and TüBa-D/Z treebanks). First we apply more general macros assigning functional annotations to each POS, syntactic category or edge label in the tree. More specific macros, such as the combination of a POS tag with the syntactic node label of the parent node or a categorial node with a specific grammatical function label, can overwrite these general macros. The order of these macros is crucial, dealing with more and more specific information. Some of the macros overwrite information assigned before, while others only add more information to the functional annotation.

To give an example, consider the POS tag ART (determiner). The first macro is triggered by this POS tag and assigns the f -structure equation $\uparrow=\downarrow, \downarrow$ det-type $=$ def. The next macro looks at combinations of POS tags and grammatical function (GF) labels and, for a determiner with the label NK (noun kernel), adds the equation $\uparrow$ spec : det $=\downarrow$, while the same POS tag gets assigned the functional equation $\downarrow \in \uparrow$ spec : number when occurring with the edge label NMC (numerical component). The annotation for the combination of POS and grammatical function label can be overwritten when a more specific macro applies, e.g. one which also considers the parent node for a particular POS-GF-combination.

The determiner with edge label NK has so far been annotated with headword, $\downarrow$ det-type $=\operatorname{def}, \uparrow$ spec $: \operatorname{det}=\downarrow$. This is overwritten with the f-structure equation $\uparrow$ obj : spec : det $=\downarrow$, if it is the child of a PP node. This is due to the fact that the annotation guidelines of the TiGer treebank analyse prepositions as the head of a PP, while the head noun (and its dependents) inside the PP is annotated as the
object of the preposition. Due to the flat annotation in the TiGer treebank, it is not helpful to use vertical context above the parent node level. The AA makes heavy use of the Special Cases module, where further annotation rules are specified for most syntactic categories. One tricky case is that of NPs, which have a totally flat structure in the TiGer treebank. There are many cases where the information about POS tag and grammatical function label is not sufficient, and neither is their relative position to the head of the phrase. In those cases the presence or absence of other nodes decides the grammatical function of the node in question.


Figure 6: NP-internal structure in TiGer ( $\mathrm{PN}=$ head)


Figure 7: NP-internal structure in TiGer (PN=apposition)
To illustrate this, consider the three examples in Figures 6-8. All three examples show an NP with a noun child node followed by a proper name (PN) node, but where the grammatical annotations differ crucially. In Figure 6, the PN is the head of the NP. In Figure 7, where we have a determiner to the left of the noun (NN), the noun itself is the head of the NP, while the PN is an apposition. The third example (Figure 8) looks pretty much like the second one, with the exception that Merkel is in the genitive case. Here the PN should be annotated as a genitive attribute. This is not so much a problem for the annotation of the original treebank trees where we have both the correct grammatical function labels as well as morphological information. For parser output, however, morphological information is not available and the grammatical functions assigned are often incorrect. In Section 4.2.1


Figure 8: NP-internal structure in TiGer (PN=genitive to the right)
we will return to this issue und discuss the reason for the missing morphological information in the parser output.

### 3.1 Differences between our AA for German and Preliminary Work

The annotation algorithm for German presented in this chapter is based on and substantially revises and extends preliminary work by Cahill et al. [2003, 2005] and Cahill [2004]. The AA by Cahill et al. provides annotations for a rather limited set of grammatical functions only ( 26 grammatical functions: 11 governable functions, 10 non-governable functions and 5 atomic features). We created a new gold standard f-structure bank containing 250 sentences from the TiGer treebank, the TIGER250, which uses a substantially extended set of grammatical functions and features ( 46 grammatical functions: 14 governable grammatical functions, 13 non-governable grammatical functions and 19 atomic features). As a result, the annotated resources contain richer linguistic information and are of higher quality and usefulness compared to the one of Cahill et al. [2003, 2005] and Cahill [2004]. Our annotation algorithm also makes use of a valency dictionary in order to distinguish between stative passive constructions and the German Perfekt with sein 'to be'.

We also adapted the AA to the feature set used in the TiGer $\mathrm{DB}^{2}$ [Forst et al., 2004] (Dependency Bank) and a hand-crafted gold standard from the TüBa-D/Z ${ }^{3}$ (TUBA100).

[^159]
## 4 LFG F-Structure Annotation and Evaluation on Two German Treebanks

For German, we adapted the AA to the node and edge labels of the two German treebanks. As described above, word order variation in German does not allow to make strong use of configurational information as in the English AA. Instead, we heavily rely on the grammatical function labels in the trees. This works well when annotating original treebank trees, but causes many problems when applied to parser output. State-of-the-art parsing results as presented in the PaGe Shared Task on Parsing German [Kübler, 2008] are in the range of $58-70 \%$ F-score for TiGer and $75-84 \%$ for TüBa-D/Z. ${ }^{4}$ The differences in annotation schemes do not allow for a direct comparison of parsing results, but the message is clear: for both treebanks automatically assigned syntactic nodes and, even more important, grammatical function labels are to a great extent error-prone, which defines an upper bound for treebank-based parsing into f-structures using the automatic annotation algorithm.

Section 4.2 presents parsing experiments with automatic LFG f-structure annotation based on TiGer and TüBa-D/Z, and evaluates the generated f-structures against hand-crafted gold standards from the TiGer treebank (TiGer DB, TIGER250) and from the TüBa-D/Z (TUBA100). However, before applying the AA to parser output we want to test its performance on gold standard syntax trees.

### 4.1 Results for LFG F-Structure Annotation on Gold Standard Syntax Trees

Table 3 shows results for automatic f-structure annotation on gold treebank trees for the sentences in the TiGer DB, the TIGER250 and the TUBA100. ${ }^{5}$ Results for

|  | Prec. | Rec. | F-Score |
| :--- | ---: | ---: | ---: |
| TiGerDB | 87.8 | 84.8 | 86.3 |
| TIGER250 | 96.8 | 97.5 | 97.1 |
| TUBA100 | 95.5 | 94.6 | 95.0 |

Table 3: Results for automatic f-structure annotation on gold treebank trees
the TIGER250 and the TUBA100 are quite good, while results for the TiGer DB are around $10 \%$ lower. This is due to mapping problems between the TiGer DB and TiGer treebank. The sentences in the TiGer DB have been converted semiautomatically into a dependency-based triple format, using a large, hand-crafted LFG grammar for German [Dipper, 2003] and then manually corrected. The TiGer DB provides a very fine-grained description of linguistic phenomena in German,

[^160]but includes additional information which is not annotated in the TiGer treebank and thus cannot be derived automatically. This means that the TiGer DB-based evaluation is biased in favour of the hand-crafted LFG grammar of Dipper [2003].

### 4.2 Parsing German with Automatically Acquired LFG Grammars

In our experiments we use the Berkeley parser [Petrov and Klein, 2008], a languageagnostic parser which automatically refines and re-annotates the training data by applying split-and-merge operations, so that the likelihood of the transformed treebank is maximised. The Berkeley parser achieved the best results in the Shared Task on Parsing German (ACL 2008).

We removed the gold standard sentences from the treebanks and extracted two training sets with 25,000 sentences each. For TiGer we persued two different ways of resolving crossing branches in the trees: (1) by attaching the non-head child nodes higher up in the tree, following Kübler [2005], and (2) by splitting discontinuous nodes into smaller "partial nodes" [Boyd, 2007], a strategy which aims at preserving local tree structure while allowing the system to recover the original dependencies after parsing. With regard to GF labels we tested two different settings: in the first setting (Atomic) we merged categorial node labels with grammatical function labels and trained the parser on the new atomic labels. In the second setting (FunTag) we removed GF labels from the training data and trained the parser on syntactic categories only. The GF labels were then assigned in a postprocessing step, using the SVM-based grammatical function labelling software by Chrupała et al. [2007]. We parsed the different test sets with the extracted grammars and, for the grammars without grammatical functions, let FunTag assign GFs to the parser output. The trees with grammatical function labels were passed over to the AA, where all nodes in the parse trees were annotated with LFG functional equations. Next we collected the equations and handed them over to a constrainst solver, which generated LFG f-structures.

### 4.2.1 Results

Table 4 shows constituent-based parsing results for the different test sets and settings (Atomic, FunTag) as well as results for f-structure evaluation. For the first setting, where we let the Berkeley parser assign the grammatical functions (Atomic), the two TiGer test sets yield constituent-based parsing results in the range of 76$79 \%$ (labelled F-score on syntactic categories) and 67-70\% (including GF labels). Results for the TüBa-D/Z are more than $10 \%$ higher, which is an artifact of the different treebank annotation schemes and does not reflect parser output quality, as can be seen in the f-structure evaluation. On the f-structure level precision is in the range of $73-81 \%$, while recall for the TüBa-D/Z f-structures is dramatically lower at around $45 \%$. For the TiGer, we achieve a recall of $73.7 \%$ for TiGer DB and of $79.7 \%$ for the TIGER250 test set.

Parsing results for the Berkeley parser trained on TiGer syntactic nodes only

| Constituent-based evaluation |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Atomic |  |  |  | FunTag |  |  |
| length<=40 | F-score | F-score GF | POS acc. | F-score | F-score GF | POS acc. |
| TiGerDB | 79.3 | 70.2 | 96.0 | 81.0 | 70.9 | 97.0 |
| TIGER250 | 76.6 | 66.9 | 95.4 | 79.3 | 68.4 | 96.5 |
| TUBA100 | 89.3 | 80.2 | 96.5 | 89.2 | 76.3 | 96.4 |
| $f$-structure evaluation |  |  |  |  |  |  |
| Atomic |  |  |  | FunTag |  |  |
|  | Precision | Recall | F-score | Precision | Recall | F-score |
| TiBerDB | 73.0 | 73.9 | 73.4 | 76.1 | 65.1 | 70.2 |
| TIGER250 | 81.4 | 79.7 | 80.5 | 87.6 | 67.5 | 76.3 |
| TUBA100 | 76.9 | 45.1 | 56.9 | 75.8 | 39.3 | 51.7 |

Table 4: C-structure parsing results (labelled F-score without and with GF) and f -structure evaluation
(FunTag) are higher than for the atomic labels. For TüBa-D/Z, however, we observe better results when training on both syntactic categories and grammatical functions. The FunTag-assigned GFs yield better evalb results and a higher precision for the TiGer f-structures. For the TüBa-D/Z, precision is slightly lower than for f -structures generated from parser output where the Berkeley parser did the function labelling. The better precision for the TiGer f-structures comes at the cost of a decrease in recall. For the TüBa-D/Z f-structures, recall is even lower than before.

There are several reasons for the low recall for the TüBa-D/Z: (1) Due to its limited size the TUBA100 does not cover all relevant grammatical phenomena and therefore is not sufficient as a test set for grammar development, which is reflected in the low recall score. (2) Phrases without a clear dependency relation to the other constituents in the tree are attached directly to the root node in the TüBa-D/Z. The resulting tree structure makes it impossible for the AA to disambiguate the sentence and find a suitable dependency relation for the highly attached node, which means that these nodes are not represented in the f-structure, further lowering recall for the TüBa-D/Z. (3) NP internal structure in the TüBa-D/Z contains less information than in TiGer, where grammatical function labels distinguish genitive attributes, dative attributes and comparative complements. The missing information can be partly retrieved from morphological annotation, but this would require an extensive treebank transformation to make this information available to the parser. The grammars extracted from the treebanks do not include morphological information, which means that the TiGer grammars encodes more specific functional information than the $\mathrm{TuBa}-\mathrm{D} / \mathrm{Z}$ grammars.

Yet another reason for the lower recall for TüBa-D/Z f-structures can be found in the design of the grammatical function labels used in the annotation. While the original treebanks use roughly the same number of grammatical functions (44 in TiGer versus 40 in TüBa-D/Z; Table 5), some of the grammatical functions in the TüBa-D/Z occur only with a very low frequency. When comparing two smaller subsets of 2,000 gold treebank trees, we still find 42 of the 44 GFs in

|  | Gold all | Gold 2000 | Atomic | FunTag |
| :--- | :---: | :---: | :---: | :---: |
| TiGer | 44 | 42 | 41 | 40 |
| TüBa-D/Z | 40 | 33 | 31 | 19 |

Table 5: Number of different grammatical functions in TiGer/TüBa-D/Z gold trees and reproduced in the different parsing settings (Atomic/FunTag)
the TiGer set, while the TüBa-D/Z subset uses only 33 of the 40 GFs. For parser output the problem gets even worse. In the TiGer-trained parser output for the same subset of 2,000 sentences we find 41 different GF labels when the Berkeley parser assigns the grammatical functions, and 40 when FunTag does the GF labelling, while in a data set of the same size from the TüBa-D/Z, only 31 different GF labels are used in the parser output (Atomic), and the FunTag approach yields only 19 different grammatical functions. This leads to a crucial difference between the type of information encoded in the GF labels for the two treebanks: while TiGer labels describe the grammatical function of one node, in TüBa-D/Z the GF labels (besides the main grammatical functions such as subject and acusative or dative object) express dependency relations between different nodes in the tree, which are often positioned in different topological fields. As pointed out, some of the grammatical functions in the TüBa-D/Z occur with a very low frequency. ${ }^{6}$ This poses a problem for machine learning methods, which rely on a sufficiently large set of training instances in order to achieve good performance on unseen data.

| GF | Atomic | FunTag | Atomic | FunTag |
| :--- | :---: | :---: | :---: | :---: |
| TiGer (2,000 sent.) |  |  | TüBa-D/Z(2,000 sent.) |  |
| DA | 52.5 | 74.9 | 56.8 | 27.2 |
| OA | 79.5 | 85.5 | 69.0 | 46.4 |
| SB | 90.0 | 88.4 | 85.2 | 72.1 |
| ALL GF | 93.1 | 94.4 | 91.9 | 88.3 |

Table 6: Evaluation of main grammatical functions in TiGer and TüBa-D/Z (dative object: DA/OD, accusative object: OA, subject: SB/ON)

Next we compare results for the main grammatical functions (subject, accusative and dative object) on 2,000 sentence test sets from TiGer and TüBa-D/Z (Table 6). For parser-assigned GFs, we observe better results for dative objects (DA/OD) for the parsing model trained on the TüBa-D/Z, while for subjects and accusative objects the TiGer-trained parser yields better results. The SVM-based FunTag shows poor performance on the TüBa-D/Z data, while for TiGer the function labeller outperformes the setting where the Berkeley parser does the GF assignment (Atomic). This divergent behaviour might be due to the different data

[^161]structures in the treebanks. The split into topological fields in the TüBa-D/Z takes away necessary context information, which is encoded in the feature set for the flat TiGer trees.

### 4.3 Different Approaches to Discontinuity and their Impact on FStructure Annotation

Boyd [2007] presents an improved method for converting the crossing branches in TiGer into context-free representations by splitting up discontinuous nodes into marked "partial" nodes. She shows that the improved conversion results in more consistent trees and improves results in a labelled dependency evaluation for accusative, dative and prepositional objects. In her experiments, Boyd used an unlexicalised PCFG parsing model (LoPar, Schmid [2000]) with gold POS tags as parser input.

We applied the split-node conversion method to the TiGer data and trained the Berkeley parser on the converted training sets. Table 7 shows parsing results for the two conversion methods: (1) raised nodes and (2) split nodes. For the TiGer DB test set, results for the split-node conversion are slightly worse, while for the TIGER250 test set there is a small improvement of $1 \%$ F-score. For both data sets, however, the number of valid f-structures decreases considerably.

|  | Precision | Recall | F-score | valid F-struc. |
| :--- | :---: | :---: | :---: | :---: |
| TiGer DB |  |  |  |  |
| raised | 73.0 | 73.9 | 73.4 | 82.4 |
| split | 71.8 | 72.0 | 71.9 | 71.0 |
| TIGER250 |  |  |  |  |
| raised | 81.5 | 80.9 | 81.2 | 88.0 |
| split | 82.7 | 81.8 | 82.2 | 84.0 |

Table 7: f-structure evaluation on converted TiGer trees (raised- vs. split-node)
Boyd's split-node conversion works well for pure PCFG parsers like LoPar. The Berkeley parser, however, makes use of horizontal markovisation, which breaks up the original grammar rules and generates new rules which have not been seen in the training set. This also admits rules with only one of the two partial nodes, which means that a reconstruction of the original tree is impossible, and often leads to clashes during f -structure generation.

## 5 LFG Parsing: Related Work

This section discusses related work and shows how our research compares to the wide-coverage hand-crafted LFG grammar of Dipper [2003], Rohrer and Forst [2006], and Forst [2007] developed in the ParGram project [Butt et al., 2002]. The ParGram German LFG uses 274 LFG-style rules (with regular expressionbased right-hand sides) and several lexicons with detailed subcategorisation information and a guessing mechanism for default lexical entries [Rohrer and Forst,

|  | ParGram |  |  | TiGerDB | DCU250 |
| :--- | :---: | :---: | :---: | :---: | :---: |
| GF | up. <br> bound | log. <br> lin. | low. <br> bound |  |  |
| $d a$ | 67 | 63 | 55 | 44 | 38 |
| $g r$ | 88 | 84 | 79 | 71 | 87 |
| $m o$ | 70 | 63 | 62 | 65 | 73 |
| oa | 78 | 75 | 65 | 69 | 63 |
| quant | 70 | 68 | 67 | 67 | 78 |
| $r c$ | 74 | 62 | 59 | 34 | 30 |
| $s b$ | 76 | 73 | 68 | 74 | 79 |
| preds | 79.4 | 75.7 | 72.6 | 72.7 | 78.6 |
| only |  |  |  |  |  |
| coverage on the NEGRA treebank $(>20,000$ sentences $)$ |  |  |  |  |  |
| 81.5 |  |  |  |  | 81.5 |
| 81.5 |  |  |  |  | 88.2 |

Table 8: F-scores for selected grammatical functions for the ParGram LFG (upper bounds, log-linear disambiguation model, lower bounds) and for two automatically acquired TiGer grammars

2006]. Preprocessing in the experiments reported in Rohrer and Forst [2006] includes modules for tokenisation, morphological analysis and manual marking of named entities, before the actual parsing takes place. An additional disambiguation component based on maximum entropy models is used for reranking the output of the parser. Forst [2007] tested parser quality on 1,497 sentences from the TiGer DB and reported a lower bound, where a parse tree is chosen randomly from the parse forest, an upper bound, using the parse tree with the highest F-score (evaluated against the gold standard), as well as results for parse selection done by the log-linear disambiguation model.

Table 8 shows results for the ParGram LFG and for the automatically induced grammars on selected grammatical relations and on all grammatical functions excluding morphological and other features (preds only). The automatically induced TiGer DB and DCU250-style grammars were trained on the full TiGer treebank ( $>48,000$ sentences, excluding the test data). We report results for the test sets from the TiGer DB and the DCU250.

The hand-crafted LFG outperforms the automatically acquired grammars on most GFs for the TiGer DB, but results are not directly comparable. The TiGer DB-based evaluation is biased in favour of the hand-crafted LFG. Named entities in the ParGram LFG input are marked up manually, while for our grammars these multiword units often are not recognised correctly and so are punished during evaluation, even if part of the unit is annotated correctly. Furthermore, the hand-crafted ParGram LFG grammar was used in the creation of the TiGer DB gold standard in the first place, ensuring compatibility as regards tokenisation and overall linguistic analysis.

F-scores for the DCU250 are in roughly the same range as the ones for the hand-crafted grammar. For high-frequency dependencies like subjects (sb) or modifiers (mo), results of the two grammars are comparable. For low-frequency depen-

|  | ParGram |  |  |  |
| :--- | :---: | :---: | :---: | :---: | :---: |
| Lig. |  |  |  |  |
| GF | $\begin{array}{c}\text { log. } \\ \text { bound }\end{array}$ | $\begin{array}{c}\text { low. } \\ \text { lin. }\end{array}$ |  |  |
| bound |  |  |  |  |$)$

Table 9: Precision for selected grammatical functions for the ParGram LFG and for the TiGer grammars
dencies like dative objects (da) or relative clauses (rc), however, the hand-crafted LFG outperforms the automatic LFG f-structure annotation algorithm by far. Coverage for the automatically acquired grammars is considerably higher than for the hand-crafted LFG grammar. Rohrer and Forst [2006] report a coverage of 81.5\% (full parses) when parsing the NEGRA treebank, which contains newspaper text from the same newspaper as in the TiGer treebank. By contrast, the automatically acquired TiGer grammars achieve close to $90 \%$ coverage on the same data. On the TiGer treebank Rohrer and Forst [2006] report coverage of $86.4 \%$ full parses, raising the possibility that, as an effect of enhancing grammar coverage by systematically extracting development subsets from TiGer, the ParGram LFG is tailored closely to the TiGer treebank.

The DCU250 test set is equally biased towards the TiGer treebank-based LFG resources, as it only represents what is encoded (directly or implicitly) in the TiGer treebank. The truth is somewhere in between: The TiGer DB evaluation of the treebank-based LFG resources attempts to a limited extent to counter the bias of the original TiGer DB resource towards the hand-crafted LFG grammar by removing distinctions which cannot be learned from TiGer data only, and by relating TiGer DB to (some of) the original TiGer tokenisation using the version prepared by Boyd et al. [2007]. The resulting resource still favours the hand-crafted LFG resources, which outperform the treebank-based resources by about $3 \%$ points absolute. Looking at precision, results for the TiGer grammars are more or less in the same range as the F-scores for the Pargram LFG (Table 9). ${ }^{7}$

### 5.1 Discussion

Our automatically extracted grammars yield better coverage than the hand-crafted LFG of Dipper [2003], Rohrer and Forst [2006] and Forst [2007], but with regard to F-score the ParGram LFG still outperforms the automatically acquired gram-

[^162]mars. The lower results for our grammars are not due to low precision: Table 9 contrasts F-scores for the Pargram LFG with results for precision as achieved by the automatically acquired TiGer grammars. Future work should therefore focus on improving recall in order to achieve results comparable with or better than hand-crafted grammars. One promising approach is the one of Seeker [2009], who describes a grammatical function labeller based on Integer Linear Programming (ILP). Seeker presents a two-step approach, consisting of a classification step and a selection step. During classification, the probability distribution over all possible labels for each node in the tree is computed, using a maximum entropy classifier. During selection, the overall probability of the whole tree is optimised, where the ILP-based approach allows the developer to implement hard constraints (e.g.: no more than one subject per local tree). First results show that global optimisation in combination with linguistically motivated constraints improves precision and coverage. F-scores for f-structure evaluation on the TiGer DB increase to more than $75 \%$, while coverage was raised from around $88 \%$ to more than $96 \%$.

An unsolved problem is the encoding of LDDs in treebank annotation schemes for (semi-) free word order languages. Currently, neither the TiGer treebank and even less so the $\mathrm{TüBa}-\mathrm{D} / \mathrm{Z}$ way of representing non-local dependencies can be learned successfully by statistical parsers. An approach to resolving LDDs at the f-structure level was described in Cahill et al. [2004] and Cahill [2004] and successfully implemented as part of the English treebank-based LFG acquisition and parsing architectures. However, the method of Cahill et al. relies on complete fstructures, which means that the recall problem must have been solved before we can reliably and profitably compute LDDs on f-structure level for German.

## 6 Conclusions

We presented two architectures for the automatic acquisiton of LFG resources, based on two German treebanks. Compared to a hand-crafted German LFG, our method yields higher coverage and comparable results for the high-frequency grammatical functions, while for the less frequent GFs the hand-crafted grammar clearly outperforms the automatic approach.

We have outlined a number of problems for treebank-based f-structure annotation for German: (1) The semi-free word order in German rules out the use of configurational information for f-structure annotation. (2) Parsing results for German, especially for GF assignment, are not reliable enough to support accurate f-structure annotation. (3) Our alternative approach to assign GF labels using an SVM-based function labeller achieves high precision, but at the cost of recall. This is due to missing context sensitivity of the function labeller, resulting in the assignment of conflicting GFs.

We showed that particular treebank encoding schemes have a strong impact on the usability of the resources. We argue that the GF label set in the TüBa-D/Z, which has been designed with the aim of expressing dependency relations between
different nodes in the tree, is less adequate for the automatic acquisition of LFG resources than the label set in TiGer. The GF labels in the TüBa-D/Z are harder to learn and also encode less specific grammatical information than the ones in TiGer.

The task of automatically inducing linguistic resources from (semi-) free word order languages is much harder than for more configurational languages like English. Future research needs to address the problem of automatic GF assignment which for German is far more important than for configurational languages (one promising line of research has been outlined in Section 5.1). Only then can we expect to automatically induce high-quality linguistic resources for languages other than English and other configurational languages.

## References

Adriane Boyd. Discontinuity revisited: An improved conversion to context-free representations. In Proceedings of the Linguistic Annotation Workshop (LAW 2007), pages 41-44, Prague, Czech Republic, 2007.

Adriane Boyd, Markus Dickinson, and Detmar Meurers. On representing dependency relations - insights from converting the German TiGerDB. In Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT07), pages 31-42, Bergen, Norway, 2007.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. The TIGER Treebank. In Erhard W. Hinrichs and Kiril Simov, editors, Proceedings of the First Workshop on Treebanks and Linguistic Theories, pages 24-42, Sozopol, Bulgaria, 2002.

Ted Briscoe and John Carroll. Robust accurate statistical annotation of general text. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02), pages 1499-1504, Las Palmas, Canary Islands, 2002.

Michael Burke, Olivia Lam, Aoife Cahill, Rowena Chan, Ruth O’Donovan, Adams Bodomo, Josef van Genabith, and Andy Way. Treebank-based acquisition of a chinese lexical-functional grammar. In Proceedings of the 18th Pacific Asia Conference on Language, Information and Computation (PACLIC-18), pages 161-172, Tokyo, Japan, 2004.

Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. The parallel grammar project. In Proceedings of COLING-02 Workshop on Grammar Engineering and Evaluation, pages 1-7, Taipei, Taiwan, 2002.

Aoife Cahill. Parsing with Automatically Acquired, Wide-Coverage, Robust, Probabilistic LFG Approximations. PhD dissertation, School of Computing, Dublin City University, Dublin, Ireland, 2004.

Aoife Cahill, Mairéad McCarthy, Josef van Genabith, and Andy Way. Automatic annotation of the penn-treebank with LFG f-structure information. In LREC-02 workshop on Linguistic Knowledge Acquisition and Representation - Bootstrapping Annotated Language Data, Third International Conference on Language Resources and Evaluation (LREC-02), post-conference workshop, pages 8-15, Paris, France, 2002.

Aoife Cahill, Martin Forst, Mairéad McCarthy, Ruth O'Donovan, Christian Rohrer, Josef van Genabith, and Andy Way. Treebank-based multilingual unificationgrammar development. In Proceedings of the Workshop on Ideas and Strategies for Multilingual Grammar Development, at the 15th European Summer School in Logic Language and Information, Vienna, Austria, 2003.

Aoife Cahill, Michael Burke, Ruth O’Donovan, Josef van Genabith, and Andy Way. Long-distance dependency resolution in automatically acquired widecoverage PCFG-based LFG approximations. In 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), pages 319-326, Barcelona, Spain, 2004.

Aoife Cahill, Martin Forst, Michael Burke, Mairéad McCarthy, Ruth O’Donovan, Christian Rohrer, Josef van Genabith, and Andy Way. Treebank-based acquisition of multilingual unification grammar resources. Journal of Research on Language and Computation; Special Issue on Shared Representations in Multilingual Grammar Engineering, pages 247-279, 2005.

Aoife Cahill, Michael Burke, Ruth O’Donovan, Stefan Riezler, Josef van Genabith, and Andy Way. Wide-coverage deep statistical parsing using automatic dependency structure annotation. Computational Linguistics, 34(1):81-124, 2008.

Grzegorz Chrupała, Nicolas Stroppa, Josef van Genabith, and Georgiana Dinu. Better training for function labeling. In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP 2007), pages 133-138, Borovets, Bulgaria, 2007.

Stefanie Dipper. Implementing and documenting large-scale grammars - German LFG, doctoral dissertation, ims, university of stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), 9(1), 2003.

Martin Forst. Filling statistics with linguistics - property design for the disambiguation of German LFG parses. In Proceedings of the ACL Workshop on Deep Linguistic Processing, pages 17-24, Prague, Czech Republic, 2007.

Martin Forst, Núria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordoni. Towards a dependency-based gold standard for German parsers - the TiGer Dependency Bank. In Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC '04), pages 31-38, Geneva, Switzerland, 2004.

Julia Hockenmaier and Mark Steedman. Generative models for statistical parsing with combinatory categorial grammar. In 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), pages 335-342, Philadelphia, PA, 2002.

Ronald M. Kaplan, Stefan Riezler, Tracy H. King, John. T. Maxwell III, Alexander Vasserman, and Richard Crouch. Speed and accuracy in shallow and deep stochastic parsing. In Proceedings of the Human Language Technology Conference and the 4th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04), pages 97-104, Boston, MA, 2004.

Sandra Kübler. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In Proceedings of the 5th International Conference on Recent Advances in Natural Language Processing (RANLP 2005), pages 293-300, Borovets, Bulgaria, 2005.

Sandra Kübler. The PaGe 2008 shared task on parsing German. In ACL Workshop on Parsing German (PaGe-08), pages 55-63, Columbus, OH, 2008.

David M. Magerman. Statistical decision-tree models for parsing. In 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95), pages 276283, Cambridge, MA, 1995.

Hiroko Nakanishi, Yusuke Miyao, and Jun'ichi Tsujii. Using inverse lexical rules to acquire a wide-coverage lexicalized grammar. In IJCNLP 2004 Workshop on Beyond Shallow Analyses - Formalisms and Statistical Modeling for Deep Analyses, Sanya City, Hainan Island, China, 2004.

Ruth O’Donovan, Michael Burke, Aoife Cahill, Josef van Genabith, and Andy Way. Large-scale induction and evaluation of lexical resources from the penn-II and penn-III treebanks. Computational Linguistics, 31(3):329-366, 2005.

Slav Petrov and Dan Klein. Parsing German with language agnostic latent variable grammars. In ACL Workshop on Parsing German (PaGe-08), pages 33-39, Columbus, OH, 2008.

Christian Rohrer and Martin Forst. Improving coverage and parsing quality of a large-scale LFG for German. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06), pages 2206-2211, Genoa, Italy, 2006.

Anne Schiller, Simone Teufel, and Christine Thielen. Guidelines für das tagging deutscher textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen, Tübingen, Germany, 1995.

Helmut Schmid. LoPar: Design and implementation. Technical report, Universität Stuttgart, Stuttgart, Germany, 2000.

Wolfgang Seeker. On the Use of Hard Linguistic Constraints in Automatic Grammar Acquisition for German. Diploma thesis, Institut für Linguistik, Potsdam, Germany, 2009.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Universität Tübingen, Germany, 2005.

Yannick Versley. Parser evaluation across text types. In Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT-05), pages 209-220, Barcelona, Spain, 2005.

# ON THE DIFFERENCE BETWEEN AUXILIARIES, SERIAL VERBS AND LIGHT VERBS 

Melanie Seiss<br>Universität Konstanz

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper I look at light verb, serial verb and auxiliary constructions crosslinguistically and try to set up criteria to distinguish these constructions. I argue that while a coherent set of properties can be found to distinguish light verbs from auxiliaries, it is more difficult to find crosslinguistic criteria which set serial verbs apart from light verbs and auxiliaries. This is because the class of serial verbs is not coherent, i.e. it is not clear which constructions should be considered serial verbs. Nevertheless, when looking at a specific language in detail, it can be established whether a construction may be considered a serial verb.

As a case study, I look at posture verbs in Ngan'gityemerri, a Northern Australian language. In Ngan'gityemerri, posture verbs can be used as simple verbs, in verb + coverb complexes and as clitics which attach to verb + coverb complexes. I show that while these constructions seem to be very similar at first glance, they behave differently when looked at in more detail. Thus, I argue that verb + coverb complexes are complex predicates while the encliticized posture verb should be best analyzed as an auxiliary.


## 1 Introduction

The study of complex predictes has received a lot of attention, both descriptive and theoretical. Butt (1995) defines a complex predicate in terms of Lexical-Functional Grammar (LFG) as follows:
(1) Definition of a Complex Predicate (Butt 1995)

- The argument structure is complex (two or more semantic heads contribute arguments).
- The grammatical functional structure (f-structure) is that of a simple predicate. It is flat: there is only a single predicate (a nuclear PRED) and a single subject.
- The phrase structure (c-structure) may be either simple or complex. It does not necessarily determine the status of a complex predicate.

Similar definitions can also be found in Mohanan $(1994,1997)$ or Alsina et al. (1997). Despite this clear definition, it is not easy to distinguish complex predicates from other syntactic constructions. Complex predicate constructions can be confused with coordinated or subordinated sentence constructions when the monoclausality of the constructions is not shown properly or they can be confused with monoclausal syntactic constructions like auxiliary verb constructions or serial verb constructions. In this paper I will focus on this second problem.

[^163]Light verbs in complex predicate constructions, auxiliaries and serial verbs are often also very similar semantically. This makes them very hard to distinguish. As an example, consider the sentences in (2). All sentences contain an inflected form of 'stand' in combination with either an uninflected or inflected "main" verb and it seems that 'stand' in these sentences conveys mostly aspectual information. Nevertheless, the constructions have been analyzed (or at least called) differently by different researchers. Lemmens (2005) treats the Dutch example in (2a) as an auxiliary construction while Aikhenvald (1999) calls the Tariana example in (2b) a serial verb construction and Bowern (2004) uses the Turkmen sentence in (2c) as an example for a complex predicate.
a. Auxiliary Construction

Ik stond te wachten.
I stood to wait-INF
'I was (standing and) waiting.' (Dutch, Lemmens, 2005, 184))
b. Serial Verb
tuiri-kere na-hwa nema.
bird-island 3PL-stay 3PL-stand
'They stayed at the Bird Island for a long time.' (Tariana, Aikhenvald, 1999, 480)
c. Complex Predicate

Ali kitabi okuyup turdu.
Ali book-ACC read-GER 'stand'-PST
'Ali kept on reading a book.
(Turkmen, Bowern, 2004, 253)
I do not want to claim here that all these constructions are the same, but I argue that a careful, detailed study to decide on the status of these multi-verbal constructions is needed. Although it it clear that there is not a clear-cut difference between the constructions, I claim that differences still exist, and that criteria can be established to decide on the status of these constructions.

Distinguishing between auxiliary, light verb or serial verb constructions goes beyond the merely terminological. A unified terminology enables linguists to compare constructions crosslinguistically and to test analyses proposed for a construction in one language against the same construction in other languages. For example, Baker's (1989) analysis of serial verbs is often criticized for only accounting for serial verbs which share their objects. Shared objecthood, however, is a defining feature for serial verbs as understood by Baker (1989). His analysis thus cannot be evaluated against serial verbs which do not share their objects.

To avoid such problems, I try to set up crosslinguistically valid criteria to distinguish between auxiliary, light verb and serial verb constructions. I briefly review the state of the art on these constructions in section 2 and propose some criteria to
distinguish the constructions. In section 3, I look at two different constructions in Ngan'gityemerri (Reid, 1990, 2000, 2002, 2003; Reid and McTaggart, 2008) as a case study. Section 4 concludes the discussion.

## 2 Establishing Crosslinguistic Criteria

### 2.1 The Problem of Serial Verbs

Serial verb constructions are an important topic in research on West African, Oceanic and Asian languages. Work on serial verbs includes among others Stewart (1963); Foley and Olson (1985); Sebba (1987); Baker (1989); Joseph and Zwicky (1990); Osam (1994); Bodomo (1996, 1997); Andrews (1997); Aikhenvald (1999, 2006); Stewart (2001); Foley (2009); Jarkey (2009); Appah (2009). In spite of this substantial body of research, still no agreed upon set of defining features of serial verbs has been established. Thus, serial verbs do not seem to be a coherent syntactic class, or, as Crowley (2002) put it, "many authors are not fully explicit about what they mean by serial verbs, with some writers simply treating any verb-verb sequence as serial verbs as long as the second verb is not obviously marked as an infinitive". For a discussion of this problem see for example also Sebba (1987) and Lord (1993).

Most importantly, researchers differ in their views on object sharing, switchsubject constructions and shared tense, aspect and polarity features, of which the issue of object sharing is the most controversial. While some researchers, e.g. Stewart (1963), Baker (1989) or Stewart (2001), require objects to be shared, others do not require objects to be shared, e.g. Crowley (2002); Aikhenvald (1999, 2006). More precisely, most researchers agree on treating sentences like (3a) as serial verbs, because èvbàré 'food', is the object of both verbs. Sentence (3b), on the other hand, is a combination of an intransitive and a transitive verb. Thus, the object cannot be shared and some researchers (e.g. Stewart 2001) would not treat this construction as an instance of serial verbs.
a. Òzó dẹ èvbàré rhié nè Ìfuẹkò

Ozo buy food give to Ifueko
'Ozo bought the food and gave it to Ifueko.' (Ẹdó, Stewart, 2001)
b. Úyì hìa lé èvbàré

Uyi try cook food
'Uyi managed to cook food.'
(Èdó, Stewart, 2001)
c. Àbié! !yúwà hión èrhán kpàán àlìmó

Abieyuwa climb tree pluck orange
'Abieyuwa climbed the tree and plucked an orange. (Ẹdó, Stewart, 2001)
(3c) is often called 'covert coordination' and is still more controversial than the examples above. Researchers who consider object sharing a defining feature clearly reject this construction as serial verbs. However, others claim that to decide whether (3c) is a serial verb construction, semantic and pragmatic features have to be taken into account. Thus, a serial verb can only be used to denote an accepted, although maybe complex, event in a culture. For example in Alamblak, an action which involves climbing a tree in order to look for insects is a reasonable event, but an action which involves climbing a tree in order to look at the moon is not (Bruce 1988, see also Durie 1997). This meaning cannot be expressed by a serial verb and (4b) is thus ungrammatical.
a. miyt ritm muh-hambray-an-m tree insects climb-search.for- $1 \mathrm{Sg}-3 \mathrm{Pl}$ 'I climbed the tree looking for insects.' (Alamblak, Bruce 1988, 29)
b. *miyt guñm muh-hëti-an-m
tree stars climb-see-1Sg-3Pl
'I climbed the tree and saw the stars.' (Alamblak, Bruce 1988, 29)
As a result of these differences, different subgroupings have been proposed by different researchers. An early distinction along with a theoretical analysis was proposed by Foley and Olson (1985) who distinguish between nuclear and core layer serialization (see also Crowley 2002), i.e. they distinguish in principle between V and VP serialization. A distinction between covert coordination and serialization of verbs which form a complex event was proposed by Osam (1994), who calls these 'clause chaining' and 'integrated serial verbs' respectively. This distinction corresponds to what other researchers have called 'linking type' and 'modifying type' (e.g. Bamgbroṣe, 1974).

Aikhenvald $(1999,2006)$ looks at the problem from a different angle and distinguishes serial verb constructions according to verb classes. In a symmetrical serial verb construction both verbs come from an open verb class while in an asymmetrical serial verb construction one of the verbs comes from a restricted verb class, e.g. from motion or posture verbs. Finally, Andrews and Manning (1999) propose formal analyses for very different serial verbs in Tariana and Misumalpan and discuss different understandings of serial verbs by different researchers.

Although researchers do not agree upon these differences, some properties are shared among all of them. Thus, Bowern (2008) lists the following concepts as properties of serial verbs in general:

## (5) <br> Properties of Serial Verbs (Bowern 2008)

- the clause contains two (or more) verbs under a single intonation contour
- the verbs must be full lexical verbs which can head simple predicates in their own right
- the verbs share at least one argument
- the verbs behave as a single unit for tense, aspect, and polarity marking

While this set may be the minimal similarities of the constructions called serial verbs in the literature, it is impossible to find a proper analysis which accounts for all constructions which may fall under this definition. In the same way, these properties make it hard to distinguish serial verbs from auxiliaries and light verbs. As a consequence, serial verbs cannot be compared as a whole class to complex predicates or auxiliaries (see also Beermann and Hellan 2002). Careful languagespecific studies are needed to decide whether certain kinds of serial verbs may be auxiliaries or complex predicates, for example serial verbs which do not share their object, like causative or aspectual serial verbs, may be complex predicates or auxiliaries.

Other serial verb constructions may be distinguished from complex predicates and auxiliary constructions, for example symmetrical serial verbs in which both verbs carry their full semantic content, i.e. when they are not "light" verbs. Additionally, morphological marking for tense, person etc. can be on just one, on more or on all verbs in a serial verb construction. On the other hand, morphological marking in complex predicates is usually just on the light verb. Finally, there seems to be a difference in the semantics of many kinds of serial verbs and complex predicates. Thus, verbs in serial verb constructions denote single events which constitute a complex event together while light verbs provide more information about the event of the main verb (Butt, 1995) and auxiliaries mainly provide information about tense, aspect and mood.

To sum up, as constructions called serial verbs vary in details such as object sharing etc., they cannot be compared as a whole syntactic class to auxiliaries or light verbs. Common properties of serial verbs as proposed by Bowern (2008) or Aikhenvald (2006) are useful for a typology of serial verbs. To decide whether a given serial verb in a specific language may be a light verb or auxiliary, a detailed study of this serial verb construction is needed. In the following, I discuss some properties of auxiliaries and light verbs which may help to decide if a serial verb may be analyzed as auxiliary or light verb.

### 2.2 Auxiliaries and their historical development

Motion and posture verbs are common sources for auxiliaries, for example the English going-to-future or the Catalan go-past (Juge, 2006). When looking at the historical development of auxiliaries, one usually finds a consensus that auxiliaries may develop from main verbs when they acquire functional properties. There also seems to be a consensus that serial verbs can be an intermediate stage on the grammaticalization cline for auxiliaries (Anderson, 2006; Heine, 1993; Lord, 1993; Delancey, 1991). However, researchers do not agree on whether light verbs are an intermediate stage between main verbs and auxiliaries. Roberts and Roussou (2003)
discuss the development of English modal auxiliaries and state that there is some evidence for assuming that the pre-modal verbs, i.e. the verbs which developed into modals, were light verbs. However, they do not discuss this in detail and do not take a definite view on the matter. Similarly, Hopper and Traugott (1993) follow Hook $(1974,1991)$ in his proposal that light verbs in Hindi and other Indo-Aryan languages are an intermediate stage between main verbs and auxiliaries. However, Hopper and Traugott (2003) revise this view and state that it is not clear that auxiliaries developed from light verbs. Thus, they follow Butt's view on light verbs. Butt and Lahiri (2002) and Butt and Geuder (2003) claim that light verbs do not develop into auxiliaries but are a dead end in the development of verb forms. They show that light verbs in Urdu have been used similarly for thousands of years. Bowern (2008) agrees with the view that light verbs are not a necessary step for the development from main verbs to auxiliaries but leaves it open if light verbs can develop into other verbal forms or inflections.

In this debate it becomes apparent that a difference in the application of the terms 'complex predicate' and 'light verb' by different researchers is the, or at least one, reason for their differing views. While for example Butt and Lahiri (2002) have a very clear, narrow definition of light verbs and complex predicates, Anderson (2006) includes various syntactic constructions, such as serial verb constructions, verb plus clausal complement sequences, clause-chained or conjunctive sequences, under the label 'complex predicates'. No evidence to my knowledge has been presented in the literature so far that Butt and Lahiri's (2002) kind of light verb developed into an auxiliary.

Independently of whether light verbs are an intermediate step in the development of auxiliaries, drawing a line between auxiliaries and other verb forms is complicated by the diachronic perspective. In general, we find two major terminological traditions: some researchers (e.g. Kuteva, 2001; Lemmens, 2005; Anderson, 2006) do not make a distinction as to how far a verb has been reanalysed as an aspect marker. As soon as a verb is used in this way, it is called an auxiliary. Others (e.g. Heine, 1993) acknowledge that there is a transition period where the distinction is not clear but for the constructions at the starting and end point of the historic development one can find distinguishing features. For example, in Heine's (1993) view, an auxiliary has reached its 'developmental end-point' when the auxiliary can be used with its corresponding main verb, in sentences like He is going to go to the cinema.

Defining auxiliaries is further complicated by the fact that auxiliaries look very different in different languages. Thus, while most researchers agree that auxiliaries in some way position the event of the main verb in context to the speech or reference time, i.e. they convey information about tense and aspect, other properties of auxiliaries differ from language to language. Thus, in some languages auxiliaries carry all morphological information relating to a predicate such as person, number, tense/aspect/modality, negation marking etc., while in other languages auxiliaries show a reduced verbal behavior.

Connected to this question is the problem whether auxiliaries can combine with inflected main verbs or if they have to carry all inflections themselves. ${ }^{1}$ One example for a combination of an auxiliary with an inflected main verb comes from Urdu ((6)). Butt and Lahiri (2002) show that in Urdu, 'be' can be used as an auxiliary marking past tense in combination with main verbs which themselves can be marked in different ways.
a. nadya=ko
mıl-e $\quad t^{h}-e$
Nadya.F=Dat letter.M.Nom receive-Perf.M.Pl be.Past-M.Pl
'Nadya had received letters.'
(Urdu, Butt and Lahiri, 2002)
$\begin{array}{llll}\text { b. nadya=ko } & \text { xat } & \text { mıl-t-e } & t^{h}-e \\ \text { Nadya.F=Dat } & \text { letter.M.Nom } & \text { receive-Impf-M.Pl } & \text { be.Past-M.Pl }\end{array}$
'Nadya used to receive letters.'
(Urdu, Butt and Lahiri, 2002)

One question on which researchers also do not agree is whether the auxiliary may still carry some of its original semantic meaning. Heine (1993), however, points out that this is not a very reliable criterion as even with accepted auxiliaries such as in the English going-to-future, it is not always clear whether is going to as used in (7b) is a grammatical or verbal element.
(7) a. He is going to town.
b. He is going to work.
c. He is going to come.

That an auxiliary still carries some of its original meaning in certain contexts is especially common of auxiliaries which developed from posture or motion verbs. For example, Lemmens (2005) looks at aspectual posture verb constructions in Dutch which are used to convey progressive, durative or habitual meaning. Examples of such constructions are given in (8).
(8) Ik zat te lezen / ik stond te wachten / ik lag te slapen. I sat to read-INF / I stood to wait-INF / I lay to sleep-INF 'I was (sitting and) reading /(standing and) waiting / (lying and) sleeping.' (Lemmens, 2005, 184)

In the examples in (8) it can be argued that the meaning of the posture verbs is still important as the meaning of the main predicate fits to their meaning. However, these constructions can also be used when the agent's posture is not an issue, or when the posture denoted by the auxiliary does not correspond to the posture of the main verb, for example as illustrated in (9).
(9) Wat zit ik hier toch rond te lopen? (pers. attestation) what sit I here (toch) around to walk?
'Why on earth am I walking (around) here?'
(Lemmens, 2005, 185)

[^164]Similar examples are also discussed in Kuteva (2001). It would be very strange to call the posture verb in (9) auxiliary but exclude the posture verbs in (8) from being an auxiliary in the Dutch verbal system. Thus, in my view an auxiliary can also carry some of the original semantics of the verb it developed from.

Summing up, in my view auxiliaries developed from main verbs and can mark tense, aspect or modality. They may also carry some of their original semantic meaning and may combine with inflected main verbs. More properties can be set up to distinguish auxiliaries from light verbs, which I will discuss in the next section.

### 2.3 Light Verbs vs. Auxiliaries

Butt (2009) states that tests to distinguish light verbs from main verbs or auxiliaries differ from language to language. However, there are also some properties which set light verbs apart from auxiliaries crosslinguistically. Butt and Lahiri (2002) name some more properties to distinguish light verbs from auxiliaries.

Properties of light verbs (Butt, 2009; Butt and Lahiri, 2002)

- light verbs are always form identical to the corresponding main verb whereas auxiliaries are usually just form identical at the initial stage of reanalysis from verb to auxiliary.
- light verbs always span the entire verbal paradigm (are not restricted to appear with just one tense or aspect form).
- light verbs do not display a defective paradigm.
- light verbs exhibit subtle lexical semantic differences in terms of combinatorial possibilities with main verbs, are thus restricted in their combinations. Auxiliaries, on the other hand, are not restricted in their combinatorical possibilities, but do not have to combine with every main verb.

When looking at complex predicates crosslinguistically, further properties of light verbs can be observed which set them apart from auxiliaries, although sometimes a very careful look is needed to distinguish the two constructions. For example, light verbs contribute semantic information about the type of event. This can sometimes include Aktionsart information, which can be confused with aspect, especially if the light verb is encoding telicity/completeness as in (11)

$$
\begin{array}{llll}
\text { nadya=ne } & \text { xat } & l k^{\mathrm{h}} & l i-y a .  \tag{11}\\
\text { Nadya.F.Sg=Erg } & \text { letter.M.Nom } & \text { write } & \text { take-Perf.M.Sg } \\
\text { 'Nadya wrote a letter (completely).' }
\end{array}
$$

(Urdu, Butt, 1995)

However, other differences also exist. Thus, light verbs can change the valency of a construction, for example in causative constructions as in (12). The light verb
faire 'make', adds an argument, the causer, to the construction. Auxiliaries are not able to add or reduce arguments. Passive auxiliaries, which may be considered as reducing the arguments at first glance, seem to be very different from light verbs when looked at in more detail. For example, passives do not change the basic argument structure, just its syntactic realization, and the agent can still be expressed as a adjunct.
(12) Jean a fait partir Marie.

Jean has made go Marie
'Jean made Marie go.'
(French, Rosen, 1990, 37)
Another property in which light verbs and auxiliaries differ is the ability to assign case. Light verbs may determine case assignment, e.g. in (13), the case of the subject depends on the choice for the light verb. Auxiliaries, in contrast, are usually not considered to be able to assign case, but may be sensitive to categories such as unaccusative vs. unergative.
a. ilaa-ko $k^{h} a a n a a$ pasand huaa

Ila-D food-N like happen-PF
'Ila liked the food.
(Hindi, Mohanan, 1997, 437)
b. ilaa-ne $k^{h} a a n a a$ pasand kiyaa

Ila-E food-N like do-PF
'Ila liked the food.
(Hindi, Mohanan, 1997, 437)

Finally, light verbs may determine theta-role assignment while auxiliaries cannot. In (14), an example from Bardi, the light verbs $m a$ 'put' or $g a$ 'carry' result in a different theta-role-assignment when combined with the coverb abarrabarr. In (14a), there is only one theta-role, a theme. In contrast, in (14b), two theta-roles are assigned, an agent and a patient.
a. abarrabarr-ma- 'to be careless'
b. abarrabarr-ga- 'to lead someone astray' (Bardi, Bowern, 2004)

To sum up, light verbs and auxiliaries may differ in their combinatorical behavior, their paradigm, their ability to change the valency of a main verb and their ability to assign case or theta roles. Both may develop from main verbs, but while auxiliaries may develop further into clitics and morphological markers, light verbs seem to be a dead end. As serial verbs are a very diverse syntactic class, no claim can be made that all serial verbs are light verbs or auxiliaries on the one hand, on the other hand it cannot be claimed that no serial verb is a light verb or auxiliary either. This has to be investigated for each serial verb construction in a language in detail. In the following section, I look at two verbal constructions in Ngan'gityemerri and show that although they look very similar at first glance, one of them behaves like a light verb while the other is best analyzed as an auxiliary.

## 3 Case Study: Ngan'gityemerri Verbal System

### 3.1 A short overview of the verbal system

Ngan'gityemerri is a non-Pama-Nyungan, polysynthetic language of the Daly River region of Northern Australia (Reid, 2003). It has 31 inflecting verbs ${ }^{2}$ with a very 'generic' meaning. In most cases, these inflecting verbs combine with a socalled coverb, an uninflecting element used to denote more specific verbal meanings. Of the 31 inflecting verbs, only twelve can be used as simple verbs in constructions without coverbs. An example of a simple verb construction is given in (15a). Of these simple verbs, seven are intransitive posture verbs like 'sit', 'stand', 'lie', 'go' etc. (Reid, 2000). These verbs will be the focus of this case study. Apart from the seven intransitive verbs, five transitive verbs can also function as simple verbs. The remaining transitive and reflexive detransitive inflecting verbs can only be used in combination with a coverb as in (15b). For more information on the verbal constructions in Ngan'gityemerri in general I refer the reader to Reid (1990, 2000, 2002, 2003); Reid and McTaggart (2008).

Apart from simple and complex verb constructions, Ngan'gityemerri has also developed a construction which Reid (2002) calls serialized posture verb construction. In this construction, posture verbs cliticize onto a light verb + coverb complex, adding aspectual information. An example of this construction is given in (15c).

## a. Inflecting Verb:

Ngirim.
1SG.S.sitPR
'I'm sitting.'
(Reid, 2002, 241)
b. Inflecting Verb + Coverb:

Ngirim-tyerrakul.
1GS.sit.PR-talk
'I'm talking.'
(Reid, 2002, 243)
$\begin{array}{ll}\text { c. } \text { Inflecting Verb + Coverb + Encliticized Inflecting Verb: } \\ \text { Nganni-batybity-tye-nginni. } \\ \text { 1PLEXS.poke.PI-sew.PAST-1PLEX.sit.PI. } \\ \text { 'We were sewing.' } & \\ \text { (Reid, 2002, 256) }\end{array}$
It has been argued successfully by different researchers (Schultze-Berndt, 2000; Wilson, 1999; Bowern, 2004) that constructions involving inflecting verbs plus coverbs like (15b) involve complex predicates, i.e. that the inflecting verb is a light verb in this case. In the next subsection I show that this is also true for inflecting verbs in Ngan'gityemerri. However, I argue that constructions like (15c) should

[^165]not be considered serial verb constructions or complex predicates. Instead, the encliticized verb should be analyzed as an auxiliary. Before going into the properties of this construction in detail, I provide some examples.

In addition to the past progressive example in (15c), these constructions can also have present or future tense where the person, number and tense markers of the encliticized verb always have to correspond to the marking on the light verb:
a. Dangim-batybity-dim.

3SG.S.poke.PR-sew-3SG.S.sit.PR
'She is sewing'
(Reid, 2002, 256)
b. Warri-batybity-pe-wirri.

3PL.S.poke.IR-sew-FUT-3PL.S.sit.IR
'They will be sewing'
(Reid, 2002, 256)
Reid claims that these constructions are used to "distinguish between present (ongoing) and habitual or between future perfective and future imperfective" (Reid, 2002).
'sit' seems to be the verb which most often cliticizes onto a verb+coverb complex and thus seems to be the most neutral one. However, other posture and motion verbs can also function in this way, e.g. 'go' is also possible as a clitic and Reid (2002) claims that 'go' is used to denote motion (17a), habitual activity (17b) or common knowledge facts (17c):
a. Werrimim-ne-tyerr-baty-wannim.

3PL.s.hands.PR.-3SG.G-mouth-hold-3.PL.s.go.PR
'They are leading him along.'
(Reid, 2002, 258)
b. Madewetimbi wa-mumu-nimbi resyin
long ago male-taboo-SRCE rations
wurrmu-wawu-tye-waddi.
3PL.S.snatch.PI-pick up-PAST-3PL.S.go.PI
'In the old days they used to collect rations from the policemen.'
(Reid, 2002, 258)
c. Detyeri-werri yenim dem-wurity-yenim
ear-ASSOC 3SG.s.gopR 3SG.s.hands.PR-fix-3SG.S.go.PR
mudiga.
car
'He knows how to fix cars’
(Reid, 2002, 258)
In the examples considered above, the encliticized verb has been bleached of its semantic content. However, there are also some, rare examples where the posture and motion verbs still carry some of their meaning:


I will argue in the next subsection that all these constructions involving encliticized motion or posture verbs are best analyzed as auxiliary constructions, no matter how semantically bleached the clitics are.

### 3.2 Discussion of the data

In this section I discuss the two different verbal complexes in Ngan'gityemerri and show that the inflecting verb + coverb construction, as exemplified in (15b), is a complex predicate, i.e. the inflecting verb is a light verb. On the other hand, the encliticized posture verb, as e.g. in (15c), according to the before established criteria, is best considered an auxiliary.

First, neither of the two constructions should be considered a serial verb. If we assume object sharing as a defining feature of serial verbs, the inflecting verb in $(15 b)$ is intransitive and thus does not share the object with the coverb. But even without this property, the inflecting verb + coverb construction is not a serial verb construction. The coverb always has to combine with an inflecting verb, thus it cannot function as a verb on its own. The inflecting verb also acts more as a classifier and does not contribute a whole "subevent" as is usual for serial verbs.

Again, if we assume shared objects as defining feature of serial verbs, the encliticized posture verb which attaches to the inflecting verb + coverb complex as in (15c) cannot be considered a serial verb because the clitic is intransitive. Additionally, it does not contribute to the event semantics, but merely acts as aspect marker. Thus, in most accounts this construction would not be considered a serial verb. It may, eventually, be included in Aikhenvald's (2006) typology of serial verbs as an asymmetrical serial verb because the clitic comes from the restricted verb class of motion and posture verbs. However, as will be shown below, it shares some features with auxiliaries. Thus, in my view it should be considered an auxiliary and not a serial verb.

### 3.2.1 Inflecting verb + coverb as light verb

The verb in the inflecting verb + coverb construction should be considered a light verb because the verb does not have a defective verbal paradigm and is also always
form identical to the corresponding full verb, thus fulfilling two criteria Butt (2003) established.

Additionally, the verb displays subtle lexical semantic differences in terms of combinatorial possibilities with coverbs, i.e. it acts as classifier (McGregor, 2002) for these coverbs. It is usually expected for auxiliaries to be able to combine with every main verb. This is not the case for the inflecting verbs which can only combine with certain coverbs and leave arbitrary gaps.

Further evidence comes from valency alternations. While auxiliaries cannot change the valency of the verb they combine with, light verbs can. In Ngan'gityemerri, the inflecting verb determines the valency of the expression jointly with the coverb, e.g. the inflecting verb can reduce the number of arguments the coverb would need. In example (19), the coverb tum 'bury' would normally take two arguments, but through combining it with the intransitive inflecting verb 'sit', the whole verbal complex becomes intransitive.

## (19) ngirim-tum.

1SG.S.sit-bury
'I'm sinking.'
(Reid, 2000, 347)
On the other hand, in (20), the coverb $d u$, 'sleep', only needs one argument. The inflecting verb dum, 'move', adds an argument which results in a causative reading.

Ngirrngirr ngu-dum-birrki-du.
Sleep 1SG.A.move-3.DU.O-sleep
'I put them to sleep.'
(Reid, 2000, 344)
Finally, the inflecting verb may carry the "main semantic information" of a sentence if it forms part of an inflecting verb + coverb complex, but not if it is used as clitic. Thus, in (21) the inflecting verb 'sit' not only contributes aspectual information, but is semantically the "main predicate" of the clause.
(21) winni-pappup-tye.
3.PL.S.sit.PI-climb-PAST
'They were sitting up on top (having climbed up).'
(Reid, 2002, 252)
In sum, this evidence shows that inflecting verb + coverb complexes should be best analyzed as complex predicates. In the remainder of this section I discuss why encliticized inflecting verbs should be considered auxiliaries.

### 3.2.2 Encliticized posture verb as auxiliary

Although the inflecting verb used as clitic still displays its full verbal paradigm, there are differences in the behavior of the inflecting verb used in the different constructions. In a complex predicate construction, the inflecting verb can combine
only with a restricted number of coverbs. On the other hand, when used as a clitic, the inflecting verb can attach to almost every inflecting verb + coverb unit. More precisely, the posture verb can cliticize onto every coverb + inflecting verb complex unless the inflecting verb is a posture verb already.

In contrast to the light verb in complex predicate constructions, the encliticized posture verb never changes the valency of the coverb + inflecting verb combination and the semantic roles of the arguments are determined by the inflecting verb or the coverb, but never by the clitic.

Additionally, the encliticized posture verb mainly provides information about tense and aspect, at most some information about the posture of the subject as could be seen in the examples in (18), repeated here in (22).

## (22) 'Yawul karrityinmade ngebem-wurity-ngirim/ngibem/ngirribem <br> spear bent 1.SG.S.bash.PR-fix-1.SG.S.sit/lie/stand.PR <br> tyatma. <br> straight

I'm sitting/lying/standing straightening this bent spear.' (Reid, 2002, 258)
This, as has been discussed above, should not be used to exclude an analysis as auxiliary. Especially because, just as with the Dutch examples above, some examples exist in which the encliticized posture verb does not correspond to the information encoded in the main verb or the context, e.g. (23) was uttered by someone standing upright, not sitting.

## (23) Nginem-purrngpurrng-nyine-ngirim!

1.SG.SheatPR-boil-FOC-1SG.S.sit.PR
'I'm boiling it right now!'
(Reid, 2002, 258)

Thus, the encliticized posture verb behaves very different from the inflecting verb in the complex predicate construction and has actually much more in common with auxiliaries crosslinguistically. Similarly to my treatment of Ngan'gityemerri cliticized posture and motion verbs, Street (1996) treats similar constructions in Murrinh-Patha, a closely related language, as auxiliaries marking continuous or habitual aspect.

## 4 Conclusion

In this paper I looked at light verb, serial verb and auxiliary constructions crosslinguistically and tried to set up criteria to distinguish these constructions. I argued that while a coherent set of properties could be found to distinguish light verbs from auxiliaries, it is more difficult to find crosslinguistic criteria which set serial verbs apart from light verbs and auxiliaries. This is because the class of serial verbs is not coherent, i.e. it is not clear which constructions should be considered
serial verbs. Nevertheless, when looking at a specific language in detail, it can be established whether a construction may be considered a serial verb.

As a case study, I looked at posture verbs in Ngan'gityemerri. In Ngan'gityemerri, posture verbs can be used as simple verbs, in verb + coverb complexes and as clitics which attach to verb + coverb complexes. I showed that while these constructions seem to be very similar at first glance, they behave differently when looked at in more detail. Thus, I argued that verb + coverb complexes are complex predicates while the clitic should be best analyzed as an auxiliary.

## References

Aikhenvald, Alexandra. 1999. Serial Constructions and Verb Compounding: Evidence from Tariana (North Arawak). Studies in Language 23(3), 469-498.

Aikhenvald, Alexandra Y. 2006. Serial Verb Constructions in Typological Perspective. In Alexandra Y. Aikhenvald and R.M.W. Dixon (eds.), Serial Verb Constructions: A Cross-linguistic Typology, pages 1-68, Oxford: Oxford University Press.

Alsina, Alex, Bresnan, Joan and Sells, Peter (eds.). 1997. Complex Predicates. Stanford: CSLI Publications.

Anderson, Gregory. 2006. Auxiliary Verb Constructions. Oxford: Oxford University Press.

Andrews, Avery. 1997. Complex Predicates and Nuclear Serial Verbs. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG97 Conference, Stanford: CSLI Publications.

Andrews, Avery D. and Manning, Christopher D. 1999. Complex Predicates and Information Spreading in LFG. Stanford: CSLI Publications.

Appah, Clement Kwamina Insaidoo. 2009. The representation of ISVC in C and F structures of LFG: A proposal. SKASE Journal of Theoretical Linguistics 6(1), 92-117.

Baker, Mark. 1989. Object Sharing and Projection in Serial Verb Constructions. Linguistic Inquiry 20, 513-53.

## Bamgbrose, Ayọ. 1974. On Serial Verbs and Verbal Status. Journal of West African Languages 9(1), 17-48.

Beermann, Dorothee A. and Hellan, Lars. 2002. VP-Chaining in Oriya. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG02 Conference, National Technical University of Athens, pages 40-57.

Bodomo, Adams. 1996. Complex Verbal Predicates: the case of serial verbs in Dagaare and Akan. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG96 Conference, Stanford: CSLI Publications.

Bodomo, Adams. 1997. A Conceptual Mapping Theory for Serial Verbs. In Miriam Butt and Tracy Holloway King (eds.), On-line Proceedings of the LFG97 Conference, Stanford: CSLI Publications.

Bowern, Claire. 2004. Bardi Verbal Morphology in Historical Perspective. Ph. D.thesis, Harvard University.

Bowern, Claire. 2008. The Diachrony of Complex Predicates. Diachronica 25(2), 161-185.

Bruce, Les. 1988. Serialization: From Syntax to Lexicon. Studies in Language 12(1), 19-49.

Butt, Miriam. 1995. The Structure of Complex Predicates in Urdu. Stanford: CSLI Publications.

Butt, Miriam. 2003. The Light Verb Jungle. Harvard Working Papers in Linguistics 9, 1-49.

Butt, Miriam. 2009. The Light Verb Jungle: Still Hacking Away. In Brett Baker, Mengistu Amberber and Mark Harvey (eds.), Complex Predicates in Crosslinguistic Perspective, Cambridge: Cambridge University Press, in press.

Butt, Miriam and Geuder, Wilhelm. 2003. Light Verbs in Urdu and Grammaticalization. In Regine Eckardt, Klaus von Heusinger and Christoph Schwarze (eds.), Words in Time: Diachronic Semantics from Different Points of View, pages 295349, Berlin: Mouton de Gruyter.

Butt, Miriam and Lahiri, Aditi. 2002. Historical Stability vs. Historical Change, unpublished Ms., http://ling.uni-konstanz.de/pages/home/butt/.

Crowley, Terry. 2002. Serial Verbs in Oceanic. New York/Oxford: Oxford University Press.

Delancey, Scott. 1991. The Origins of Verb Serialization in Modern Tibetan. Studies in Language 15, 1-23.

Durie, Mark. 1997. Grammatical Structures in Verb Serialization. In Alex Alsina, Joan Bresnan and Peter Sells (eds.), Complex Predicates, pages 289-354, Stanford: CSLI Publications.

Foley, William. 2009. Events and Serial Verb Constructions. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), Complex Predicates: A Crosslinguistic Perspective, Cambridge: Cambridge University Press, in press.

Foley, William A. and Olson, Mike. 1985. Clausehood and Verb Serialization. In Johanna Nichols and Anthony C. Woodbury (eds.), Grammar Inside and Outside the Clause, pages 17-60, Cambridge: Cambridge University Press.

Heine, Berndt. 1993. Auxiliaries: Cognitive Forces and Grammaticalization. New York/Oxford: Oxford University Press.

Hook, Peter Edwin. 1974. The Compound Verb in Hindi. Center for South and Southeast Asian Studies: The University of Michigan.

Hook, Peter Edwin. 1991. The Emergence of Perfective Aspect in Indo-Aryan Languages. In Elizabeth Traugott and B. Heine (eds.), Approaches to Grammaticalization, pages 59-89, Amsterdam: Benjamins.

Hopper, Paul J. and Traugott, Elizabeth C. 1993. Grammaticalization. Cambridge: Cambridge University Press.

Hopper, Paul J. and Traugott, Elizabeth C. 2003. Grammaticalization. Cambridge: Cambridge University Press, second edition.

Jarkey, Nerida. 2009. Cotemporal Serial Verb Constructions in White Hmong. In Mengistu Amberber, Brett Baker and Mark Harvey (eds.), Complex Predicates: A Cross-linguistic Perspective, Cambridge: Cambridge University Press, in press.

Joseph, Brian D. and Zwicky, Arnold M. (eds.). 1990. When Verbs Collide: Papers from the 1990 Ohio State Mini-conference on Serial Verbs. Columbus: Ohio State working papers in Linguistics.

Juge, Matthew L. 2006. Morphological Factors in the Grammaticalization of the Catalan 'go' Past. Diachronica 32(2), 313-339.

Kuteva, Tania. 2001. Auxiliation: An Enquiry into the Nature of Grammaticalization. Oxford: Oxford University Press.

Lemmens, Maarten. 2005. Aspectual Posture Verb Constructions in Dutch. Journal of Germanic Linguistics 17(3), 183-217.

Lord, Carol. 1993. Historical Change in Serial Verb Constructions. Amsterdam/Philadelphia: John Benjamins.

McGregor, William B. 2002. Verb Classification in Australian Languages. Berlin: Mouton de Gruyter.

Mohanan, Tara. 1994. Argument Structure in Hindi. Stanford: CSLI Publications.
Mohanan, Tara. 1997. Multidimensionality of Representation: NV Complex Predicates in Hindi. In Alex Alsina, Joan Bresnan and Peter Sells (eds.), Complex Predicates, pages 431-471, Stanford: CSLI Publications.

Osam, Emmanuel Kweku. 1994. From Serial Verbs to Prepositions and the Road Between. Language Typology and Universals 47(1), 16-36.

Reid, Nicholas. 1990. Ngan'gityemerri: A language of the Daly River Region, Northern Territory of Australia. Ph. D.thesis, The Australian National University, Canberra.

Reid, Nicholas. 2000. Complex Verb Collocations in Ngan'gityemerri: A Nonderivational Strategy for Encoding Valency Alternations. In R.M.W. Dixon and Alexandra Y. Aikhenvald (eds.), Changing Valency: Case Studies in Transitivity, pages 333-359, Cambridge/Melbourne: Cambridge University Press.

Reid, Nicholas. 2002. Sit Right Down the Back: Serialized Posture Verbs in Ngan'gityemerri and other Northern Australian Languages. In John Newman (ed.), The Linguistics of Sitting, Standing and Lying, pages 293-268, Amsterdam/Philadelphia: John Benjamins.

Reid, Nicholas. 2003. Phrasal Verb to Synthetic Verb: Recorded Morphosyntactic Change in Ngan'gityemerri. In Nicholas Evans (ed.), The non-Pama-Nyungan Languages of Northern Australia, pages 95-123, Canberra: Pacific Linguistics.

Reid, Nicholas and McTaggart, Patricia Marrfurra (eds.). 2008. Ngan'gi Dictionary. Armidale: Australian Linguistics.

Roberts, Ian and Roussou, Anna. 2003. Syntactic Change: A Minimalist Approach to Grammaticalization. Cambridge: Cambridge University Press.

Rosen, Sara Thomas. 1990. Argument Structure and Complex Predicates. New York/London: Garland Publishing.

Schultze-Berndt, Eva. 2000. Simple and Complex Predicates in Jaminjung. Nijmegen: University of Nijmegen.

Sebba, Mark. 1987. The Syntax of Serial Verbs. Amsterdam: Benjamins.
Stewart, John. 1963. Some Restrictions on Objects in Twi. Journal of African Languages 2(2), 145-149.

Stewart, Osamuyimen. 2001. The Serial Verb Construction Parameter. New York/London: Garland.

Street, Chester S. 1996. Tense, Aspect and Mood in Murrinh-Patha. In William McGregor (ed.), Studies in Kimberley Languages in Honour of Howard Coate, pages 205-225, M"unchen/Newcastle: Lincom Europa.

Wilson, Stephen. 1999. Coverbs and Complex Predicates in Wagiman. Stanford: CSLI Publications.

# HEBREW FLOATING QUANTIFIERS 

Ilona Spector<br>The Hebrew University of Jerusalem<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>\section*{CSLI Publications}<br>http://csli-publications.stanford.edu/


#### Abstract

This paper addresses the issue of Floating Quantifier phenomena in Hebrew, focusing on the case of the universal quantifier kol ('all') and proposes a new, non-derivational analysis in the LFG framework in which the floating and the non-floating quantifier constructions are treated separately. This is based on semantic evidence showing that the quantifiers in these constructions have non-identical semantic effects. Thus, there is no need to assume a derivational relation between the two constructions.

I propose to analyze the Floating Quantifier construction in Hebrew as an instance of Topicalization, accompanied by Triggered Inversion. The incorporated pronoun on the floating quantifier is explained by the Extended Coherence Principle, when the pronoun is anaphorically bound by the topic. The pragmatic markedness of this construction naturally follows from this analysis, it being an instance of Topicalization. ${ }^{1}$


## 1. Introduction

### 1.1 Floating Qunatifier phenomenon

The phenomenon of Floating Quantifiers (henceforth FQ) is demonstrated in (1) for French:
(1) a. Tous les enfants ont $v u$ ce film. all the children have seen this movie 'All the children have seen this movie.'
b. Les enfants ont tous vu ce film. the children have all seen this movie 'The children have all seen this movie.'
c. Les enfants verront tous ce film. the children see.fut all this movie 'The children will all see the movie.' (Sportiche 1988:426-7)

What is particularly interesting in these constructions is the relation between the quantifier tous and the DP it modifies, les enfants in (1b,c), where it seems that the quantifier has floated rightwards from its DP. Similar constructions exist in Hebrew as well:
a. kol ha-yeladim halxu la-yam
all the-children.mASC.PL went to-the-sea
'All the children went to the sea

[^166]b. ha-yeladim kul-am halxu la-yam the-children.MASC.PL all.3.MASC.PL went to-the-sea 'The children went all to the sea.'
c. ha-yeladim halxu kul-am la-yam the-children.MASC.PL went all.3.MASC.PL to-the-sea 'The children went all to the sea.'

This paper focuses on the universal quantifier kol ('all'). The quantifier may appear adjacent to the NP, forming a $\mathrm{QP}^{2}((2 \mathrm{a}))$, thus it will be referred to as NP-adjacent Q. ${ }^{3}$ Alternatively, it can appear in FQ constructions, as in ( $2 \mathrm{~b}, \mathrm{c}$ ). However, the difference between French and Hebrew is that in Hebrew the quantifier must appear in its inflected form with an incorporated pronoun, ${ }^{4}$ thus agreeing with the subject in number and gender (and is phonologically realized as kul and not kol). Moreover, the FQ construction is discourse marked in Hebrew.

In the sections below I will provide an account for the FQ phenomena in Hebrew (which can probably be extended to related Semitic languages as well). Some particular facts require an explanation: first, the incorporated pronoun on the FQ, which agrees with the antecedent NP; second, the nature of the relation between the NP-adjacent Q and the FQ constructions. Moreover, an analysis of FQ phenomenon in Hebrew must also take into account the pragmatic markedness of this construction and the interaction of these pragmatic factors with the syntax of the construction. And finally, the important question to ask is what accounts for the quantifier 'float', or alternatively, why does the quantifier surface in a different position.

## 2. Previous Accounts

### 2.1. Adverbial analyses

Adverbial accounts treat FQs as adverbs. As was originally noted by Kayne (1975), and mentioned by Pollock (1989), Baltin (1995) and Hurst (2007), FQs occupy positions in which adverbs canonically surface, namely

[^167]to the left of V and to the right of verbal elements, such as auxiliaries and modals.
(3) Les soldats ont \{tous les deux\} été \{tous les deux\} présentés \{tous the soldiers have \{all the two \} been\{all the two\} introduced \{all les deux\} à Anne par ce garcon. the two \} to Anne by this boy 'Both soldiers were introduced to Anne by this boy.' (Kayne 1975)

This holds for both English and French. Moreover, the possibilities for the position of adverbs in these languages correspond to the possible positions of placing FQs. While English allows an adverb or an FQ to immediately follow the subject, French does not:
(4) a. My friends all/probably will leave.
b. *Les enfants tous/bientôt vont partir. the children all/soon will leave 'The children will all/soon leave'
(Pollock 1989)

Moreover, it was observed by Sag (1978) that FQs pattern with adverbs, and not with negation, in the case of VP-ellipsis:
(5) a. Otto has read this book, and my brothers have (all/certainly) read it, too.
b. Otto has read this book, and my brothers have (*all/*certainly) $\qquad$ , too.
c. Otto has read this book, but my brothers have (n't/not/*all)
$\qquad$ _.

However, Bobaljik (2003) points that it was noted by Kayne (1981) and Beletti (1982) that the dependence between an FQ and an NP obeys in essence the same locality constraints (in terms of c-command) as those holding between an anaphor and its antecedent. Thus, the DP must ccommand the FQ in (6), and no finite clause boundary or specified subject may intervene between them, as shown in (7):
(6) a. *[The mother of my friends $\left.{ }_{i}\right]$ has all $_{i}$ left.
b. *La mère de mes amis est tousi $_{i}$ partie. the mother of my friends is all left
intended: ‘The mother(s) of all my friends left.’ (Kayne 1981)
a. $\quad$ *My friendsi think that I have alli left.
b. *Mes amis pensent que je suis tous i parti. $^{\text {b }}$ my friends think that I am all left
intended: ‘My friends all think that I have left.’ (Kayne 1981)

Thus, FQs were treated as anaphoric adverbs, related to their hosts via binding. Baltin (1995), on the other hand, argues that FQs are preverbs, a class of adverbs adjoined to the left edge of a predicate.

When trying to formalize these analyses for Hebrew, one encounters a problem. Hebrew exhibits an incorporated pronoun on the floated Q, which is not a property of adverbs. This type of pronoun is usually attached to NPs and to Qs. Thus, we would not want to claim that there exists a special category of adverbs in Hebrew which can host incorporated pronouns. While it is true also for Hebrew that there is a locality constraint between the antecedent NP and the FQ, I believe this can be explained by other means, as we shall further see, taking into account the presence of an incorporated pronoun on the FQ and the pragmatic markedness of the FQ construction.

### 2.2. Derivational analyses

The most influential account of FQs in the literature is Sportiche (1988) for French. This analysis was eventually taken as an argument for the Subject Internal VP-Hypothesis, based on the distribution and the proposed structural position of FQs in French. Several known properties of FQs served as a background for Sportiche's analysis. First, it has been assumed that FQs and NP-adjacent Qs modify their related NP in the same way, e.g. tous in (1a) and (1b,c) universally quantifies over the set denoted by the NP. Second, in some languages (e.g. Romance), the quantifier and its NP agree in number and gender, pointing out determiner-like properties of the quantifier. Third, FQs tend to appear on the left periphery of VP. And finally, there is an anaphoric locality condition on FQs and their NP antecedent (Kayne 1981, Beletti 1982).

All the aforementioned facts led Sportiche to assume that the quantification in (1a) and ( $1 \mathrm{~b}, \mathrm{c}$ ) is identical, that is to say that the floating and the non-floating quantifiers are of the same logical type. Thus, there is a derivational relation between them, i.e. they have the same underlying syntactic structure. FQ forms a constituent with the NP at D-structure and the phenomenon of Q -float is actually the stranding of the Q in situ, in a position adjacent to the trace of the NP (cf. (8) for (1a,b)):


## Sportiche (1988)

The innovation of this analysis is the claim that the quantifier itself originates VP internally, even in the NP-adjacent NP cases, where on the surface the quantifier precedes its NP (cf.(1a)). That is to say that the original position of the quantifier is the 'floated' one. The quantifier then stays in situ, while its NP-associate is the one that moves to SPEC IP to get case, leaving a trace to the right of Q . Thus, even when the quantifier is stranded from its NP, the antecedent-anaphor relations still hold. This analysis captures the observation that even the floated Q is able to modify its NP and in some languages to agree with it, since at D-structure [Q NP] form a single constituent (Bobaljik 2003).

Shlonsky (1991) extends this analysis for Hebrew. While Hebrew quantifiers do not agree with their NPs in number and gender as in Romance languages, FQs in Hebrew host an incorporated pronoun (cf. 2b,c). Shlonsky treats this pronoun as an agreement clitic, which licenses movement of Q to SPEC QP and assumes a QP projection. The NP then moves to SPEC IP, creating a FQ configuration:
(9) $[\mathrm{NP}]_{\mathrm{i}} \ldots\left[\mathrm{QPP}[\mathrm{e}]_{\mathrm{i}} \mathrm{Q}[\mathrm{e}]_{\mathrm{i}}\right]$

There are a few basic premises of the derivational accounts which seem problematic. To begin with, Sportiche (1988), followed by Shlonsky (1991) base their analyses on the assumption that the NP-adjacent Q and the FQ are semantically identical. In the next section I will show that in fact, they differ substantially.

Secondly, based on the alleged semantic identity of these two types of quantifiers, derivational accounts also propose a syntactic derivational relationship between the floating and the non-floating quantifier
constructions. However, I will claim that there is no justification for this assumption and will propose a different account of Hebrew FQs.

Finally, while the derivational accounts rely on the constituency of [Q NP/t] in both constructions, I will show that the quantifier forms a constituent with its NP only in the non-floating version, i.e. [Q NP].

## 3. Semantic Differences ${ }^{5}$

Hebrew kol is polysemous. ${ }^{6}$ This paper deals with one interpretation of kol, namely that of English plural all. The kol we are dealing with takes a plural, definite NP or a plural incorporated pronoun as its complement.

As was already mentioned, one of the motivations for derivational accounts is the alleged semantic identity between NP-adjacent Qs and FQs (Sportiche 1988). In this section I will show that this is not so and claim that these quantifiers differ semantically on several counts:

### 3.1 Type of Predication:

NP-adjacent Q and Floating Q differ in type of predication, i.e. whether the sentence has a collective or distributive reading, depending on the position of the quantifier:
a. kol ha-yeladim herimu even all the-children picked up stone 'All the children picked up a stone.'
b. ha-yeladim herimu kul-am even the-children picked up all3.MASC.PL stone 'The children all picked up a stone.'
(10a) has both collective and distributive readings. If there is a group of six children, the sentence means either that each of the six children picked up one stone (six stones in total) on the distributive reading, or that the six children as a group picked up one stone (one stone in total) on the collective reading. (10b), on the other hand, is understood collectively. If there are six children, the most salient reading is that the six children as a group picked up one stone (one stone in total). Consider (11):
(11) a. kol ha-yeladim herimu even ve-Dani herim even. all the-children picked up stone and-Dani picked up stone 'All the children picked up a stone and Dani picked up a stone.

[^168]b. ??? ha-yeladim herimu kulam even ve-Dani the-children.MASC.PL picked up all3.MASC.PL stone and-Dani herim even.
picked up stone
'The children all picked up a stone and Dani picked up a stone.'

Assuming that there are six children and Dani is one of them, (11a) is acceptable. The fact that sentence (11b) is odd shows that the distributive reading is less appealing when FQ is used: if the six children picked one stone as a group, it is infelicitous and redundant to claim that Dani, a group member, also picked up a stone.

### 3.2 Type of quantification:

NP-adjacent Q and FQ impose different readings in terms of sets vs. members of sets and presupposition of existence.

As a universal quantifier, FQ must range over the whole set, each and every member of it. It is as if the quantifier refers to each member of the set, so even in the case of collective predication, each member is counted in the group effort. This is not the case with kol:
a. kol ha-feyot blondiniyot

All the-fairies3.FEm.PL blonde3.FEM.PL
'All the fairies are blonde.'
$\forall x(F x \rightarrow B x)$
b. ha-feyot kul-an blondiniyot
the-fairies3.FEM.PL all3.FEM.PL blonde3.FEM.PL
'The fairies are all blonde.'
$\forall x\left(x_{1 \ldots n} \text { is a fairy } \rightarrow B x\right)^{7}$

The Q in (12a) ranges over sets and reflects a relation between the set of fairies and the set of blondes. In particular, it says that the set of fairies is a subset of the set of blondes. The Q in (12b) ranges over members of sets and reflects a relation between individual fairies and the set of blondes.

Kol is also a strong quantifier (Milsark 1977). Strong quantifiers, unlike weak ones, such as numerals and kama -'several', presuppose existence of a background set. Thus, though kol as a logically universal quantifier does not entail existence, in the language it does presuppose existence.

[^169]a. kol ha-parot ha-sgulot notnot xalav all the-cows the-purple give milk 'All the purple cows lactate.'
b. ??? ha-parot ha-sgulot notnot kul-an xalav the-cows.FEM.PL the-purple.FEM.PL give all3.FEM.PL milk 'The purple cows all lactate.'

Although both Qs presuppose existence, it seems that the presupposition is stronger in the case of FQ. This explains why it quantifies over individuals, as opposed to NP-adjacent Q which may quantify over an empty set. In (13), 'purple cows' denotes an empty set. The fact that it can appear with kol in (13a), but not with FQ in (13b), supports the claim that FQ presupposes the existence of the set denoted by the predicate it quantifies over. Since there are no purple cows, there are no members for FQ to range over, thus (13b) is odd. ${ }^{8}$

### 3.3. Scope ambiguities

The interaction of NP-adjacent Q and FQ with modality and/or negation results in scope ambiguities (Dowty and Brodie 1984).
(14) a. kol ha-mitxarim yexolim lenatzeax $\quad \diamond>\forall, \forall>\diamond$ all the-contestants can win 'All the contestants can win.'
b. ha-mitxarim yexolim kul-am lenatzeax $\diamond>\forall$ the-contestants.3.mASC.PL can all.3.MASC.PL win 'The contestants can all win.'
(15) a. kol ha-mitxarim lo nitzxu $\quad \neg>\forall, \forall>\neg$ all the-contestants not won 'All the contestants did not win.'
b. ha-mitxarim kul-am lo nitzxu $\quad>\forall$ the-contestants.3.MASC.PL all.3.MASC.PL not won 'The contestants did not all win.'
(14a) and (15a) have ambiguous readings such that the quantifier may take wide or narrow scope relative to a modal or negation. In the (b) sentences the

[^170]Q takes only narrow scope, below the modal or negation. Bobaljik (2003) also notes that NP-adjacent Q may undergo scope changing operations (QR, Reconstruction), while FQ is "frozen" in situ in terms of scope.

### 3.4 Genericity:

Bobaljik (2003) notes that only FQ constructions allow a generic interpretation: ${ }^{9}$
(16) a. All lions, tigers and bears are scary.
b. Lions, tigers and bears are all scary.
(16a) asserts that every lion is scary, every tiger is scary, and every bear is scary, that is, all quantifies over [lions, tigers and bears]. (16b) allows this reading as well. However, (16b) can also assert that lions are generally scary, and tigers are generally scary, and bears are generally scary, i.e. the predicate be scary is true of all of the terms in the subject NP, but it allows for the individual plural nouns to be interpreted as generics. This generic reading is unavailable in (16a). This is also true for Hebrew:
(17) a. kol ha-arayot, ha-nemerim ve-ha-dubim mafxidim all the lions the-tigers and-the-bears scary 'All lions, tigers and bears are scary.'
b. arayot, nemerim ve-dubim kulam mafxidim
lions tigers and-bears all.3.MASC.PL scary
'Lions, tigers and bears are all scary.'
All these semantic differences lead to the conclusion that NPadjacent Q and FQ are not semantically identical quantifiers. This, of course, bears on the reading of the sentence in which they appear; constructions with NP-adjacent quantifiers thus differ semantically from constructions with floating quantifiers. These findings lead to the conclusion that it would be inaccurate to posit a derivational relation between the two, based on semantic identity. In other words, contrary to derivational accounts, I maintain that since there is no semantic identity, no syntactic identity follows.

[^171]
## 4. The Analysis

### 4.1. Constituency

Not everybody agrees that the two constructions should receive a unified account (Bobaljik 2003, Benmamoun 1999, inter alia), based on syntactic evidence such as reconstruction, Case and agreement. Another crucial piece of evidence comes from constituency tests. While one of the main ingredients of derivational accounts is the assumption that the quantifier and its modified NP form a constituent at all levels of representation, and thus both in the floating and the non-floating quantifier versions, ${ }^{10}$ application of several additional constituency tests shows that this is not the case in the FQ construction (cf. Spector 2008 for more tests):
(18) Adverb insertion
ha-tapuzim kim'at/vaday kul-am nirkevu
the-oranges almost/certainly all.3.MASC.PL got rotten
'The oranges all almost/certainly got rotten.'
(19) Preposing as a unit
*et ha-yeladim kul-am ani raiti
acc the-children all.3.mASC.PL I saw
'The children all I saw.'
(20) Sentence fragment
ha-yeladim ku-lam halxu la-yam
the-children all.3.MASC.PL went to-the-sea
A: $\quad$ mi halax la-yam? Who went to-the-sea?
B: * ha-yeladim kul-am The-children all.3.MASC.PL
(21) Relative Clause/PP modification
a.* ha-yeladim kulam še ohavim lisxot halxu la-yam the-children all.3.mASC.PL that like swim went to-the-sea 'The children all who like swimming went to the sea.'
b. * ha-yeladim kulam me-ha-gan šeli halxu le-tiyul the-children all.3.mASC.PLfrom-the-kindergarten my went to-trip 'The children all from my kindergarten went for a trip.'

[^172]a. John likes [ice cream], but not [vegetables].
b. * Dani axal et ha-tapuzim kul-am aval lo et ha-bananot Dani ate ACC the-oranges all.3.MASCPL but not ACC the-bananas rub-an
most.3.FEM.PL
Lit: Dani ate the oranges all but not the bananas most
'Dani ate all the oranges but not most the bananas.'

### 4.2. The proposal

I propose to analyze the FQ constructions in Hebrew ${ }^{11}$ ( $(2 \mathrm{~b}, \mathrm{c})$, repeated in (23)), as instances of Topicalization. When the Q appears post verbally in (23b), it involves Triggered Inversion:
[ha-yeladim-TOPIC], [kulam-SUBJ halxu la-yam]
the-children.mASC.PL all3.MASC.PL went to-the-sea
'The children went all to the sea.'

I assume that the NP ha-yeladim has an overlay discourse function TOPIC, while the real syntactic subject is $k u l+$ incorporated pronoun which agrees with the topic in number and gender. Aside from subjecthood and topichood tests which will be provided in the next section to support this analysis, FQ constructions of the above type ${ }^{12}$ exhibit a unique intonational pattern, where there is a "comma"-break after ha-yeladim, separating it from the rest of the clause.

According to the Extended Coherence Condition (ECC) in LFG, overlay functions must be linked or associated with arguments. In (23) ECC is observed by associating the topic with the incorporated pronoun on the Q , which functions as a subject. The incorporated pronoun on the quantifier is anaphorically bound by the TOPIC and the identification takes place via coindexation. This is in line with Bresnan and Mchombo (1987): "A topic is bound whenever it is functionally identified with, or anaphorically binds a bound function". This enables us to explain the obligatoriness of the pronoun on the floating Q .

[^173]The alternation in word order in (23a,b) is accounted for by Triggered Inversion (cf. next section). And finally, Topicalization makes the FQ construction discourse-marked in Hebrew, which is normally a non-topic prominent language.

### 4.3. Basic assumptions

### 4.3.1. Topichood of ha-yeladim

Various definitions of topic exist in the literature. In this section I review a few of them to motivate the proposed Topicalization analysis.

For Chafe (1976), "the topic sets a spatial, temporal or individual framework within which the main predication holds". According to Dik (1978), "the topic presents the entity 'about' which the predication predicates something in the given setting". And indeed, halxu kulam la-yam predicates about ha-yeladim, by saying that 'as for the children - they all went to the sea'. Furthermore, topic represents old or given information (Chafe 1976). Ha-yeladim in (23) is the old information, while kulam is new. The new information presented in this sentence is that it is all children and not just some that went to the sea, while the set of children is assumed to be known or has already been identified in the discourse, in line with Bresnan and Mchombo (1987), who maintain that "the topic designates what is under discussion, whether previously mentioned or assumed in discourse".

In addition, topics are usually definite and clause initial (Lambrecht 1981), and this is the case here. Notice that ha-yeladim in this construction cannot be indefinite:

* yeladim halxu kulam la-yam
children.3.MASC.PL went all.3.MASC.PL to-the-sea.
'children went all to the sea.'
Another piece of evidence is adopted from Bresnan (2001) for Chichewa. Bresnan claims that topics cannot be questioned, and subsequently, be focused. The common view is that in questions, the wh-word bears the FOCUS function. From this it follows that one may ask about the subject ((25)), but not about the topic ((26)):
a. [ha-yeladim SUBJ] halxu la-yam The-children went to-the-sea
b. [mi FOCUS] ata amarta še __ halax la-yam? who you said that went to-the-sea 'Who did you say that ___ went to the sea?'
a. [ha-yeladim TOPIC] halxu kul-am la-yam The-children went all.3.MASC.PL to-the-sea
b. * [mi $\mathbf{F O C} / \mathbf{T O P}]$ ata amarta še __ halxu kul-am la-yam? who you said that went all.3.MASC.PL to-the-sea 'Who did you say that __all went to the sea?'

The ungrammaticality of (26b) follows from the fact that something cannot at the same time be both TOPIC (old information) and FOCUS (new information): it results in function clash. This also shows that ha-yeladim is not the subject of (23). Since it refers to the same entity as kul-am, the only option left for ha-yeladim is to be a topic.

### 4.3.2. Subjecthood of kul-am

Trying to convincingly show that kul-am is a subject is not easy. All the usual subject properties discussed in Falk (2006) distinguish primarily between subjects and objects and are often applicable to both subjects and topics, since cross-linguistically, the subject is usually the discourse topic. Therefore, the argumentation and the evidence have no choice but to be negative, i.e. since we established that ha-yeladim is not a subject, but a topic, the only other NP that can be the subject is kul-am. This hypothesis is supported by the following:

The governable grammatical functions can be divided into semantically restricted and semantically unrestricted functions (Bresnan 1982). The claim that $k u$-lam functions as a subject in this construction is supported by Fillmore (1986), who argues that "semantically unrestricted functions like SUBJ and OBJ can be associated with any semantic role". And indeed, in the examples below, kul-am exhibits a wide range of semantic roles:
a. Ha-yeladim halxu kulam la-yam the-children went all.3.MASC.PL to-the-sea 'The children went all to the sea.'
b. Ha-yeladim kiblu kulam matanot the-children received all.3.mASC.PL presents 'The children all received presents'.
c. Ha-yeladim ohavim kulam et ha-mora the-children love all.3.MASC.PL ACC the-teacher Experiencer 'The children all love the teacher.'

Although semantically unrestricted functions can be either OBJ or SUBJ, it is clear that in this construction, kul-am is definitely not an OBJ, since e.g in
(27a) it is not selected by the verb, given that $g o$ is intransitive. This leaves kul-am with only one possible grammatical function, namely SUBJ.

### 4.3.3. Identification of the topic

According to the Extended Coherence Condition, overlay functions must be identified with arguments or adjuncts. In FQ constructions, the agreement features of the incorporated pronoun on the subject are identified with the same features on the topic. Thus, the topic is bound by the core function SUBJ. The proposed analysis enables us to explain the ungrammaticality of (28a,b):
a. *ha-yeladim halxu kul-an la-yam the-child.pl.masc went.3.PL. all-3.PL.FEM to-the-sea 'The children went all(fem.) to the sea.'
b. *ha-yeladim halxu kol la-yam the-child.pl.mASC went.3.PL. all to-the-sea 'The children went all(fem.) to the sea.'

The incorporated pronoun on Q in (28a) is 3.PERS.PL.FEM -an, and it needs to provide identification for the topic by the Extended Coherence Condition: the agreement features that sit on -an need to be co-referential with the same features on the topic. In (28a) they are not, thus the sentence is ungrammatical. In (28b), the FQ lacks the incorporated pronoun, thus leaving the topic unidentified with the subject.

At the same time, the topic serves as the antecedent for the anaphorically bound incorporated pronoun. Bresnan (2001) argues that "a pronominal inflection will be in complementary distribution with a headed syntactic phrase of the same function. Independent (headed) NPs that cooccur with these pronominal inflections must then have non-argument functions, like the dislocated topics. The incorporated pronoun will agree with such nominals anaphorically, in just the way a pronoun agrees with its antecedent... When dislocated topics are anaphorically linked to a pronominal element within the clause, what is identified is ...the referential index of the two functions". Thus, ha-yeladim (the topic) binds the pronoun -an on the Q (the subject). This way, the topic is identified with the pronoun, which is anaphorically bound by it, in the sense of Dalrymple (1993). ${ }^{13}$

[^174]
### 4.3.4. Triggered Inversion

Triggered Inversion in Hebrew has been discussed by Borer (1995), Shlonsky and Doron (1992), and Shlonsky (1997). In LFG it has been discussed by Falk (2004) for the following:
a. beyalduto, Eli patar targilei matematika in childhood.3.SG.MASC, Eli solved exercises mathematics 'In his childhood, Eli used to solve exercises in mathematics.'
b. beyalduto, patar Eli targilei matematika in childhood.3.SG.MASC, solved Eli exercises mathematics 'In his childhood, Eli used to solve exercises in mathematics.'
(29a) and (29b) are free variants, when the sole difference between them is the position of the verb and the subject. While (29a) has the regular SVO order, in the presence of a trigger, the order can be manifested as VSO as in (29b). According to Shlonsky (1997), "in Triggered Inversion, the verb moves... in the presence of a non subject initial elements". For Falk (2004), a trigger is "an element with discourse prominence [that] can be placed at the beginning of a Hebrew clause". Thus, a non subject initial element with discourse prominence in FQ constructions is the topic ha-yeladim.

Topicalized constructions can be manifested as either SV or VS ((29a,b)), namely, the inversion is optional. In the same fashion Floating Quantifier constructions can be either SV or VS ((23a,b)). This accounts for the preverbal and post-verbal site of the quantifier 'float', since triggered subject-verb inversion is optional.

### 4.3.5 Contrasting NP-adjacent $Q$ with FQ

I. Lexical Entry of NP-adjacent $Q$

$$
\begin{array}{ll}
\text { kol } \mathrm{Q}: & \\
& \\
& \\
& \left(\uparrow \text { OBJ NUM } \quad \operatorname { ~ } \operatorname { k o l } \left\langle(\uparrow \mathrm{OBJ}){ }_{\mathrm{c}} \mathrm{PL}\right.\right. \\
& \\
& (\uparrow \text { OBJ DEF })=\mathrm{c}+
\end{array}
$$

$C$-Structure of NP-adjacent $Q$

$F$-Structure of NP-adjacent $Q$


Despite the convention that the standard categories in LFG which take ( $\uparrow$ OBJ) are prepositions and verbs, I believe that since Q functions as the head of QP and takes a complement, this should be expressed in the fstructure as well by allowing Q to take ( $\uparrow \mathrm{OBJ})$. For supporting argumentation see Fassi Fehri (1988) for an analysis of Qs taking the complement NP as object in Arabic, and see Maling (1983) for kinds of adjectives that take OBJs.
II. Lexical entry of $F Q$

> kul[_] Q: PRED $\quad \operatorname{kul}\langle(\uparrow \mathrm{OBJ})\rangle '$
> $(\uparrow$ OBJ $\quad \mathrm{PRED})={ }^{\prime} \mathrm{PRO}^{\prime}$

The notation kul[_] stands for underspecified incorporated pronoun's features.

C-Structure of FQ:

$F$-Structure of FQ:

| TOP | $\left[\begin{array}{l} \text { PRED 'yeled' } \\ \text { DEF + } \\ \text { NUM PL } \end{array}\right]$ |
| :---: | :---: |
| SUBJ | $\left[\begin{array}{ll} \text { PRED } & \operatorname{kul}\langle(\uparrow \mathrm{OBJ})\rangle ' \\ \text { OBJ } & {\left[\begin{array}{ll} \mathrm{PRED} & \text { 'pro' } \\ \text { PRS } & 3 \\ \text { GEND } & \text { MASC } \\ \text { NUM } & \text { PL } \end{array}\right]} \end{array}\right]$ |
| PRED 'halax $\left\langle(\uparrow\right.$ SUBJ $\left.)\left(\uparrow \mathrm{OBL}_{\mathrm{DIR}}\right)\right\rangle$ ' |  |
|  | PRED 'le $\langle(\uparrow$ OBJ $)\rangle$ ' |
| OBLdir | OBJ $\left[\begin{array}{l}\text { PRED 'yam' } \\ \text { DEF } \\ \text { + } \\ \text { NUM } \\ \text { SGL }\end{array}\right]$ |

The anaphoric binding is indicated by the co-indexation of TOPs and SUBJs f -structures.

## 5. Conclusion

This paper presents a new LFG approach to the Floating Quantifier phenomenon in Hebrew by showing that there is no actual floating involved, but these are rather two different syntactic constructions. The proposed analysis accounts both for the markedness of the FQ construction and for the obligatory inflection of the quantifier in the "floated" preverbal and postverbal positions. The different position of Q-float site is explained by different positions of the verb and the subject via Triggered Inversion. When the uninflected Q appears in these positions (i.e., when the Q does not contain an incorporated pronoun whose function is to provide an anaphoric identification for the topic), the TOPIC function remains unidentified with an argument function (SUBJ), thus violating the Extended Coherence Principle, rendering these sentences ungrammatical. The markedness of FQ construction is explained by appealing to Topicalization. Topicalized constructions in a non-topic prominent language like Hebrew would always be discourse-marked.

Several directions are to be pursued in follow-up work, and which were not taken into account in the present work. First, prosodic aspects of FQ constructions at the syntax-phonology interface can shed light on the correctness of the proposed analysis. Second, other floated positions were not examined in this work, e.g. the right periphery of VP in object position ((30)) and the sentence-final constructions (cf. Hurst 2007), such as (31):
axalti et ha-tapuzim kul-am
ate.I ACC the-oranges all3.MASC.PL
'I ate the oranges all $=\mathrm{I}$ ate all the oranges'
ha-yeladim halxu la-yam kul-am
the-children.mASC.PL went to-the-sea all.3.mASC.PL
'The children went to the sea all (of them).'
Another possible direction for further work might be to examine other Topicalization constructions in Hebrew. This would provide a broader view on discourse-marking strategies in Hebrew and would locate the FQ constructions in a proper and wider context.

## 6. References

Abney, S. (1987). The English Noun Phrase in its Sentential Aspect. Doctoral Dissertation, MIT, Cambridge, MA.
Baltin, M. (1995). Floating Quantifiers, PRO and Predication. Linguistic Inquiry, 26, 199-248.
Belletti, A. (1982). On the Anaphoric Status of the Reciprocal Construction in Italian. The Linguistic Review 2: 101-138.

Benmamoun, E. (1999). The Syntax of Quantifiers and Quantifier Float. Linguistic Inquiry 30(4), 621-642.
Bobaljik, D. (2003). Floating Quantifiers: Handle with Care. The Second Glot International State-of-the-Article Book. Lisa Cheng and Rint Sybesma, eds. 107-148., Mouton de Gruyter.
Borer, H. (1995). The Ups and Downs of Hebrew Verb Movement. Natural Language and Linguistic Theory 13: 527-606.
Bresnan, J. (1982). "Control and Complementation." in Joan Bresnan, ed., The Mental Representation of Grammatical Relations. Cambridge, Mass.: MIT Press.
Bresnan, J. (2001). Lexical-Functional Syntax (Blackwell Textbooks in Linguistics). Oxford: Blackwell.
Bresnan, J., \& Mchombo, S.(1987). Topic, Pronoun, and Agreement in Chichewa. Language, 63(4), 741-782.
Chafe, W.L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In C.N.Li ed. Subject and Topic. New York: Academic Press.
Cirillo, R.J. (2009). The Syntax of Floating Quantifiers: Stranding Revisited. Utrecht: LOT.
Dalrymple, M.(1993). The Syntax of Anaphoric Binding. Standford: CSLI
Dalrymple, M.(2001). Lexical-Functional Grammar (Syntax and Semantics,34). New York: Academic Press.
Dik, S. (1978). Functional Grammar. North-Holland Linguistic Series, 37, Amsterdam: North-Holland.
Dowty, D. R. \& Brodie, B.(1984). The Semantics of "Floated" Quantifiers in a Transformationless Grammar. In Proceedings of the Third West Coast Conference on Formal Linguistics. Stanford.
Falk, Y.N. (2001). Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax. Standford: CSLI
Falk, Y.N. (2004). The Hebrew Present-Tense Copula as a Mixed Category. Proceedings of the LFG O4 Conference, Butt M. and T. Holloway King eds., 226-246, University of Canterbury.
Falk, Y.N. (2006). Constituent Structure and Grammatical Functions in the Hebrew Nominal Phrase. Architectures, Rules, and Preferences: A Festschrift for Joan Bresnan, Grimshaw J., Maling J., Manning C., Simpson J. and Zaenen A. eds.185-207. CSLI Publications.
Fassi Fehri, A. (1988). Agreement in Arabic, Binding and Coherence. Agreement in Natural Language: Approaches, Theories, Descriptions, Barlow M. and C. Ferguson eds., 107-159. Chicago: The University of Chicago Press.
Fillmore, C. J. (1968). The Case for Case. In Universals in Linguistic Theory. E.Bach and R. T. Harms, eds., 1-210. Holt, Rinehart, and Winston: New York.
Hurst, P. (2007). A Sketch Analysis of "Each" in English. Unpublished ms.,

University of Melbourne, Australia.
Kayne, R. (1975). French Syntax: The Transformational Cycle (Current Studies in Linguistics). Cambridge, Mass.: The MIT Press.
Kayne, R. (1981). Binding, Quantifiers, Clitics and Control. In: Frank Heny ed., Binding and Filtering, 191-211. Cambridge, MA: MIT Press.
King, T.(1995). Configuring Topic and Focus in Russian. Stanford, CA: CLSI Publications.
Lambrecht, K. (1981). Topic, Antitopic, and Verb Agreement in Nonstandard French. Amsterdam: John Benjamins.
Maling, J. (1983). Transitive Adjectives: A Case of Categorial Reanalysis. Linguistic Categories: Auxiliaries and Related Puzzles vol.1, Heny F. and B. Richards eds., 253-289, D.Reidel Publishing Company.

Milsark, G. (1977). Toward an Explanation of Certain Peculiarities of the Existential Construction in English. Linguistic Analysis, 3, 1-29.
Pollock, J.Y. (1989). Verb Movement, Universal Grammar and the Structure of IP. Linguistic Inquiry 20: 365-424.
Rullmann, H. (2003). Bound-Variable Pronouns and the Semantics of Number. Proceedings of the Western Conference on Linguistics: WECOL 2002, Agbayani B., P. Koskinen and V. Samiian eds.243254, California State University, Fresno.
Sag, I. (1978). Floating Quantifiers, Adverbs and Extraction Sites. Linguistic Inquiry 9: 146-150.
Shlonsky, U. (1991). Quantifiers as Functional Heads: A Study of Quantifier Float in Hebrew. Lingua, 84,159-180. North-Holland.
Shlonsky, U. (1997). Clause Structure and Word Order in Hebrew and Arabic: An Essay in Comparative Semitic Syntax. Oxford Studies in Comparative Syntax, New York: Oxford University Press.
Shlonsky, U. \& Doron, E. (1992).Verb Second in Hebrew. In D. Bates ed., Proceedings of the Tenth West Coast Conference on Formal Linguistics, 431-435.
Spector, I. (2008). Hebrew Floating Quantifiers: A Non Derivational Approach. M.A. Thesis. Hebrew University of Jerusalem. http://pluto.huji.ac.il/~msyfalk/Spector_MA.pdf
Spector, I. and Moldovano, P. (2007). Hebrew Floating Quantifiers in LFG. Ms., Hebrew University of Jerusalem.
Sportiche, D. (1988). A Theory of Floating Quantifiers and its Corollarie for Constituent Structure. Linguistic Inquiry, 19(3), 425-449.

# OUTSIDE-IN BINDING OF REFLEXIVES <br> IN INSULAR SCANDINAVIAN 

Tania E. Strahan<br>The University of Iceland NORMS (Nordic MicroComparative Syntax)

Proceedings of the LFG09 Conference
Miriam Butt and Tracy Holloway King (Editors)

2009
CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

This paper looks at the standard approach to long-distance reflexives within the Lexical-Functional Grammar framework, which defines the binding relation between a reflexive and its non-local antecedent by prescribing the types of syntactic elements which must and must not occur along the path from the reflexive to its antecedent. Evidence from the Insular Scandinavian languages suggests that the binding relation should be expressed at least partially as a constraint on the path from the antecedent to the reflexive. In other words, I suggest that long-distance reflexives in Icelandic and Faroese are governed by outside-in functional uncertainty, not purely inside-out functional uncertainty, as is standardly assumed.


## 1. Long-distance reflexives - from the inside out ${ }^{1}$

Following Dalrymple (1993) and Bresnan (2001), anaphoric binding, in particular long-distance reflexivisation (LDR), is viewed in LexicalFunctional Grammar as a kind of inside-out functional uncertainty. LDRs are those where the reflexive and its antecedent are not in the same clause, as illustrated in (1). The antecedent in both Icelandic and Faroese here must be John - it cannot be Maria.
(1)

| Jón | segir [ a ${ }^{\text {d }}$ | María | elski | sig]. | ICELANDIC |
| :---: | :---: | :---: | :---: | :---: | :---: |
| b. Jógvan | sigur, [ at | Maria | elskar | seg]. | Faroese |
| J | says that | M | love | $\mathrm{R}^{2}$ |  |
| 'John says that Maria loves self' |  |  |  |  |  |

The standard functional uncertainty rule for Icelandic LDR uses inside-out functional uncertainty. It looks something like (2), which says that a reflexive

[^175]has a SUBJect antecedent which is found by looking outwards in the fstructure through a series of COMPlement clauses.
(2) $\left(\left(\mathrm{COMP}^{+} \mathrm{GF} \uparrow\right) \mathrm{SUBJ}\right)_{\sigma}=\uparrow_{\sigma}$

The f-structure in (3) illustrates this. The reflexive has the object function in the embedded complement clause. The path to its antecedent may pass through COMPlement f -structures, as indicated by the heavy lines, to be linked to the same semantic structure as the subject of the higher f -structure.
(3) Simplified f-structure for (1a) Jón segir [að María elski sig].


In this paper, I will present evidence that more information about the antecedent is needed in order to establish coreference than just its grammatical function. In addition, data from Insular Scandinavian (i.e. Icelandic and Faroese) suggests that LDR should be viewed as a kind of antecedent-based, outside-in functional uncertainty, rather than a reflexivebased inside-out functional uncertainty, as in the standard view. Bresnan (2001 249) suggested that LDR must be licenced simultaneously by fstructure and the 'extended indirect discourse', but is not explicit about how to do this. Here, I will make a suggestion as to how this might be accomplished, specifically by arguing that the role of perspective-holder is crucial to establishing the link between an LDR and its antecedent, and that this role is calculated from the outside-in. I will argue for my analysis by assuming the standard inside-out approach, pointing out where this breaks down, and showing how an outside-in approach is better.

## 2. The Icelandic data

There is a contrast between the minimal pair of Icelandic sentences in (4), in that the reflexive is not permitted (a), only a pronoun is (b). That is, the reflexive and the pronoun are in complementary distribution here.
(4) a. *Hann kemur ekki nema pú bjóðir sér.
b. Hann kemur ekki nema pú bjóðir honum. he comes not unless you invite $\mathrm{R} / \mathrm{him}$ 'He won't come unless you invite self/him'

Given the inside-out constraint for LDRs in Icelandic in (5) (repeated from (2)), a simplified f-structure for (4a) is given in (6). The dotted line indicates that this object cannot be linked to the same semantic structure object as the higher subject.
(5) $\quad\left(\left(\mathrm{COMP}^{+} \mathrm{GF} \uparrow\right) \mathrm{SUBJ}\right)_{\sigma}=\uparrow_{\sigma}$
(6) f-structure for (4a) Hann kemur ekki nema pú bjóðir sér.


Clearly the f-structure in (6) violates the inside-out binding constraint in (5), as the anaphor is within an ADJunct, which the functional uncertainty equation does not allow it to bind out of. This sentence is therefore correctly predicted by the standard binding theory to be ungrammatical.

The next example illustrates that embedding a sentence like (4) under a 'perspectivising predicate' such as segja 'say' or halda 'believe/think' renders an LDR reading possible (Thráinsson, 1976).

| (7) Jón | segir | að | hann | kæmi | ekki |
| :--- | :---: | :---: | :--- | :---: | :--- |
| J | says | that | he | comes.SBJN | not |
|  | nema | pú | bjóðir | sér. |  |
| unless | you | invite. SBJN | R |  |  |

'John says that he won't come unless you invite him'
(8) f-structure for (7) Jón segir að hann kcmi ekki nema pú bjóðir sér.


The outermost predicate in this f-structure is segja 'say', which takes a nominative subject, and a COMP where the predicate must be in the subjunctive mood. The f-structure of this COMP is identical to that in (4) above, except that its main predicate is in the subjunctive mood (as required by the verb segja 'say'). The intended coreference is illustrated by the identical semantic-structure referred to by the antecedent, reflexive, and intervening pronoun. As above, the dotted line indicates that this object cannot be linked to the same semantic structure object as the higher subjects according to the binding constraint in (5), which disallows the intended binding in (7)/(8).

However, following Bresnan (2001), we can stipulate that the lexical features of sig here allow binding, due to the 'logophoric' nature of the construction. Thus, sig has the lexical features as given in (9).
(9) Lexical features of $\operatorname{sig}[+\mathrm{LOG},+\mathrm{SBJ}]$

On Bresnan's (2001) account, while sig must be bound to a subject [+SBJ], it is the [+LOG] ('logophoric') feature which allows the LDR binding to occur in (7). I presume that this implies that the antecedent must also be labelled in the f-structure, with something like [+LOG] or [+LOGANTE]. In fact, I will argue that the relevant feature of the antecedent is that of PERSPECTIVE-HOLDER, and that this feature resides, not in the $f$ structure, but in some other structure. Possible candidate locations for this feature are discourse-structure, or the apparently abandoned anaphoricstructure, both of which may be mapped from (or to?) the f-structure (Kaplan, 1995). In the next section, I briefly discuss the well-known links between LDR and logophoricity/perspective.

### 2.1. Logophoricity

Logophoricity was first identified and defined by Hagège (1974), to describe a context in which a third person's thoughts, feelings or emotions are expressed, and presented as though from their perspective. Logophoric pronouns are found in several African (Niger-Congo) languages, including Ewe (Clements, 1975) and Gokana (Hyman and Comrie, 1981). (10a) and (10b) contrast the logophoric pronoun in Ewe with a normal pronoun. The logophoric pronoun must be coreferential with the perspective-holder (10a), while the normal pronoun must be disjoint with this referent (10b).

| a. Kofi | be | yè-dzo |
| :--- | :--- | :--- |
| K. | say | LOG-leav |

'Kofi said that he (Kofi) left.'
b. Kofi be e-dzo
K. say PRO-leave
'Kofi said that he/she (not Kofi) left.'
Logophoric pronouns typically occur embedded under a verb meaning 'say'. Stirling (1993: 259) suggested a hierarchy of 'logocentric predicates', and it has been shown that these predicates are typically the ones which also occur with LDRs, with verbs to the left in the hierarchy clearly occurring more frequently with LDRs than those towards the right. ${ }^{3}$
(11) Communication $>$ Thought $>$ Psychological State $>$ Perception

LDRs do occur with non-logocentric predicates, and Reuland and Sigurjónsdóttir (1997) suggested that this is due to a difference between logophoric/discourse LDR on the one hand, and non-logophoric/syntactic

[^176]LDR. The discussion here will be restricted to the logophoric/discourse type, aka 'true LDR', rather than 'middle-distance' LDR over a non-finite clause boundary.

Sigurðsson (1986) specifically links point-of-view (POV) with Icelandic LDRs, illustrating that a proposition that is presented from a third person's POV and refers to that referent will be referred to with a reflexive, while the use of a pronoun signals that the referent is not the perspective-holder, cf (12) and (13). Notably, the verbs which are used most often in presenting a third person's perspective are those which are ranked more highly in Stirling's logocentric hierarchy.
(12) a. Jón segir að María elski sig. (= from Jón’s POV)
b. Jón segir að María elski hann. (= from someone else's, J says that M loves.SBJN R/him not Jón's, POV) 'John says that Maria loves self/him'
(13) a. Jón heldur að María elski sig. (= from Jón's POV)
b. Jón heldur að María elski hann. (= from someone else's, J thinks that M loves. SBJNR/him not Jón's, POV) 'John thinks that Maria loves self/him'

The link between LDR and logophoricity thus has to do with perspective, or point-of-view. Kuno's (1987) empathy is clearly also a related topic. Oshima (2007) argues that these three aspects of linguistic 'point-of-view' should be kept distinct, however for the purposes of this paper I am assuming these concepts are closely enough related that I may refer to them all under the rubric of 'perspective'. Also related to perspective is grammatical mood, where the subjunctive mood typically implies that the speaker does not vouch for the reliability of the proposition, instead assigning it to some other, mentioned party. This is discussed next.

### 2.2. Subjunctive mood, perspective and logocentricity in Icelandic

The correlation between the use of the subjunctive mood in Icelandic and the acceptability of LDR is often used as the basis for defining Icelandic LDR in terms of grammatical mood (eg, Anderson, 1986, Holmberg and Platzack, 1995). However, this is wrong. While the difference between (14a) and (b) could be due to the presence of the subjunctive mood in (a), and its absence in (b), Sigurðsson (1986) showed that this cannot be the case. Firstly, some Icelandic speakers accept (14b)/(15a). Secondly, those who accept (15a) do not accept (15b), where the higher subject Jón cannot be a perspective-holder/logophoric antecedent.
(14) a. Jón segir að María elski sig. (= from Jón’s POV) J says that M love.SBJN R 'John says that Maria loves self'
$\begin{array}{lllll}\text { b. * Jón veit að María } & \text { elskar sig. } & \text { (= ?not from } \\ \text { J } & \text { knows that } \mathrm{M} & \text { love.IND R } & \text { Jón's POV) }\end{array}$
'John knows that Maria loves self'
(15) a. Jón veit að María elskar sig. (= from

J knows that M love.IND R Jón's POV)
'John knows that Maria loves self'
$\begin{array}{clllll}\text { b. * Jón } & \text { veit ekki að María elskar sig. } & \text { (= not from } \\ \text { J } & \text { knows not that } \mathrm{M} & \text { love.Ind R } & \text { Jón's POV) }\end{array}$ 'John doesn't know that Maria loves self'

In addition, Thráinsson (1976) showed that the match between LDR and the subjunctive mood in Icelandic is not perfect. As well as the examples in (15), where LDR is permitted without the subjunctive mood, there are also examples like (4), which have the subjunctive mood in the embedded clause, but which do not permit LDR.
a. * Hann kemur ekki nema bú bjóðir sér.
he comes not unless you invite R
'He won't come unless you invite self'

However, Icelandic does not allow LDR out of adjunct clauses generally, so this example does not prove the lack of LDR/subjunctive correlation. I do not know of any subjunctive complement clause that does not allow LDR.

Still, the conclusion that Thráinsson and Sigurðsson have reached is that the subjunctive mood does not 'license' LDR in Icelandic, although the two often co-occur. This is a clear case of 'correlation $\neq$ causation'. LDR, logophoricity and the subjunctive mood all seem to have in common an involvement with perspective. Rather than LDR being a purely syntactic phenomenon, it seems more reasonable to assume that there are several linguistic features in Icelandic that co-occur with LDRs, and that it is the build-up of all of these that licence binding. As Thráinsson and Sigurðsson have shown, alone the subjunctive mood is neither sufficient nor necessary to licence Icelandic LDR.

## 3. Intriguing questions about the Icelandic examples

My main question is, given that the LDR rule is defined standardly as inside-out functional uncertainty, what is it that changes, from the point-of-
view of the reflexive and the constraint upon it, between (4) and (7)? Both are within ADJ clauses, both are OBJs of verbs that are in the subjunctive mood. Why does the constraint rule (7) in, but (4) out, given that the path from the reflexive is the same, at least initially, in both cases? Why does the constraint not break in (7), since it does break in (4)?

We know the reflexive can be bound to a perspective-holder, but how does the perspective-holder get this label? What allows the reflexive in (7) to get the [ + LOG] feature, but not the reflexive in (4), assuming that it is the [LOG] feature that allows the perspective-binding?
a. * Hann kemur ekki nema pú bjóðir sér.
he comes not unless you invite $\quad$ R
'He won't come unless you invite self'
(7) Jón segir að hann kæmi ekki J says that he comes.SBNJ not nema bú bjóðir sér. unless you invite.SBNJ R
'John says that he won't come unless you invite him'

There are at least two approaches to a solution to this problem.
Firstly, we could say that segja (and other logocentric verbs) licences a subjunctive chain, linking the reflexive's f-structure to the outside f-structure, which allows the reflexive to 'bypass' the ADJ, or makes the ADJ 'more COMP-like', for the purposes of the binding rule.

This constraint could be written such that there is a disjunction between either requiring a COMP or a subjunctive mood with say at the top on the path from the R to its antecedent, as shown in (16). Notice that we cannot just say 'require the subjunctive mood' alone, since this would incorrectly rule in (4). Requiring a chain of subjunctive moods, and the specification of the predicate segja 'say' are both off-path constraints.

$$
\begin{align*}
& \left(\left(\mathrm{COMP}^{+} \quad \mathrm{GF} \uparrow\right) \mathrm{SUBJ}\right)_{\sigma} \mid  \tag{16}\\
& \begin{array}{l}
(\mathrm{GF})^{+} \\
\rightarrow \mathrm{MOOD}=\text { subjnctv }
\end{array}=\uparrow_{\sigma} \\
& \text { Go through at least one comp TO a subject OR } \\
& \text { go through at least one f-structure } \\
& \text { each f-structure containing the subjunctive mood }
\end{align*}
$$ FROM some GF ( R can be anything), TO a subject of the PRED 'say'.

This constraint actually restricts the antecedent of the reflexive to the subject of the predicate say, but is does not use the logophoric label. We could use the constraint in (17) which does. Again, indicating that the SUBJ must be logophoric is an off-path constraint.
(17) ((COMP ${ }^{+}$GF $\left.\left.\uparrow\right) \mathrm{SUBJ}\right)_{\sigma} \mid$

$$
\begin{array}{cc}
\left((\mathrm{GF})^{+}\right. & \mathrm{GF} \uparrow) \mathrm{SUBJ})_{\sigma}=\uparrow_{\sigma} \\
\rightarrow \mathrm{MOOD}=\text { subjnctv } & \rightarrow[+\mathrm{LOG}]
\end{array}
$$

Go through at least one comp TO a subject OR
go through at least one f-structure
each f-structure containing the subjunctive mood
FROM some GF (R can be anything),
TO a subject
that is logophoric.
Independent rules will assign the feature [LOG] to the correct NP, which will then be able to be chosen as an antecedent for the reflexive. Yet this still does not explain which NP this will be - this task still remains.

Alternatively, we could assume that segja and its subjunctive mood cooccurs with the subject being labelled as [PERSPECTIVE-HOLDER] ([LOGANTE], [LDR-ANTE]). Then, as long as this perspective chain continues, the influence of the [PERSPECTIVE-HOLDER] continues. Based on findings by Thráinsson, Maling, Strahan, and others, Asudeh (2009 slide 51) suggested a rule (18) that would have this effect. This rule assigns the role of perspectiveholder, or 'logocentre', to the subject of segja 'say', and also passes this logophoricity down through subsequent embedded clauses.

$$
\begin{align*}
& \text { segja }(\uparrow \text { PRED })=\text { 'say }\langle\mathrm{SUBJ}, \mathrm{COMP}\rangle  \tag{18}\\
& \begin{array}{l}
\left((\uparrow \mathrm{SUBJ})_{\varpi} \text { logocentre }\right)=+ \\
(\uparrow \text { logophoric })=+ \\
\quad\left(\uparrow \mathrm{GF}^{+}\right)
\end{array} \\
& \quad \begin{array}{r}
\text { assigns role of logocentre } \\
\text { to subject of 'say' }
\end{array} \\
& \quad(\uparrow \text { mood })=_{c} \text { subjunctive }
\end{align*}
$$

A similar rule would presumably apply to vita 'know' for those speakers who allow LDR out of its (indicative mood) complement.

Notice that both of these possibilities correctly constrain the choice of antecedent to the subject of segja in (7), and never the subject of koma. Hann is never recognised as a perspective-holder/logocentre, and therefore is never recognised as a possible antecedent for an LDR.

Therefore, I suggest that the realisation of the anaphor as either the reflexive sér or the pronoun honum here relies crucially on the creation of a logocentric context, which is created by the predicate (as indicated in the two suggestions here), and also by features of the antecedent such as animacy (eg Thráinsson, 2007).

Conclusion: Both of these approaches will work to constrain LDR in Icelandic, using inside-out functional uncertainty. However, the use of insideout functional uncertainty still leaves open the problem of how to assign the role of perspective-holder or logocentre.

After considering the Faroese data, I will suggest that the perspectiveholder in both Faroese and Icelandic is assigned to a particular NP for reasons independent of anaphora, and that there is a constraint on LDR requiring its antecedent to be a perspective-holder.

## 4. Faroese

Examples of Faroese LDR are given in (19). The Icelandic equivalents are also given, for comparison. (19a, b) have only third person nominals, while (19c, d) have a second person pronoun as the subject of the embedded clause (i.e. the clause containing the reflexive).
(19) a. Jógvan sigur, [ at Maria elskar seg]. FAROESE b. Jón segir [ að María elski sig]. ICELANDIC $J$ says that M 'John says that Maria loves self'

| c. * Jógvan | sigur, | $[$ at | tú | elskar | seg $].$ | FAROESE |
| :--- | :--- | :--- | :--- | :--- | :--- | ---: |
| d. Jón | segir <br> J. | $[$ að | bú | elskir | sig $].$ | ICELANDIC |
|  | J | says | that | you | love | R | 'John says that you love self'

(20) gives the f-structure of (19a). Notice that the reflexive can bind out of the COMP to the SUBJ, like in Icelandic. (Faroese does not have grammatical mood.)
(20) f-structure for Jógvan sigur, at [Maria elskar seg]


In (21) is the f-structure for the version of this sentence with a second person pronoun.
(21) f-structure for Jógvan sigur, at [tú elskar seg]


The f-structure in (21) is identical to that in (20), except that the subject of the embedded COMP clause is second and not third person. This causes the sentence to be unacceptable.

Native speakers, when asked why (21) is bad, invariably say there is a problem with the second person pronoun - it appears to make the sentence direct speech. Most people laugh and shake their heads and apologise for the badness of (21), especially when they are reminded that they said that (20) was fine! Intriguingly, very few Faroese speakers change their mind about the ungrammaticality of (21) when its similarity to (20) is pointed out to them - the presence of non-third person has a strong confounding effect on the acceptability of LDR in Faroese, for most (but not all) speakers

Notice that this restriction against the presence of non-third person pronouns holds even (or especially) out of ADJunct clauses, as well as out of COMPs, as shown by the examples in (22) and (23). Notice also that the equivalent Icelandic sentences are very (22a, 23), or at least rather (22b),
ungrammatical. $(22 \mathrm{c}, \mathrm{d})$ give the percentage of speakers who reported that this sentence sounded 'completely natural' in the large syntactic overview projects ongoing in Iceland and the Faroe Islands. These percentages are based on results from around 1,000 Icelandic speakers and around 250 speakers of Faroese. The other judgements are from my own fieldwork.
(22)
a. Zakaris lesur ikki bókina, FAROESE Z reads not book.DEF.3sg.F
[ tí að hon keðir seg]. because 3sg.NOM.F bore R
'Zakaris doesn't read the book, because it bores self'
b. ?* Jón les ekki bókina, $\quad$ ICELANDIC J reads not book.DEF.3sg.F [ pví að hún ergir sig]. because 3sg.NOM.F annoy R 'John doesn't read the book, because it irritates self'
c. Hann brúkar tað, [sum passar sær]. FAR. (60\%)
d. *Hann notar pað, [sem passar sér]. ICEL. (25\%) he uses that which suits R 'He uses that which suits self'
(23)

| Magnus | dámar | Beintu, |
| :--- | :--- | :--- |
| M | likes | [ tí at |
| because |  |  |

hon hjálpir sær við heima arbeiðinum]. she help R with house work
'Magnus likes Beinta because she helps him with the housework'
b. *Magnus dámar meg, [ tí at FAROESE M likes me because eg hjálpi sær við heima arbeiðinum]. I help R with house work
'Magnus likes me because I help him with the housework'
c. Olaf ivast í, [um Maria vil

FAROESE
O doubts in if M want
hjálpa sær við heima arbeiðið]. help R with house work
'Olaf doubts whether Maria want to help R with the house work'

| d. * Olaf | ivast | í, | $[$ um | tú | vil | FAROESE |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| O | doubts | in | if | you want |  |  |
|  | hjálpa | sær | við | heima | arbeiðið $].$ |  |

Faroese LDR appears to have a very straight-forward binding restriction, namely that the presence of a non-third person pronoun causes LDR to be ungrammatical. This can be very easily expressed in an OFF-PATH CONSTRAINT (Dalrymple, 1993), restricting the path's journey through any fstructure that itself contains a first or second person pronoun. There does not appear to be a difference between COMP or ADJ paths. ${ }^{4}$

Furthermore, at least some Faroese speakers allow an LDR to have a nonsubject antecedent, even with a first-person pronoun present. The percentages are those who find the sentence 'completely natural', based on 10 speaker judgements. The figure of $43 \%$ comes from two of these speakers selecting 'almost completely natural' as their judgement instead of 'completely natural'.

'I showed Maria the book which was written about self'
b. Eg vísti Mariu bókina, I showed M book.DEF sum var skrivað um sín abba [43\%] which was written about R's grandfather 'I showed Maria the book which was written about self's grandfather'

Faroese speakers who accept LDRs also prefer them to a pronoun.
We could postulate the regular expression governing LDR in Faroese as in (25).

$$
\begin{align*}
& \left(\mathrm{GF}^{+} \mathrm{GF} \uparrow\right)_{\sigma}=\uparrow_{\sigma}  \tag{25}\\
& \neg(\rightarrow \mathrm{PS}=1 \vee 2)^{5}
\end{align*}
$$

[^177]This rule says that the antecedent is not restricted to any grammatical function (GF), nor to following any particular path through the f-structure to the antecedent. It does have an off-path constraint, restricting the path's journey through any f-structure that itself contains a person feature of 1 or 2.

However, I am not satisfied with the rule in (25) for three reasons. Firstly, not all speakers have the off-path constraint requirement. Secondly, many speakers do in fact have a preference for a path through COMPs and not ADJs between the reflexive and its antecedent, and for those speakers who have a person restriction associated with LDR (for whom the off-path constraint applies), it tends to be stronger out of adjunct clauses than out of complement clauses (Strahan, in press). That is, there is an interaction effect between person and clause type, which is not captured by the suggested constraint.

Thirdly, this off-path constraint is stipulative, although the motivation is straight-forward. Intuitively, if the antecedent of an LDR is a perspectiveholder, which is passed down through subsequent f-structures, we can appeal to the fact that first and second person pronouns outrank third person pronouns in perspective-holding-ability. This would mean that a first or second person pronoun will always (for many speakers) become the perspective-holder, ruling out (third person) LDR. This observation itself provides direct motivation for the identification of the antecedent for the LDR, namely that, in a general text containing and about third persons, the perspective-holder, and thus LDR antecedent, is a third person nominal, unless a non-third person pronoun appears.

If we assume that speakers and hearers are always aware of which discourse referent is the perspective-holder, then the off-path constraint is redundant, since it falls out of the need to identify the highest-ranked perspective-holder. I will describe the general principles behind how to calculate this in the following section. Notice that this means that the calculation of the perspective-holder must take place before any binding constraints apply. Furthermore, this calculation of perspective-holder necessarily applies from the outside in.

## 5. Outside-in or inside-out functional uncertainty?

I have pointed out some problems for the standard inside-out view of anaphoric binding, in particular with respect to Icelandic perspectivising LDR and Faroese LDR in conjunction with non-third person pronouns. However, suggesting an outside-in view of anaphoric binding clearly poses a rather large problem. As pointed out by an anonymous reviewer, a strict view of the LDR constraint as simply '[outside-in functional uncertainty] would mean that each possible perspective-holder would launch a search for possible LDRs, which does not seem plausible'.

I agree with this sentiment. However, the inside-out functional uncertainty suffers from the same problem, namely, how can the binding constraint rule in a sentence with a perspective-holder antecedent, if it does not know what this perspective-holder looks like? That is, the perspective-holder needs to be labelled as such, somehow.

It could be argued that all reflexives must launch a search for an antecedent anyway, thus it is more economical to leave it to the reflexive. However, the question remains as to how the antecedent is to be identified, and that can only be satisfied if the antecedent actually exists, and is already identified as somehow being 'available' to be the antecedent for an LDR. That is to say, the inside-out LDR binding constraints suggested for Icelandic, which relies on the AntecedentFunction being labelled as [+LOG], or the Faroese constraint which essentially says 'bind to anything you like, so long as there is no non-third person around', still do not answer the question of which NP will be the antecedent. Both the [+LOG] label and the function of perspective-holder must be calculated or assigned using some other tool, which must be top-down/outside-in.

An antecedent-based, outside-in, account of binding also deals neatly with 'discourse' binding, where the antecedent is not even in the same sentence, as in (26). The antecedent of sér here is not even mentioned in this excerpt.


What (26) clearly shows, is that the perspective-holder of each given domain is already calculated, for reasons independent of LDR. The first sentence of (26) should be interpreted as, not that María is an objectively nasty person, but that she is subjectively nasty, in particular, she is nasty to the owner of the narrative, to the perspective-holder. Even with no reflexive, a perspective-holder is calculated. This perspective-holder is carried through, not only subsequent clauses, but also subsequent sentences. An overt indication of this is the use of the subjunctive mood in the second sentence. The conjunction pegar 'when' does not itself require the subjunctive mood, only continuing domain of the perspective-holder does.

As McCready (2007: 41) shows and says, 'subordinating discourse relations enable point of views established in one discourse segment to be retained into later segments'.

To implement the passing down of perspective, I suggest that the perspective-holder, here labelled ' P ', is initially assumed to be the speaker. At a complement clause boundary in both Icelandic and Faroese, the P either changes to the animate, subject NP of the preceding clause, or remains/changes to the speaker, as in (27) and (28). For an LDR to be used, the P at the point in the utterance of the anaphor must be the appropriate referent in the sentence or discourse.


In Icelandic, and for some speakers of Faroese, at an adjunct clause boundary, the P either changes to the speaker or remains the same - there is no option to change to the preceding subject/animate NP, as in (29). This means that ordinarily only a pronoun may have an antecedent on the other side of an adjunct clause boundary, as in (30), while a reflexive may not. For other speakers of Faroese, the P change follows the rule for a complement clause, compare (31) in Faroese with (32) in Icelandic.

|  | Hann kemur ekki <br> he comes.IND not | nema pú bjóðir sér. <br> unless you invite.SBJN R <br> ADJ:speaker, or same <br> (i.e., must be speaker, thus *sér) | ICELANDIC |
| :--- | :--- | :--- | :--- |
|  | speaker | Hann kemur ekki | nema pú bjóðir honum. <br> he comes.IND not |
| $P:$ | speaker | unless you invite.SBJN him <br> ADJ:speaker, or same |  |
|  |  | (i.e., must be speaker) |  |

(31) Zakaris lesur ikki bókina, tí at hon keðir seg. FAROESE Zakaris reads not the.book because it bores R
$P$ : speaker
(32)
?* Jón les ekki bókina, John reads not the.book
$P$ : speaker

ADJ:speaker, or Zakaris
pví að hún ergir sig. ICELANDIC because it bores R
ADJ:speaker, or same
(i.e., must be speaker)

For most speakers of Faroese, at the use of a first or second person pronoun, the P changes to the referent of that pronoun, as indicated in the last line of (33). This means then that third person reflexives are ruled out since the P at the instant of the anaphor is a second person referent, thus there is no 'available' third person $P$ for the anaphor to bind to.

| (33) | * Jógvan sigur, | at | tú elskar seg. |
| ---: | :--- | :--- | :--- | :--- |
| John says | that | you love self |  | FAROESE

Calculation of the $P$, given that it is calculated for reasons independent of binding, probably occurs in the discourse-structure although a dedicated anaphoric-structure is also plausible (Kaplan, 1995). Full details of how to implement this are beyond the scope of this paper, as the relevant structures in LFG are not yet stable enough to implement this analysis without a substantial amount of architectural explanation. I leave this issue for future work. The important point here is that the calculation of the antecedent of an LDR is now reduced to the calculation of the perspective-holder, and it occurs from the outside-in. For at least Faroese and Icelandic, f-structure factors are important in this calculation (COMPs versus ADJs, person). Non-f-structure factors are also relevant, given the evidence of discourse reflexives as in (26), where the use of the subjunctive mood in Icelandic continues throughout a paragraph across sentences boundaries, which is probably some kind of i - or d-structure phenomenon.

The idea that there is a single, simple negative constraint on LDR that applies to the syntax, namely that these reflexives cannot be bound to a coargument, coupled with the single positive constraint that the reflexive must be bound to the perspective-holder, is highly appealing. Unfortunately, this position ignores the clause-bound uses of reflexives, such as with inherently reflexive predicates, reflexives in locative PPs, and the fact that the possessive reflexive can definitely have a local antecedent. I leave it for another paper to explore whether these kinds of reflexives are also bound to the perspective-holder, and make the claim for outside-in binding here only for LDRs. ${ }^{6}$

[^178]
## 6. Summary and final remarks

In Icelandic, reflexives may be bound out of ADJuncts, and out of sentences, when the antecedent is a perspective-holder. I suggest that the use of an ADJ in Icelandic normally reduces the prominence of the current perspective-holder, but that when embedded within a strong third person perspective-holder as is the case for a proposition embedded under the verb segja, then ADJ f-structures are no boundary. I also suggest that the use of the subjunctive mood with complement clauses increases the likelihood that the subject is a perspective-holder, and thus a potential LDR antecedent. In Faroese, reflexives and their antecedents may be bound across an ADJ clause boundary if the speaker can construe the sentence as being 'about' the intended antecedent.

Given the fact that the perspective-holder is calculated for reasons independent of anaphora resolution, it seems sensible to have the outside-in constraint apply in the appropriate structure. This structure would be something like information-structure, discourse-structure, anaphoric-structure or pragmatic-structure.

Asher and Wada (1988) have already had some success in implementing a multi-faceted, top-down/outside-in algorithm which could correctly predict whether a discourse referent was going to be referred to with a pronoun or a full NP. Their success in accounting for the distribution of pronouns versus full NPs using an antecedent-based rule is a good indication that a similar approach could work for reflexives.

In conclusion, evidence from Icelandic and Faroese points to an online, cognitive model of LDR, where discourse referents are evaluated for their perspective-holding ability, in each relevant domain. This is probably calculated in the d-structure or anaphoric-structure. The single simple rule given in (35) applies, where P represents the perspective-holder, and $\uparrow$ represents the reflexive.

$$
\begin{equation*}
P_{\sigma}=\uparrow_{\sigma} \tag{34}
\end{equation*}
$$

When a reflexive needs an antecedent, it therefore binds to the DP/NP which is already indicating its availability, passing this information down, from the outside in.

## References

Anderson, Stephen R. 1986. The Typology of Anaphoric Dependencies: Icelandic and other Reflexives. In Topics in Scandinavian Syntax,
work for Faroese, nor for discourse reflexives as in (26), which are otherwise unified under the account suggested here.
eds. Lars Hellan and Kirsti Koch Christensen, 65-88: D. Reidel Publishing Company.
Asher, Nicholas, and Wada, Hajime. 1988. A computational account of syntactic, semantic and discourse principles for anaphora resolution. Journal of Semantics 6:309-344.
Asudeh, Arshia. 2009. Reflexives in the correspondance architecture. Proceedings of Research seminar, University of Iceland.
Bresnan, Joan. 2001. Lexical-Functional Syntax. Great Britain: Blackwell Publishers.
Clements, George M. 1975. The Logophoric Pronoun in Ewe: Its Role in Discourse. Journal of West African Languages 10:141-177.
Dalrymple, Mary. 1993. The Syntax of Anaphoric Binding. Stanford: CSLI Publications.
Hagège, Claude. 1974. Les pronoms logophoriques 'Logophoric pronouns'. Bulletin de la Société de la Linguistique de Paris 69:287-310.
Holmberg, Anders, and Platzack, Christer. 1995. The Role of Inflection in Scandinavian Syntax. New York: Oxford University Press.
Hyman, Larry M., and Comrie, Bernard. 1981. Logophoric Reference in Gokana. Journal of African Languages and Linguistics 3:19-37.
Kaplan, Ronald M. 1995. The formal architecture of Lexical-Functional Grammar. In Formal issues in Lexical-Functional Grammar, eds. Mary Dalrymple, Ronald M. Kaplan, John T. III Maxwell and Annie Zaenen, 7-27: Stanford University.
Kuno, Susumo. 1987. Functional Syntax: anaphora, discourse and empathy. Chicago: The University of Chicago Press.
McCready, Eric. 2007. Discourse subordination and logophoric binding. Research on Language and Computation 5:37-48.
Oshima, David. 2007. On Empathic and Logophoric Binding. Research on Language and Computation 5:19-35.
Reuland, Eric, and Sigurjónsdóttir, Sigríður. 1997. Long Distance 'Binding' in Icelandic: Syntax or Discourse? In Atomism and Binding, eds. Hans Bennis, Pierre Pica and Johan Rooryck, 323-340. Dordrecht: Foris Publications.
Sigurðsson, Halldór Ármann. 1986. Moods and (Long-distance) Reflexives in Icelandic. Ms. Kiel.
Stirling, Lesley. 1993. Switch-reference and Discourse Representation. Great Britain: Cambridge University Press.
Strahan, Tania E. 2003. Long-distance reflexives: a quantitative study. München: Lincom-Europa.
Strahan, Tania E. in press. Faroese long-distance reflexives face off against Icelandic long-distance reflexives. In Nordlyd, eds. Peter Svenonius et al. Tromsø: Universitet i Tromsø. http://www.ub.uit.no/baser/ nordlyd/.
Thráinsson, Höskuldur. 1976. A semantic reflexive in Icelandic. Ms.

Thráinsson, Höskuldur. 2007. The Syntax of Icelandic: Cambridge syntax guides. Cambridge: Cambridge University Press.

# IRISH CLEFTING AND INFORMATION-STRUCTURE 

Sebastian Sulger<br>Universität Konstanz

Proceedings of the LFG09 Conference Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

This paper presents an analysis of Irish clefting couched within LexicalFunctional Grammar. In Irish, cleft sentences are formed using two syntactic permutations. First, a copula is introduced, taking the clefted phrase or word as a predicate. Second, a relative clause is formed containing the remaining material of the original sentence. This basic pattern is valid across a variety of languages. I present different approaches towards copula predication taken within Lexical-Functional Grammar and discuss my analysis of Irish copula constructions. Based on the insights from simple copula predication, I derive my analysis for Irish clefting, claiming that in principle, a parallel syntactic approach for both simple copula clauses and clefting can be assumed. The syntactic analysis of copula clefting has been implemented in a computational LFG grammar using the XLE software

Next, pragmatic aspects of clefting are discussed. Cleft sentences are not only interesting from a syntactic point of view, but also are used to separate new information from old information. In the modular architecture of Lexical-Functional Grammar, additional levels of representation may be added to allow for extra-syntactic analysis. I use the projection of i(nforma-tion)-structure to map strings in the sentence to discourse functions. Applying this type of analysis to clefting, I arrive at a more complete picture of the form and function of Irish clefts.


## 1 Introduction

In this paper, I present an analysis of Irish clefting. The analysis extends not only over the syntactic aspects of clefting, but also presents ideas related to the information-structural consequences of clefting. In Modern Irish, as in other languages, the pattern for clefting relies on the use of the copula verb. The basic pattern for clefting in Irish is copula - XP - relative clause (Stenson, 1981). I argue that the relative clause in fact resembles a clausal subject, in which the nominal head is understood, though not overtly expressed. The clefting pattern in that case is identical to the general copula sentence pattern; my conclusion is that one syntactic analysis can and should account for both general copula predication and clefting constructions. The difference between the two constructions are to be found in the information-structure of the sentences. I give ideas on how to analyze discourse functions for cleft sentences using a separate projection within LFG architecture called i-structure (King, 1997; Butt and King, 1998; Andréasson, 2007).

Section 2 presents different copula analyses within the LFG literature and discusses why I chose a PREDLINK analysis for Irish copula. In section 3, I provide clefting data from Irish and show the connection between simple copula sentences and cleft sentences. Section 4 discusses consequences of clefting for the

[^179]information-structure of a sentence. The syntactic analysis of Irish copula and clefting has been implemented in a computational LFG grammar of Irish using the XLE grammar writing platform (Crouch et al., 2008).

## 2 Copula Analyses in LFG

I argue that understanding copula predication is essential to understanding clefting patterns. I therefore first take a look at different copula analyses in the LFG literature. There are three different copula analyses across the literature: a singletier analysis, an open-complement double-tier analysis (XCOMP) and a closed complement double-tier analysis (PREDLINK) (Attia, 2008).


Figure 1: Copula construction analyses in LFG

### 2.1 Single-Tier Analysis

The single-tier analysis involves the copula predicate (i.e., the adjective in the example below) functioning as a sentential head, meaning that it selects for a subject. Dalrymple et al. (2004) note that this is a preferable analysis for cases in which the copula is optional, such as with Japanese predicative adjectives.
(1) a. hon wa akai.
book TOP red
'The book is red.'
(Dalrymple et al., 2004, p. 190)
b. sono hon wa akai desu.
this book TOP red is
'This book is red.'
(Dalrymple et al., 2004, p. 191)

$$
\left[\begin{array}{lll}
\text { PRED } & \text { 'red }\langle(\uparrow \text { SUBJ })\rangle \\
\text { SUBJ } & {\left[\begin{array}{cc}
\text { PRED } & \text { 'book' }
\end{array}\right]}
\end{array}\right]
$$

Figure 2: Single-tier analysis in Japanese
A single-tier analysis is problematic for two reasons. First, it has to be assumed and shown that the copula predicate can in fact select for a subject. Second, the copula in Japanese is not optional, but has to surface with nominal predicates. Dalrymple et al. (2004) assume that adjectives are able to select for subjects, while
nouns are not. A different type of analysis would have to be assumed for nominal predicates. This is not ideal since the predication in both cases is the same the optionality of the copula is no sufficient motivation for assuming two separate analyses of the copula (Attia, 2008).

### 2.2 Open-Complement Double-Tier Analysis (XCOMP)

In the open-complement double-tier analysis, the copula predicate does not function as a sentential head. The head is the copula itself, when it is present, or a null element, when the copula does not surface. The copula clause predicate selects for a subject which is functionally controlled by the main clause. This means that the subject of the main clause (i.e., the subject of the copula) is unified with the subject of the copula predicate. In LFG, a control relation of this type is constructed using an XCOMP function. The partial XCOMP f-structure alone does not contain a subject value, i.e., is not complete; the subject function is filled with a clause-external subject via functional control through the linking verb (Bresnan, 1982).

It has been argued that this a preferable analysis for cases in which the predicate shows agreement with its subject. Dalrymple et al. (2004) give the example of French adjectives in predicate position, agreeing with the subject of the main clause in the same way as verbs do. The following examples are taken from Dalrymple et al. (2004).
(2) a. Il est petit.
he.Masc.Sg COP.3P.Sg.Pres small.Masc.Sg
'He is small.'
(Dalrymple et al., 2004, p. 195)
b. Elle est petite.
she.Fem.Sg COP.3P.Sg.Pres small.Fem.Sg
'She is small.'
(Dalrymple et al., 2004, p. 195)


Figure 3: Open double-tier analysis of French copula

Dalrymple et al. (2004) argue that the agreement is a strong indication for a control relation between the subject and the predicate. They also state that if one assumes an XCOMP type of analysis, one can write basic lexical entries for the predicate as in (3).

```
(3) petite ( \(\uparrow\) PRED) \(=^{\prime}\) small \(<(\uparrow\) SUBJ \()>^{\prime}\)
    ( \(\uparrow\) SUBJ NUM) \(=\mathrm{c}\) sg
    ( \(\uparrow\) SUBJ GEND) \(=\mathrm{c}\) fem
```

Attia (2008) counters this analysis. He makes the point that in French, the agreement of the copula predicate with the subject is not the same type of agreement verbs show with their subjects. He maintains that the agreement is not enough evidence to assume that the copula predicate subcategorizes for a subject.

Another important argument contra the XCOMP open complement analysis comes from a predicational perspective. The analysis in Figure 3 is not suitable for copula constructions according to Attia (2008) because this is exactly the way normal subject raising verbs (such as seem, appear) are analyzed in LFG. Assuming an f-structure like Figure 3 would therefore mean that there is no difference between copula constructions and subject raising verbs.

Dalrymple et al. (2004) provide the most compelling evidence against the opencomplement XCOMP-type analysis. In cases where the post-copular complement already has a subject which is different from the subject of the main clause, the closed complement PREDLINK analysis is the preferred analysis. See the examples in (4).
(4) a. The good thing is that he did not throw the snowball.
b. The main goal is (for the student) to succeed in the exam.

If we assume an XCOMP analysis for sentences like these, the result is a clash of PRED values, i.e., because of the control equations, the XCOMP f-structure would contain two subjects which are not unifiable (Attia, 2008; Dalrymple et al., 2004; Butt et al., 1999). See the illformed f-structure in Figure 4 for sentence (4a).


Figure 4: Clashing open double-tier analysis with divergent subjects

### 2.3 Closed-Complement Double-Tier Analysis (PREDLINK)

Another alternative is the so-called closed complement analysis; "closed" because the PREDLINK function is a closed function and therefore does not allow functional control. The PREDLINK analysis models the fact that a particular property is predicated of the subject in a syntactically reasonable way. The main PRED of the f -structure expresses that a specific property is predicated of the subject. Exactly this is captured by the PREDLINK function (Butt et al., 1999).

The closed complement double-tier analysis is a universal LFG analysis for copula constructions according to both Attia (2008) and Butt et al. (1999). The main advantage of this approach is that since it does not necessarily rely on control equations, it does not have any issues with sentences such as in (4). On the other hand, when control equations become mandatory (i.e., when there is obvious agreement), these can also be formulated in a closed complement PREDLINK analysis (Attia, 2008; Butt et al., 1999), although the equations are slightly more troublesome in cases of long-distance agreement. Therefore, the main arguments for an open complement XCOMP type of analysis presented by Dalrymple et al. (2004) are not enough to motivate a pluralist approach to copula constructions in terms of analyses, since all of the advantageous properties of the XCOMP analysis can be reproduced within a PREDLINK closed complement analysis. Attia (2008) maintains that syntactic features such as the presence and absence of the copula form and the presence and absence of agreement on the predicate do not affect the syntactic function of the predication. Only because languages like French show agreement on the predicate, copula predication does not necessarily require diverging syntactic analyses.

There are several advantages to the double-tier closed complement PREDLINK analysis. First, it does not matter what kind of constituent the copula complement is; this analysis seems to be the only one that succeeds in providing valid representations for all constituent types, which can take different semantic roles; see Attia (2008) for an overview. Other approaches seem to have problems with this unified approach. Bresnan (2001), assuming that adjectives can subcategorize for subjects, also assumes that nouns and prepositional phrases can do so. To account for this, she proposes to manipulate the PRED of the noun or preposition by means of lexical rules; see the sentences in (5) and the corresponding rules in (6), cited by Lodrup (2008).
(5) a. The pills made him a monster.
(Lodrup, 2008, p. 22)
b. She seems in a bad mood.
(Lodrup, 2008, p. 22)
(6)

```
a. 'monster' =>
    \(`\) be-a-monster<( \(\uparrow\) SUBJ) \(>^{\prime}\)
b. 'in<( \(\uparrow\) OBJ) \(>\) ' =>
        'be-in-a-state-of<( \(\uparrow\) SUBJ), ( \(\uparrow\) OBJ) >' (Lodrup, 2008, p. 22)
```

Both Attia (2008) and Lodrup (2008) find this approach problematic, since it not only results in artificial and complex annotation, but also presupposes that any PP or NP in a given language can in principle subcategorize for a subject. Dalrymple et al. (2004) and also Rosén (1996) in an earlier paper maintain that this type of analysis is certainly not desirable. Within the closed complement analysis, these problems vanish, since there is no XCOMP f-structure, hence we do not need any subject that is functionally controlled. I give the f-structure for (6b) in Figure 5 , assuming a closed complement double-tier analysis.


Figure 5: Well-formed closed double-tier analysis of She seems in a bad mood.
In sentences where there is no copula, the PREDLINK analysis has intuitive appeal, since, at the f-structure level, it mirrors the juxtaposition of constituents when the copula is missing (Attia, 2008). As many languages contain copula-less sentences (Carnie, 1995), the analysis has cross-linguistic appeal. Attia (2008) further argues that the presence vs. absence of the copula itself is a parameter of variation. Since the copula is generally considered as semantically empty, there is no functional distinction to be made between sentences containing the copula and sentences without the copula. The predication in the absence of the copula is modeled using a null-be predicator in the LFG rule notation. See the sentence in (7) and the rule in (8); the resulting f-structure is shown in Figure 6.
(7) hwa tālibun
he student
'He is a student.'
(Attia, 2008)
(8) S --> NP: ( $\uparrow$ SUBJ) $=\downarrow$;
$\varepsilon: \quad(\uparrow$ PRED $)=$ 'null-be< $(\uparrow$ SUBJ $),(\uparrow$ PREDLINK $)>^{\prime}$
( $\uparrow$ TENSE) $=$ pres;
$\{\mathrm{NP} \mid \mathrm{AP}\}: \quad(\uparrow$ PREDLINK $)=\downarrow$
$(\downarrow$ GEND $)=(\uparrow$ SUBJ GEND $)$
$(\downarrow$ NUM $)=(\uparrow$ SUBJ NUM)


Figure 6: Closed double-tier analysis without surface copula in Arabic
Note that the tense feature is provided by the empty element $\varepsilon$, which captures the insight that the copula in Arabic can only be omitted in present tense. If a copula was present in the sentence, then the tense feature would be provided by the copula itself.

To sum up the discussion about the different approaches towards copula constructions in LFG, I stress that each one of the possible analyses has its advantages and disadvantages. While we have to assume a subcategorization frame for predicate elements in the single-tier and open-complement types of analysis which might be less appropriate for some languages than for others, the PREDLINK approach is more neutral in this respect. I think in this discussion it is important to see LFG in the context of parallel (i.e., cross-linguistic) grammar designing. As long as there are no serious reasons to object to the PREDLINK analysis (e.g., the case of Abkhaz discussed above), I maintain that a universally applicable analysis should be favored to increase the cross-linguistic appeal of LFG.

### 2.4 Towards an Analysis of Irish Copula in LFG

In this section, I present my own analysis of Irish copula predication in LFG. This analysis provides the basis for the approach to Irish cleft constructions since simple cleft constructions in Irish rely on copula predication.

Consider the examples in (9). They all contain a copula, a predicate and a subject, nothing more and nothing less, as indicated by the bracketing.
(9) a. Is [le Pól $]_{P R E D}[\text { an carr. }]_{S U B J}$

COP.Pres with Paul ART.Def.Sg car
'The car belongs to Paul.' [lit. 'The car is with Paul.']
b. Ba [dhuine deas] $]_{P R E D}[$ é.] SU $B J$

COP.Past man nice he
'He was a nice man.'
c. Is [maith li-om $]_{P R E D}[\text { tae. }]_{S U B J}$

COP.Pres good with-me tea
'I like tea.'
Consider the arguments given by Dalrymple et al. (2004) in favor of a divergent analysis for copula constructions across languages and within a certain lan-
guage. First, agreement is given as an argument for an XCOMP (open complement double-tier) analysis. In Irish, however, the predicate does not show any agreement whatsoever with the subject. Therefore, I maintain that there is no reason concerning agreement to choose XCOMP as a possible copula analysis for Irish.

Second, in cases where the copula is absent, Dalrymple et al. (2004) argue for a special analysis: the single-tier analysis. They present the case of Japanese, where the occurrence of the copula is governed by the category of the predicate. They propose that different analyses have to be assumed depending on the presence or the absence of the copula. Attia (2008) however, as we have seen, maintains that this is merely a case of stylistic variation. In principle, the predication is the same, no matter if the copula is present or not.

In Irish, the factor governing the occurrence of the copula is tense. The copula may be overt or dropped in the present tense, but its occurrence is mandatory when tensed for future or past. Still, the predication in the sentences does not change in principle, whether the copula is present or not. Therefore, I maintain that the presence of the copula is a means of stylistic variation in the present tense, but must be present in clauses with future or past tense. The presence vs. absence of the copula does not lead to semantic differences (see also Ó Siadhail, 1989; Stenson, 1981). Additionally, the present tense copula may not be deleted when the negative form is used (Ó Siadhail, 1989). This can also be modeled via c-structure rule annotations.

I follow Butt et al. (1999) and Attia (2008) in claiming the closed complement double-tier analysis as a universally applicable analysis for copula constructions. The possible variation in the choice of the predicate constituent is immediately reflected by the variation in the PREDLINK f-structure. Different lexical categories can be head of the predicate, hence head of the PREDLINK f-structure. A sample analysis for sentence (9c) is given in Figure 7. In (9c), the head of the predicate is an adjectival construction (i.e., an adjective head maith 'good' with a prepositional phrase adjunct liom 'with me').


Figure 7: Sample copula analysis using PREDLINK

## 3 Irish Clefting

This section provides some data from Irish clefting. I compare a special kind of Irish copula construction, called "identification sentence" in the literature (e.g., Stenson, 1981; Ó Siadhail, 1989), to Irish clefts, suggesting that these constructions are very similar. As a consequence, the syntactic analysis of simple Irish clefts becomes straightforward and fits into the closed complement double-tier PREDLINK analysis introduced in the LFG literature perfectly.

### 3.1 The Irish Data

In this section, I present data illustrating simple Irish clefting. The data shows the general form of clefts and the variability in the choice of the clefted constituent. All of these sentences have the same basic structure. A constituent of the basic sentence appears after the copula as its predicate and a relative clause is formed containing the rest of the sentence, including the main verb. These are the basic syntactic facts given throughout relevant literature (Stenson, 1981; Ó Siadhail, 1989; The Christian Brothers, 1960).

In the following examples, I provide the basic (i.e., non-clefted) sentence and variations of that sentence. That means, for example, (10b) and (10c) are permutations of (10a), and so on. When forming a cleft sentence, the NP can appear in a position before the main verb of the sentence, violating the basic VSO word order of Irish. It is possible to front subject NPs and object NPs. (10c) is another permutation of (10a) whereby the adverbial inné 'yesterday' occurs in the cleft.

> a. Léigh an aninteoir leabhar inné.
> read.Past ART.Def.Sg teacher book yesterday
> 'The teacher read a book yesterday.'
b. Is é an múinteoir a léigh

COP.Pres AGR.Masc.Sg ART.Def.Sg teacher COMP.Rel read.Past
leabhar inné.
book yesterday
'It is the teacher who read a book yesterday.'
c. Is inné a léigh an múinteoir leabhar COP.Pres yesterday COMP.Rel read.Past ART.Def.Sg teacher book 'It is yesterday that the teacher read a book.'

In (11b), the aspectual verb phrase ag péinteáil cathaoir 'painting a chair' in (11a) occurs in fronted position. The aspectual phrase here expresses progressive action; see Ramchand (1997) for a discussion of Scottish Gaelic aspectual phrases. It also takes on the function of a predicate of the substantive verb ${ }^{1}$ in the matrix

[^180]sentence (i.e., it functions as the main verb). In (11c), the object NP cathaoir 'a chair' is fronted.
(11) a. Bhí an fear ag péinteáil cathaoir inné. be.Past ART.Def.Sg man ASP paint.VN chair yesterday 'The man was painting a chair yesterday.'
(Ó Siadhail, 1989, p. 236)
b. Is ag péinteáil cathaoir a bhí an fear
COP.Pres ASP paint.VN chair COMP.Rel be.Past ART.Def.Sg man
inné.
yesterday.
?'It is painting a chair that the man was (doing) yesterday.'
(Ó Siadhail, 1989, p. 236)
c. Is cathaoir a bhí an fear ag péinteáil

COP.Pres chair COMP.Rel be.Past ART.Def.Sg man ASP paint.VN inné.
yesterday
'It is a chair that the man was painting yesterday.'
(Ó Siadhail, 1989, p. 236)
(12) and (13) show the possibilities of adjective fronting. In (12b) the predicative adjective te 'hot' is fronted; it is not possible to front an attributive adjective out of an NP, as is shown by (13b).
(12) a. Tá sé fuar.
be.Pres it cold
'It's cold.'
b. (Níl sé fuar.) Is te a-tá sé.
be.Pres.Neg it cold COP.Pres hot COMP.Rel-be.Pres it.
'(It's not cold.) It is hot.' [lit. 'It is hot that it is.']
(13)
a. Léigh sí an leabhar dearg. read.Past she ART.Def.Sg book red 'She read a red book.'
b. * Is dearg a léigh sí an leabhar.

COP.Pres red COMP.Rel read.Past she ART.Def.Sg book

[^181]As is obvious from examples like (11b) and (12b), English is more rigid concerning the choice of the clefted material. While the English equivalent of (11b) is at least questionable, even with the insertion of the verbal residue doing, the literal translation presented in (12b) is completely unacceptable.
(13) shows that, as in English and other languages, it is not possible to front an attributive adjective, while (12) shows that it is possible, on the other hand, to front predicative adjectives. The clefting of adjectives often has a marked connotation as a result, as it is the case in (12b). Both Ó Siadhail (1989) and Stenson (1981) note that an exclamatory or question form in these cases is often preferred, as in (14).
(14) Nach te a-tá sé!

COP.Pres.NegQ hot COMP.Rel-be.Pres it 'Isn't it hot!'

The structure of clefts seems, at first glance, not to have much in common with normal copula constructions. In the next section, however, I argue along the lines of Stenson (1981) that cleft sentences in fact have much in common with copula constructions - especially with identification sentences.

### 3.2 Copula Constructions, Questions, Clefting: Some Similarities

Looking at certain types of copula constructions and constituent questions suggests that there are structural similarities between these and cleft sentences. Consider the case of constituent questions, which can be seen as a device for distinguishing presupposed from new (i.e., focused) information. If someone asks the question in (15), then the anticipated referent of the WH-phrase is in the focus, and the remainder of the question is presupposed (see also Stenson, 1981; Sornicola, 1996).
(15) What cleared the road? (The snowplow cleared the road.)

Interestingly, in Irish, the structure of constituent questions is very similar, if not identical to the structure of simple clefts. Consider (16).
a. Cé a léigh leabhar?
who COMP.Rel read.Past book
'Who read a book?'
b. Is é an múinteoir a léigh COP.Pres AGR.Masc.Sg ART.Def.Sg teacher COMP.Rel read.Past leabhar.
book
'It is the teacher who read a book.'
The syntactic similarity between (16a) and (16b) is obvious. While Ó Siadhail (1989) notes that the copula is mostly dropped in the context of clefts, Stenson (1981) mentions:

Indeed, one of the commonest situations for copula deletion is the clefted answer to a Wh-question, suggesting that the opposition between old and new information that the copula expresses has already been established by the question itself. (Stenson, 1981, p. 108)

So, a more appropriate answer to the question in (16a) would be (17), where the copula form is deleted from the surface. The similarity between the question phrasing and the cleft becomes even more obvious then. ${ }^{2}$
(17) An múinteoir a léigh leabhar.

ART.Def.Sg teacher COMP.Rel read.Past book
'It is the teacher who read a book.'
Finally, and most importantly, consider the similarities between copula sentences identifying two NPs with one another and clefts with an NP as the clefted constituent. If we try to distinguish these structures from one another by simply looking at the surface, we might conclude that simple cleft sentences differ from identification sentences only in that they seem to lack one of the NPs which are identified with each other. (18a) is a copula sentence, equating the two NPs as indicated by the bracketing. (18b) is a cleft sentence.
a. Is é [mo dheartháir] [an fear a
COP.Pres AGR.Masc.Sg my brother ART.Def.Sg man COMP.Rel
bhí tinn inné.]
be.Past sick yesterday
'My brother is the man who was sick yesterday.
b. Is é [an fear][a bhí tinn

COP.Pres AGR.Masc.Sg ART.Def.Sg man COMP.Rel be.Past sick inné.]
yesterday
'It is the man who was sick yesterday.'
Stenson (1981) claims that there is no surface head to the relative clause in cleft sentences such as (18b). Now consider the sentence in (19), which is in fact a pseudo-cleft sentence, using the same type of paraphrase as English pseudo-clefts.
(19) Is é [mo dheartháir] [an té a

COP.Pres AGR.Masc.Sg my brother ART.Def.Sg one COMP.Rel
bhí tinn inné.]
be.Past sick yesterday
'My brother is the one who was sick yesterday.'

[^182]In pseudo-clefts, the relative clause does have a nominal head; therefore, they resemble run-of-the-mill copula sentences such as (18a) even more. In fact, (19) and (18a) are identical in structure. The nominal head inserted in the pseudo-clefts does not add any semantics, but serves as a syntactic head for the relative clause. Stenson (1981) notes that when the nominal head is not overt, then the head of the relative clause in a cleft sentence like (18b) is understood to be referring to a human referent. In other sentences, the nominal head might not refer to a human referent; then, another provisory head noun (e.g., rud 'thing' instead of té 'one') is overt in pseudo-clefts.

The nominal heads in these sentences are understood, even when they do not surface. The conclusion is that cleft sentences such as the one in (18b) are derived from pseudo-clefts such as the one in (19). Since (19) in turn resembles copula sentences like (20a), the structure in (20) is proposed for cleft sentences with an NP in predicate position. Compare this to the structure of the copula sentence in (18a), which is given in (21).

$$
\begin{array}{lll}
\mathrm{COP} & {[\mathrm{NP}]_{\text {predicate }}} & {[\text { relative clause }]_{\text {subject }}}  \tag{20}\\
& & \\
\mathrm{COP} & {[\mathrm{NP}]_{\text {predicate }}} & {[\mathrm{NP}-(\text { relative clause })]_{\text {subject }}}
\end{array}
$$

In cleft sentences, the subject nominal which is overt in pseudo-clefts does not surface, leaving the relative clause alone as a sentential subject to the matrix clause. The subcategorization frame of the copula is filled by the NP in predicate position which is assigned the PREDLINK function, and the relative clause, analyzed as a sentential subject, which is assigned the SUBJ function. An example of the overall syntactic analysis is given in the f-structure in Figure 9 for the sentence in (22). ${ }^{3,4}$

## (22) Is é an múinteoir a léigh

COP.Pres AGR.Masc.Sg ART.Def.Sg teacher COMP.Rel read.Past leabhar.
book
'It is the teacher who read a book.'
The main predicate of the whole sentence is supplied by the copula, which in turn has two positions to be filled in its subcategorization frame: subject (SUBJ) and predicate (PREDLINK). The relative clause, the head of which is léigh 'read', is assigned the topmost SUBJ function, filling one place in the copula's frame - as indicated by the indices on the f-structures. The NP in predicate position is assigned the PREDLINK function, thus filling the other place in the copula's frame, while at the same time supplying the SUBJ function within the relative

[^183]

Figure 8: Cleft sentence with NP in predicate position: f-structure
clause - again, indices show how that the NP in predicate position is mapped to different f-structures: the PREDLINK f-structure and the SUBJ f-structure inside the relative clause.

### 3.3 Variability in Choosing the Clefted Constituent

Note that (21) applies to only those clefts that have an NP in predicate position. It does not explain why other constituents - prepositional phrases, adverbial phrases, adjectives - can also appear in predicate position. I therefore extend Stenson's (1981) view, arguing that not only clefts with NPs in predicate position have the same structure as so-called "identification sentences", but clefts with other constituents in predicate position can be analyzed accordingly. Evidence for this comes from sentences such as in (23). In these examples, the substantive verb bí 'be' - the stage level predicate, to go with Ramchand's (1997) terminology - is used to link adjective phrases/prepositional phrases/adverbials and a noun phrase. Note that although word order is different - the subject comes before the predicate, rather than after it - the main syntactic circumstances in these sentences are the same as in the copula examples: the verb is used to link the predicate and the subject. It is the semantics of this predication that is different from the one in copula sentences (Ramchand, 1997; Stenson, 1981). In (24), I give cleft sentences corresponding to the sentences in (23). The fact that the clefts in (24) are formed from sentences with the substantive verb follows naturally from the differentiation between stage level and individual level predication - predicates like tinn 'sick', faoin mbord 'under the table' or amárach 'tomorrow' cannot be inherent to the subject, but refer to a situation (Ramchand, 1997).
(23) a. Tá mo dheartháir tinn.
be.Pres my brother sick
'My brother is sick.'
b. Tá an cat faoi-n mbord.
be.Pres ART.Def.Sg cat under-ART.Def.Sg table
'The cat is under the table.'
c. Beidh an fhleá amárach.
be.Fut ART.Def.Sg party tomorrow
'The party will be tomorrow.'
(24)
a. Is tinn a-tá mo dheartháir.

COP.Pres sick COMP.Rel-be.Pres my brother
'My brother is sick!' [lit. 'It is sick that my brother is!']
b. Is faoi-n mbord a-tá an cat. COP.Pres under-ART.Def.Sg table COMP-Rel-be.Pres ART.Def.Sg cat 'It is under the table where the cat is.'
c. Is amárach a beidh an fhleá.

COP.Pres tomorrow COMP.Rel be.Fut ART.Def.Sg party
'The party will be tomorrow!' [lit. 'It is tomorrow that the party will be!']

The clefting pattern stays the same across the examples: the relative clause containing the main verb ( $b i$ 'be' in theses cases) follows the sentence-initial copula and the predicate; see (25). The template for simple stage level predication using the substantive verb bí is given in (26).
(25)

COP $[\{\mathrm{AP}|\mathrm{PP}| \mathrm{ADVP}\}]_{\text {predicate }} \quad[\text { relative clause }]_{\text {subject }}$
(26) bí $[\mathrm{NP}]_{\text {subject }}[\{\mathrm{AP}|\mathrm{PP}| \mathrm{ADVP}\}]_{\text {predicate }}$

Below I give two sample analyses that show how the PREDLINK analysis can easily be extended to deal with stage level predication in (27) and a clefted version of the same sentence in (28). The corresponding f-structures are given in Figures 9 and 10 .
(27) Beidh an fhleá amárach.
be.Fut ART.Def.Sg party tomorrow
'The party will be tomorrow.'

```
"Beidh an fhleá amárach."
```



Figure 9: Stage level predication with ADVP as predicate: f-structure
(28) Is amárach a beidh an fhleá.

COP.Pres tomorrow COMP.Rel be.Fut ART.Def.Sg party
'The party will be tomorrow!' [lit. 'It is tomorrow that the party will be!']


Figure 10: Cleft sentence with ADVP in predicate position: f-structure

Having established the variability in the choice of the constituent appearing in predicate position, the clefting pattern can be modified from (22) and (27) to (29).

$$
\begin{equation*}
\text { COP } \left.[\mathrm{XP}]_{\text {predicate }} \text { [relative clause }\right]_{\text {subject }} \tag{29}
\end{equation*}
$$

I conclude that syntactically, copula predication in simple copula sentences such as (20a) on the one hand and stage level predication in sentences such as (27) on the other hand are not different from cleft sentences such as (20b) or (28). A single syntactic analysis - the PREDLINK analysis - has been shown to be able to account for both constructions in a straightforward way. The advantage of such an analysis is that it is kept in parallel for constructions which involve the same type of syntactic processes, demonstrating the universal applicability of the PREDLINK analysis, as mentioned by Attia (2008) and Butt et al. (1999).

## 4 Information-Structure and Clefting

Much recent work in LFG has focused on integrating discourse functions like topic and focus in the grammar architecture (e.g., King, 1997; Butt and King, 1998; O'Connor, 2004; King and Zaenen, 2004; Andréasson, 2007). Discourse functions can be used to encode extra-syntactic cues within a sentence - cues indicating how the information in the sentence is structured and fits inside a speaker-hearer dialogue. Discourse functions (DF) hence encode Information-Structure (IS), providing an extra level of markup, which can be useful in further semantic processing or other applications that need to access IS, such as anaphora resolution or summarization (e.g., King and Zaenen, 2004).

Traditionally, DFs have been encoded directly in the f-structure (e.g., Bresnan and Mchombo, 1987; Bresnan, 2001). Since, however, there may be mismatches between grammatical functions and DFs, information structure can also be encoded using a separate projection in LFG, called i-structure (e.g., Butt and King, 1998; King and Zaenen, 2004; O'Connor, 2004; Andréasson, 2007) and mapped on top of c -structure. In the remainder of the paper, I present my ideas on how an analysis of cleft sentences can be extended to encode information structure; I use the istructure projection to do so. The discourse functions I use for the purpose of this
paper, i.e., for describing information-structure within simple Irish clefts, are based on Vallduví (1993a). A simple two-way distinction for structuring information is proposed - focus vs. ground. Focus is the "informative, newsy, dominant, or contrary-to-expection" part of the sentence, while ground is the "noninformative, known, or expected" part of the sentence (Vallduví, 1993a). In a cleft sentence, the "newsy" focus part of the information appears in the cleft as the predicate, while the "known" ground part of the information appears in the relative clause this generalization holds cross-linguistically (Halvorsen, 1977; Declerck, 1988). Figure 11 presents such a mapping for the cleft sentence in (28).

$$
\left[\begin{array}{l}
\text { FOC }\{\text { amárach 'tomorrow' }\} \\
\text { GROUND }\{\text { fhleá 'party' }\}
\end{array}\right]
$$

Figure 11: Cleft sentence with focus NP: i-structure
In cases like this, it is easy to assign focus, as there is only one candidate PRED for focus in each sentence - the adverbial phrase amárach 'tomorrow'. In other cases, the mapping between c- and i-structure is not as straightforward, and the focus-ground-division is not as easy. The clefted constituent may contain multiple PREDs and/or contain elements with contrastive stress. In such cases, elements that are part of the clefted constituent, but do not carry contrastive stress are analyzed as being part of the ground (Vallduví, 1993b); see also Halvorsen (1977) and Declerck (1988). (30) is an example.
(30) Is faoi-n mbord a bhí an cat.

COP.Pres under-ART.Def.Sg table COMP.Rel be.Past ART.Def.Sg cat 'It is under the table where the cat was.'

From the c-structure of this sentence, three different i-structures can be projected, since it is not clear whether the whole PP faoin mbord 'under the table' or just the preposition faoin 'under' or just the preposition's object mbord 'table' should be assigned the DF FOC. See Figures 12-14.

$$
\left[\begin{array}{l}
\text { FOC }\{\text { faoin mbord 'under the table' }\} \\
\text { GROUND }\{\text { cat 'cat' }\}
\end{array}\right]
$$

Figure 12: Cleft sentence with focus PP: i-structure

$$
\left[\begin{array}{l}
\text { FOC }\{\text { faoin 'under' }\} \\
\text { GROUND }\{\text { cat 'cat', mbord 'table' }\}
\end{array}\right]
$$

Figure 13: Cleft sentence with focus PP: i-structure

$$
\left[\begin{array}{l}
\text { FOC }\{m b o r d \text { 'table' }\} \\
\operatorname{GROUND}\{\text { cat 'cat', faoin 'under' }\} \text { ' }
\end{array}\right.
$$

Figure 14: Cleft sentence with focus PP: i-structure

Here, syntax clearly reaches its limits, and prosodic phonology takes over. Clefting, as a syntactic focusing device, is only able to select the whole PP as the clefted constituent. It is not possible to have just the preposition or the preposition's object in this position. So if a speaker intends to focus just these, they have to employ contrastive focus, which is a prosodic device. Such prosodic cues are, however, not normally present in written text - which is why a grammar operating on written text can and should merely produce multiple i-structures in cases like (30). Note in this respect the work by Bögel et al. (2009), who have experimented with parsing prosodic information using a computational LFG grammar. Given the appropriate markup in the text, such a grammar could be used to disambiguate between structures such as Figures 12-14.

## 5 Conclusion

The paper has presented a syntactic analysis of Irish clefting. It started out by giving alternatives for analyzing copula constructions in LFG and showed why the PREDLINK analysis is a favorable one for Irish. The paper discussed data from Irish clefting and the similarities between simple copula predication and stage level predication using the substantive verb $b i$ as well as clefting. It was argued that in fact one syntactic analysis can be used for both types of constructions. Section 4 extended the analysis to include Information Structure by encoding discourse functions in a separate projection of LFG, i-structure.

The similarities between run-of-the-mill copula sentences and clefts strongly suggest that a single analysis should be chosen, since the syntactic predication
is identical across constructions. With the approach presented in this paper, one does not have to rely on different syntactic analysis - the PREDLINK analysis models the parallelism. Another advantage of the proposed analysis is that it can be extended to other languages in a straightforward way, since several languages use copula predication in cleft sentences.

## References

Andréasson, Maia. 2007. The Architecture of I-Structure. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG07 Conference, CSLI Publications.

Attia, Mohammed. 2008. A Unified Analysis of Copula Constructions in LFG. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG08 Conference, CSLI Publications.

Bögel, Tina, Butt, Miriam, Kaplan, Ronald M., King, Tracy Holloway and Maxwell III, John T. 2009. A treatment of clitics via prosodic phonology in LFG. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG09 Conference.

Bresnan, Joan. 1982. Control and Complementation. In Joan Bresnan (ed.), The Mental Representation of Grammatical Relations, MIT Press.

Bresnan, Joan. 2001. Lexical-Functional Syntax. Blackwell Publishing.
Bresnan, Joan and Mchombo, Sam A. 1987. Topic, pronoun, and agreement in Chicheŵa. In William Bright (ed.), Language, volume 63, Waverly Press.

Butt, Miriam and King, Tracy Holloway. 1998. Interfacing Phonology with LFG. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG98 Conference, Stanford: CSLI Publications.

Butt, Miriam, King, Tracy Holloway, Niño, María-Eugenia and Segond, Frédérique. 1999. A Grammar Writer's Cookbook. CSLI Publications.

Carnie, Andrew. 1995. Non-verbal predication and head movement. Ph. D.thesis, MIT.

Crouch, Dick, Dalrymple, Mary, Kaplan, Ronald M., King, Tracy Holloway, Maxwell III, John T. and Newman, Paula. 2008. XLE Documentation. Palo Alto Research Center.

Dalrymple, Mary, Dyvik, Helge and King, Tracy Holloway. 2004. Copular Complements: Closed or Open? In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG04 Conference.

Declerck, Renaat. 1988. Studies on Copular Sentences, Clefts and Pseudo-Clefts. Leuven University Press.

Halvorsen, Per-Kristian Giaever. 1977. Syntax and Semantics of Cleft Constructions. Ph. D.thesis, University of Texas at Austin.

King, Tracy Holloway. 1997. Focus Domains and Information-Structure. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG97 Conference, Stanford: CSLI Publications.

King, Tracy Holloway and Zaenen, Annie. 2004. F-structures, Information Structure, and Discourse Structure. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG04 Conference.

Lodrup, Helge. 2008. Functional Structure. In Robert D. Borsey and Kersti Börjars (eds.), Nontransformational Syntax: A Guide to Current Models, Blackwell Publishing, to appear.

O'Connor, Robert. 2004. Information Structure in Lexical-Functional Grammar: The Discourse-Prosody Correspondence in English and Serbo-Croatian. Ph. D.thesis, University of Manchester.

Ó Siadhail, Mícheál. 1989. Modern Irish - Grammatical Structure and Dialectal Variation. Cambridge University Press.

Ramchand, Gillian Catriona. 1997. Aspect and Predication. Oxford University Press.

Rosén, Victoria. 1996. The LFG Architecture and "Verbless" Syntactic Contructions. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG96 Conference.

Sornicola, Rosanna. 1996. Topic, Focus, and Word Order. In Keith Brown and Jim Miller (eds.), Concise Encyclopedia of Syntactic Theories, Elsevier.

Stenson, Nancy. 1981. Studies in Irish syntax. Gunter Narr Verlag.
The Christian Brothers. 1960. Graiméar Gaeilge na mBräithre Críostaí (The Christian Brothers' Irish Grammar). M. H. Mac an Ghoill agus a mhac Tta.

Vallduví, Enric. 1993a. Information packaging: A survey. Technical Report, Centre for Cognitive Science \& Human Communication Research Centre.

Vallduví, Enric. 1993b. The Informational Component. Ph. D.thesis, University of Pennsylvania at Philadelphia.

# PARSING ARABIC USING TREEBANK-BASED LFG RESOURCES 

Lamia Tounsi Mohammed Attia Josef van Genabith NCLT, School of Computing<br>Dublin City University

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009
CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In this paper we present initial results on parsing Arabic using treebank-based parsers and automatic LFG f-structure annotation methodologies. The Arabic Annotation Algorithm ( $\mathrm{A}^{3}$ ) (Tounsi et al., 2009) exploits the rich functional annotations in the Penn Arabic Treebank (ATB) (Bies and Maamouri, 2003; Maamouri and Bies, 2004) to assign LFG f-structure equations to trees. For parsing, we modify Bikel's (2004) parser to learn ATB functional tags and merge phrasal categories with functional tags in the training data. Functional tags in parser output trees are then "unmasked" and available to $\mathrm{A}^{3}$ to assign f-structure equations. We evaluate the resulting f-structures against the DCU250 Arabic gold standard dependency bank (Al-Raheb et al., 2006). Currently we achieve a dependency f-score of $77 \%$.


## 1 Related Work

Arabic parsing systems have been reported in (Ditters, 2001; Zabokrtsky and Smrz, 2003; Othman et al., 2003; Ramsay et Mansour, 2007). (Attia, 2008) gives an overview of an LFG rule-based analysis of Arabic using XLE (Xerox Linguistics Environment). He concentrated on short sentences and used robustness techniques to increase the coverage. All of these use hand-crafted grammars, which are time-consuming to produce and difficult to scale to unrestricted data. More recently, the Penn Arabic Treebank (ATB) has been employed to acquire wide-coverage parsing resources. The best-known Arabic statistical parser was developed by Bikel (Bikel, 2004). Bikel reports parse quality "far below" English and Chinese (Kulick et al., 2006). The main reasons cited were a significant number of POS-tag inconsistencies (in the version of the ATB available at the time) and the considerable differences between Arabic and English sentence structure. (Dieb et al., 2004) and (Habash and Rambow, 2005) present knowledge- and machine-learning-based methods for tokenisation, basic POS tagging with a reduced tagset and base phrase chunking. Bikel's parser produces phrase-structure trees (c-structures). The main objective of our research is to automatically enrich the output of Bikel's parser with more abstract and "deep" dependency information (in the form of LFG f-structures), using the Arabic A ${ }^{3}$ annotation algorithm (Tounsi et al., 2009), extending the approach of (Cahill et al., 2004), originally developed for English.

## 2 The Penn Arabic Treebank (ATB)

Arabic is a subject pro-drop language. It has relatively free word order: mainly S (ubject) V (erb) and O (bject), with VSO and VOS also possible. Arabic is a highly inflectional and cliticizing language. The ATB consists of 23,611 parse-annotated sentences (Bies and Maamouri, 2003; Maamouri and Bies, 2004) from Arabic newswire text in Modern Standard Arabic (MSA). The ATB annotation scheme involves 497 different POS-tags with morphological information (reduced to 24 basic POS-tags by Bikel e.g. NN, NNS, JJ), 22 phrasal tags e.g. NP, VP, PP and 20 functional tags e.g. SBJ, OBJ, TPC (52 combined functional tags, as functional tags can stack).

## 3 The Arabic Annotation Algorithm ( $\mathbf{A}^{3}$ )

The Arabic Annotation Algorithm (Tounsi et al., 2009) is constructed adapting and revising the methodology of (Cahill et al., 2004) for English as follows:

1. Automatic extraction of the most frequent rule types from the treebank ${ }^{1}$.
2. Head lexicalisation of ATB trees to identify local heads.
3. Default f -structure equations are assigned to ATB functional tags. In addition, lexical macros exploits the rich morphological information provided by the ATB.

[^184]4. Left/right annotation principles for COMPs, XCOMPs, ADJUNCTs, etc ${ }^{2}$.
5. Coordination
6. Traces to handle non-local dependencies.
(Tounsi et al., 2009) report an f-score of $95 \%$ on automatically annotated gold ATB trees against the DCU250 Arabic Dependency Bank.

## 4 Adapting the Parser

We use Bikel's implementation of Collins' Model 1 as our c-structure engine (Bikel, 2004). As the $\mathrm{A}^{3}$ of (Tounsi et al., 2009) heavily relies on ATB function tags, we modify the Bikel parser to learn ATB tags. We "mask" ATB function tags in the training data by merging phrasal with function tags and adjust the head-finding rules in Bikel's Arabic language pack accordingly. For example, the functional and phrasal tag NP-OBJ are stuck together as NP_OBJ which makes the shallow parser interpret it as one phrasal tag during training and parsing (NP-OBJ $\Rightarrow$ NP_OBJ). After parsing, we unmask ATB function tags and make them available to $\mathrm{A}^{3}$.

## 5 Experiments and Evaluation

250 of the 23,611 parse-annotated sentences in ATB were randomly selected as test set (Dieb et al., 2004). The DCU 250 gold standard dependency bank for Arabic (Al-Raheb et al., 2006) is semiautomatically constructed using $\mathrm{A}^{3}$ and manual correction and extension. We use gold-POS-tagged ATB text and the lexical morphological information from ATB in the results reported below:

| Precision | Recall | F-score |
| :---: | :---: | :---: |
| 70.40 | 72.38 | 71.37 |

Table 1: C-structure evaluation (Evalb).

| Precision | Recall | F-score |
| :---: | :---: | :---: |
| 74.75 | 81.07 | 77.78 |

Table 2: F-structure evaluation.

## 6 Discussion and Further Work

Compared to similar results for English which have a dependency f-score of $87 \%$ against DCU 105 (Cahill et al., 2002), initial results (dependency f-score of $77 \%$ ) for Arabic are somewhat disappointing. The most likely reason is the explosion in the size of the phrasal category set with 22 ATB phrasal categories as opposed to 150 (masked) categories (fusing ATB phrasal and functional tags) to be learnt by Bikel's parser, resulting in substantial data-sparseness. However, the result provides a base-line for what, to the best of our knowledge, is the first treebank-based LFG parsing approach to Arabic. In an effort to improve on the baseline presented in this paper, our current experiments use a two-stage architecture with a simple probabilistic phrase-structure parser, followed by a machine-learning-based ATB function labeller, to provide input to $\mathrm{A}^{3}$.

[^185]
## References

Y. Al-Raheb, A. Akrout, J. van Genabith, J. Dichy. 2006. DCU 250 Arabic Dependency Bank: An LFG Gold Standard Resource for the Arabic Penn Treebank The Challenge of Arabic for NLP/MT at the British Computer Society, UK, pp. 105-116.
M. Attia. 2008. Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation. Ph.D. Thesis. The University of Manchester, Manchester, UK.
A. Bies and M. Maamouri. 2003. Penn Arabic Treebank Guidelines URL: http://www.ircs.upenn.edu/arabic/Jan03release/guidelines-TB-1-28-03.pdf.
D. Bikel. 2004. Intricacies of Collins' parsing model Computational Linguistics, 30(4), 2004.
A. Cahill, M. McCarthy, J. van Genabith and A. Way. 2002. Automatic Annotation of the Penn-Treebank with LFG $F$-Structure Information. LREC 2002 workshop on Linguistic Knowledge Acquisition and Representation, pp. 8-15.
A. Cahill, M. Burke, R. ODonovan, J. van Genabith, A. Way. 2004. Long-Distance Dependency Resolution in Automatically Acquired Wide-Coverage PCFG-Based LFG Approximations. Meeting of the Association for Computational Linguistics ACL 2004, pp. 319-326.
M. Diab, K. Hacioglu, D. Jurafsky. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04).
E. Ditters. 2001. A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic. Workshop on Arabic Processing: Status and Prospects at ACL/EACL, Toulouse, France.
N. Habash, O. Rambow. 2005. Arabic Tokenization, Morphological Analysis, and Part-of-Speech Tagging in One Fell Swoop. In Proceedings of the Conference of American Association for Computational Linguistics ACL 2005.
S. Kulick S, G. Ryan, M. Mitchell. 2006. Parsing the Arabic Treebank: Analysis and improvements. In Proceedings of TLT 2006. Treebanks and Linguistic Theories..
M. Maamouri and A. Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004.
E. Othman, K. Shaalan, A. Rafea. 2003. A Chart Parser for Analyzing Modern Standard Arabic Sentence MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches, USA.
A. Ramsay and H. Mansour. 2007. Towards including prosody in a text-to-speech system for modern standard Arabic Computer Speech and Language 22:84-103.
L. Tounsi, M. Attia, J. van Genabith. 2009. Automatic Treebank-Based Acquisition of Arabic LFG Dependency Structures European Chapter of the Association for Computational Linguistics, EACL 2009, Workshop Computational Approaches to Semitic Languages, pp 45-52.
Z. Zabokrtsky, O. Smrz. 2003. Arabic syntactic trees: from constituency to dependency European Chapter of the Association for Computational Linguistics EACL 2003, pp. 183-186.

# LFG AND DYNAMIC SYNTAX: TWO NON-DERIVATIONAL THEORIES 

Nigel Vincent

The University of Manchester

Proceedings of the LFG09 Conference

Miriam Butt and Tracy Holloway King (Editors)
2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

This paper provides a brief report of the workshop convened and chaired at LFG09 by the author under the title 'Blurring Component Boundaries: Levels of analysis or growth of information?'. The purpose of the workshop was to introduce the LFG community to the system developed by Ruth Kempson and a number of co-workers under the name Dynamic Syntax (DS), and to promote discussion and comparison of LFG and DS and the thinking that lies behind them. The paper explains the theme of the workshop, summarises some of the points made in the presentations, only one of which is published here in full, and comments briefly on some issues that emerged in or arose from the discussion.


## 1. Introduction ${ }^{1}$

One of the (many) highlights of LFG 2009 was a workshop entitled 'Blurring Component Boundaries: Levels of analysis or growth of information?', which aimed to promote comparison of and interaction between LFG and the system recently developed by Ruth Kempson and a

[^186]number of co-workers known as Dynamic Syntax (DS). ${ }^{2}$ The full programme of the workshop included the following papers, succeeded by a period of general discussion:

Louise Mycock "What do you do?" Variation in interrogative predicates
Ruth Kempson \& Jieun Kiaer Narrowing the competenceperformance gap: Syntax as time-linear growth of semantic representation
Miriam Bouzouita \& Stergios Chatzikyriakidis Clitics as calcified processing strategies: The case study of Spanish clitic placement and the PCC as a tree-logic restriction
Joan Bresnan The dynamics of syntax: Implications for LFG
The contributors were chosen so as to provide a balance both in terms of approach (Bresnan and Mycock for LFG and Kempson \& Kiaer and Bouzouita \& Chatzikyriakidis for DS) and experience (Bresnan and Kempson are senior scholars whose work has been fundamental in developing the respective systems while the others are recent PhDs: Kiaer (2007), Bouzouita (2008), Chatzikyriakidis (in prep.) at King's College London with Kempson and Mycock (2006) at Manchester with Vincent). Unfortunately, for a variety of reasons, only the paper by Bouzouita \& Chatzikyriakidis is published in full in these proceedings, while the others are represented by their abstracts.

In the account of the workshop that follows I will explain the thinking behind convening it, summarise some of the points made in the presentations and comment briefly on some issues that emerged in the discussion.

[^187]
## 2. The theme

Our idea in organizing the workshop was that it would be of interest to compare LFG and DS as systems or architectures which share a commitment to both non-derivationality and formalization. At the same time we felt that a simple point-by-point comparison of the two approaches would have taken up more time than was available within the confines of a half-day workshop, and in any case was not necessarily the most illuminating way to proceed. We therefore sought to identify a theme that would highlight the differences and similarities while also advancing general linguistic debate at a level beyond the parochialities, necessary but sometimes overly constraining, of individual theories and notations. The eventual theme was proposed by Ruth Kempson since, as she noted, the process of theory construction and elaboration inevitably involves drawing distinctions and creating boundaries, which other theories may feel the need to break down. The boundaries here are of two kinds, which we may for convenience dub 'internal' and 'external'. By 'internal' I mean the relation between the components or levels within a given theory - within LFG for example the decision as to whether certain information is best represented in c-structure or f-structure or s-structure. Internal issues in this sense are particularly characteristic of a parallel correspondence architecture such as that of LFG. 'External' refers to the way the contents of the grammar relate to other parts of the language processing and production systems or to the types and sources of linguistic data. This kind of boundary has been of especial significance in the single level structure of DS which brings together in its formal metalanguage properties traditionally associated with the grammar (or competence) and with the parser (i.e. related to performance). Boundaries of all kinds were addressed in the workshop.

## 3. The papers

In this section I will briefly review the papers, seeking to relate them, as appropriate, to the external and internal interpretations of the theme. Louise Mycock's paper introduces a new class of data into the extensive theoretical discussion of interrogative constructions, namely languages in
which a single, and in the limiting case synchronically unanalysable, verb expresses the semantic content of English what happened? or what did you do (to $X$ )? (for the relevant data see the survey in Hagège 2008). Whereas interrogative constructions are usually analysed through a combination of c-structure and f-structure properties, Mycock seeks to show that the best analysis for this new data by-passes f-structure and relies instead on a combination of i -structure and s -structure, thus challenging standard assumptions about the interaction of levels within LFG, and more generally pushing the boundaries of our understanding when it comes to the analysis of the full range of cross-linguistically available interrogative constructions. The paper by Kempson \& Kiaer, the full version of which is now in press (Kempson \& Kiaer 2010), deals with the more widely studied body of data brought to light by (multiple) scrambling and long distance dependency patterns in Japanese and Korean and argues for an account in which the grammar directly reflects the constraints and needs of the human language processing system. In this sense, as their title indicates, it represents a move in the direction of narrowing the gap between performance and competence, an external boundary in our terms, and thus views natural language structures as constrained, and explained, by the dictates of performance. This explanatory strategy is built directly into the formalism of DS and is accordingly characteristic of all work within this framework. The insight that is pursued here is akin to that developed over a number of years by John Hawkins (see for example Hawkins 2004 and references there).

The same logic of explanation is explored in relation to changes in pronominal systems by Bouzouita \& Chatzikyriakidis, thereby probing and potentially blurring - the traditionally clear boundary between synchrony and diachrony. They investigate two connected phenomena, namely first the emergence in the history of Spanish of morphosyntactically fixed clitic combinations from the pragmatically conditioned distribution of the cognate items in Latin, and second the Person Case Constraint whereby accusative and dative clusters involving one or more first or second person pronouns are blocked or restricted in their distribution. Both of these are argued to follow from a general condition of DS that allows the parser to hold only one item unplaced (in their terms an 'unfixed node') at a time. The diachronic change is that the pragmatic principles of pronoun choice are resolved through the freezing
of the relevant information in complex lexical entries by a process named 'routinization'. Bouzouita \& Chatzikyriakidis thereby offer a new take on the kind of dataset that has figured extensively in the literature on grammaticalization, in which pragmatics, traditionally an aspect of performance, becomes grammar, or competence. That literature has tended to eschew formal approaches, though as noted in Vincent \& Börjars (forthcoming) there is no fundamental conflict between formal methods and the evidence of grammaticalization (see also $\S 4.5$ below). Interestingly, too, the account of case involved here (and also in Cann \& Kempson 2008) draws heavily on the theory of constructive case developed within LFG by Nordlinger (1998).

Whereas work in DS emphasises the processing dimension, Joan Bresnan's contribution focussed on evidence, both corpus-based and experimental, demonstrating how the requirements of incremental production influence linguistic structure and preferences. The presentation drew on the data of English dative and genitive alternations, and the circumstances of production which favour give Mary the money over give the money to Mary, or the woman's shadow over the shadow of the woman. Bresnan concluded her paper with some reflections on what the data she had discussed imply for LFG. She noted the openness of the LFG architecture to various developments such as competition-based (OT) and stochastic interpretations, the latter in particular allowing for a move towards a different part of the functionalist community than the processing type of explanation favoured by DS and by Hawkins. She thus opens the door to probabilistic models of grammar, including Data-Oriented Parsing, which have traditionally been eschewed within most if not all formalist traditions. The competence/performance boundary is once again under challenge, though from a different direction.

## 4. Some issues

In this section I focus on some of the issues that emerged from the workshop. It goes without saying that this is a personal take on the occasion and others may well have differing interpretations. I hope, however, by formulating the issues in a general way to provide ground for further conversations and debates of this kind.

### 4.1 The data

One topic that took up a good deal of the discussion time concerned the nature of linguistic data. Of note in this connection was Bresnan's reliance on corpus and attested examples or data elicited under controlled experimental conditions as opposed to the traditional appeal to the native speaker's intuition. Lively debate arose from her suggestion that once preference is given to such data the nature of models necessarily changes. Her opening of the door to the exemplar-based approach (see for example Bod 2009 and references there) represents in many ways a more radical divergence from the traditional view of the relation between theory and data than anything else in the domain of theoretical syntax, DS included.

Two further data-related points that emerge from the other papers are the fruitfulness of building on typologically inspired research in developing theoretical issues (Mycock) and the need for careful attention to the accuracy and reliability of historical evidence in formal as well as in philological work (Bouzouita 2008).

### 4.2 Theories, architectures and programs

Inevitably, given the workshop's conception, attention focussed on a number of points of detail about differences and similarities between DS and LFG as theories of natural language (morpho)syntax and semantics. A larger issue that hangs behind such discussions is what it means to talk about a theory and to compare one theory with another. LFG is regularly referred to as a theory but it is also common to say that LFG provides an 'architecture' for grammar (cf. Bresnan 2001, the first two parts of which are entitled 'On the architecture of universal grammar' and 'Formally modelling the architecture'), and Bresnan's presentation alluded at various points to the 'non-procedural architectural design of LFG' (quotation from her workshop abstract). What, we may then ask, is the difference between an architecture and a theory? A third term to add to the mix here is 'program' (the choice of the American spelling in this context is deliberate!). Advocates of Minimalism are especially insistent that what
they are pursuing is a program and not a theory (Chomsky 1995, Boeckx 2008: 3-4). Again the question arises: is an architecture different from a program, and if so, how?

In this connection, Hornstein (2009: 15) writes: 'There are many analyses that fly under the minimalist flag and many different ways of understanding the goals of the program, often embodied in different technologies.' The same could be said of LFG. To take an instance from the workshop, Mycock's analysis of interrogative predicates shifts the burden of accounting for these constructions and unifying them with other modes of interrogation from c-/f-structure to i -/s-structure, yet either account is clearly consistent with the overall parallel correspondence architecture of LFG and both differ materially from any derivationalist version.

One positive answer to the question about the difference between a program and an architecture is that a program has a particular conceptual goal or ambition which guides the types of analysis that are formulated. An architecture on the other hand defines a conceptual space within which analyses and goals may be formulated, but does not constrain those who inhabit that space to a single vision of what they are doing there. In the case at hand, LFG does not enforce or endorse a particular understanding of how language relates to the mind in the way that Minimalism does. A less charitable answer would be that a program - or at least the Minimalist Program! - is vaguer and more open to inconsistencies since practitioners have freedom to redefine almost at will crucial concepts and constructs (Hornstein's 'different technologies'). In these terms, DS appears more like a program since it has an overall vision of where and how to locate explanations for linguistic phenomena and all the analyses so far published, whether of clitics, scrambling or whatever, tend towards the same general processing-related conclusion. Unlike Minimalism, however, the completeness and consistency of its formal definition make it harder if not impossible for individual researchers to use the same metalanguage but mean different things by it, as all too often happens to notions like 'Case', 'Agree' and the like within Minimalist writings.

The term 'theory', by contrast, is capable of a wide or a narrow use. We can, and people frequently do, speak of overarching systems and notations for morphosyntactic analysis like LFG and DS as theories. But we also talk for instance of Bouzouita's theory of Spanish clitic formation
or Mycock's theory of long-distance dependencies, where what we mean is an account which models the data in terms of a set of formal constructs that are drawn from and depend on the theory in the first sense. The key notions here are 'model' and 'construct', which are fundamental to scientific explanation of any kind (cf. the quotation from Smolensky \& Dupoux in section 4.6 below).

### 4.3 Derivational vs. non-derivational

The relative merits of derivational and non-derivational theories (in the broad sense) of the structure of natural language are regularly debated (see Johnson \& Lappin 1997 for excellent discussion and further references, and Sag \& Wasow, forthcoming). There is also a long tradition from the derivational side of dismissing other models as notational equivalents (see already Chomsky 1970/76: 69ff). It is by contrast unusual to find explicit comparison of different non-derivational approaches such as emerged in the papers from the workshop and the ensuing discussion. Suffice it to say that here we will take the arguments against derivational or multi-stratal approaches for granted, noting only that this debate is not to be confused with the formalist vs. functionalist debate (cf. §4.5).

### 4.4 Theory reduction

In a variant of the notational equivalence argument, Hornstein (2009) at various points treats a range of theories as essentially the same and as equivalent to GB. Thus, he writes (cf. also his discussion on pp. 155ff):

I say 'GB style' for I include in this GB's cousins including LFG, GPSG, HPSG and RG. Though the particulars of GB are what I concentrate on, all the above mentioned approaches cut grammars along more or less the same joints. (Hornstein 2009: viii, note 2)
Minimalism, by contrast, is argued to be a step beyond all of these, GB included, in its capacity for generalization and insight. And in Hornstein's version the core construct of Minimalism is taken, not unreasonably, to be Move, since Merge, understood as a variety of concatenation, is simply the default operation for composing parts into larger syntactic entities. What is
at stake here is the more general philosophical question of theory reduction, that is to say the strategy by which a more particular theory is subsumed, and hence explained by, a more general one. This is a longstanding matter of debate in psychology, where the concern has been whether psychological explanations always reduce to biophysical ones. ${ }^{3}$

This issue - not, it has to be said, explicitly discussed in our workshop - arises in connection with a comparison between LFG and DS, since LFG as traditionally understood is vulnerable to Hornstein's argument that Minimalism operates at a higher - and by implication more explanatory - level of abstraction than other approaches in a way that DS, with its basis in processing, is not. Put another way, a system that overtly links itself to external constraints, whether due to processing or production, is able to anchor itself against the winds of reductionism which can buffet the free-standing, speaker-hearer neutral, architecture (as opposed to program) that LFG traditionally is. In this sense, as Bresnan reminded us in her presentation, LFG should not forget its roots in the search for a psychologically realistic mode of syntactic description (cf. the papers in Part III of Bresnan 1982).

### 4.5 Formalism and functionalism

One often discussed issue that the juxtaposition of LFG and DS brings into new relief is the contrast between formalist and functionalist approaches to the description and explanation of natural language phenomena. This is commonly treated as a contrast akin to that between political parties: someone is thought to be either a functionalist or a formalist just as they might be either a Democrat or a Republican or vote Labour or Conservative. Indeed it is sometimes even assumed that individuals

[^188]read/write for different journals according to their stance on this issue. Thus, Croft (2007: 411), discussing a paper by Bas Aarts, writes:
"I hope that Aarts will succeed in bringing this fact [that there is variation in grammatical categorization: NV] to greater prominence in the formalist research tradition. However, the audience of this journal [Studies in Language: NV] is largely functionalist ..."
Yet DS in particular is avowedly functionalist in inspiration while meeting the highest standards of formal completeness and consistency. LFG, by contrast, in keeping with its neutrality in relation to processing/production and its respect for the traditional performance/competence distinction (and despite the use of the word 'functional' in two other senses!), has certainly been formally explicit but has not historically been committed to any position on functionalist explanations for natural language phenomena. In this respect, Bresnan's work over the last few years, outlined and summarised in her presentation at the workshop, has marked a departure from LFG orthodoxy. There are, it is true, hints in this direction already in the reference in Bresnan (2001: 92) to a 'principle of functionality of cstructure' and an allusion to Haiman's work on the economy of expression, but it is in the papers on the stochastic implementation of the model in more recent years that this line of work is most developed. To judge by the exchanges in the workshop, this is still a contentious issue within LFG. ${ }^{4}$

### 4.6 The role of Universal Grammar (UG)

Whereas the Chomskyan tradition has always firmly adopted the formalist stance that true scientific explanation derives from within a theoretical edifice via the postulation of principles of increasing scope and generality (cf. our discussion of Hornstein 2009 above), LFG has remained more agnostic. As we have noted, its origin lies in an attempt to develop a

[^189]psychologically more real (and realistic) account of natural language syntax, and it comes with no innatist baggage. It has therefore always been as much concerned with 'external' issues of computational implementability as with 'internal', reductionist modes of explanation. The tensions between the two are sidestepped within Chomskyan accounts through on the one hand the adoption of a strongly realist stance on the relation between theories and the objects they purport to describe and explain and on the other the postulation of an object - Universal Grammar (UG) - which is inaccessible to independent observation, with the attendant risk of falling into vicious circularity.

The status of universals and UG takes on renewed relevance in the context of Evans \& Levinson's (2009) polemical target article in Behavioral and Brain Sciences (henceforth E\&L) and the extensive discussion to which it has given rise. ${ }^{5}$ There are, as it happens, no proponents of LFG or DS within the published discussants of E\&L, but the issues addressed are of a piece with those that emerged in the course of our workshop and which I have tried to sketch here. One of those discussants, Mike Tomasello, unequivocally entitles his contribution 'Universal Grammar is dead' and writes: 'To make progress in understanding human linguistic competence, cognitive scientists must abandon the idea of an innate universal grammar and instead try to build theories that explain both linguistic universals and diversity and how they emerge.' (2009: 470). ${ }^{6}$ I will conclude this paper by briefly summarising E\&L and the main points that arise in the discussion before suggesting how research in LFG and DS can respond to Tomasello's exhortation. In so doing I am moving things on from issues that explicitly arose in our workshop in the belief a) that there are many common threads between the discussions in Cambridge and those that appear in the pages of $B B S$, and b ) that it is of value to link work

[^190]within more specialized research communities such as LFG and DS to these broader debates.

The essence of E\&L's argument is that the case for exceptionless linguistic universals has been hugely overstated, and that in consequence there are no grounds to postulate an innate, modular and autonomous UG to account for them. Rather, what are attested are recurrent statistical patterns which are to be explained through a combination of the general properties of human cognition and the particular circumstances of culturalhistorical change. They discuss a wide range of claimed universals, including both substantive universals like CV syllable structure and a basic categorial distinction between nouns and verbs, and formal universals such as subjacency and the principles of Binding Theory, and show that in all cases there are counterexamples to any claim for absolute universality. In place of a Chomskyan innate UG they postulate 'an evolutionary model with attractors (e.g. the CV syllable, a color term "red", a word for "arm"), "canals", and numerous local peaks or troughs in an adaptive landscape. Some of the attractors are cognitive, some functional (communicational), some cultural-historical in nature' (2009: 446). Perhaps predictably, the responses oscillate between complete agreement (cf. Tomasello's remarks quoted above) and haughty dismissal as when Friedin (2009: 455) concludes his response thus: 'Data alone cannot speak to the validity of explicit proposals about the content of UG. What is required is an explicit analysis of data that follows from a precisely formulated fragment of a grammar ... The discussion of UG in this article misses the mark entirely.'

There is not room to go into detail here on the range of arguments and examples the authors and their respondents provide, but I would note two things. First, in their contribution Smolensky \& Dupoux distinguish two types of what they call $\operatorname{cog}$ (nitive)-universals: architectural and specific universals, a distinction which relates to the discussion in section 4.2 above about the nature of an architecture. They argue that 'architectural universals do not yield falsifiable predictions regarding typology, but they yield falsifiable predictions regarding language learnability ... specific universals are tied to particular formal theories specifying in detail the architecture's levels, structures, and operations, thus yielding falsifiable predictions regarding language typology' (2009: 468).

Second, at various points in their paper E\&L allude to and compliment LFG as being a model which permits formally testable claims, is responsive to typological diversity, does not involve the postulation of considerable amounts of empty structure, and allows for both constituency and dependency relations to be expressed. DS, perhaps understandably since it is less widely known and discussed, does not get a mention. Yet there are clearly aspects of the DS stance on the nature of grammar and the way it can change over time which are also consistent with both E\&L's take on the relation between cognition and culture and Smolensky \& Dupoux's underscoring of the need for formal, falsifiable theories. It is of some interest, moreover, that both LFG and DS are able to express on the one hand formal universals indicative of the architecture of grammar, and on the other hand the variability intrinsic to words as reflections of the diachronic changes that have given rise to their particular form and/or interpretation. This confirms once again that formal and functional generalisations do not have to be seen as being in conflict with each other. There is a profound misunderstanding of the role of language change evident in Nevins's (2009: 461) observation that 'integration with the cognitive sciences ... will come from mechanistic explanations, not from handwaving at diachronic contingencies'. There is no more room for handwaving in diachronic linguistics than there is in synchronic work, but to ignore the evidence of change is to discard much of what makes language language.

In short, models like LFG and DS can only gain from E\&L's refocussing of the nature of the debate towards the interaction of linguistic structure, cognition and history and away from an obsession with an innate but untestable UG. The door is open for researchers from within these communities to establish even more strongly than hitherto the relevance and importance of their research on the international scene.

## 5. Conclusions

The first conclusion, to judge by the number of questions and contributors to the discussion period and by informal comments afterwards, is that the workshop was certainly a success. This in turn suggests that further systematic comparison between the assumptions and consequences of
work within LFG and those of other frameworks might be fruitful. There has, it is true, been some work seeking to compare LFG and HPSG. For example, in 2000 the annual conferences of the two groups were held back-to-back with a day of overlap devoted to topics of common interest, but there is certainly room for more such events. ${ }^{7}$ It is an interesting and disappointing reflection of the sociology of the field that the little work that exists on comparing systems tends to lie at one of two extremes. On the one hand there is a long tradition of research into the mathematical power of grammars stretching back to Chomsky's seminal work in the 1950's. On the other there are informal comparisons that arise en passant while the main focus of attention and thrust of the argument lies in another direction. Thus, many papers at LFG conferences and elsewhere depart from a dataset or a theoretical point drawn from the large body of literature that simply takes derivationality (formerly in its GB guise and now in a Minimalist one ${ }^{8}$ ) for granted. Such papers implicitly accord Minimalism the status of the yardstick by which other work should be judged, whereas in fact it is simply (pace Hornstein) one among many theoretical systems currently available. I hope therefore that future LFG conferences will see more attempts to compare and reflect on work from systems such as RRG, SBCG, exemplar-based grammar and the like, and thereby to pursue the larger goal of understanding the complex phenomenon that is natural language.

[^191]
## References

Bod, Rens (2009) From exemplar to grammar: a probabilistic analogybased model of language learning. Cognitive Science 33: 752-793.
Boeckx, Cedric (2008) Understanding Minimalist syntax. Lessons from locality in long-distance dependencies. Oxford: Blackwell.
Bouzouita, Miriam (2008) The diachronic development of Spanish clitic Placement. PhD Thesis, King's College London.
Bresnan, Joan (ed.) (1982) The mental representation of grammatical relations. Cambridge, Mass: MIT Press.
Bresnan, Joan (2001) Lexical-functional syntax. Oxford: Blackwell.
Cann, Ronnie, Ruth Kempson \& Lutz Marten (2005) The dynamics of language. Oxford: Elsevier.
Cann, Ronnie \& Ruth Kempson (2008) Production pressures, syntactic change and the emergence of clitic pronouns. In Cooper \& Kempson, pp. 179-220.
Chatzikyriakidis, Stergios (in prep.) Clitics in four dialects of Modern Greek: A dynamic Account. PhD Thesis, King's College London.
Chomsky, Noam (1970/76) Deep structure, surface structure and semantic interpretation. Reprinted in Studies on semantics in generative grammar, The Hague: Mouton. [page references to 1976 reprint]
Chomsky, Noam (1995) The Minimalist program. Cambridge, Mass: MIT Press.
Cooper, Robin \& Ruth Kempson (eds) (2008) Language in flux. Dialogue co-ordination, language variation, change and evolution. Kings College London: College Publications.
Croft, William (2007) Beyond Aristotle and gradience. A reply to Aarts. Studies in Language 31: 409-430.
Evans, Nicholas \& Stephen Levinson (2009) The myth of language universals: Language diversity and its importance for cognitive science. Behavioral and Brain Sciences 32: 429-492.
Fodor, Jerry (1974) Special sciences (or: The disunity of science as a working hypothesis). Synthese 28: 97-115.
Fodor, Jerry (1995) Special sciences: still autonomous after all these years (a reply to Jaegwon Kim's "Multiple realization and the metaphysics of reduction"). Reprinted in Fodor (1998), pp. 9-24 [page references to the reprinted version].

Fodor, Jerry (1998) In critical condition. Polemical essays on cognitive science and the philosophy of mind. Cambridge, Mass: MIT Press.
Hagège, Claude (2008) Towards a typology of interrogative verbs. Linguistic Typology 12: 1-44.
Hawkins, John (2004) Efficiency and complexity in grammars. Oxford: Oxford University Press.
Hornstein, Norbert (2009) A theory of syntax. Minimal operations and universal grammar. Cambridge: Cambridge University Press.
Johnson, David \& Shalom Lappin (1997) A critique of the Minimalist Program. Linguistics \& Philosophy 20: 273-333.
Kempson, Ruth, Wilfred Meyer-Viol \& Dov Gabbay (2001) Dynamic syntax: The flow of language understanding. Oxford: Blackwell.
Kempson, Ruth \& Jieun Kiaer (2010) Multiple-long-distance scrambling: Syntax as reflections of processing. Journal of Linguistics 46.
Kiaer, Jeun (2007) Processing and interfaces in syntactic theory: The case of Korean. PhD thesis, King's College London.
McCauley, Robert N. (ed.) (1996) The Churchlands and their critics. Oxford: Blackwell.
Mycock, Louise (2006) The typology of constituent questions: A LexicalFunctional Grammar analysis of 'wh'-questions. PhD thesis, University of Manchester.
Nordlinger, Rachel (1998) Constructive case. Evidence from Australian languages. Stanford: CSLI.
Sag, Ivan (forthcoming) Sign-Based Construction Grammar: An informal synopsis. To appear in H. Boas and I. A. Sag (eds.) Sign-Based Construction Grammar. Stanford: CSLI Publications.
Sag, Ivan \& Thomas Wasow (forthcoming) Performance-Compatible Competence Grammar. In Robert Borsley \& Kersti Börjars (eds) Nontransformational syntax. Oxford: Blackwells.
Vincent, Nigel \& Kersti Börjars (forthcoming) Grammaticalization and models of language. In Elizabeth Closs Traugott \& Graeme Trousdale (eds) Gradience, Gradualness, and Grammaticalization. Amsterdam: John Benjamins.

# UDI PERSON MARKERS AND LEXICAL INTEGRITY 

Michael T. Wescoat

Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)

2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

In Udi, person markers occur either within the verb, in positions determined by internal morphological structure, or else at the right edge of a syntactic constituent in focus. Harris reasons that since describing the distribution of these person markers requires reference to both morphological and syntactic entities, this phenomenon challenges the lexicalist separation of morphology and syntax advocated by the Lexical Integrity Hypothesis. I argue that one can maintain a lexicalist approach by adopting lexical sharing, which allows words to instantiate multiple terminals. Udi person markers are positioned within the word by purely morphological alignment constraints; words containing a person marker instantiate an additional terminal, which is positioned by purely syntactic alignment constraints. This analysis preserves the separation of morphology and syntax, showing that the data surrounding Udi person markers are not, in fact, damaging to lexicalist theory.


## 1 The problem

In her richly detailed study of the morphosyntax of Udi, ${ }^{1}$ Harris (2002) claims to have found counterevidence to the lexicalist separation between syntax and morphology, such as is assumed in the Lexical Integrity Hypothesis (Bresnan, 2001, p. 92). She examines person markers (PMs), a class of clitics that agree in person and number with the subject; (1) shows the paradigm of general PMs. ${ }^{2}$ For Harris, describing the distribution of PMs requires an intermingling of morphological and syntactic constraints that she feels renders lexicalist theories untenable. I begin with a brief overview of her analysis and conclusions, before going on to suggest ways in which a lexicalist approach could overcome Harris's objections.

|  | SG | PL |
| :--- | :--- | :--- |
| 1 | zu, z | yan |
| 2 | nu, n, lu, ru | nan, lan, ran |
| 3 | ne, le, re | q'un |

### 1.1 The distribution of person markers in Udi

Harris's treatment of Udi PMs is couched within Optimality Theory (OT), using the notion of Generalized Alignment (McCarthy and Prince, 1993). This employs the Align relation, defined in (2), to state violable constraints governing the relative positioning of edges, left or right, of entities designated by prosodic or grammatical categories. I present Harris's main constraints with just enough data to illustrate her proposed domination hierarchy. See Harris 2002, chaps. 6-7 for full details.
(2) $\operatorname{Align}\left(\mathrm{Cat}_{1}, \mathrm{Edge}_{1}, \mathrm{Cat}_{2}, \mathrm{Edge}_{2}\right) \leftrightarrow$
$\forall \mathrm{Cat}_{1} \exists \mathrm{Cat}_{2}$ such that Edge ${ }_{1}$ of $\mathrm{Cat}_{1}$ and Edge ${ }_{2}$ of $\mathrm{Cat}_{2}$ coincide, where $\mathrm{Cat}_{1}, \mathrm{Cat}_{2} \in \mathrm{PCat} \cup \mathrm{GCat}$ (prosodic and grammatical categories), Edge $_{1}$, Edge $_{2} \in\{\mathrm{~L}($ eft $), \mathrm{R}($ ight $)\} . \quad$ (McCarthy and Prince, 1993, p. 2)

[^192]One place where PMs may occur is following a subset of tense-aspect-mood (TAM) suffixes, namely the FUT(ure)II, s(u)BJ(uncti)V(e)I, and SBJVII, as well as the IMP(ererative) when plural. These suffixes are almost always realized as al or a. Constraint (4) aligns the left edge of the PM with the right edge of the suffix.
(3) a. b-al-le ${ }^{3}$
b. b- $a$-ne
c. b-a-ne-y
d. b-a-nan do-FUTII-3SG do-SBJVI-3SG do-SBJVII-3SG-PAST do-IMP-2PL
(4) ALIGN-PM-al/a (abbreviated al/a) Align(PM,L, al/a, R)
(Harris, 2002, pp. 27, 149)

A PM may also arise on a focus constituent (FocC). Thus, ne ' 3 SG ' is attached to AdvP in (5), PP in (6), and NP in (7); each is interpreted as sentence focus and appears immediately before the verb, where FocC usually occurs in Udi. Constraint (8) aligns the left edge of the PM with the right edge of FocC.
(5) irähät-en mya-ne bist'a cil-l-ux peasant-ERG here-3SG sow.PRES seed-OBL-DAT 'The peasant sows seeds here.'
(6) äyel-en k'uč'an-ne bey-sa child-ERG puppy.ABSL-3SG watch-PRES
'The child is watching a puppy.'
(7) xe-n-en-k-ne tay-sa
water-OBL-ERG-for-3SG thither-PRES
'She went for water.'
(8) Align-PM-FocC (abbreviated FocC) Align(PM,L, FocC, R)
(Harris, 2002, pp. 95, 120, 150)
When an al/a TAM suffix is present, like al 'FUTII' in (9), the PM cannot attach to FocC; thus, shifting the PM in (9) to k'uč'an 'puppy' would result in ungrammaticality. This motivates the domination hierarchy in (10).
(9) äyel-en k'uč'an bȩ̣-al-le
child-ERG puppy.ABSL watch-FUTII-3SG
'The child will watch a puppy.'
(10) $a l / a \gg$ FOCC
(Harris, 2002, pp. 120, 150)
In Udi, the majority of verb stems are complex, combining an incorporated element (IncE) with a light verb. One sort of IncE is an infinitive marked with es; see eč-es 'bring-INF' and cip-es 'spread-INF' in (11), for example. An IncE may also be a noun, such as $a \check{s}$ 'work, business, matter' in (12), an adjective, an adverb, or a simplex verb stem, among other things. A PM may occur between the IncE and the light verb, as seen in (11) and (12); constraint (13) favors this positioning, aligning the left edge of the PM with the right edge of the IncE.

[^193](11) me pasčay-en eč-es-ne-st'a
this king-ERG bring-INF-3SG-CAUS.PRES
kul cip-es-ne-st'a pak-i
earth.ABSL spread-INF-3SG-CAUS.PRES garden-DAT
'This king has earth brought. .. ; he has it spread in the garden.'
(12) zavod-a $a \check{s}$-ne-b-sa
factory-DAT work-3SG-do-PRES
'She works in a factory.'
(13) ALIGN-PM-IncE (abbreviated IncE)

Align(PM, L, IncE, R)
(Harris, 2002, pp. 122, 151)
When an al/a TAM suffix is present, that suffix attracts the PM in preference to an IncE; thus, le '3SG' aligns with al 'FUTII' rather than $a s$ ' 'work, business, matter' in (14). A FocC is also a more powerful attractant than an IncE; hence, in (15), $z$ ' 1 SG ' aligns with zavod- $a$ 'factory-DAT' in preverbal focus position, rather than with $a \check{s}$. These facts support the domination hierarchy in (16).
(14) bez vič-en aš-b-al-le zavod-a
my brother-ERG work-do-FUTII-3SG factory-DAT
'My brother will work in a factory.'
(15) zavod-a-z ${ }^{4}$ aš-b-sa
factory-DAT-1SG work-do-PRES
'I work in a factory.'
(16) al/a> FOCC $\gg$ INCE
(Harris, 2002, pp. 123, 151)
One of the more noteworthy places where a PM may occur is inside of a simplex verb stem. In (17), ne ' 3 SG ' occurs between the penultimate and final segments of the monomorphemic form bey 'look.' The two part glossing with numeric subscripts, 'look ${ }_{1}-\ldots-$ look $_{2}$,' is meant to represent the interruption of the simplex verb stem. Similarly, $z$ '1 SG' falls inside of $a q$ ' 'receive, take' in (18). Constraint (19) stands out from previous rules; it aligns the right edge of the PM with the right edge of the simplex verb stem. Tableau (20) illustrates the application of (19). One violation-mark is assessed for each segment separating the two edges; the intervening segments are used for violation-marks in place of asterisks, to enhance clarity. The optimal place for the PM falls one segment before the end of the verb stem.
(17) pasčay-un yar-en gölö be-ne- $\gamma$-sa met'a-laxo
king-GEN boy-ERG much look $1_{1}$-3SG-look 2 -PRES this.GEN-on
'The prince looks at this for a long time.'
(18) kayuz-ax $a-\mathbf{z}-q$ '-e
letter-DAT receive ${ }_{1}-1$ SG-receive $_{2}$-AORII
'I received the letter.'

[^194](19) Align-PM-Verb stem (abbreviated Vstem) Align(PM, R, Verb stem, R)
(Harris, 2002, pp. 125, 153)

| 'looks' | ALIGN-PM-VERB STEM |
| ---: | :---: |
| ne- $b e \boldsymbol{e} \boldsymbol{\gamma}$-sa | be! y |
| $b$-ne- $e \boldsymbol{\gamma}$-sa | e $!!$ |
| $b e-$ ne- $\gamma$-sa | X |
| $b e \gamma$-ne-sa | ne! |

The position inside of the simplex verb stem becomes unavailable for the PM if the option of aligning with an al/a TAM suffix, a FocC, or an IncE is available; thus, the PM is aligned with al 'FUTII' in (21), with muqluy-en 'pleasure-INST' in preverbal focus position in (22), and with the incorprated xabar 'ask' in (23). In contrast to (18) above, nowhere in (21)-(23) does the PM occur inside of the monomorphemic form $a q$ ' 'receive, take.' These observations motivate the domination hierarchy in (24).
(21) sa xinär-en...aq'-al-le k'alpesun-un p'iz-ax
one girl-ERG receive-FUTII-3SG read.MAS-GEN prize-DAT
'A girl... will receive the prize for studying.'
(22) ...muqluy-en-zu aq'-sa
pleasure-INST-1SG receive-PRES
'[Whoever comes to me as a guest] I receive with pleasure.'
(23) xabar-re-aq'-sa šot-uxo
ask-3SG-take-PRES him-ABL
'He asks him.'
(24) al/a> FOCC $\gg$ INCE $\gg$ VSTEM
(Harris, 2002, pp. 126f, 164)

### 1.2 Harris's conclusions and the goals of this study

Harris (2002, chap. 5) argues at length that Udi PMs are clitics, exhibiting many characteristics typically attributed to that class. She observes, for instance, that PMs are promiscuous, attaching to a variety of forms; the foregoing data show that the PM may associate not only with the verb, but also with adverb, preposition, and noun-see (5)-(7). Also, PMs may occur 'outside' of (i.e. farther away from the root than) other clitics; such is the case in (25), where $l e$ ' 3 SG ' trails the clitic al 'and.' Harris offers these and numerous other arguments.
(25) abaz-in-al-le aš-b-esa
[coin]-INST-and-3SG work-do-PRES
'And he works for an abaza [twenty kopeks].'
Harris's examination of Udi PMs also reveals some theoretical surprises. If PMs are clitics, then the fact that they occur inside of a word means that they fall into the class of endoclitics, assumed by some to be nonexistent (e.g., Klavans, 1979). A further, closely related point is of particular interest here; in Harris's
account of the distribution of Udi PMs, the alignment constraints refer both to syntactic elements (FocC) and to verb-internal morphological items (al/a TAM suffixes, IncE, verb stem). This point poses a challenge for lexicalist theories:

The problem is that the rules that position... PMs must be in part syntactic rules, given that PMs may occur on words outside the verb... But if the rules are syntactic, the Lexical Integrity Hypothesis claims that they do not have access to the internal structure of a word and therefore cannot position the PM inside the verb... (Harris, 2002, p. 3)
In this study, I consider the problem of describing the distribution of Udi PMs in a lexicalist framework and conclude that it is feasible, provided one uses the right tools. More particularly, my goal is to provide an analysis that avoids the intermixing of morphological and syntactic constraints, while reusing as much of Harris's original treatment as possible. The approach I sketch draws on Optimal Paradigm Theory (McCarthy, 2005) and Optimality-Theoretic Lexical Functional Grammar (OT-LFG, Bresnan, 2000; Sells, 2001), augmented with the mechanism of lexical sharing (Wescoat, 2002), to which I turn my attention next.

## 2 Lexical sharing

In traditional thinking, a clitic is a form that is syntactically free, but phonologically bound to a host. Thus, this view recognizes the existence of two separate elements of $c$ (onstituent)-structure, corresponding to clitic and host. At the same time, Harris's discussion of Udi PMs suggests that clitic and host may be more than just phonologically bound; the clitic may reside inside of the host, its precise position there determined by the host's internal morphological composition. If the domain of morphology is the word, then it seems logical to say that the host-clitic amalgam functions as one word, at least as far as the Udi data are concerned. Lexical sharing is a mechanism designed to accommodate phenomena with these very characteristics, one word corresponding to multiple constituents in c-structure.

In this section, I exemplify lexical sharing with English data, mostly involving non-syllabic auxiliary contractions, such as ' $l l$ for will, when it is not pronounced as a syllable unto itself. I have argued (Wescoat, 2005) that the behavior of these contractions vis-à-vis their hosts resembles types of phenomena that occur within a word. For instance, non-syllabic auxiliary contractions are selective; they occur only with pronouns and question words. Thus, in (26a), non-syllabic ' $l l$ [1] accompanies the pronoun $I$, but only syllabic ' $l l[1]$ occurs with so in (26b). Moreover, (26a) illustrates that non-syllabic ' $l l$ can trigger an idiosyncratic alternation for $I$, [aI] $\sim[a]$; in at least my dialect, the [a] variant can occur nowhere else. While I'll [arl/al] is one word, the conjunction in (27) suggests that it corresponds to two constituents, a D and an I; the auxiliary I resides in the left-hand conjunct, while the pronominal D lies outside of the conjunction and takes scope over both conjuncts. Thus, non-syllabic auxiliary contractions are good examples of lexical sharing.
(26) a. I'll [arl/al] help. b. So'll [so.l/*sol] I. ["." = syllable boundary]
(27) I'll [arl/al] be there on Sunday and am looking forward to seeing you.

### 2.1 Basic concepts

To provide a formal model of lexical sharing, I exploit the fundamental architecture of LFG, which assumes parallel structures related by structural correspondences (Kaplan, 1995). One may conceptualize lexical sharing in a series of steps. First, think of a traditional c-structure, as in (28a). A c-structure is a set of nodes N , labeled with syntactic categories or words, and related by a mother function $\mathrm{M}: \mathrm{N} \rightarrow \mathrm{N}$ and a precedence relation $<\subseteq \mathrm{N} \times \mathrm{N}$. Second, remove from c-structure all nodes labeled with words, as in (28b). Note that this changes the set of terminals T , which comprises all non-mother nodes $(\mathrm{T}=\mathrm{N}-\operatorname{ran}(\mathrm{M})$, where $\operatorname{ran}(\mathrm{M})$ is the range of M ); in (28b), T consists of the nodes labeled D , I, and V. Third, put the words into a separate representation called l(exical)-structure, as in (28c). An l-structure, like $\langle\mathrm{I}$, will, help $\rangle$, consists of a linearly ordered set of words W. ${ }^{5}$ Fourth, introduce a structural correspondence between c- and l-structure, in the form of the lexical exponent mapping $\lambda: \mathrm{T} \rightarrow \mathrm{W}$, as in (28d). If $\lambda$ maps a terminal X to a word w , then one may say that w instantiates X , or that w is the lexical exponent of X . The domain of $\lambda$ is all of T , and the range of $\lambda$ is all of W . The graphic in (28d) employs the sort of curving lateral arrows most often employed for depicting structural correspondences in LFG; however, I believe it is more perspicuous to represent $\lambda$ with vertical arrows descending from terminals to words lined up in order below c-structure, suppressing the l-structure's brackets and commas to avoid clutter, as in (28e). The $\lambda$ mapping permits a straightforward representation of lexical sharing; $\lambda$ may be one-to-one, as in (28e), or it may map two or more terminals into a single word, as in (28f), where the D and I 'share' I'll.
(28) a.

b.

c.


d.

e.

f.


One must next restrict the relative ordering of c - and 1 -structure, to avoid such absurdities as (29a), where I slept is linked to a c-structure in verb-subject or-

[^195]der. Consistency in ordering between c- and l-structure is established by the order preservation axiom: For all terminals X and Y , if $\lambda(\mathrm{X})$ precedes $\lambda(\mathrm{Y})$, then X precedes Y. Simply put, this axiom prevents the arrows of the $\lambda$ mapping from crossing. An order-preserving mapping, like $\lambda$, between linearly ordered sets, like T and W , is technically a homomorphism, so I call structures with crossing arrows homomorphism violations and I label them 'Ill-formed!' to emphasize that they are not countenanced by the theory. The homomorphic nature of $\lambda$ also entails the homomorphic lexical integrity theorem: Only sequences of adjacent terminals may share a lexical exponent. By way of proof, note that if two terminals, X and Z , share a lexical exponent v , then an intermediate terminal Y with a distinct lexical exponent w inevitably causes a homomorphism violation, as suggested by (29b).
(29) a.

b.


The homomorphic lexical integrity theorem leads to a class of empirical predictions that I call edge attraction effects. For ease of expression, I limit my attention to words that instantiate no more than two terminals. If terminals $X$ and $Y$ share a lexical exponent, and X resides in a phrase Z , while Y stands outside of Z , then X occurs at the edge of $Z$ nearest $Y$, in one of the patterns $[\mathrm{Z} \ldots \mathrm{X}] \mathrm{Y}$ or $\mathrm{Y}[\mathrm{ZX} \ldots]$. For example, suppose one analyzes the English possessive in 's with lexical sharing (Wescoat, 2002), assuming that the word marked with 's instantiates two terminals, one of them being the D that takes the possessor as its specifier. It then follows that the word bearing 's falls at the right edge of the possessor, as in (30a); otherwise, a homomorphism violation would result, as suggested by (30b).
(30) a

b.


As a grammar formalism, I use context-free rewriting rules, as in (31a), to describe c-structure, and a lexicon consisting of lexical-exponence rules, as in (31b), to describe $\lambda$. A lexical-exponence rule $\mathrm{w} \leftarrow \mathrm{X}_{1} \cdots \mathrm{X}_{n}$ (with a leftward arrow) allows $\lambda$ to map $n$ adjacent terminals labeled from left to right $\mathrm{X}_{1}, \ldots, \mathrm{X}_{n}$ into w .

## Constituent-structure rules

a. $\mathrm{IP} \rightarrow \mathrm{DP} \mathrm{I}^{\prime}$
b. $\mathrm{I}^{\prime} \rightarrow \mathrm{I}$ VP
c. $\mathrm{DP} \rightarrow \mathrm{D}$
d. $\mathrm{VP} \rightarrow \mathrm{V}$

## Lexical-exponence rules

e. $I \leftarrow D$
g. help $\leftarrow \mathrm{V}$
f. will $\leftarrow$ I
h. I'll $\leftarrow$ D I

### 2.2 LFG and lexical sharing

To integrate lexical sharing into LFG, one must establish a relationship between 1 -structure and $f$ (unctional)-structure, LFG's representation of grammatical functions. This may be accomplished in three steps. First, one must include elements of l-structure in the domain of the structural correspondence $\varphi$, which was originally conceived as a mapping from c- to f-structure (Kaplan, 1995); henceforth, $\varphi: \mathrm{N} \cup \mathrm{W} \rightarrow \mathrm{F}$ is a mapping from nodes and words to members of the set F of f -structures. Second, one may define a new metavariable for convenient reference to the f-structures of lexical exponents; $\Downarrow$ abbreviates $\varphi(\lambda(*))$ 'the f-structure of the lexical exponent of the current node $[=*]$.' Finally, one must provide the righthand sides of lexical-exponence rules with functional annotations, as in (32).
(32) a. I'll $\leftarrow$
$\begin{array}{cc}\text { D } & \text { I } \\ (\downarrow \text { PRED })=\text { 'PRO' } & (\downarrow \text { TNS })=\text { FUT } \\ \Downarrow=\downarrow & (\downarrow \text { SUBJ })=\mathrm{c} \downarrow\end{array}$
b. help $\leftarrow$
$(\downarrow$ PRED $)=\begin{gathered}\left.\quad \begin{array}{c}\text { HELP } \\ \Downarrow=\downarrow \\ \Downarrow \\ \downarrow\end{array}(\downarrow \text { SUBJ })\right\rangle\end{gathered}$

If one assumes that c-structure rules receive functional annotations in accord with universal principles of structure-function mapping (Bresnan, 2001), then the rules shown above provide the c-, l-, and f-structure in (33) for I'll help.


Consider this grammar with lexical sharing in connection with the relation between syntax and morphology assumed under a lexicalist theory, as outlined here:

There are undeniably observable interactions between morphemes and syntax;. . however, $\ldots$. the interactions are such that it is not necessary to intermix the terms and rules of syntax and morphology. Rather the two theories share a small theoretical vocabulary, including the parts of speech and certain features (such as 'tensed'), and the interpenetration that exists is channeled through this shared vocabulary. (Di Sciullo and Williams, 1987, p. 47)

The 'shared vocabulary' assumed in the present grammar with lexical sharing is almost identical to that employed in traditional LFG; a word transmits categorial and functional information to the syntax. However, rather than convey such information in reference to just one terminal, lexical sharing relays information about multiple terminals. The only qualitatively new type of information passed to the syntax concerns the ordering among the terminals being instantiated. Thus, I believe lexical sharing to be a moderate extension of previous practice, which remains true to the spirit of lexicalist theory.

## 3 The morphology of Udi person markers

### 3.1 Introducing and aligning person markers within the word

Maintaining Harris's analysis of Udi PMs as clitics, one may treat words containing a PM as instances of lexical sharing. The difference between a word with a PM, such as xabar-re-aq'-sa 'ask-3SG-take-PRES,' and the corresponding word without a PM, such as xabar-aq'-sa 'ask-take-PRES,' is then that the former instantiates an extra terminal, to which I assign the syntactic category PM, for lack of any better classification. The lexical-exponence rules in (34) illustrate the difference.


The treatment sketched in (34) rests on certain expectations about morphology. Alongside derivation and inflection, I assume there is also instantiationaltering morphology; the presence of an instantiation-altering morpheme increases the number of terminals that a word instantiates. Udi PMs, like re ' 3 SG ' above, are therefore instantiation-altering morphemes. Just as inflectional morphemes tend to occur 'outside' of derivational ones, I assume that in most cases instantiationaltering morphemes tend to lie 'outside' of inflection; thus, possessive 's follows plural en in ox-en-'s, for instance. Udi is unusual in not adhering to this tendency. Traditionally, linguists have regarded derivation as producing lexemes, inflection as producing word-forms, and the addition of clitics as producing clitic groups, the last not being a word. To the degree that a given clitic is analyzable with lexical sharing, ${ }^{6}$ I assume that it is an instantiation-altering morpheme and that forms in which it occurs are in fact words. Hence, xabar-re-aq'-sa 'ask-3SG-take-PRES,' with the PM re, is a word. Finally, I assume that the derivational, inflectional, and instantiation-altering morphology each defines its own domain within the word, so the present scheme can accommodate any morphophonological phenomenon that

[^196]is thought to be sensitive to the difference between a 'clitic group' and a 'word,' where the latter is understood in the traditional sense, i.e., without clitics.

I next consider alignment of PMs within the word. For this purpose, I redeploy all of Harris's alignment constraints that are sensitive to morphological categories; these are repeated in (35)-(37). (I will, however, modify constraint (35) shortly.) Note that only constraint (8), ALIGN-PM-FOCC, is excluded from this list. Alignment with FocC is nonetheless significant with respect to morphological alignment constraints; when a PM is associated with FocC, it may be attached to a non-verb, in which case the PM occurs word-finally, as in (5)-(7). To allow for this possibility, one may posit constraint (38). This constraint, which is violated when the PM is not in word-final position, is dominated by all of the constraints that mention morphological categories in the verb, as indicated in hierarchy (39).
(35) ALIGN-PM-al/a (abbreviated al/a)
$\operatorname{Align(PM,L,~al/a,~R)~[=(4)]~}$
(36) Align-PM-IncE (abbreviated IncE)

Align(PM,L,IncE, R) [=(13)]
(37) Align-PM-Verb stem (abbreviated Vstem)

Align(PM, R, Verb stem, R)
[=(19)]
(38) Align-PM-Final (abbreviated Final)

Align(PM, R, Word, R)
(39) al/a $\gg$ INCE $\gg$ VSTEM $\gg$ FINAL

The morphological alignment constraints in (35)-(38) need not look outside of the word. Sample tableaux are given in (40) and (41). (I omit constraint (35), because it requires special consideration.) Listed in (40) are some of the candidates that would arise in connection with the combination of the IncE xabar 'ask,' the light verb $a q$ ' 'take,' the tense sa 'PRES,' and the 3 SG PM. ${ }^{7}$ Tableau (41) features the same PM with the noun $k$ 'uč'an 'puppy.' The asterisks in the columns labeled IncE and Vstem arise because the definition of the Align relation in (2) is such that constraints (36) and (37) turn out to assert the existence of an IncE and a verb stem, respectively; obviously these elements of verbal morphology are absent from non-verbs. This does not affect the selection of the optimal candidate, though.

| 'asks' | INCE | VsTEM | FINAL |
| ---: | :---: | :---: | :---: |
| ne-xabar-aq'-sa | n!exabar | xabaraq' | xabaraq'sa |
| xabar-re-aq'-sa |  | aq' | aq'sa |
| xabar-a-ne-q'-sa | a! | q’ | q'sa |
| xabar-aq'-sa-ne | a!q'sa | sane |  |


| 'puppy' | INCE | VSTEM | FINAL |
| ---: | :---: | :---: | :---: |
| ne-k'uč'an | $*$ | $*$ | k'!uč'an |
| k'uč'an-ne | $*$ | $*$ |  |

[^197]
### 3.2 The $a l / a$ tense-aspect-mood suffixes and paradigm gaps

Recall that Harris assumes hierarchy (42), in which Align-PM-al/a dominates all other PM alignment constraints. It follows that a PM must accompany all TAM suffixes to which ALIGn-PM-al/a is sensitive. ${ }^{8}$ A lexicalist approach can model this generalization directly by predicting a gap in Udi verbal paradigms; words containing the relevant TAM suffixes but no PMs are lacking.
(42) al/a $\gg$ FOCC $\gg$ INCE $\gg$ VSTEM

Optimal Paradigm Theory provides a theoretical foundation for this analysis. Relevant parts of McCarthy's (2005, p. 173) summary are reproduced in (43).
(43) a. Candidates consist of entire inflectional paradigms, where an inflectional paradigm contains all and only the words based on a single lexeme...
b. Markedness and Input-Output faithfulness constraints evaluate all members of the candidate paradigm. The violation-marks incurred by each paradigm member are added to those incurred by all the others.

On the foregoing foundation, Rice $(2005,2007)$ builds a theory of defective paradigms that contain optimal gaps. In OT, one naturally expects words to accrue violation-marks from a variety of constraints. On this basis, the optimal paradigm would seem to be one that contains only gaps and thereby accumulates no violation-marks. To avoid empty paradigms, Rice reasons that there must be $\operatorname{MAX}\{\mathrm{CAT}\}$ constraints, which enforce expression of a morphological category CAT. A gap, symbolized $\odot$, would leave CAT unexpressed and thus would fall afoul of $\operatorname{MAX}\{\mathrm{CAT}\}$. The presence of $\operatorname{MAX}\{\mathrm{CAT}\}$, which punishes gaps, creates tension with constraints that punish actual words. For instance, if a constraint Const dominated $\operatorname{MAX}\{\mathrm{CAT}\}$, then a gap would be more harmonic than a word $w$ that violates Const, as suggested in (44)—provided that no higher ranking constraint favors w. When the violation-marks for paradigms are summed, between any two paradigms that are identical except for the choice between including w or allowing a gap, the paradigm containing the gap will be more harmonic, as suggested by (45).

Returning to Udi, one may assume that there are constraints of the MAX $\{\mathrm{CAT}\}$ type for all TAM suffixes. For instance, in Udi MAX\{FUTII\} would be satisfied if the tense feature FUTII is expressed by the suffix al. Additionally, I propose to alter the statement of constraint (35), ALIGN-PM-al/a, by changing the order of

[^198]the arguments to the Align relation; the revised constraint is given in (46). Recall that the definition of Align in (2) employs both a universal and an existential quantifier, which are distributed in such a way that the revised constraint in (46) may be paraphrased as 'for any al/a TAM suffix, there exists a PM such that the right edge of the TAM suffix and the left edge of the PM coincide.' Thus, constraint (46) is violated any time an al/a TAM suffix arises in a word that contains no PM; additionally, when a PM is present, the constraint requires it to be adjacent to the TAM suffix. Next I propose to make constraint (46) dominate MAX\{FUTII\}, as indicated in (47), as well as all similar constraints requiring expression of TAM features. Consider a form like be $\gamma$-al 'watch-FUTII,' which one might have expected to be a part of the bey paradigm. Since bey-al contains the al suffix, which expresses FUTII, but has no PM, it violates (46). Therefore, bey-al is less harmonic than a gap and will thus be absent from the optimal bey paradigm, as suggested by (48). This scheme predicts a set of systematic gaps throughout the verbal paradigms of Udi, wherever al/a TAM suffixes occur without an accompanying PM.
(46) ALIGN-PM-al/a (abbreviated al/a)

Align (al/a, R, PM,L)
[cf. (4), (35)]
(47) $a l / a \gg \operatorname{MAX}\{$ FUTII $\}$

| bey paradigm | al/a | MAX\{FUTII $\}$ |
| ---: | :---: | :---: |
| $\langle\ldots$, be $\gamma-$ al,$\ldots\rangle$ | $* \times(m+1)$ | $* \times n$ |
| $\langle\ldots, \ldots\rangle$ | $* \times m$ | $* \times(n+1)$ |

## 4 The syntax of Udi person markers

### 4.1 The place of the person marker in constituent structure

I now return to the matter of formalizing a lexical-sharing analysis of Udi PMs, focusing here on the syntax. Recall that a PM is an instantiation-altering morpheme that causes words to instantiate an additional terminal; to this terminal I assign the syntactic category PM. These points are illustrated in the lexical-exponence rules in (49) for be $\gamma$-al-le 'watch-FUTII-3SG' and $k$ 'uč'an-ne 'puppy-3SG.' In order to construct the syntactic analysis, one must situate the PM constituent within c-structure.



When the PM is associated with the verb, a logical place for the PM constituent is right-adjoined to V , as illustrated in (50). The association of the adjoined PM with a grammatical function, viz. SUBJ(ect), is specified lexically; see
the $(\uparrow$ SUBJ $)=\downarrow$ annotation on the PM node in (50), which is provided by (49a). Being adjoined to V puts the PM in the right place for $\uparrow$ to pick out the f -structure of V; thus, ( $\uparrow$ SUBJ) names the f-structure for the clause's subject.
(50)


Next consider cases in which the PM is associated with the FocC. In Udi, the FocC usually falls immediately before the verb (Harris, 2002, chap. 3). In these circumstances, the PM constituent may once again be adjoined to V, though to its left, as shown in (51). The functional annotations on PM work as in the last case.


Note that the structural differences between (50) and (51) have no effect on the functional analysis; (50) and (51) yield almost identical f-structures, differing only in the value for TNS (tense). The common elements and the single difference may be seen in (52a). Also, (50) and (51) yield the same $i$ (nformation)-structure (albeit underspecified in this example) where details about discourse functions are recorded (King, 1997); this is shown in (52b).

$$
\text { a. }\left[\begin{array}{ll}
\begin{array}{ll}
\text { PRED } & \text { 'WATCH }\left\langle f_{1}, f_{2}\right\rangle \prime \\
\text { TNS } & \text { FUT/PRES }
\end{array}  \tag{52}\\
\text { SUBJ } & f_{1}\left[\begin{array}{ll}
\text { PRED } & \text { 'CHILD' } \\
\text { CASE } & \text { ERG } \\
\text { PERS } & 3 \\
\text { NUMB } & \text { SG }
\end{array}\right] \\
\text { OBJ } & f_{2}\left[\begin{array}{ll}
\text { PRED } & \text { 'PUPPY' }
\end{array}\right]
\end{array}\right] \quad \text { b. }\left[\begin{array}{ll}
\text { FOC } & \{\text { PUPPY }\}
\end{array}\right]
$$

Although FocC usually occurs in the position immediately before the verb, this is not invariably the case. Example (53) demonstrates that question words, which assume the role of focus, may sometimes occur sentence-initially, presumably in complementizer position. Note that the sentence-initial question word in (53) bears the PM nu '2SG.' By the assumptions outlined above, this means that there is a PM constituent. However, the placement of the adverb mya 'here' makes it unlikely that this PM constituent could be adjoined to V. I therefore assume that in this case the PM constituent is adjoined either to VP or to S. In sum, it appears that the PM constituent can be adjoined anywhere in the clause's functional domain, including $\mathrm{V}, \mathrm{VP}$, and S , all of which are mapped into the same f-stucture by $\varphi .^{9}$ Thus, one might generalize that there are no category-based constraints on the adjunction of the PM constituent; it can arise wherever the metavariable $\uparrow$ will pick out the clausal f-structure.
(53) ek'aluy-nu mya are?
why-2SG here come
'Why have you come here?'
(Harris, 2002, p. 49)
The assumption that the PM constituent is adjoined to a node in the clause's functional domain gives an immediate explanation of the fact that the word bearing the PM morpheme must fall at the right edge of the FocC, as exemplified by $p^{\prime}$ a $\dot{e} \check{s}$-ne 'two apple-3SG' in (54). This is the sort of edge attraction effect discussed in §2.1. Consider the c- and l-structure in (55). The PM constituent, which lies outside of the FocC, can share a lexical exponent with the adjacent N , as is shown, but it cannot share a lexical exponent with Q , since to do so would result in a homomorphism violation. Moreover, if one assumes that PM morphemes cause a new PM terminal to be instantiated to the right, yielding 'éš-ne $\leftarrow \mathbf{N}$ PM' and not 'és-ne $\leftarrow \mathrm{PM} \mathrm{N}$,' then it is necessarily to the right edge of the FocC that the word bearing the PM morpheme will be attracted.
(54) äyel-en p'a èš-ne aq'-e
child-ERG two apple-3SG take-AORII
'The child took two apples.'
(Harris, 2002, p. 55)

[^199]

### 4.2 Aligning the PM constituent

Next I consider alignment in the syntax, for which purpose I employ the framework of OT-LFG. This rests on the background assumptions of OT, which makes use of two components, GEN(eration), which enumerates a set of candidates, and EVAL(uation), which compares candidates to a hierarchy of violable constraints. In OT-LFG, GEN is an LFG; for my purposes, I assume that the LFG in question is one that incorporates lexical sharing, as described in §2.2. The candidates enumerated by this GEN are quadruples of $\mathrm{c}-, \mathrm{l}-\mathrm{f}-\mathrm{f}$, and i -structures.

Recall that the morphological alignment constraints discussed in §3 did not include Harris's Align-PM-FocC. This will now be repurposed as a syntactic alignment constraint that tracks the PM constituent rather than the PM morpheme. FocC can be identified as a phrase whose f-structure PRED(icate) FN (function) is a FOC(us) in i-structure. This state of affairs is illustrated in the $\mathrm{c}-, \mathrm{f}-$, and i -structures in (50)-(52), where the FocC is the NP in c-structure whose f -structure has as PRED the value 'PUPPY', the FN of which is PUPPY, which is in turn a member of the set that is the value of FOC in i-structure. I will continue to use 'FocC' as a convenient shorthand with this interpretation. I revise AlIGN-PM-FOCC by reordering the arguments of Align, in the manner described in $\S 3.2$ in connection with AlIGN-PM-al/a; the result, visible in (56), may be interpreted as saying 'for any FocC, there exists a PM constituent such that the right edge of the FocC and the left edge of the PM constituent coincide.' Since the elements of c-structure are not laid out in a linear fashion, it is hard to evaluate (56) as a gradient constraint, which may be violated to varying degrees depending on distances between constituents. I therefore treat syntactic alignment constraints like (56) as non-gradient, counting one violation if any constituent is ordered between aligned elements. Alongside (56), I provide a corresponding constraint that aligns PM constituents with V; this is shown in (57). The former constraint dominates the latter, as indicated in hierarchy (58).
(56) Align-PM-FocC

Align(FocC,R,PM,L)
(57) Align-PM-V

Align(V,R,PM,L)
(58) Align-PM-FocC $\gg$ Align-PM-V

Tableau (59) illustrates the functioning of the foregoing constraints. The predictions are straightforward. If there is a FocC, ALIGN-PM-FocC will require the PM constituent to be aligned with it. In the first candidate, the PM constituent immediately follows the FocC; with no intervening constituents, AlIGN-PM-FOcC is satisfied. The opposite is true of the second candidate, where V stands between the PM constituent and FocC. Thus, the first candidate is more harmonic. In cases where there is no FocC, however, Align-PM-FocC will be vacuously satisfied, and ALIGN-PM-V will come into play, requiring that the PM constituent be aligned with V.

| 'I work in a factory.' | Align-PM-FocC | Align-PM-V |
| :---: | :---: | :---: |
|  |  | * |
|  | *! |  |

The responsibility of the syntactic alignment constraints in (56)-(57) is strictly limited to placing the PM constituent in the proximity of one or the other of the FocC and V. Beyond that, the location within the word of the PM morpheme is left entirely to the morphological alignment constraints.

There remains one unresolved issue. Recall that in Harris's original analysis, the constraint aligning PMs with al/a TAM suffixes dominates the one aligning PMs with FocC. This predicts that (60a) is more harmonic than (60b). In contrast, under the system of syntactic alignment constraints set forth in this section, (60b) satisfies the dominant constraint, ALIGn-PM-FocC, while (60a) does not; thus, the system advocated here seems to favor the ungrammatical (60b) over the grammatical (60a). The solution to this dilemma may be found in the discussion in §3.2, where it is posited that optimal verbal paradigms in Udi do not contain forms in which an al/a TAM suffix occurs without an accompanying PM morpheme. This implies that bey-al 'watch-FUTII' in (60b) is not an available word-form. Under these circumstances, (60a) prevails by default, despite the fact that it violates ALIGN-PM-FocC.
(60) a. äyel-en k'uč'an bȩ̣-al-le [=(9)] child-ERG puppy.ABSL watch-FUTII-3SG
'The child will watch a puppy.'
b. *äyel-en k'uč'an-ne bey-al
child-ERG puppy.ABSL-3SG watch-FUTII

## 5 Conclusion

Harris (2002) proposes an OT analysis of Udi PMs that combines constraints that make reference to both morphological and syntactic categories. She reasons that this analysis is at odds with lexicalist theories of grammar, such as that embodied in the Lexical Integrity Hypothesis, since it seems to defy the lexicalist separation of morphology and syntax. This study considers the main constraints employed in Harris's analysis and recasts them in a lexicalist approach. This is facilitated by the assumption of lexical sharing, which allows a single word to instantiate multiple elements of c-structure. Adding a PM morpheme to a word causes it to instantiate a new terminal, the PM constituent. Those of Harris's constraints that are sensitive to morphological categories are slightly modified and applied within the word to determine the position of the PM morpheme. One of Harris's constraints that makes reference to syntactic categories is slightly reworked and applied in the syntax to position the PM constituent. Working in parallel, the independent morphological and syntactic alignment constraints make the same empirical predictions as do the constraints posited by Harris that are the point of departure for this study. Thus, it appears that Udi PMs are not an obstacle to a lexicalist theory of grammar.

## References

Bresnan, Joan. 2000. Optimal syntax. In J. Dekkers, F. van der Leeuw, and J. van de Weijer, eds., Optimality Theory: Phonology, Syntax, and Acquisition, pages 334-385. Oxford: Oxford University Press.
——. 2001. Lexical-Functional Syntax. Oxford: Blackwell.
Clifton, John M., Deborah A. Clifton, Peter Kirk, and Roar Ljøkjell. 2002. The sociolinguistic situation of the Udi in Azerbaijan. In Studies in Languages of Azerbaijan, pages 107-123. Baku: Academy of Sciences of Azerbaijan; St. Petersburg: SIL International.
Di Sciullo, Anna-Maria and Edwin Williams. 1987. On the Definition of Word. Cambridge, MA: MIT.
Harris, Alice C. 2002. Endoclitics and the Origins of Udi Morphosyntax. Oxford University Press.
Kaplan, Ronald M. 1995. Three seductions of computational psycholinguistics. In M. Dalrymple, R. M. Kaplan, J. T. Maxwell, III, and A. Zaenen, eds., Formal Issues in Lexical-Functional Grammar, pages 339-367. Stanford: CSLI.

King, Tracy Holloway. 1997. Focus domains and information structure. In M. Butt and T. H. King, eds., Proceedings of the LFG97 Conference. Stanford: CSLI.
Klavans, Judith. 1979. On clitics as words. In P. R. Clyne, W. F. Hanks, and C. L. Hofbauer, eds., The Elements: A Parasession on Linguistic Units and Levels, pages 68-80. Chicago: Chicago Linguistics Society.

McCarthy, John. 2005. Optimal paradigms. In L. J. Downing, T. A. Hall, and R. Raffelsiefen, eds., Paradigms in Phonological Theory, pages 295-371. Oxford: Oxford University Press.
——and Alan Prince. 1993. Generalized alignment. Technical report no. 7, Rutgers University Center for Cognitive Science, New Brunswick, NJ.
Rice, Curt. 2005. Optimal gaps in optimal paradigms. Catalan Journal of Linguistics 4:155-170.
2007. Gaps and repairs at the phonology-morphology interface. Journal of Linguistics 43:197-221.
Sells, Peter. 2001. Formal and Empirical Issues in Optimality Theoretic Syntax. Stanford: CSLI.

Wescoat, Michael T. 2002. On Lexical Sharing. Ph.D. thesis, Stanford University.
—_. 2005. English nonsyllabic auxiliary contractions: An analysis in LFG with lexical sharing. In M. Butt and T. H. King, eds., Proceedings of the LFG05 Conference, pages 468-486. Stanford: CSLI.

Zwicky, Arnold M. 1977. On Clitics. Bloomington: Indiana University Linguistics Club.

# CASE AND GRAMMATICAL FUNCTIONS IN IMBABURA QUECHUA: AN LFG APPROACH 

Beatriz Willgohs and Patrick Farrell<br>University of California, Davis<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

In Imbabura Quechua, accusative case occurs on core arguments that are patient-like to some degree, including patient, theme, causee, goal, and experiencer. There are double-accusative causative and transfer-ofpossession constructions that have the kind of typical asymmetrical object properties that are handled straightforwardly in Lexical-Functional Grammar (LFG) by the distinction between primary object (OBJ) and secondary object $\left(\mathrm{OBJ}_{\theta}\right)$. The accusative case marker can be analyzed as going on both kinds of object because it is constrained to occur on NPs with the GF feature specification $[+o]$. In addition, there is a desiderative construction that can have no apparent subject and the experiencer argument realized with accusative case, possibly in addition to another pa-tient-like accusative argument. In this case, the more patient-like accusative argument behaves like an $\mathrm{OBJ}_{\theta}$ and the experiencer like an OBJ in some ways and like a subject in others. In earlier analyses (Jake 1985, Hermon 1985), the experiencer is analyzed as an object at some level and a subject at another. The properties of this construction can be accounted for in LFG by analyzing the experiencer as OBJ and attributing its subject-like properties to its status as pivot (PIV) in the sense of Falk (2006).


## 1 Basic Case System

Imbabura Quechua (IQ) is an SOV language with flexible word order and a mixture of head-marking and dependent-marking properties, in the sense of Nichols (1986). As illustrated by the following examples (adapted from Jake 1985), ${ }^{1}$ subjects and oblique agents are not case-marked; other dependents are marked with a variety of case suffixes, including accusative, dative, benefactive, ablative, locative, and instrumental; verbs show agreement with subjects and optionally with a $1^{\text {st }}$ singular object; and pronominal subjects and objects can be pro-dropped.
(1) a. kan-ga kuchillu-wan (ñuka-ta) kuchu-wa-rka-ngui

2-TOP knife-INSTR 1SG-ACC cut-1SG.OBJ-PST-2SBJ
'You cut me with a knife.'
b. quitsa jari-man aswa-ta kara-rka-mi
girl man-DAT beer-ACC serve-3SBJ.PST-VAL
'The girl served beer to the man.'

[^200]c. chugri-manda-ka mana trabaja-sha-chu
wound-ABL-TOP NEG work-1SBJ.FUT-NEG
'I won't work because of the wound.'
d. alku-kuna ñuka-nchi-ka kani-shka-mi ka-rka-nchi
dog-PL 1-PL-TOP bite-PASS-VAL be-PST-1PL.SBJ
'We were bitten by the dogs.'
The topic marker - $k a$ (with phonetically-conditioned allomorph -ga) typically appears on the clausal subject, but need not appear at all, as illustrated by (1b), and may appear on non-subjects, as illustrated by (1c). Passive clauses generally have a copular auxiliary inflected to agree with a non-agent argument, the main verb in a participial form, and both the subject and the oblique agent without morphological case, as shown by (1d).

In this paper, we take the view that agreement morphology on the verb and case marking on dependents are morphological reflexes of grammatical function (GF) categories, like subject and object, or a combination of GF and semantic role categories, like goal and instrument. We assume that the GFs of arguments are defined in terms of the binary features $[ \pm r]$ (restricted) and $[ \pm 0]$ (object), as in standard LFG mapping theory (e.g., Bresnan \& Kanerva 1989, Bresnan \& Moshi 1990):

Argument GF features

|  | $[-\mathrm{r}]$ | $[+\mathrm{r}]$ |
| :--- | :--- | :--- |
| $[-\mathrm{o}]$ | SUBJ | $\mathrm{OBL}_{\theta}$ |
| $[+\mathrm{o}]$ | OBJ | $\mathrm{OBJ}_{\theta}$ |

In addition, there are various other GFs, including ADJ (adjunct), TOP (topic), FOC (focus), and PIV (pivot) (Falk 2001, Falk 2006). The overlay/discourse GFs, including TOP and PIV, are associated with dependents that also bear argument or ADJ GFs. The argument GFs are each associated with one and only one argument of a given predicate, as dictated by a principle of Function-Argument Biuniqueness (Bresnan 1982), with the understanding that $\mathrm{OBL}_{\mathrm{GOAL}}$ and $\mathrm{OBL}_{\mathrm{INSTR}}$, for example, are distinct GFs. The topic marker in IQ, in general, marks discourse topics, following a grammar that we make no attempt to elucidate here.

An overt manifestation of subject agreement appears on a tensed verb or auxiliary when there is a $1^{\text {st }}$ or $2^{\text {nd }}$ person SUBJ. Otherwise, the verb is in a default or $3^{\text {rd }}$ person SUBJ agreement form, which usually means that there is a tense affix that can be interpreted as also marking $3^{\text {rd }}$ person by virtue of the absence of $1^{\text {st }}$ or $2^{\text {nd }}$ person morphology, as in (1b). Dependents with GF $\mathrm{OBL}_{\theta}$ or ADJ (adjunct) are marked with a case that is related to a semantic role. OBJ and $\mathrm{OBJ}_{\theta}$ are marked with accusative case. Elsewhere, case is not expressed. The case marking conventions need to ensure that the "semantic" cases are restricted to obliques and adjuncts, since an instrument, for exam-
ple, that happens to have the subject GF cannot be marked with instrumental case, as in the following passive example:
(3) SUBJ OBL ${ }_{\text {AGENT }}$
pala-ka ñuka alla-shka ka-rka
shovel-TOP 1SG dig-PASS be-3SBJ.PST
'The shovel was dug with by me.'
There is no morphological case restricted specifically to SUBJ or OBJ. Although SUBJ is not case-marked, this is not a circumstance restricted to SUBJ or any particular GF. Although it might be possible to say that there is an abstract nominative case that is associated with SUBJ and that this case has no phonological manifestation, no benefit appears to accrue from this, as no aspect of the grammar is sensitive specifically to what would be designated nominative case (as opposed to SUBJ). We assume here that overt morphological case and its presence and absence are all that need to be accounted for. With this in mind, we use Butt's (2006) general approach to case, without the assumption that there are default rules to assign (possibly abstract) case to SUBJ and OBJ. All case morphemes can be treated as lexical items that are constrained, by inside-out functional uncertainty, to occur on dependents with compatible GF and semantic role: ${ }^{2}$
(4) Lexical entries for selected case-marking suffixes

```
-man: \(\quad(\uparrow \mathrm{CASE})=\mathrm{DAT}\)
    (GOAL \(\uparrow l c-s t r)\)
    \((\neg[-\mathrm{r}] \uparrow)\)
\(-m a n d a:(\uparrow \mathrm{CASE})=\mathrm{ABL}\)
        (SOURCE \(\uparrow l c-s t r)\)
        \((\neg[-\mathrm{r}] \uparrow)\)
-wan: \(\quad(\uparrow\) CASE \()=\) INSTR
    (INSTRUMENT \(\uparrow l c-s t r)\)
    ( \(\neg-\mathrm{r}] \uparrow\) )
\(-t a: \quad(\uparrow \mathrm{CASE})=\mathrm{ACC}\)
    ([+o] \(\uparrow\) )
```

[^201]The GF notation $\neg[-\mathrm{r}]$ simply precludes the semantic cases from occurring on SUBJ or OBJ, while allowing them to occur where they do, i.e., on either ADJ or OBL. Ability to be used with both oblique arguments and adjuncts is characteristic of semantic cases in IQ and other languages. The ablative case marker, for example, occurs on an adjunct in (1c). But, this morpheme can also be used on an oblique argument:
OBL ${ }_{\text {SOURCE }} \quad$ SUBJ
urku-manda-ka supai-kuna shamu-nga-chari
mountain-ABL-TOP devil-PL come-3SBJ.FUT-DUB
'Maybe the devils will come from the mountains.'

The accusative suffix is constrained to appear on members of the class of GFs specified by the feature $[+o] .{ }^{3}$ This accounts for the fact that it occurs on both OBJ and $\mathrm{OBJ}_{\theta}$, as in the following example.

$$
\begin{equation*}
 \tag{6}
\end{equation*}
$$

Just in case a semantic goal has the GF OBJ (or SUBJ), the dative case marker is not used, since it is incompatible with a [-r] GF specification.

The standard idea that a Case Filter requires all NPs to bear case is not useful for IQ (under the assumption that overt morphological case is what is at issue), since even leaving aside the case of SUBJ and OBL ${ }_{\text {AGENT }}$, there are conditions under which case morphology need not occur (Jake 1985:21-23). For example, the locative case suffix is optional on dependents that are interpreted as being temporally rather than spatially locative:
(7) $\mathrm{ADJ}_{\text {LOC }}$

$$
\begin{array}{ll}
\text { Lunis(-pi) } & \text { ri-sha-mi } \\
\text { Monday-LOC go-1SBJ.FUT-VAL } \\
\text { 'I'll go (on) Monday.' }
\end{array}
$$

[^202]Similarly, instrumental case is not required in cases where the interpretation is comitative. Preceding a nominalized embedded verb, accusative case marking is also optional. Still, in most cases, a case suffix must appear if it is lexically licensed. A set of language-specific well-formedness conditions such as the following, which spell out the default or elsewhere condition and the exceptions, can be assumed to account for the appearance of case morphology, in conjunction with the lexical specifications of the case suffixes.
(8) Well-formedness conditions on morphological case in IQ
a. Locative case may be omitted on temporally locative adjuncts.
b. Instrumental case may be omitted on comitatively instrumental adjuncts.
c. Accusative case may be omitted on a dependent preceding a nominalized verb.
n. Elsewhere, nominal phrases that are compatible with a lexical case must be case-marked.

This set of conditions and the non-existence of case suffixes for SUBJ and OBL $_{\text {AGENT }}$ account for the necessary absence of case morphology in sentences such as (3).

The main concern of this paper is the analysis of the following accusative experiencer construction (Hermon 1985:1).
(9) ñuka-ta aycha-ta miku-naya-wa-n-mi

1SG-ACC meat-ACC eat-DESID-1SG.OBJ-3SBJ-VAL
'I want to eat meat.'
This is a syntactically monoclausal construction with a complex verb consisting of a verb stem, ${ }^{4}$ the desiderative suffix -naya, and the appropriate inflectional morphemes that occur on all verbs. We assume here that the semantic argument of -naya binds the most prominent semantic argument of the verb stem (as in Farrell 1995) at the level of lexical conceptual structure and that GF linking treats the semantically complex "experiencer" dependent as a single element in argument structure, functional structure, and constituent structure. The problem, in a nutshell, is that this experiencer argument has both various object properties and some subject properties. Previous analyses in Relational Grammar and Government-Binding Theory (Jake 1985, Hermon 1985) treat the experiencer as the object of a complex predicate at some level and the subject of the same predicate at another. Moreover, the argument in question is generally characterized as a non-canonical subject, i.e., basically a subject with unexpected object marking (Cole \& Jake 1978, Cole \& Hermon 1991, Hermon 2001). Our goal here is to show that the accusative experi-

[^203]encer may be better characterized as a non-canonical object. Consistent with the principle of Function-Argument Biuniqueness, it bears only the OBJ GF of the complex predicate. However, whereas most languages require an alignment of SUBJ with the overlay GF PIV (Falk 2006), which is what some "subject-sensitive" phenomena in some languages are keyed to, IQ aligns the PIV function with the OBJ in the construction type in question.

## 2 Double-Object Constructions

The main observation underlying the proposed analysis of the desiderative construction is that, abstracting away from the absence of an agent subject, its syntactic properties closely parallel those of other double-accusative constructions: ${ }^{\text {5 }}$
(10) DOUBLE-ACCUSATIVE CAUSATIVE CONSTRUCTION
a. taita-ka churi-ta ruwana-ta awa-chi-rka-mi
father-TOP son-ACC poncho-ACC weave-CAUS-3SBJ.PST-VAL
'The father made his son weave a poncho.'
DOUBLE-ACCUSATIVE TRANSFER-OF-POSSESSION CONSTRUCTION
b. warmi-ka jari-ta aswa-ta ku-rka-chu
woman-TOP man-ACC beer-ACC give-3PST-Q
'Did the woman give the man beer?'
DOUBLE-ACCUSATIVE DESIDERATIVE CONSTRUCTION
c. jari-ta-ka aswa-ta ufya-naya-n
man-ACC-TOP beer-ACC drink-DESID-3SBJ
'The man wants to drink beer.'
To begin with, all three constructions are systematically related to an alternative construction with only a single accusative dependent; and in all three cases the lone accusative dependent is the theme:
a. taita-ka churi-man ruwana-ta awa-chi-rka-mi
father-TOP son-DAT poncho-ACC weave-CAUS-3PST-VAL
'The father let his son weave a poncho.'
b. warmi-ka jari-man aswa-ta ku-rka-chu
woman-TOP man-DAT beer-ACC give-3SBJ.PST-Q
'Did the woman give beer to the man?'

[^204]c. jari-ka aswa-ta ufya-naya-n
man-TOP beer-ACC drink-DESID-3SBJ
'The man wants to drink beer.'
In the case of the causative, the choice between the alternative constructions is correlated with a semantic distinction, i.e., coercive vs. permissive causation. There may be a semantic difference between (10b) and (11b)—perhaps something along the same lines as whatever difference there may be in the English glosses. It is also unclear whether (10c) has a slightly different meaning than (11c), ${ }^{6}$ possibly corresponding to a more literal (but unavailable) English gloss such as 'It wants to the man to drink beer'.

In any case, the hypothesis to be entertained is that, since accusative case marking indicates objecthood of some kind, as shown in the proposed lexical entry for $-t a$ in (4), and there can be only one OBJ in any given clause, there is both an OBJ and an $\mathrm{OBJ}_{\theta}$ in all of the examples in (10) and the semantic argument that is realized as $\mathrm{OBJ}_{\theta}$ in (10) is realized as the single OBJ in the examples in (11). Initial support for this hypothesis comes from the way these constructions interact with passive. There is a passive version of each of the single-accusative constructions and, in each case, the semantic argument that is marked accusative in the construction types illustrated by (11) is the SUBJ in the passive clause and the other argument cannot be marked accusative:
(12) a. ruwana-ka taita churi-man/*ta awa-chi-shka ka-rka poncho-TOP father son-DAT/ACC weave-CAUS-PASS be-3SBJ.PST 'The poncho was let/*made to be woven by his son by the father.'
b. aswa-ka jari-man/*ta quitsa kara-shka-mi ka-rka
beer-TOP man-DAT/ACC girl serve-PASS-VAL be-3SBJ.PST
'The beer was served *(to) the man by the girl.'
c. wawa-ka ñuka(*-ta) wajta-naya-shka ka-rka
child-TOP 1SG-ACC hit-DESID-PRES be-3SBJ.PST
'I wanted to hit the child.'
(literally: 'The child was wanted-to-be-hit by me')
It is also possible to have the causee, goal, or experiencer realized as SUBJ in a passive clause, in which case the other argument, if expressed, is marked accusative: ${ }^{7}$

[^205]a. ñuka-ka Maria papa-ta yanu-chi-shka ka-rka-ni 1SG-TOP Maria potato-ACC cook-CAUS-PASS be-PST-1SG.SBJ 'I was made to cook potatoes by Maria.'
b. quitsa-ka mishqui-ta mama kara-shka-mi ka-rka girl-TOP candy-ACC mother serve-PASS-VAL be-3SBJ.PST 'The girl was served candy by her mother.'
c. ñuka-ka mishqui-ta miku-naya-shka ka-rka-ni 1SG-TOP candy-ACC eat-DESID-PASS be- PST-1SG.SBJ 'I wanted to eat candy.'
(literally: 'I was wanted-to-be-eaten candy (by it)')
The standard way to handle facts like these in LFG mapping theory is to treat the possibility of alternation as a reflex of different inherent GF classifications of specified semantic arguments. The causee, goal, and experiencer can either be inherently [-r], and therefore active OBJ and passive SUBJ, or not. We assume here a version of mapping theory like that articulated in Kibort (2004). Specifically, argument structure consists of a set of ranked arguments (aligned with generalized semantic roles in ways that can vary across and within languages). These are constrained to have only certain inherent GF classifications and to map to GFs by a general mapping principle:
(14) Argument structure and inherent GF classifications
\[

$$
\begin{array}{cccccc}
\arg _{1} & \arg _{2} & \arg _{3} & \arg _{4} & \arg _{\mathrm{n}} \\
{[-\mathrm{o}] \text { or }[-\mathrm{r}]} & {[-\mathrm{r}]} & {[+\mathrm{o}]} & {[-\mathrm{o}]} & & {[-\mathrm{o}]}
\end{array}
$$
\]

## Mapping Principle

Arguments are mapped onto the highest (i.e., least marked) compatible function on the markedness hierarchy.

```
GF Markedness Hierarchy
[-o,-r]=SUBJ > [-r, +o]=OBJ > [-o, +r]= OBL 
```

The way that this works for transfer-of-possession predicates in the active voice is as follows.
(15) DOUBLE-OBJECT CONSTRUCTION, as in (10b)

| lc-str: | AGENT | GOAL | THEME |
| :--- | :---: | :---: | :---: |
| arg-str: | $\arg _{1}$ | $\arg _{2}$ | $\arg _{3}$ |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[+\mathrm{o}]$ |
| GFs: | SUBJ | OBJ | $\mathrm{OBJ}_{\theta}$ |
| case $:$ |  | ACC | ACC |

ing to eat candy' = 'I wanted to eat candy'). More importantly, as Jake notes, a $1^{\text {st }}$ person non-accusative experiencer subject of an active voice desiderative clause has the exceptional property of not triggering subject agreement. The agreement shown in (13c) is therefore unexpected on the resultative analysis, but not on the passive analysis.

| SINGLE-OBJECT CONSTRUCTION, as in (11b) |  |  |  |
| :--- | :---: | :---: | :---: |
| lc-str: | AGENT | THEME | GOAL |
| arg-str: | $\arg _{1}$ | $\arg _{2}$ | $\arg _{4}$ |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ |
| GFs: | SUBJ | OBJ | OBL $_{\theta}$ |
| case: |  | ACC | DAT |

Passive voice is the outcome of an override of the default mapping of $\arg _{1}$, such that it maps to $\mathrm{OBL}_{\theta}$, rather than SUBJ, which gives rise to a mapping to SUBJ of the least-marked remaining choice among arguments that are inherently [-r] or [-o]. Given this, the only possible passive realization for the argument structure underlying (10b) is (13b), with the goal argument as SUBJ and the patient as $\mathrm{OBJ}_{\theta}$, and the only possible passive realization for the argument structure underlying (11b) is (12b), with the patient as SUBJ and the goal as $\mathrm{OBL}_{\theta}$ :
(16) PASSIVE OF DOUBLE-OBJECT CONSTRUCTION, as in (13b)

| lc-str: | AGENT | GOAL | THEME |
| :--- | :---: | :---: | :---: |
| arg-str: | $\arg _{1}$ | $\arg _{2}$ | $\arg _{3}$ |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[+\mathrm{o}]$ |
| passive: | $[+\mathrm{r}]$ |  |  |
| GFs: | $\mathrm{OBL}_{\theta}$ | SUBJ | $\mathrm{OBJ}_{\theta}$ |
| case: |  |  | ACC |

PASSIVE OF SINGLE-OBJECT CONSTRUCTION, as in (12b)
lc-str: AGENT THEME GOAL

| arg-str: | $\arg _{1}$ | $\arg _{2}$ | $\arg _{4}$ |
| :--- | :--- | :---: | :---: |
|  | $[-\mathrm{o}]$ | $[-\mathrm{r}]$ | $[-\mathrm{o}]$ |
| passive: | $[+\mathrm{r}]$ |  |  |
| GFs: | $\mathrm{OBL}_{\theta}$ | SUBJ | $\mathrm{OBL}_{\theta}$ |
| case: |  |  | DAT |

In essence, the goal alternates between the $\arg _{2}$ and the $\arg _{4}$ positions, with the patient alternating correlatively with the $\arg _{3}$ and $\arg _{2}$ positions. This is a common pattern across languages and is, of course, found in English. The general schema is that semantic arguments with sufficient patient-like properties vie for the $\arg _{2}$ slot and only if it loses out for this, a theme is necessarily an $\arg _{3}$, and therefore $\mathrm{OBJ}_{\theta}$. The causative construction has essentially the same analysis, with the causee being treated like the goal. With the desiderative construction, the key difference, we propose, is that the experiencer, which is patient-like enough to have a [-r] inherent classification, alternates between the $\arg _{1}$ and $\arg g_{2}$ positions, rather than the $\arg _{4}$ and $\arg g_{2}$ positions. The reason for this, plausibly, is that an experiencer argument is both agentlike and patient-like and can therefore alternate between $\arg _{1}$ and $\arg _{2}$ across and within languages. The related effects of this alternation on a theme argument, if present, are the same as with the transfer-of-possession construc-
tion. In order to get the mapping to GFs to work according to the schema outlined in (14), we assume that when the experiencer is in the $\arg _{2}$ position, the $\arg _{1}$ is a null expletive, i.e., it is a syntactic argument with a GF that corresponds to no role at all in lexical conceptual structure. Abstracting away from the effects of the dative alternation, to be discussed below, the mappings for the active-voice alternative desiderative constructions are as follows, beginning with the construction in which the experiencer is realized as OBJ.


The embedded lexical conceptual structure of the stem to which the desiderative suffix attaches is indicated by bracketing. The assumption is that the experiencer argument of -naya binds the highest semantic role of the stem verb in lexical conceptual structure. Since the stem can have any number of arguments, the general schema has to allow for this. Parentheses indicate optional expression. In the case of $(10 \mathrm{c})$, the stem is monotransitive, with only agent and theme. (18a) and (18b) illustrate manifestations of the same construction built on an intransitive verb stem and on a transitive verb stem with an additional oblique argument.

```
a. SUBJ OBJ
    ñuka-ta-ka puñu-naya-rka
    it 1SG-ACC-TOP sleep-DESID-3SBJ.PST
    'I wanted to sleep.'
```

| b. SUBJ | OBJ $\mathrm{OBL}_{\text {GOAL }}$ <br> wawa-ta-ka $\mathrm{OBJ}_{\text {THeme }}$ <br> kan-man parlu-ta villa-naya-n |
| :--- | :--- | :--- |

    it child-ACC-TOP 2-DAT story-ACC tell-DESID-3SBJ
    'The child wants to tell a story to you.'
    The construction with the experiencer as subject has the following mapping:


Again, this general mapping schema can be employed with various stem types. (11c) is an example with a simple transitive stem.

As expected, each of desiderative constructions has a single passive voice realization. Passivization of the experiencer-object construction works as follows:

(13c) exemplifies the passive of a desiderative built on a simple transitive stem. (21) illustrates the passive of (18a), i.e., a desiderative built on an intransitive stem.

```
OBLø SUBJ
        ñuka-ta puñu-naya-shka ka-rka-ni
it 1SG-TOP sleep-DESID-PASS be-PST-1SG.SBJ
'I wanted to sleep.' (literally: 'I was wanted-to-be-slept (by it)')
```

The passive version of the experiencer-subject desiderative construction employs the following mapping schema:


We are aware, of course, that some theories of grammar avoid positing either or both null expletives and oblique expletives and that we posit both in the case of (21), for example. Without being committed to its ultimate correctness, we take this stance here for several reasons. First, the basic architecture of LFG mapping theory, as outlined in (14), makes it such that there must be an expletive $\arg _{1}$ in a clause type in which the highest GF borne by any of the semantic arguments of the verb is OBJ. It is the presence of a null expletive SUBJ in (18b), for example, that makes it possible to account in a technically straightforward way for the fact that neither the experiencer nor the goal is mapped to SUBJ. Second, oblique expletives appear to exist as a natural language possibility, as in such cases as You should see to it that
nothing happens and I'm not bothered by it that they're winning (see Postal \& Pullum 1988). Third, since the passive construction in IQ is otherwise characterized by a mapping to OBL of $\arg _{1}$, which is what opens up the possibility of another argument mapping to SUBJ, there is no reason not to use this same characterization for a passive clause with a null expletive $\arg _{1}$, although an analysis with suppression of GF-mapping for the null expletive in this case would also account for the facts.

## 3 Supporting Evidence

The analysis of the experiencer-object desiderative construction summarized in (17) is supported by the fact that it accounts for its case marking and subject agreement properties straightforwardly by simply applying general principles for mapping arguments to GFs. The experiencer is marked with accusative case because it is OBJ and dependents with this GF are marked accusative in IQ. It does not trigger subject agreement because only SUBJ does this. The verb is necessarily in its default subject agreement form (not $1^{\text {st }}$ or $2^{\text {nd }}$ person) because there is a null expletive SUBJ. If there is a theme argument present, it is $\mathrm{OBJ}_{\theta}$ because, just as in the case of transfer-ofpossession and causative constructions with goal or causee as primary object, the only possibility for a theme argument that cannot be $\arg _{2}$ is $\arg _{3}$, which must map to $\mathrm{OBJ}_{\theta}$. The theme is marked accusative because any kind of object (i.e., an argument mapped to a [ +o ] GF) bears accusative case in IQ. The way that passive voice works with the proposed argument structure follows without stipulation. The remainder of this section is devoted to providing additional supporting evidence for the analysis.

### 3.1 Accusative Experiencer is not SUBJ

## Argument One: Object Agreement

So-called object agreement in IQ is restricted to the optional $1^{\text {st }}$ singular suffix -wa. It cross-references only the OBJ, as in (23a), or some range of hu-man-referring OBL $_{\theta}$ arguments (Jake 1985:30), as illustrated by (23b-c).
a. Maria-ka (ñuka-ta) maka-wa-rka-mi Maria-TOP 1SG-ACC hit-1SG.OBJ-3SBJ.PST-VAL 'Maria hit me.'
b. Maria-ka (ñuka-paj) trabaja-wa-rka-mi

Maria-TOP 1SG-BEN work-1SG.OBJ-3SBJ.PST-VAL
'Maria worked for me.'
c. Maria-ka (ñuka-wan) parla-wa-rka-mi

Maria-TOP 1SG-INSTR talk-1SG.OBJ-3SBJ.PST-vAL
'Maria talked with me.'

If the accusative experiencer in the desiderative construction is OBJ and not SUBJ, it follows that it can determine object agreement, as shown in (9), since $-w a$ is constrained to cross-reference only non-SUBJ $1^{\text {st }}$ person arguments.

## Argument Two: Switch Reference

There is switch reference morphology on the verb of an adverbial clause indicating the status of its subject vis-à-vis that of the main clause with respect to matters of coreference (Jake 1985:35). When the adverbial clause and the main clause do not have subjects with disjoint reference, the switch-reference marker is either -shpa (temporal) or -ngapaj (purposive) (glossed CREF); when they do not have coreferential subjects, the marker is -jpi or -chun (glossed SWR) as illustrated by the following examples.
a. [wasi-man ri-shpa/*-jpi] miku-ngui-chu
house-DAT go-CREF/-SWR eat-2SG.SBJ-Q
'When you go home, will you eat?'
b. [mama tigra-mu-*shpa/-jpi] miku-sha-mi
mother return-CISLOC-CREF/-SWR eat-1SBJ.FUT-VAL
'When mother returns, I'll eat.'
c. ñuka pani kaya-wa-rka [parla-wa-ngapaj/*-chun]

1SG sister call-1SG.OBJ-PST talk-1SG.OBJ-CREF/-SWR 'My sister called me to talk to me.'
d. [chai jari kalpa-*ngapaj/-chun] kaya-rka-ni-mi that man run-CREF/-SWR call-PST-1SG.SBJ-VAL 'I called for that man to run.'

When the verb of one or both of the clauses has an expletive subject or nonreferential subject, as in the case of weather verbs, either switch-reference marker can appear:
a. [tamya-ju-shpa/-jpi] wawa-kuna mana shamu-nga-chu rain-PROG-CREF/-SWR child-PL NEG come-3SBJ.FUT-NEG 'If it's raining, the children won't come.'
b. [tamya-gri-ngapaj/-chun] waira fuku-shka-mi rain-INCH-CREF/-SWR wind blow-PERF-VAL 'The wind blew (enough) for it to rain.'

When a weather verb is involved, both subjects with disjoint reference and subjects with coreference are lacking, which is consistent with the meanings of both kinds of switch-reference marking. Desiderative verbs with an object experiencer behave like weather verbs with respect to switch-reference marking (Hermon 2001:163-4): ${ }^{8}$

[^206]a. [(ñuka-ka) trabaja-jpi-ka] miku-naya-wa-rka-mi 1SG-TOP work-SWR-TOP eat-DESID-1SG.OBJ-3SBJ.PST-VAL 'While I was working, I had a desire to eat.'
b. [trabaja-shpa-ka] miku-naya-wa-rka-mi work-CREF-TOP eat-DESID-1SG.OBJ-3SBJ.PST-VAL 'While I was working, I had a desire to eat.'

If the experiencer in the main clause, which is registered by object agreement, is OBJ and there is no referential subject, the switch-reference marking is exactly as expected: -shpa is possible because there are not subjects with disjoint reference and -jpi is possible because there are not coreferential subjects. Under the proposed analysis, the non-referential expletive SUBJ in the object-experiencer desiderative construction guarantees both lack of coreference and lack of disjoint reference.

### 3.2 Accusative Experiencer is OBJ

## Argument One: Tough Movement

$\mathrm{OBJ}_{\theta}$ only occurs in clauses in which there is an OBJ, as shown in section 2. In the double-object scenario, the OBJ has certain syntactic privileges that $\mathrm{OBJ}_{\theta}$ lacks. One of these is that the OBJ can be the target (i.e., elided argument) in the embedded clause of the tough-movement construction (Jake 1985:136-142), as illustrated by the following examples.
a. wawa-ka mana sinchi-chu ka-rka [mama mishki-ta
child-TOP NEG tough-NEG be-3SBJ.PST mother candy-ACC
kara-chun-ga]
serve-SWR-TOP
'The baby wasn't difficult for the mother to give candy.'
b. * libru-ka facil-mi ka-nga [ñuka kan-da apa-chun-ga]
book-TOP easy-VAL be-3SBJ.FUT 1SG 2-ACC carry-SWR-TOP
'The book will be easy for me to bring you.'
(27b) shows that the secondary accusative argument (the theme) in the dou-ble-accusative transfer-of possession construction cannot be the target in the tough-movement construction. The experiencer-object desiderative construction shows the same pattern, i.e., the theme argument cannot be a toughmovement target:

> *aswa-ka ali-mi $\quad$ [kan-da ufya-naya-chun] beer-TOP good-VAL 2-ACC drink-DESID-SWR 'Beer is good for you to want to drink.'
which there can be no overt expression of subject. The condition on control, which we discuss below, is a partially separate matter.

Since the theme argument of a verb such as 'drink' is otherwise an OBJ, its status as $\mathrm{OBJ}_{\theta}$ in the desiderative construction can only be attributed to the unavailability of the OBJ GF. Under the proposed analysis, the OBJ GF is unavailable for the theme because it is taken by the experiencer.

## Argument Two: Desiderative Transfer-of-Possession Verbs

In a desiderative construction formed on a verb with a goal and theme, there cannot be three accusative arguments (Jake 1985:216):

> * jari-ta ñuka-ta kafi-ta kara-naya-rka-chu man-ACC 1 SG-ACC coffee-ACC serve-DESID-3SBJ.PST-Q 'Did the man want to serve me coffee?'

The ungrammaticality of this construction is explained on the proposed analysis as follows. The experiencer is marked accusative because $\arg _{1}$ is a null expletive and the experiencer is, therefore, $\arg _{2}$ (classified [-r] and realized as OBJ). No other argument can be OBJ. The only possibility for a theme that is not $\arg _{2}$ is $\arg _{3}$, which must be mapped to $\mathrm{OBJ}_{\theta}$. The inherent classifications for a goal are limited to [-r] $\left(\arg _{2}\right)$ and $[-\mathrm{o}]\left(\arg _{4}\right)$. For an experiencer they are limited to $[-\mathrm{r}]\left(\arg _{2}\right)$ and $[-\mathrm{o}]\left(\arg _{1}\right)$. The triple-accusative desiderative construction is ruled out by the fact that there is only one available $\arg _{2}$ slot and the experiencer and goal cannot both occupy it simultaneously. There are two repairs. One is to have the goal show up as a dativemarked oblique, by having the desiderative attach to a verb with an argument structure that yields an accusative + dative construction, as in the case of (30b), which is the (declarative) desiderative form of (30a).
a. wawa-ka kan-man parlu-ta villa-rka-chu child-TOP 2-DAT story-ACC tell-3SBJ.PST-Q 'Did the child tell a story to you?'
b. wawa-ta-ka kan-man parlu-ta villa-naya-n $=(18 \mathrm{~b})$ child-ACC-TOP 2-DAT story-ACC tell-DESID-3SBJ 'The child wants to tell a story to you.'

The other repair is to have the experiencer show up as a subject, with no case marking, in the alternative desiderative construction:

```
jari ñuka-ta kafi-ta kara-naya-rka-chu
man 1SG-ACC coffee-ACC serve-DESID-3SBJ.PST-Q
'Did the man want to serve me coffee?'
```

The key point is that the ungrammaticality of (29) follows from the analysis of the accusative experiencer as OBJ. If it were an exceptionally marked SUBJ or OBL, nothing would preclude a goal and theme from being realized as OBJ and $\mathrm{OBJ}_{\theta}$, respectively, and bearing accusative case.

## 4 "Subject" Properties of Accusative Experiencer

## PIV-Sensitive Phenomena

It is well known that the so-called that-trace effect in English, i.e., the ungrammaticality of subject extraction from an embedded clause headed by complementizer that, is a phenomenon that occurs in some way in many languages (Falk 2006). In IQ, it so happens that non-subjects, in general, can be extracted from an embedded clause (the verb of which is nominalized), but subjects cannot and in the experiencer-object desiderative construction, the accusative experiencer, unlike the accusative theme, behaves like a subject, as illustrated by the following examples (Cole \& Hermon 1991:13-14)

```
a. * pi-taj Maria kri-n
                    [
                                    aicha-ta miku-shka-ta]
    AN-WH Maria believe-3SBJ meat-ACC eat-PST.NOM-ACC
    'Who does Maria think that ate the meat?'
    b. ima-ta-taj Maria kri-n [Juzi _ miku-shka-ta]
    INAN-ACC-WH Maria believe-3SBJ Jose eat-PST.NOM-ACC
    'What does Maria believe that Jose ate?'
```

a. pi-ta-taj Maria Juzi-man ni-rka [Juan-da AN-ACC-WH Maria Jose-DAT say-3SBJ.PST Juan-ACC riku-naya-j-ta]

```
        see-DESID-PRES.NOM-ACC
    'Who did Maria say to Jose that Juan wants to see?'
b. * pi-ta-taj Maria Juzi-man ni-rka
    AN-ACC-WH Maria Jose-DAT say-3SBJ.PST
    [ _ miku-naya-j-ta]
        eat-DESID-PRES.NOM-ACC
    'Who did Maria say to Jose that wants to eat?'
```

The accusative experiencer in the desiderative construction also behaves like a subject with respect to control of adverbial clauses (Hermon 1985:124125). As shown by the following examples, the SUBJ but not the OBJ of a typical main clause can be interpreted as the same as the missing (controlled) subject of the adverbial clause, represented here as pro, and yet the experiencer OBJ in the desiderative construction, which determines object agreement on the verb, can control like a subject.
a. [pro $\mathrm{i}_{\mathrm{i} /{ }_{\mathrm{j}}}$ miku-ju-shpa] Juan ${ }_{j}$-da riku-rka-ni ${ }_{i}$ eat-PROG-CREF Juan-ACC saw-PST-1SG.SBJ
'I saw John, when eating.' (NOT: when he was eating)
b. [ $\operatorname{pro}_{i}$ trabaja-shpa-ka] miku-naya-wai-rka-mi work-CREF-TOP eat-DESID-1SG.OBJ-3SBJ.PST-VAL
'While I was working, I had a desire to eat.' (= (26b))

## The Analysis

It might be possible to account for the subject-like properties of the accusative experiencer by treating it as a SUBJ with exceptional (or "quirky") case, i.e., an accusative subject in the sense of Icelandic, for example (Zaenen, Maling, \& Thráinsson 1985, Van Valin 1991). However, not only would such an analysis lose the generalization about accusative case that the analysis proposed here makes possible (i.e., objects of any kind bear accusative case), but it would have to treat all the coding properties of the accusative-experiencer desiderative construction as exceptional and would not account for the facts discussed in section 3.2. Moreover, it is unclear how one might explain why, unlike in Icelandic, accusative case does not occur on the experiencer when it is "raised" to a higher clause (Hermon 1985:114):
(35) kan-ga [puñu-naya-y] yari-ngui

2-TOP sleep-DESID-INF seem-2SG.SBJ
'You seem to want to sleep.'
With these things in mind, it is preferable to treat the accusative experiencer as a non-canonical OBJ. Its subject properties are, in fact, properties that are characteristic of dependents with the overlay GF PIV, in languages that provide evidence for factoring the traditional SUBJ GF into two potentially independent GFs: highest argument GF (= SUBJ in this paper) and PIV (Falk 2006). What is routine about IQ is that the conditions on the controller of adverbial clauses and on extraction, for example, are sensitive to the PIV GF. What is exceptional is only that, although the default scenario is SUBJ = PIV, in the experiencer-object desiderative construction OBJ = PIV. The generalization might be that in each clause the highest argument GF associated with a semantic role is aligned with the overlay function PIV.

## Residual matters

Hermon $(1985,2001)$ notes that there are a few other subject properties of the desiderative experiencer, including ability to be controlled and ability to raise to subject, as illustrated by (35). Since the relevant evidence comes from infinitival desideratives in which there is no overt coding of the GF of the experiencer, it is possible that these are actually SUBJ-sensitive phenomena and the infinitival clauses instantiate the experiencer-subject argument structure of desiderative verbs (see (17)). Evidence for this interpretation comes from the so-called lexical experiencer construction, which typically consists of a verb of physical experience with a single OBJ argument that does not have an alternative SUBJ realization and therefore necessarily bears accusative case when expressed as an NP. The experiencer in this construction generally has the same applicable properties as the experiencer in the experiencer-object desiderative construction. It differs, however, in not being able to be controlled and not being able to raise to subject (Hermon 1985:114):
(36) *kan-ga [yarja-y] yari-ngui

2-TOP hunger-INF seem-2SG.SBJ
'You seem to be hungry.'
Necessarily left unresolved here is the question of dialectal variation in the coding of the alternative desiderative constructions. The speakers whose variety of IQ the description in Hermon (1985) is based on are said to mark the -naya-suffixed verbs with the causative morpheme -chi (see footnote 6). The analysis suggested here entails that this morphological coding would have to be context-sensitive, such that it need not (perhaps could not) appear on infinitival verb forms, as in (35), for example. Whether this is a viable analysis is a question that requires further investigation and leads well beyond the scope of this paper.

## 5 Conclusion

We began with the observation that there seems to be a generalization concerning the occurrence of the accusative case suffix -ta in IQ, as it appears, in general, on what is clearly either a primary object (OBJ) or a secondary object $\left(\mathrm{OBJ}_{\theta}\right)$, as in the following typical double-accusative example:

```
PIV
SUBJ OBJ OBJ 
quitsa jari-ta aswa-ta kara-rka-mi
girl man-ACC beer-ACC serve-3SBJ.PST-VAL
'The girl served the man beer.'
```

This generalization can be captured by lexically constraining -ta to occur on NPs in the category of GFs specified by the feature [ +o ]. For this approach to be viable, it has to be the case that the double-accusative desiderative construction has accusative OBJ and $\mathrm{OBJ}_{\theta}$ as well:

PIV
SUBJ OBJ $\mathrm{OBJ}_{\theta}$
Ø ñuka-ta aycha-ta mkiu-naya-wa-n-mi
1SG-ACC meat-ACC eat-DESID-1SG.OBJ-3SBJ-VAL
'I want to eat meat.'
Such an analysis turns out to be well motivated. The claim that the experiencer is OBJ entails that a theme argument be mapped to $\mathrm{OBJ}_{\theta}$. This is supported by parallels with the theme in (37). Neither can be the target of toughmovement or the passive SUBJ in a clause in which the goal or experiencer is marked accusative. Moreover, this analysis accounts for the impossibility of a triple-accusative clause, which might be expected when -naya is suffixed to a transfer-of-possession verb which itself has a double-object valence. The explanation is that there can only be one each of OBJ and $\mathrm{OBJ}_{\theta}$ in IQ, given the principle of Function-Argument Biuniqueness and the impossibility of a
mapping of goal to $\mathrm{OBJ}_{\theta}$. The coding properties of the experiencer in this construction (case, lack of subject agreement, possibility of object agreement, and switch-reference marking) follow from the analysis of the experiencer as OBJ. The so-called subject properties that it has are actually properties that typify the PIV function cross-linguistically. The experiencer is a non-canonical OBJ by virtue of its alignment with PIV, which is otherwise constrained to align with SUBJ in IQ.

## References

Alsina, Alex. 1996. The Role of Argument Structure in Grammar: Evidence from Romance. Stanford: CSLI.
Bresnan, Joan. 1982. The Mental Representation of Grammatical Relations. Cambridge, MA: MIT Press.
Bresnan, Joan, and Jonni M. Kanerva. 1989. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. Linguistic Inquiry 20:1-50.
Bresnan, Joan, and Lioba Moshi. 1990. Object Asymmetries in Comparative Bantu Syntax. Linguistic Inquiry 21:147-185.
Butt, Miriam. 2006. Theories of Case. Cambridge: Cambridge University Press.
Cole, Peter. 1982. Imbabura Quechua. Amsterdam: North-Holland.
Cole, Peter, and Gabriella Hermon. 1991. Subjecthood and Islandhood: Evidence from Quechua. Linguistic Inquiry 12:1-30.
Cole, Peter, and Janice Jake. 1978. Accusative Subjects in Imbabura Quechua. Studies in the Linguistic Sciences 8:72-96.
Falk, Yehuda. 2001. Lexical-Functional Grammar. Stanford, CA: CSLI Publications.
Falk, Yehuda. 2006. Subjects and Universal Grammar: An Explanatory Theory. Cambridge: Cambridge University Press.
Farrell, Patrick. 1995. Lexical Binding. Linguistics 33:939-980.
Guerrero, Lilián, and Robert D. Van Valin, Jr. 2004. Yaqui and the Analysis of Primary Object Languages. International Journal of American Linguistics 70:290-319.
Hermon, Gabriella. 1985. Syntactic Modularity. Dordrecht: Foris.
Hermon, Gabriella. 2001. Non-Canonically-Marked A/S in Imbabura Quechua. In Non-Canonical Marking of Subjects and Objects, A.Y. Aikhenvald, R.M.W. Dixon, and M. Onishi (eds.), 149-176. Amsterdam: John Benjamins.
Jake, Janice. 1985. Grammatical Relations in Imbabura Quechua. New York: Garland.
Kibort, Anna. 2004. Passive and Passive-like Constructions in English and Polish. Cambridge University dissertation.
Lee, Keon Soo. 1991. Multiple Accusative Constructions in Korean and the Stratal Uniqueness Law. University of Hawaii dissertation.

Nichols, Johanna. 1986. Head-Marking and Dependent-Marking Grammar. Language 62:56-119.
Perlmutter, David M. 1982. Syntactic Representation, Syntactic Levels, and the Notion of Subject. In The Nature of Syntactic Representation, P. Jacobson and G. Pullum (eds.), 283-340. Dordrecht: D. Reidel.
Postal, Paul M., and Geoffrey K. Pullum. 1988 Expletive Noun Phrases in Subcategorized Positions. Linguistic Inquiry 19:635-670.
Van Valin, Robert D., Jr. 1991. Another Look at Icelandic Case Marking and Grammatical Relations. Natural Language and Linguistic Theory 9:145-194.
Wunderlich, Dieter, and Renate Lakämper. 2001. On the Interaction of Structural and Semantic Case. Lingua 111:377-418.
Willgohs, Beatriz. 2009. Ditransitive and Causative Constructions in Quechua, Ms., University of California, Davis.
Zaenen, Annie, Joan Maling, and Hoskuldur Thráinsson. 1985. Case and Grammatical Functions: The Icelandic Passive. Natural Language and Linguistic Theory 3:441-483.

# OBLS HOBBLE COMPUTATIONS 

Annie Zaenen and Dick Crouch PARC<br>MS-Powerset

Proceedings of the LFG09 Conference
Miriam Butt and Tracy Holloway King (Editors)
2009

CSLI Publications
http://csli-publications.stanford.edu/


#### Abstract

We discuss the problems created by the distinction between oblique arguments and adjuncts in general and in the XLE-based ParGram English grammar implementation. We argue that it is better to do away with the distinction for semantically marked obliques.


## 1 Introduction

One of the wonders of Natural Language is the way it manages to repurpose limited resources. Unfortunately for linguistic computations the way it does this is by relying heavily on non-linguistic context. This means that given the incomplete information that we have when we try to analyze text relying only on syntactic structure and limited lexical resources, our analyses are most of the time ambiguous in multiple ways. This problem can be compounded by the incompleteness of our understanding of syntax and of the structure of the lexicon. In this paper we discuss a case where ambiguities are created by incomplete syntactic and contextual knowledge and where attempts to remedy the explosion of ambiguities that they allow in fact make the problem worse. The problem we focus on is how the interaction between some obliques and adjuncts in the XLE-based ParGram English grammar (henceforth ParGramEnglish) and the interaction between its idea of subcategorization and VerbNet (http://verbs.colorado.edu/ mpalmer/projects/verbnet.html) information together create computational bottlenecks. We propose to solve part of these problems by assimilating one class of obliques to adjuncts.

## 2 Problem 1: The theoretical notion of oblique in LFG

Like most modern linguistic theories, LFG makes a distinction between syntactic arguments and adjuncts, or, in theory internal parlance, governable and nongovernable grammatical functions (GFs). Within the governable GFs a further distinction is made between semantically unrestricted ones, subject (SUBJ), object (OBJ), and restricted ones, object theta (OBJ-TH) and obliques (OBL-TH) and within the OBL-THS we further distinguish between idiosyncratically marked ones and semantically marked ones. The topic of this paper is the status of semantically marked OBL-THS in computational embodiments of LFG. We argue that it would be better not to encode these as such but to assimilate them to ADJs.

Semantically marked obliques are marked by prepositions in English. The preposition is meaningful and indicates what the semantic role of the oblique is. A standard example is the to-phrase in sentences such as
(1) Mary gave the book to Bill.

[^207]As any textbook about prepositions for non-native speakers will tell you, to is an indicator of goals. The marking is not specific to the verb give and, depending on how general one thinks the notion of goal is, it can be argued it is not even specific to verbs of transfer of possession. This contrasts with the use of on in a sentence such as (2).
(2) John relied on Mary.

Here the use of the preposition is completely determined by the verb and there is no need to give it any independent meaning of its own. These obliques are classified as idiosyncratic marked (Bresnan, 1982a) or quirky case-marked (Butt and King, 2005).

### 2.1 Arguments and adjuncts

The main criterion that LFG uses to distinguish arguments from adjuncts is uniqueness as discussed in (Bresnan, 1982b). In a sentence arguments are unique, whereas adjuncts can be multiply specified. The example given in (Bresnan, 1982b) is
(3) Fred deftly [Manner] handed a toy to the baby by reaching behind his back [Manner] over lunch [Temp] at noon [Temp] in a restaurant [Loc] last Sunday [Temp] in Back Bay [Loc] without interrupting the discussion [Manner].

In this example we have italicized the arguments and given the adjuncts in bold. This criterion at first seems reasonably straightforward but it requires careful syntactic analysis and an a priori agreement on what counts as the same or a different adjunct or argument. For instance, what is the difference in analysis between
(4) I count on you, on your kindness.
and
(5) He lives in France, in a small village.

Most people would take you in (4) to be an argument. This forces one to assume that on your kindness is a kind of parenthetical whereas in France and in a small village in (5) can be analyzed as separate locative adjuncts. There is no agreed upon list of either (oblique) arguments or adjuncts, and the same prepositions can introduce either. This means that in the absence of careful analysis it is often impossible to determine whether a prepositional phrase is an adjunct or an argument and careful analysis is lacking for most cases. For example, in
(6) He drove from Paris to Venice via Milan.
we could analyze from Paris to Venice via Milan as three adjuncts of the type directional or as three arguments of three different types or as a mixture with one or two arguments and one or two directional adjuncts. Maybe linguistic theory will eventually clarify these issues but at this point an implementation of LFG cannot
make the required distinctions for all predicates, with the result that sentences like the one above will typically get several analyses.

As we will discuss in section 4, the distinctions that one wants to make in natural language processing are of a semantic nature. A uniqueness criterion is not particularly relevant in that respect. Other theories use more semantic criteria to make the distinction between adjunct and argument. For instance, (Dowty, 1989) proposes to use semantic entailment: a semantic argument is entailed by the meaning of its verb. For instance in John walks, John is an obligatory participant because there can be no walking without there being a walker. As formulated by Dowty, the criterion does not correspond to the pre-theoretic distinction between arguments and adjuncts because it would make arguments out of the elements that are most often classified as adjuncts, viz, locative and temporal elements. In
(7) John worked in the kitchen.
or
(8) Mary worked at noon.
in the kitchen and at noon are in general not seen as arguments although all working takes place somewhere and at some time. (Koenig et al., 2003) improve upon the criterion by stipulating that semantic entailment is a necessary but not a sufficient condition. They add to it a specificity condition: arguments are required only by a restricted set of verbs. This excludes immediately the locative and temporal elements mentioned above. (Koenig et al., 2003) are interested in psycholinguistic evidence for the argument adjunct distinction and argue that their criteria pick out classes that correlate with processing differences. This seem plausible but specificity in this sense is rather difficult to pin down as a crisp criterion for each verb. It seems then that in the current state of affairs no linguistic theory is developed enough to give criteria that allow us to straightforwardly distinguish arguments from adjuncts in many cases. So, even in the cases where we can hope one day to make the distinction based on syntactic and lexical criteria we are not able to do it now.

## 3 Problem 2: The implementation of obliques in the ParGramEnglish syntax

Subcategorized grammatical functions in LFG are unique. The theory assumes that different types of OBLs will be distinguished through different names. But the ParGramEnglish implementation chooses to allow only one oblique, OBL ${ }^{1}$, which is treated like all other non-ADJ functions as being unique. This was done because the theoretical situation sketched above and the lack of contextual information would

[^208]allow for too many ambiguities: for all semantically marked obliques of verbs that also have simple transitive or intransitive subcategorization frames, we would also get an analysis that would treats these obliques as ADJs. For instance a verb like drive, has a subcategorization frame where it takes only a subject as in (9) as well as one where it takes a from-PP and a to-PP. Given this, a sentence such as (10) would at first blush get four analyses: (OBL,OBL), (OBL, ADJ), (ADJ,OBL), (ADJ, ADJ).
(9) John drove.
(10) John drove from the house to the school.

The restriction to one OBL eliminates one of these readings, the double OBL one.

Another analysis, the (ADJ,ADJ) one is eliminated by a feature of the ParGramEnglish implementation, called OT marks for Optimality Theory Marks (because it is in spirit related to Optimality Theory). The OT subsystem is described and motivated in (Frank et al., 1998). The XLE system allows the grammar writer to attach preference and dispreference marks to rules. These preferences and dispreferences can be further ordered in the configuration files, which are grammar specific. In the grammar under consideration OT marks are used to regulate OBLs, ADJs and PP attachment preferences. One OT mark says that, when the same c-structure span can cover either an OBL or an ADJ, the OBL is preferred. This excludes the (ADJ, ADJ) reading for the sentence above. But, in fact, the situation is more complex: without further information, the to- PP in the sentence above can also be attached as an NP adjunct (NADJ) to house and indeed that reading is not excluded on the syntactic level. These possibilities start to multiply when we consider a sentence such as:
(11) John drove the car from the house to the school.

Here the NP attachments can be be to car and to house.
If we look at the two PPs we see we can analyze both of them in three ways: as an OBL, as an ADJ and as a NADJ (nominal adjunct). A local OT mark tells us to prefer the OBL to the ADJ in each case, whereas there are no constraints on the NADJ combinations. The result is that we end up with the following possibilities: (OBL, NADJ), (NADJ, OBL), (NADJ,NADJ). This last possibility leads to two parses: one where the two NADJs are attached to the same noun and one where the second in embedded under the first one, but what is important for us is that the possibilities that we would want (OBL,OBL) or (ADJ, ADJ) or (OBL, ADJ) have all disappeared for various reasons. We could go back and reconsider how OT marks are assigned but assigning OT marks is a delicate balancing act and in this case it is not clear that we can improve the system as a whole.

This unfortunate result leads us to look more closely at whether there is an overriding reason to have the OBL-ADJ ambiguity for optional semantically marked obliques. As the discussion in section 2.1 shows, the status of these OBLs is very theory dependent and the LFG classification is very sui generis, which in itself leads us to think that we should not be too attached to it when it gets in our way.

This impression is reinforced when we consider how this class fares in further processing.

## 4 Problem 3: Constraining interpretations and combining lexical resources

A large coverage NLP system needs extensive lexical information. The systems that we are developing ((Bobrow et al., 2007)aim at a rather deep normalization of natural language text. We are developing a level, called AKR (Abstract Knowledge Representation) on which texts that mean the same thing are represented in the same way regardless of the variation in the surface string and texts that have different meanings are represented differently regardless of the similarity in the surface string. ${ }^{2}$ For this deeper analysis, we definitely need information about how the meaning of one item in a sentence can constrain the meaning of other items. A subcase of this is the way the meaning of verbs constrain the meaning of their dependents. This information is typically encoded in the lexicon.

### 4.1 Is subcategorization information what computational approaches need?

An important subset of these constraints on verbal dependents are often talked about as semantic or thematic roles: the subject of a verb like work is the agent or the worker depending on the level of generalization/abstraction one wants to use for this type of information. This is useful information because it helps with paraphrases or entailments (for instance, the difference between transitive and intransitive sink) and, more generally, with the very necessary narrowing down of lexical choices: for instance, if we know that something has to be an agent, we know it has to have independent force. ${ }^{3}$ If we are given this information we know, for example, that the meaning of pilot will not be the pilot light one in
(12) The pilot smiled at the passenger.

Computational lexicons, such as VerbNet, that encode these restrictions often give information about alternations: what is listed with a verb is the dependents that can be expressed in more than one way. This clearly doesn't represent the theoretical distinction between arguments and adjuncts. For instance, in the following alternation from (Levin, 1993), one can argue that, in the first variant, we have three arguments but in the second nobody will analyze (the) horse's as an argument of touch.

[^209](13) Selina touched the horse on the back.
(14) Selina touched the horse's back.

The alternation information is the information that computational lexicons need and try to cope with. Because it often looks like the information that pre-theoretically can be thought of as subcategorization information, one has the tendency to assimilate it completely to this. But as the example above shows, this is not warranted. In fact the conflation of lexical constraints with subcategorization information has led computational linguists to neglect important lexical information that can constrain the interpretation of adjuncts.
(15) He left for three days. $\Rightarrow$ The period of three days is after the leaving.
(16) He worked for three days. $\Rightarrow$ The period of three days is the period of the working.

Here the choice of the verb determines the interpretation of the temporal phrase. This is not seen as a subcategorization restriction and hence there is much less knowledge about the verb classes involved in this and similar phenomena than there is about phenomena that are considered to be part of subcategorization.

### 4.2 Lexical resources

For lexical information, all systems are dependent on resources that have been created outside of the system because no one enterprise can do it all. Specifically our implementation based on the ParGramEnglish syntax also relies on VerbNet. VerbNet is based on (Levin, 1993) verb classes. It intends to describe the alternations for a large subclass of verbs. The alternation information about the verbal dependents is expressed in syntactic categories such as NP and PP that can be found in the immediate environment of a verb. These syntactic categories are mapped onto thematic roles such as agent, patient, and the like. These in turn are associated with a semantic frame that spells out the event structure of a verb argument combination in terms of semantic predicates such as cause, manner, directed motion, etc.

Our implementation combines the information from its own lexicon with information from VerbNet to create the Unified Lexicon (UL). ${ }^{4}$ The VerbNet information is combined with the ParGramEnglish subcategorization information because that is the only information that the system has about the dependents of its verbs. If one interprets the information contained in VerbNet also as subcategorized information one has to figure out what type of notion of subcategorization VerbNet uses to ascertain whether this mapping is warranted. VerbNet does not tell us this. As far as one can derive from looking at what is in VerbNet, the notion used is a mix of the entailment and the alternation approach. ${ }^{5}$ There is no reference whatsoever to a uniqueness criterion.

[^210]As we discussed in section 4.1, seeing the information in VerbNet as subcategorization information might be the wrong idea but whether one assumes that the information in VerbNet is subcategorization information or not, it is clear that ParGramEnglish and VerbNet have a different view on which dependents should be listed in the lexicon with each verb. Given this, the discussion in section 2 and the OBL restriction noted in section 3, it comes as no surprise that the syntactic frames of VerbNet do not correspond well to either LFG or ParGramEnglish subcategorization frames. At the very least, the system has to provide a way to handle the multiple PP complements that VerbNet allows. This is currently done by allowing ParGramEnglish ADJs to function as VerbNet arguments. But, apart from the restrictions to one OBL, LFG and ParGramEnglish associate c-structure components with grammatical functions and VerbNet associates c-structure components with thematic roles. There is no one-to-one mapping between these two. This has as a consequence that the mapping rules often give incorrect results. Most of the errors are in the mapping of OBLs because that is where the two lexicons differ the most. ${ }^{6}$

Our UL then contains unreliable information about PP arguments. Moreover, it is unwieldy: it follows VerbNet in treating each possible subcategorization frame for each predicate as a separate lexical entry. Thus, all the possible combinations of PPs that can be associated with each argument taking predicate need to be spelled out. The spelling out can be done by rule but the result is still a list of lexical items. Given that there is no agreement on what belongs to a subcategorization frame, there is no end to the number of PPs that can be proposed as parts of a subcategorization frame. More importantly, given that, for normalization, our interest is in fact in alternations or in meaning restrictions, the notion of subcategorization frame is not the most relevant and possibly more combinations are relevant than anybody would put in a subcategorization frame.

## 5 Towards a solution

### 5.1 Proposal for the elimination of semantically marked obliques: A (partial) solution to the ambiguity problem

We have seen that treating semantically marked obls as complements leads to a proliferation of ambiguity in parsing. This is especially so in the case of multiple obliques, but is also true for single obliques, e.g.
(17) John sent flowers to Mary.
(18) John sent flowers.

Given the existence of subcategorization frames for send both with and without the to oblique, when a to- PP is present, the grammar has the option of treating it either

[^211]as an oblique or as an adjunct. Deft placement of OT marks is required to eliminate the ambiguity.

We have also seen that the system has a restriction to single obliques, in an effort to reduce syntactic ambiguity. This makes the integration of VerbNet information more complex when in the VerbNet frame there is more than one PP argument. For a parse assigning a single syntactic subcategorization frame but with multiple PP adjuncts, one needs to look through a variety of VerbNet frames to see if any of the PP adjuncts can be treated as a semantically marked oblique.

From a parsing perspective, the adjunct/complement distinction for semantically restricted obliques is, as we have seen, hard to draw determinately; and for multiple obls, all but one are in any case treated as adjuncts. So why not treat all semantically marked obliques as PP adjuncts?

The effects of this on parsing would be threefold. First, there would be a reduction in the degree of ADJ/OBL ambiguity, with concomitant gains in parsing speed. Second, OT marks controlling this ambiguity could be simplified. Third, the number of verb/subcategorization-frame pairs in the lexicon would be reduced, simplifying the task of lexical maintenance.

From a semantic point of view, there is no loss of information if syntax treats semantically marked OBLS as ADJs. This is because mechanisms already exist for finding some of the VerbNet specified semantic obliques amongst the adjunct set; one merely has to apply this mechanism to find all such obliques amongst the adjunct set. But the more interesting question is whether lexical semantic processing could, like syntactic processing, be made simpler.

The processing complexity at issue arises from the fact that an (OBLless) syntactic subcategorization frame like V-SUBJ-OBJ for a verb such as send may have to be compared to a variety of semantic frames like v-SUBJ-OBJ-OBL(TO), V-SUBJ-OBJ-OBL(FROM), or V-SUBJ-OBJ-OBL-OBL(FROM,TO). Comparison would be facilitated if instead there was just one semantic frame, v-SUBJ-OBJ, with an additional specification that from- and to-PPs were candidates for mapping to semantically marked roles like source and destination. Another way of looking at this conceptually is that from and to do not introduce obliques at any level, merely PP adjuncts whose range of interpretation is constrained by the verb to which they apply.

Computationally, therefore, eliminating semantically marked obliques appears to lead to gains all round.

### 5.2 Efficient encoding of adjunct role restrictions: a (partial) solution to the mapping problem

By eliminating semantically marked obls, there is a reduction in the number of syntactic and semantic verb frames that need to be encoded. But a naive implementation would still record, for each frame with obliques eliminated, the role restrictions on the prepositional adjuncts. This is still more verbose than it needs to be. Most of the role restrictions are not dependent on the particular verb, at least
not for the level of granularity at which VerbNet assigns roles. ${ }^{7}$
Indeed, most of the role assignments in VerbNet are based around alternation classes of verbs. For example, in verbs undergoing the instrumental alternation (John broke the window with a hammer vs. A hammer broke the window) with-PPs have a restricted Instrument interpretation open to them. Rather than recording this separately for each verb frame tagged as undergoing the instrumental alternation, the information can be encoded just once as a property of all instrumental verbs. ${ }^{8}$. This can be done in a similar way for the other verb classes that VerbNet encodes. This approach also allows us to specify the mapping of prepositional ADJs to thematic roles per verb class and in that way alleviate the problem of wrong mappings mentioned in section 4.2

## 6 Conclusion and outstanding problems

One of the main reasons for using lexical semantic information from VerbNet is to capture the kinds of paraphrase that are opaque if one looks at grammatical roles alone, e.g.
(19) John $_{\text {SUBJ }}$ broke the window ${ }_{\text {OBJ }} . \Rightarrow$ The window ${ }_{\text {SUBJ }}$ broke.
(where window maps onto the theme role in both cases). The elimination of semantically marked obliques does not interfere with the representation and recognition of alternation paraphrases. Both complements and semantically restricted adjuncts continue to receive their semantic role assignments, albeit through slightly different means of specification.

The proposal sketched above alleviates the ambiguity problem and the mapping problem that semantically marked obliques pose for the system. It does not eliminate all unwanted ambiguities: many are due to the lack of semantic information about the verbal dependents themselves. The proposal also does not deal with idiosyncratically marked obliques. Idiosyncratically marked obliques are ones that either have to be syntactically present (e.g. rely on), or are syntactically optional but their presence substantially alters the meaning of the verb (e.g. answer vs. answer for). Both kinds of idiosyncratically marked oblique need to be explicitly marked in verb frames.

The syntactically obligatory obliques are easy to identify, and clearly have to be encoded in syntactic subcategorization frames (or else their obligatory nature will not be reflected in parsing). It is less clear that optional obliques need to be recorded in syntactic subcategorization frames (since they are optional, they don't constrain parsing), and their identification relies on judgments of sense differences between verbs with and without the oblique. One can use WordNet to produce an

[^212]initial list of verb-preposition pairs that have different senses than verbs alone, and hence where the prepositions are candidates for idiosyncratic obliques; but the list is liable to be both incomplete and error-prone. However, since the oblique controls the sense of the verb, and not just to the role assignment to the prepositional argument, the oblique does at least need to be explicitly recorded in the semantic verb frame. For the moment we opt for leaving them as subcategorized.

## References

Bobrow, Daniel G., Bob Cheslow, Cleo Condoravdi, Lauri Karttunen, Tracy Holloway King, Rowan Nairn, Valeria de Paiva, Charlotte Price, and Annie Zaenen. 2007. PARC's Bridge and Question Answering System. In T. King and E. Bender, eds., The Proceedings of the GEAF07 Workshop. CSLI.

Bresnan, Joan. 1982a. Control and Complementation. In J. Bresnan, ed., The Mental Representation of Grammatical Relations, pages 282-390. MIT Press.

Bresnan, Joan. 1982b. Polyadicity. In J. Bresnan, ed., The Mental Representation of Grammatical Relations, page 164. MIT Press.

Butt, Miriam and Tracy Holloway King. 2005. The Status of Case. In V. Dayal and A. Mahajan, eds., Clause Structure in South Asian Languages, pages 153-198. Kluwer.

Crouch, Dick and Tracy King. 2005. Unifying lexical resources. In Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes. http://www-csli.stanford.edu/ thking/verb05crouchking.pdf.gz.

Dowty, David. 1989. Grammatical Relations and Montague Grammar. In P. Jacobson and G. Pullum, eds., The Nature of Syntactic Representation, pages 69-129.

Frank, Anette, Tracy Holloway King, Jonas Kuhn, and John Maxwell. 1998. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. Proceedings of the LFG98 Conference .

Koenig, Jean-Pierre, Gail Mauner, and Breton Bienvenue. 2003. Arguments for Adjuncts. Cognition 89:67-103.

Levin, Beth. 1993. English Verb Classes and Alternations. University of Chicago Press.

# PARAPHRASES IN LFG-BASED BROAD-COVERAGE SEMANTICS 

Sina Zarrie 3 and Jonas Kuhn<br>Universität Potsdam Universität Potsdam<br>Proceedings of the LFG09 Conference<br>Miriam Butt and Tracy Holloway King (Editors)<br>2009<br>CSLI Publications<br>http://csli-publications.stanford.edu/


#### Abstract

This paper adresses the problem of modelling paraphrases in a deep linguistic processing framework where the meaning construction component is based on an LFG grammar. We present a syntax-based approach to paraphrase extraction that operates on shallow dependency analyses in a parallel corpus. By means of an XLE-based conversion routine, we generate transfer rules for the automatically acquired semantic correspondences. These rules can be used as an additional component in the rule-based process of meaning construction which will augment the meaning representation with entailments that hold for complex phrasal units.


## 1 Introduction

This paper ${ }^{1} 2$ deals with the induction of a paraphrase lexicon for a rule-based meaning construction component which is based on a wide-coverage LFG grammar. We describe a technique for extracting paraphrases from a parallel corpus that exploits several broad-coverage analysis tools. The output of the paraphrase extraction is then fed to an XLE-based conversion routine that automatically derives meaning representations for phrasal expressions. The resulting paraphrase lexicon is implemented in the framework of LFG-based meaning construction outlined in Crouch and King (2006). The lexicon can be used as an additional module in the process of meaning construction.

Crouch and King's meaning construction system makes use of XLE's term rewrite engine to derive semantic representations from LFG F-structures. In addition to a hand-crafted rule component, the system integrates modules that augment the representation with lexical entries obtained from external resources. For instance, the meaning representation of a sentence containing the verb see would be enriched with a semantic predicate which asserts the meaning equivalence between see and its synonyms watch, perceive, and notice. This strategy of explicitly augmenting the meaning representation with all possible entailments can be considered as a process that derives the "deductive closure" of a given semantic analysis of a sentence. Given the "deductive closure" of two meaning representations, the computation of entailment between them boils down to a matching problem and no inference module is required. This strategy of "deductive closure" makes the system particularly suitable for semantic applications that need to deal with the problem of textual entailment - see Bobrow et al. (2007) for a question answering application that is built on top of Crouch and King's meaning construction.

The effectiveness of the strategy of "deductive closure" depends on the quality and the coverage of the captured semantic correspondences. However, whereas the coverage of currently available resources is often limited to single lexical items,

[^213]real-world semantic applications (like search or textual entailment) need to capture complex, phrasal correspondences. As an example, the following pair of sentences illustrates a synonymy relation between the phrase put obstacles in the way of and the simplex verb impede.
(1) The European Union puts obstacles in the way of importing geneticallymodified products.

The European Union impedes the import of genetically-modified products.

State-of-the-art approaches to paraphrase extraction usually do not focus on the further use of the resulting paraphrase resources in the framework of deep linguistic processing systems. ${ }^{3}$ For instance, the extraction methods presented in Bannard and Callison-Burch (2005) would represent the semantic equivalence between phrases as the correspondence of their surface strings. As we will discuss in section 2, this simplifies the problem of semantic equivalence of phrases in an inappropriate way, since syntactic or semantic argument relations cannot be captured. Moreover, it is not clear how to integrate knowledge about surface string correspondence into a meaning representation that abstracts away from surface strings.

Regardless of the final use of the paraphrase resource, one could argue that the paraphrase extraction itself should not be exclusively based on deep processing provided by a particular linguistic formalism. A reasonable recall is essential for the paraphrase resources to be of any practical relevance for the already mentionned real-word applications. Therefore, our approach to the induction of a paraphrase lexicon aims to combine shallow and deep linguistic processing techniques: (i) The paraphrase extraction exploits shallow dependency analyses in addition to word alignments. In section 2, we show that a minimum of syntactic information is needed in order to establish well-formed semantic correspondences later. (ii) The derivation of deep meaning representations for paraphrases is carried out in an LFG setting. In the framework of Crouch and King (2006), we can map complex phrasal expressions that may relate argument slots at various levels in a hierarchical embedding structure to a simple semantic predicate with corresponding argument slots. Our proposal includes a routine for generating such transfer rules (which do get unwieldy for larger phrasal units) automatically from text instances, exploiting the XLE parsing system.

The rest of the paper is structured as follows: In section 2, we will discuss some examples of verb paraphrases, found in the Europarl corpus. These examples motivate our approach to paraphrase extraction, presented in section 3. In the first part of section 4, we will briefly introduce the LFG-based semantic framework proposed by Crouch and King (2006), which constitutes the underlying formalism for the implementation of our paraphrase representation. In the second part of

[^214]section 4, we describe the implementation of the conversion routine that produces semantic transfer rules for arbitrary types of semantic paraphrases.

## 2 Semantic Correspondences in Parallel Corpora

This section gives a general, non-technical overview of the strategy we exploit to induce meaning representations for phrasal expressions from parallel data.

The idea to acquire lexical semantic knowledge from translational data has been particularly pursued in the field of word sense disambiguation and acquisition (Resnik and Yarowsky, 1999; Ide et al., 2002; Dyvik, 2004). Crosslingual models of word sense inventories mainly exploit the fact that a lexical item in a source language usually has a (large) set of possible translations since its different senses are likely to translate to different words in a target language.

The main idea we propose in this paper is to extend this view to translational correspondences where a single lexical item in the source language corresponds to a complex expression in the target language. Zarrie $ß$ and Kuhn (2009) use these complex translational correspondences to identify multiword expressions. They assume that a phrase which has a simplex translation in another language can be considered a (at least partially) non-compositional multi-word. The semantic compositionality of phrases is also highly relevant for application-oriented semantic systems that need to account for inference relations. As an example, consider the German-English sentence pair in (3)-(4) from Europarl and the corresponding meaning representations derived by the transfer semantics where the English analysis corresponds to Crouch and King (2006) and the German analysis is produced as described in Zarrie $ß$ (2009). The meaning representations can basically be seen as flat, DRT-style analyses; for further detail, see section 4.
(3) Mit dem Gesetz wurde die Lage verschlimmert.

With the law was the situation aggravated.

```
HEAD (verschlimmern)
PAST (verschlimmern)
ROLE (Agent, verschlimmern,pro)
ROLE (Theme, verschlimmern,Lage)
ROLE (prep(mit),verschlimmern,Gesetz)
```

(4) The law made the situation even worse.

| HEAD (make) |
| :--- |
| PAST (make) |
| ROLE (Cause,make,law) |
| ROLE (Experiencer,make,situation) |
| ROLE (Pred,make,bad) |
| COMPARATIVE-DIFF (bad,situation,unspecified) |

The meaning representation for the English sentence in (4) would not permit the inference that there is an aggravate-relation between the Cause and the Experiencer since the predicative construction has been assigned a compositional meaning. The fact that the make worse construction can also be assigned a noncompositional meaning can be directly read off the meaning representation of its German translation where an aggravate relation holds between the corresponding Agent and Theme. On the other hand, the German meaning representation would not permit the inference that the instrumental with-PP acts as a Cause in the sentence, information that is explicit in the English meaning representation. The English representation also makes explicit the fact that the mentioned situation is compared to some previous, presupposed situation. This information remains implicit in the German representation and, therefore, could not be inferred. One could argue that the predicative construction (4) explicitly decomposes the lexical semantics of the German main verb whereas the German verb reflects the semi-compositional status of the predicative construction. Thus, both sides of the paraphrase inform each other.

In Zarrie $ß$ and Kuhn (2009), we find that complex translations of simplex words actually occur very frequently in Europarl. This observation can be directly exploited for paraphrase extraction, i.e. the extraction of monolingual semantic correspondences. Bannard and Callison-Burch (2005) use the source language of a parallel text as a pivot providing contextual features for identifying semantically similar expressions in the target language. Following Bannard and Callison-Burch (2005), we relate the meaning of some English expressions if they can translate to an identical German expression and vice versa. This line of reasoning is illustrated in figure 1 that exemplifies two translation instances of the German main verb verschlimmern ('aggravate'). From the fact that the verb has been translated by two different English phrases we make the assumption that their meanings correspond to each other. This means that the representation of a make worse predication can be enriched by the semantics obtained for an exacerbate predication, which results in the representation at the bottom of the figure.

To demonstrate the contrast between the deep and surface-based semantic correspondence extraction, figure 2 shows an example output of the system described in Bannard and Callison-Burch (2005) when paraphrases for the English verb exacerbate are looked for in the English-German Europarl section. ${ }^{4}$ First of all, it can be noted that (at least in this particular case) the system has problems with phrasal correspondences as it proposes paraphrase pairs like exacerbate - worse or exacerbate - made. Moreover, for the pair exacerbate - deteriorate, it is unclear how the arguments or roles of exacerbate correspond to the arguments of deteriorate. It might be possible that the subject Experiencer of the latter corresponds to the object Patient of the former. The form in which the correspondences in figure 2 are given does not allow us to derive deep meaning representations for them.

[^215]Das Gesetz verschlimmert die Lage. The law makes the situation worse.

Die Ereignisse haben die Lage verschlimmert. The situation was exacerbated by the events.


Figure 1: Inducing deep, monolingual meaning equivalences from translations

| exacerbated | worse |
| :--- | :--- |
| exacerbated | increased |
| exacerbated | intensified |
| exacerbated | more |
| exacerbated | undermined |
| exacerbated | better |
| exacerbated | stepped |
| exacerbated | worsened |
| exacerbated | made |
| exacerbated | reinforced |
| exacerbated | waived |
| exacerbated | deteriorated |
| exacerbated | improved |
| exacerbated | increased |
| exacerbated | triggered |
| exacerbated | compounded |
| exacerbated | aggravated |

Figure 2: Examples of paraphrases of exacerbate found in Europarl by the method proposed in Bannard and Callison-Burch (2005)

In contrast to surface-based paraphrase extraction, deep accounts of complex meaning equivalences need to capture correspondences between argument slots. These correspondences can be easily derived if we have some information about the syntactic or semantic parallelism between the two sides of the paraphrase. In figure 1, we can establish a correspondence relation between the Cause of make worse and the Agent of exacerbate, because they are aligned to the same German verb's Cause. In general, we extend the concept of a pivot, defined as a surface string in Bannard and Callison-Burch (2005), to a semantic relation that occurs in a particular argument frame. These frame correspondences can serve as lexical entries which can be integrated into a deep representation of sentence meaning.

The "pivot approach" to paraphrase induction has several limitations. First of all, it ignores the problem that the meaning equivalence of two expressions might be very context dependent. Thus, paraphrases that are valid in some specific context do not necessarily have exactly the same set of senses and selectional preferences such that, in other contexts, their substitution might not be possible. A related question is the directionality of paraphrase rules (one expression might have a more general meaning than the other) which cannot be captured by the naive pivot approach. Unfortunately, the automatic modelling of the context-dependence of paraphrase relations can be considered an unsolved problem (see e.g. Erk and Pado (2009)) and reliable systems are not yet available.

Being aware of the limitations of a paraphrase acquisition method that does not deal with the context dependence of meaning and is, therefore, noisy to a certain extent, we consider the corpus-based extraction of meaning correspondences an attractive way to supplement existing, hand-coded resources. To assure the quality of the paraphrase lexicon in the context of a high-precision semantic representation,
one could manually inspect the automatically acquired lexical entries, still benefiting from the improved recall offered by corpus-based paraphrase acquisition.

To summarize, this section first discussed examples of translations which suggest that translational correspondences on the phrase level can decompose the semantics of a lexical item and make explicit some of its inferences in the meaning representation. Second, by means of the pivot approach, translations can be exploited for the acquisition of monolingual correspondences. We proposed an extension to the pivot approach used by Bannard and Callison-Burch (2005) to discover inference relations between two complex meanings. In order to capture correspondences between semantic relations for deep meaning representation, the paraphrase extraction has to capture correspondences between argument slots.

## 3 Extraction of Syntactic Correspondences from Parallel Corpora

This section will describe the implementation of the paraphrase extraction that has been discussed theoretically in the last section.

Crucially, our approach only relies on flat dependency analyses that can be obtained from currently available, statistical state-of-the-art parsers. This shallow syntactic information is used to approximate information about argument slot correspondences needed by the meaning representation derivation. Moreover, the extraction method exploits the syntactic information as an indicator of the reliability of the translation candidate. The final conversion from syntactic to semantic correspondences is treated in section 4.

The section is structured as follows: The data preprocessing is described in section 3.1. Section 3.2 deals with the extraction of crosslingual syntactic correspondences from shallow dependency analyses. In 3.3, we discuss the mapping of crosslingual correspondences onto monolingual ones.

### 3.1 Parallel Data for Paraphrase Extraction

We base our investigations on the German and English portion of the Europarl corpus (Koehn, 2005) which is available in a sentence-aligned, tokenized format. To produce word-alignments for the German-English parallel text, we used the wide-spread, open-source GIZA++ tool (Och and Ney, 2003). We employed the standard settings for alignments in both directions (viterbi alignments, IBM model 4) and the refined alignment heuristics for bidirectional alignment

To obtain robust syntactic analyses for the two portions of the parallel corpus, we used MaltParser (Nivre et al., 2007), a data-driven dependency parsing system which is freely available. ${ }^{5}$ In comparison to other statistical state-of-the-art parsing systems, MaltParser has especially proven succesful for a broad range of languages. The English version of the parser was trained on the Penn treebank. The

[^216]German version of the parser was trained on the Tiger treebank. In future work, we might use a more recent model of the German parser which was trained on the Tiger treebank enriched with features from deep LFG parses ( $\emptyset$ vrelid et al., 2009).

Technically, the resulting resource of parallel, word-aligned dependency parses is stored as a relational database where, for each token, its monolingual properties, like lemma, POS, and syntactic head, as well as its crosslingual relations, i.e. the aligned tokens of the target language, can be efficiently represented. The extraction procedures described in the following section are basically implemented as a cascade of queries on the database.

### 3.2 Syntax-based Paraphrase Extraction

Given the parallel dependency trees obtained from EuroParl, we now want to extract German-English paraphrasing translations that involve the correspondence between a simplex lexical item on the source (German) side and a complex phrasal expression on the target (English) side. As an example, consider the sentence pair given in (5)-(6) where the German verb behindern corresponds to the English expressions constitute an obstacle to. The extraction of such complex translational correspondences involves the major challenge that, typically, only certain parts of the target phrase can be reliably aligned to the source item due to the low occurrence correlation of the other parts. For instance, in sentences (5)-(6), GIZA++ is not likely to be able to capture the correspondence between the German main verb and the whole English phrase, but instead to find only an alignment link between the German main verb and the noun obstacle. For further detail on this alignment problem see Zarrieß and Kuhn (2009).

The general intuition is that the alignment of phrasal correspondences somehow needs to relax the requirement of high cooccurrence correlation while still detecting reliable translation instances. The main idea of our paraphrase detection approach is to relax the cooccurrence correlation based on leveraging syntactic information. For instance, consider the pair of parallel configurations in figure 3 for the sentence pair given in (5) and (6). Although there is no strict one-to-one alignment for the German verb, the basic predicate-argument structure is parallel: The verb's arguments directly correspond to each other and are all dominated by a verbal root node.

We propose a generate-and-filter strategy for our translation detection which extracts partial, largely parallel dependency configurations. The input to the candidate generation is a source lexical item in a predefined syntactic configuration that exhibits two or more argument slots, e.g. a verb with its subject and object argument dependency relations. The output of the candidate generation is a set of translation instances where the German verb occurs in the predefined argument frame and the English translation exhibits argument slots that can be consistently aligned with the source slots. To filter noise due to parsing or alignment errors, we further introduce a filter on the length of the path that connects the target root and its dependents and a filter that excludes paths crossing sentence boundaries.
(6) Corruption constitutes an obstacle to development.


Figure 3: Example of a phrasal translational correspondence configuration

By allowing target dependency paths to be aligned to source single dependency relations, we admit configurations where the source item is translated by more than one word. Thus, we propose to use the aligned arguments as anchors of the configuration. Then, the search basically selects all the items lying on a path between the target root item and the target arguments, very similar to Lin and Pantel (2001) who pursue a similar approach for paraphrase extraction on monolingual corpora. For instance, given the configuration in figure 3, we allow the German verb, which connects the argument $X_{1}$ and $X_{2}$, to be aligned to the path connecting $Y_{1}$ and $Y_{2}$.

We evaluate the accuracy of our translation detection approach, especially for the accuracy of phrasal translations, by manually classifying 300 random, exclusively phrasal translations that our system detects for 50 German verbs (all selecting for a nominative subject and accusative object). We extract 50 random, transitive verbs from the German LFG grammar lexicon. We supply these verbs in their desired syntactic configuration to the translation search method described in this section and recover the reliable alignments that the search detects. Out of the resulting correspondences, we select 300 random instances from the set of phrasal configurations. The results of the classification are displayed in table 1. We observe that $17.1 \%$ of the translations detected by the system do not correspond to semantic correspondences, whereas more than $80 \%$ show different patterns of complex translation patterns. For a more detailed discussion of the translation search and an analysis of the patterns see Zarrieß and Kuhn (2009).

| Trans. type | Proportion |  |  |
| :--- | :---: | :--- | :---: |
|  |  | MWE type | Proportion |
|  |  | V Part | $8.2 \%$ |
| MWEs | $57.5 \%$ | V Prep | $51.8 \%$ |
|  |  | LVC $^{6}$ | $32.4 \%$ |
|  |  | Idiom | $10.6 \%$ |
| Paraphrases | $24.4 \%$ |  |  |
| Alternations | $1.0 \%$ |  |  |
| Noise | $17.1 \%$ |  |  |

Table 1: Classification of 300 types sampled from the set of one-to-many translations for 50 verbs

### 3.3 From Crosslingual to Monolingual Correspondences

In section 2, we have explained the general method allowing for the deduction monolingual semantic correspondences from translation pairs. The paraphrase extraction produces pairs of given source-configurations $s_{i}$ and their corresponding set of translations $T_{i}$, where for each argument slot on the source side $a_{s_{n}} \in s_{i}$, each target configuration $t_{i} \in T_{i}$ contains a corresponding argument slot $a_{t_{n}} \in t_{i}$. From a particular set of target configurations $T_{i}$, we would obtain the resulting set of target correspondences by taking the Cartesian product of this set $T_{i} \times T_{i}$. In practice, taking the product of the set of target configurations would result in a huge set of meaning postulates since the number of translations to be found in parallel corpora is usually very high (Dyvik, 2004). Moreover, the set product would replicate a major part of the already existing rules for word level correspondences since our target correspondence might not necessarily be phrasal expressions.

A theoretical issue brought up by paraphrase or entailment induction is the directionality of the relation. As Basili et al. (2007) point out, automatically assembled paraphrase resources usually lack a notion of directionality and capture entailment as a bidirectional equivalence relation. The authors also remark, however, that the problem of context dependence is usually more serious in practice than the problem of uni-directional entailment. As with the treatment of contextdependence, the work presented in this paper does not deal with directionaly either.

The current implementation of the paraphrase derivation proceeds as follows: the set of target translations $T_{i}$ is separated into a set of configurations that exhibit a word-level correspondence to the source item, $T_{w_{i}}$, and a set of configurations that exhibit a phrasal expression corresponding to the source item, $T_{p_{i}}$. From the set of simplex translations $T_{w_{i}}$, we select the most frequent translation and relate it to all elements of complex translations $T_{p_{i}}$. Future work might implement more sophisticated methods for the selection of the actual monolingual correspondences.

[^217]
### 3.4 Discussion

First of all, this syntax-based extraction of translational correspondences has the advantage that the alignment is supplied with additional cues to coccurrence and can thus extract configurations that do not have to be very frequent. Moreover, we can control for the syntactic configuration on the source side of the translation such that we are likely to find, for a given pivot configuration, instances of that configuration that all share the same argument frame. In section 2, we have seen that consistent argument frames among the source language pivots are essential for establishing the correct correspondence between the argument slots. In this way, our syntax-based search inherently controls for the voice of the source verb which is also crucial for establishing the correct argument correspondences.

The syntax-based approach also has certain drawbacks, for instance, the fact that it relies, to a certain extent, on syntactic parallelism, i.e. on the parallel realization of predicate arguments. However, research in syntax projection has shown that the divergence of dependency structure across languages might be quite drastic such that, without additional information sources, straightforward projection of dependency relations between aligned pairs of words yields relatively poor results (Bouma et al., 2008; Hwa et al., 2005). Similarly, Pado (2007) observes that crosslingual parallelism on the level of predicate argument relations still shows considerable variation. For our method which presupposes the parallelism of the argument slots, this means that it cannot take account of the many translation instances that do not exhibit syntactic parallelism (e.g. target passive translations of a source active verb where the agent of the source verb is omitted).

A further limitation of the syntax-based approach lies in the fact that the expressions we want to extract need to be "syntactically anchored". In the case of transitive verbs where the pair of arguments can naturally serve as syntactic anchor, this does not pose a problem. But lexical items like adjectives or intransitive verbs which do not have more than one argument position do not offer the possibility of finding their translational equivalents by looking for the path that connects the translation of their arguments. Future work on the extraction method might investigate a more general way to take the syntactic configuration of a translation into account, adressing the partial parallelism as well as the anchoring problem.

## 4 Induction of Paraphrase Meanings from Syntactic Correspondences

The paraphrase extraction described in the previous section produces pairs of dependency graph configurations. In each of the configurations, a verb and its arguments on the source side correspond to a target phrasal expression that realizes the same argument slots somewhere in its dependency configuration. This section deals with the method we employ to map these parallel dependency configurations onto semantic correspondence rules that can be applied in an LFG-based meaning
construction component. In section 4.1, we will first describe the basic properties of the transfer semantic representation and the architecture performing the meaning construction. Section 4.2 then deals with the induction of the paraphrase lexicon in the context of the transfer semantics.

### 4.1 An LFG-based Transfer Component for Meaning Construction

The meaning construction described in Crouch and King (2006) converts LFG Fstructures produced by the English ParGram grammar to flat representations in a Neo-Davidsonian style. Since the ParGram initiative (Butt et al., 2002) has particularly focussed on crosslingual parallelism on the level of syntactic analyses, this symbolic conversion routine can be easily ported to other LFG grammars, as has been done e.g. for German (Zarrieß, 2009).

The main idea of the system is to convert the surface-independent, syntactic relations and features encoded in an F-structure to normalized semantic relations. The semantic conversion was implemented by means of the XLE platform, used for grammar development in the ParGram project. It makes use of the built-in transfer module to convert LFG F-structures to semantic representations. The idea to use transfer ordered rewriting rules to model a semantic construction has also been pursued by Spreyer and Frank (2005) who use the transfer module to model a Minimal Recursion Semantics construction for the German treebank TIGER.

### 4.1.1 The Meaning Representation

To begin with an example, a simplified F-structure analysis for the following sentence and the corresponding semantic representation are given in figure 4.
(7) Where was Peter seen?

First, the interrogative pronoun induces a semantic context that embeds the proposition headed by the main verb. For the sake of readability, we visualize the semantic contexts as DRT-style boxes. The syntactic arguments and adjuncts of the main predicate are represented in terms of semantic roles of the context introduced by the main predicate or some higher semantic operator. Thus, the grammatical roles of the main verb in sentence (7) are semantically normalized such that the subject of the passive is assigned the Stimulus role and an implicit Experiencer is introduced; see figure 4. This type of semantic representation is inspired by Neo-Davidsonian event semantics (in the style of Parsons (1990)). Other semantic properties of the event introduced by the main verb such as tense or nominal properties such as quantification and cardinality are explicitly encoded as conventionalized predications.

The contexts can be thought of as propositions or possible worlds. They are headed by an operator that can recursively embed further contexts. Context embeddings can be induced by, e.g. negation, conditionals or clause-embeddings.
$\left[\begin{array}{lll}\text { PRED } & \text { 'see }<\text { NULL, }(\uparrow \text { SUBJ })>{ }^{\prime} \\ & {\left[\begin{array}{ll}\text { PRED } & \text { 'Peter' } \\ \text { SUBJ } & \text { sg } \\ \text { NTYPE } & \text { propername }\end{array}\right]} \\ \text { ADJ } & \{[\text { PRED } & \text { 'where' }]\} \\ \text { CLAUSE } & \text { int } & \\ \text { PASSIVE } & + \\ \text { TENSE } & \text { past }\end{array}\right]$

| HEAD (SEE) |
| :--- | :--- |
| PAST (SEE) |
| CARDINALITY (PETER,SG) |
| PROPER-NAME (PETER) |
| ROLE (EXPERIENCER,SEE,PRO) |
| ROLE (STIMULUS,SEE,PETER) |
| ROLE (PREP(WHERE),SEE,WHERE) |
| QUERY-TERM (WHERE) |
| WORD (SEE,[UNDERSTAND-2,WITNESS-2, VISUALIZE-1,...]) |
| VN-SEMANTICS (SEE, PERCEIVE(EXPERIENCER,STIMULUS)) $)$ |

Figure 4: LFG F-structure analysis and corresponding semantic representation

### 4.1.2 Representation of Semantic Correspondences on the Word-Level

The semantic representation in figure 4 illustrates how the lexical meaning of the individual words that make up the sentence is represented. The predication WORD links the word see to a WordNet index that contains the corresponding synsets of the predicate. Thus, all synonyms of a word are enumerated in the representation and, by this means, directly available to entailment components. Therefore, if the WordNet entry for see contains the predicate spot, the correspondence between the sentence (7) and e.g. the sentence Where was John spotted? boils down to a matching of the semantic representation of these two sentences. Likewise, the predication VN-SEMANTICS states a VerbNet entry for the given verb. These entries capture verb-class equivalences and some deeper alternations. Generally, the meaning representation of a sentence explicitly and exhaustively asserts the lexical entries of its individual words, a strategy we referred to as "deductive closure" in section 1. For a more detailed description of the lexical resource interface implemented for the English transfer semantics system see Crouch and King (2005).

### 4.1.3 The Meaning Construction Component

The XLE transfer module, which is used for the implementation of the conversion of F-structures to semantic representations, is a term rewrite system that applies an ordered list of rewrite rules to a given F-structure input and yields an output transfer structure. Depending on the rewrite mode and on the definition of the rule, the output can be a fully-fledged F-structure again or else a set of (recursively embedded)
prolog clauses whose format is not further constrained. An example rewrite rule which yields F-structure output is given in figure 5. It applies to F-structures that have a passive and vtype feature as well as an oblique agent, mapping the oblique agent to a subject. The F-structure scope or embedding of the features $(\% \mathrm{~V}$ in this case ) is given as the first variable of the fact.

```
VTYPE(%V, %%), PASSIVE(%V,+),
OBL-AG(%V, %LogicalSUBJ)
==>
SUBJ(%V, %LogicalSUBJ).
```

Figure 5: Example rewrite rule for passive normalization

The transfer system comes as a generic rewrite system and does not only apply to XLE F-structures. Therefore, it can be generally used to formulate mappings between clausal structures (given in the Prolog-format currently used by XLE). This flexible rewrite architecture makes it possible to organize the semantic construction or conversion in a modular way since rules can also apply to semantic transfer structures. This architecture substantially eases the integration of lexical knowledge. An example for an exemplary semantic lexicon and its integration in the semantic conversion is given in figure 6 . The fact marked with $\mid$ - first asserts that aristocracy is a collective noun. The following rule then matches all input meaning representations that contain a singular collective noun and rewrites their cardinality to plural.

```
|- collective_number(aristocracy).
collective_number(%NounForm),
in_context(%C, cardinality(%NounForm, sg))
==>
in_context(%C, cardinality(%NounForm, pl)).
```

Figure 6: Example rewrite rule for semantic rewrite

Essentially, the induction of the lexicon entries presented in section 4.2 makes use of the transfer system by automatically generating transfer rules that map partial semantic representations onto some semantically equivalent representation.

### 4.2 Deriving Transfer Rules for Semantic Correspondences

Our ultimate goal in this section is to define lexicon entries (or transfer rules) for paraphrases that match sentences that contain instances of these paraphrases. The lexicon entries will augment the transfer meaning representation with entailments which hold for larger units than single words. Thus, we extend the strategy of "deductive closure" from simplex lexical items to complex phrases.

The main requirement for the procedure of lexicon entry derivation is that it has to be independent of the semantic pattern of the paraphrase. As can be seen in the classification of the extracted paraphrase types in table 1, phrasal correspondences in parallel corpora yield very different types of semantic correspondences. As an example, consider the paraphrases in (9) and (10) and their semantics (as derived by the LFG-based transfer semantics) which have been found corresponding to the hinder-configuration in example (8). Item (9) exemplifies a light verb construction with the complex preposition in the way of. The paraphrase in (10) is even more complex because it exhibits a coordination, mapped to a complex event operator, where the argument slots of the dependency configuration have to be mapped to several semantic roles in the meaning representation.

These examples make clear that the definition of the lexicon entries cannot be done by handwritten templates. In order to match the paraphrase meaning representation with possible input sentences, the lexicon entries need to anticipate the analysis that is assigned to the paraphrase by the core meaning construction.

| HEAD (hinder) |
| :--- |
| ROLE (Agent,hinder,X) |
| ROLE (Patient,hinder,Y) |

(9) $x$ puts obstacles in the way of $y$

| HEAD (put) |
| :--- |
| ROLE (Agent,put,X) |
| ROLE (Theme,put,obstacle) |
| ROLE (Destination,put,way) |
| ROLE (prep(of),way,Y) |


| HEAD (Complex Event) |  |
| :---: | :---: |
| OR ( | HEAD (obstruct) |
|  | ROLE (Agent,obstruct,X) <br> ROLE (Experiencer,obstruct,Y) |
|  | HEAD (fire) |
|  | $\begin{aligned} & \hline \text { ROLE (Agent,fire,X) } \\ & \text { ROLE (Patient,fire,pro) } \\ & \text { ROLE (prep(on),fire, } \mathrm{Y} \text { ) } \\ & \hline \end{aligned}$ |

Figure 7: Examples for complex semantic correspondences

### 4.2.1 The conversion routine

To convert the pairs of dependency configurations, produced by the paraphrase extraction, to pairs of meaning representations, we make use of the XLE analysis pipeline. The basic method of paraphrase conversion works as follows:

1. Map the set of dependency configurations, as well as the target verb configuration, to a set of surface sentences parsable by the XLE engine. Replace the argument slots of the dependency configurations by some dummy pronouns that will be uniquely identifiable later.
2. Parse all the paraphrase sentences, as well as the target meaning sentence, with XLE. Run the F-structure output through the transfer semantics construction.
3. Run the output semantic representations through a sequence of transfer rules that removes the context-specific clauses from the semantic representation.
4. In each of the stripped down semantic representations, replace the scope identifiers by prolog variables. Replace the dummy pronouns, corresponding to the aligned argument slots, by prolog variables such that one argument slot corresponds to a unique variable.
5. Generate a set of transfer rules that has a paraphrase meaning representation as its left-hand side and the meaning representation of the target simplex verb as its right-hand side.

The mapping of the extracted dependency configurations onto surface sentences (point 1) is simply done by linearizing the lexical items in their original word order. We normalize the inflection of the verbal root node to third person singular such that the subject dummy pronoun can be parsed as a third person pronoun. This first step of the routine is simplified by the fact that we only investigate verbs in a well-defined argument frame such that the surface sentences can receive a full syntactic (and semantic) analysis. In so doing, we avoid having to identify parts of the meaning representation since (almost) the entire semantic analysis of the paraphrase sentence actually corresponds to the paraphrase meaning.

Note, however, that the complete meaning representation of the surface paraphrase sentence obtained from the meaning construction is not yet exactly the lexical entry we want to include in the paraphrase lexicon. If we parse an example sentence like $x$ puts obstacles in the way of $y$, we necessarily obtain a semantic analysis that includes non-general, context-specific features like tense for verbs or cardinality for nominals. Therefore, we define an additional set of transfer rules which is applied on top of the usual meaning construction and which deletes these non-general components of the paraphrase meaning. Besides tense and cardinality, this list of deletion rules contains the pronoun-specific facts, specifier predications and clauses that keep the original F-structure attributes. ${ }^{7}$

[^218]
### 4.2.2 Example

The conversion routine described in the preceding section yields transfer rules that map an arbitrarily complex predication to a simplex semantic relation, assigning a non-compositional meaning to the paraphrase. Coming back to the example sentences (8) and (9), we obtain a lexical entry for the light verb construction put obstacles in the way of as given in figure 8. In contrast to the representations shown so far, the rule is given in its internal specification where the contexts do not correspond to boxes, but to context predicates.

The right hand side of the rule introduces a new word hinder and maps the subject of put to the subject of hinder and the object of the preposition to the object of hinder. Note that the rule does not delete the original analysis of the paraphrase, but just augments the representation with an additional relation that holds between the involved referents (the + in front of the left hand side facts tells the rewrite mechanism not to delete them from the set of input clauses). This is consistent with the strategy of deductive closure that is already implemented for semantic equivalences on the word-level. After the application of the lexical entry in figure 8 , the respective meaning representation also matches all further lexical rules that hold for the hinder-relation. By this means, the paraphrase put obstacles in the way of can also be related to all synonyms of hinder.

```
+context_head(%ctx,put:%put),
+in_context(%ctx,role(Theme,put:%p,obstacle:%o)),
+in_context(%ctx, role(Agent,put:%p,%X)),
+in_context(%ctx, role(prep(of), way:%w,%Y)),
+in_context(%ctx,role(Destination, put:%p,way:%w))
==>
context_head(%ctx,hinder:%p),
skolem_info(hinder:%p,hinder,verb,verb,%p,%ctx),
in_context(%ctx, role(Agent,hinder:%p,%x)),
in_context(%ctx, role(Patient,hinder:%p,%Y)).
```

Figure 8: Example of a paraphrase representation as a transfer rule.
In contrast to the surface string representation of paraphrases we discussed in section 2, the lexical entry given in figure 8 subsumes a large number of possible surface realizations of the paraphrase. The following sequence of example sentences illustrates a number of surface phenomena that the lexical entry abstracts from.

1. X is putting obstacles in the way of Y .
2. X is putting some major obstacles in the way of Y .
3. A huge obstacle was put in the way of $Y$ by $X$.
4. X puts an obstacles in Y's way.

For instance, the application of the lexical entry is independent of the tense of the construction and possible modifications of the nominal obstacle since the non-general clauses where deleted during the conversion routine. The entry is also independent of the voice that is instantiated by the paraphrases. Also, since the semantic constructions maps genitive and of-possessives onto the same representation, the lexical entry abstracts even further from specific surface realizations. This is a very desirable property of a paraphrase lexicon since, apart from syntactically rather fixed multiword expressions, paraphrases can occur in a wide range of syntactic specifications.

A further issue illustrated by the rule in figure 8 is the treatment of embeddings by means of variables. The new hinder-relation is not necessarily added to the main context of the input semantic representation, but it is made dependent on the embedding of the clauses which match the left hand side of the rule. If the left-hand side of the rule exhibits any context-embeddings (e.g. the representation in (10)), the context of the right-hand side will be the root context of the left-hand side. This treatment of context ensures that the lexical entry does not change the inferential properties of the input meaning, but applies to embedded paraphrases (e.g. through negation) as well.

## 5 Conclusion

In this paper, we have presented a way to augment a rule-based, hand-crafted meaning construction with semantic knowledge that has been automatically acquired from a large corpus. In particular, we have addressed the problem that surface string correspondences, as they are found by most corpus-based paraphrase extraction tools, cannot be easily integrated into deep meaning representations of sentences. The main reason for this problem is that the deep meaning representations have to be anchored via their semantic roles. However, surface-string correspondence does not allow us to induce this anchoring. We have outlined a way to approximate the correspondence of semantic roles via aligned syntactic arguments in a parallel corpus considered at the stage of paraphrase extraction.

Our syntax-based paraphrase extraction operates on wide-coverage, shallow dependency analyses. Technically, this involves the limitation that the method currently only works for expressions that have at least two argument positions which can serve as syntactic anchors. The syntax-based approach also suffers from the drawback of relying on the partially parallel realization of the predicate arguments in the target language. However, the translation search which is based on syntactic anchors performs better than raw word-alignment for transitive verbs.

The implementation of our XLE-based conversion routine produces lexical entries for paraphrases by deriving transfer rules, as defined in the XLE transfer module. These rules capture correspondences of complex phrasal expressions and contribute information about their inferential properties (e.g. their compositionality, implicit presuppositions). We have further shown that deep meaning representa-
tions of paraphrases has the practical advantage that the resulting lexical entries can capture a wide range of surface realizations of that paraphrase.

The main challenge for future work will be the treatment of the context dependence of paraphrases. It is crucial for the strategy of "deductive closure" of meaning representations that all asserted entailments hold in the given text or sentence pair. One possible way to assure the quality of the paraphrase lexicon would be a manual post-processing step that removes overly context-specific paraphrases from the list of transfer rules. But finally, it seems indispensable to have a statistical disambiguation component integrated in the process of meaning construction that discards invalid entailments from the representation, based on a context-sensitive model of meaning.

## References

Bannard, Colin and Callison-Burch, Chris. 2005. Paraphrasing with Bilingual Parallel Corpora. In ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pages 597-604, Morristown, NJ, USA: Association for Computational Linguistics.

Basili, Roberto, Cao, Diego De, Marocco, Paolo and Pennacchiotti, Marco. 2007. Learning Selectional Preferences for Entailment or Paraphrasing Rules. In Proceedings of RANLP 2007.

Bobrow, Daniel G., Cheslow, Bob, Condoravdi, Cleo, Karttunen, Lauri, King, Tracy Holloway, Nairn, Rowan, de Paiva, Valeria, Price, Charlotte and Zaenen, Annie. 2007. Precision-focused Textual Inference. In ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, pages 28-29.

Bouma, Gerlof, Kuhn, Jonas, Schrader, Bettina and Spreyer, Kathrin. 2008. Parallel LFG Grammars on Parallel Corpora: A Base for Practical Triangulation. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG08 Conference, pages 169-189, Sydney, Australia: CSLI Publications, Stanford.

Butt, Miriam, Dyvik, Helge, King, Tracy Holloway, Masuichi, Hiroshi and Rohrer, Christian. 2002. The Parallel Grammar Project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation, Taipei, Taiwan.

Crouch, Richard and King, Tracy Holloway. 2005. Unifying Lexical Resources. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes, pages 32-37.

Crouch, Richard and King, Tracy Holloway. 2006. Semantics via F-Structure Rewriting. In Miriam Butt and Tracy Holloway King (eds.), Proceedings of the LFG06 Conference.

Dyvik, Helge. 2004. Translations as Semantic Mirrors. From Parallel Corpus to WordNet. Language and Computers 1, 311-326.

Erk, Katrin and Pado, Sebastian. 2009. Paraphrase Assessment in Structured Vector Space: Exploring Parameters and Datasets. In Proceedings of the EACL Workshop on Geometrical Methods for Natural Language Semantics (GEMS).

Hwa, Rebecca, Resnik, Philip, Weinberg, Amy, Cabezas, Clara and Kolak, Okan. 2005. Bootstrapping Parsers via Syntactic Projection across Parallel Texts. Natural Language Engineering 11(3), 311-325.

Ide, Nancy, Erjavec, Tomaz and Tufis, Dan. 2002. Sense Discrimination with Parallel Corpora. In Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. ACL2002, July Philadelphia 2002, pages 56-60.

Koehn, Philipp. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In MT Summit 2005.

Lin, Dekang and Pantel, Patrick. 2001. DIRT - Discovery of Inference Rules from Text. In In Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323-328.

Nivre, Joakim, Hall, Johan, Nilsson, Jens, Chanev, Atanas, Eryigit, Gülsen, Kübler, Sandra, Marinov, Svetoslav and Marsi, Erwin. 2007. MaltParser: A Language-Independent System for Data-Driven Dependency Parsing. Natural Language Engineering 13(2), 95-135.

Och, Franz Josef and Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics 29(1), 19-51.

Øvrelid, Lilja, Kuhn, Jonas and Spreyer, Kathrin. 2009. Improving data-driven dependency parsing using large-scale LFG grammars. In Proceedings of the Annual Meeting for the Association for Computational Linguistics (ACL) (Short Paper).

Pado, Sebastian. 2007. Translational Equivalence and Cross-lingual Parallelism: The Case of FrameNet Frames. In Proceedings of the NODALIDA Workshop on Building Frame Semantics Resources for Scandinavian and Baltic Languages, Tartu, Estonia.

Parsons, Terence. 1990. Events in the Semantics of English, volume 19 of Current studies in linguistics series. Cambridge, Mass.: MIT Press.

Resnik, Philip and Yarowsky, David. 1999. Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation. Nat. Lang. Eng. 5(2), 113-133.

Spreyer, Kathrin and Frank, Anette. 2005. The TIGER 700 RMRS Bank: RMRS Construction from Dependencies. In Proceedings of LINC 2005, pages 1-10.

Zarrieß, Sina. 2009. Developing German Semantics on the basis of Parallel LFG Grammars. In Tracy Holloway King and Marianne Santaholma (eds.), Proceedings of the ACL Workshop on Grammar Engineering Across Frameworks (GEAF), Singapore.

Zarrieß, Sina and Kuhn, Jonas. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In Proceedings of the ACL Workshop on Multiword Expressions, Singapore.


[^0]:    ${ }^{\dagger}$ We are grateful to Tracy Holloway King and the audience at LFG09 for comments and suggestions (in particular Ash Asudeh and Ron Kaplan) and to members of the Essex Arabic Syntax Workshop for discussion of contemporary work on MSA and the Arabic vernaculars.
    ${ }^{1}$ Note: glossing is morphological, reflecting the standard morphosyntactic desrciption of MSA. Where examples have been taken from sources, transliterations have been standardized to the DIN31635 format (and some randomly omitted case marking has been reinserted in some examples from Benmamoun (2000)).

[^1]:    ${ }^{2}$ The ordering of functional heads is critical to Benmamoun's proposal, but Shlonsky (1997) takes Neg to be higher than T in the hierarchical structure in Arabic (Shlonsky, 1997, 103-4).

[^2]:    ${ }^{3}$ The issue here is perhaps only one of unclarity of presentation, making the resultant analysis opaque to those less than totally familiar with the assumptions of the framework.

[^3]:    ${ }^{5}$ but not ruling out an aspectual sa-Imperfective in V appearing as part of a periphrastic verbal expression in a negative clause.
    ${ }^{6}$ The subjunctive is the same in the 3 MPL, as shown below.

[^4]:    ${ }^{7}$ See below for short discussion of alternative analyses. For example, an approach in terms of form selection (at cstructure) might be more appropriate (Falk, 2008), in which case we would not use the MOOD feature at f-structure at all.
    ${ }^{8}$ We assume that the combination of negative future with the prospective is ruled out on semantic grounds.

[^5]:    ${ }^{\dagger}$ We are grateful to Tracy Holloway King and the audience at LFG09 for comments and suggestions and to members of the Essex Arabic Syntax Workshop for discussion of contemporary work on MSA and the Arabic vernaculars.
    ${ }^{1}$ Examples are taken from these sources but normalised to a single transliteration system, DIN 31635.

[^6]:    ${ }^{2}$ An allied assumption is that the presence of full functional material would prevent the formation of a complex predicate.

[^7]:    ${ }^{3}$ The question arises as to how to deal with the fact that kinship terms are excluded without stipulating that kinship terms are somehow alienable. Kim speculated that there may be some sort of 'part-of' requirement, or that alternatively it may be that what the construct state adjective wants is a property (rather than an entity) and a kinship term "inherently denotes an entity". This question is left open.
    ${ }^{4}$ Hazout (2000) sees this process to be a side effect of compound formation in the morphology.

[^8]:    ${ }^{5}$ The function associated with the complement noun is given here as OBJ , but could as well be $\mathrm{OBJ}_{\theta}$ : the important point is that it is both a direct function and not the SUBJ.
    ${ }^{6}$ In the examples above, $f y r$ is the soft mutated form of byr.
    ${ }^{7}$ The attribute SIND indicates the semantic INDEX in the semantic structure.

[^9]:    a. imrar ªt-un taw̄̄l-at-un ğamīl-at-u -l-wağh-i
    woman-F-NOM.INDEF tall-F-NOM.INDEF beautiful-F-NOM the-face-GEN a tall woman with a beautiful face

[^10]:    ${ }^{8}$ We follow King and Dalrymple (2004) in distinguishing INDEX and CONCORD (agreement) features, and express NP internal agreement in MSA in terms of CONCORD features.
    ${ }^{9}$ This does not, of course, explain this curious restriction, but it does capture it. It seems to be some sort of low level matter of realization more than anything else.

[^11]:    ${ }^{10} \mathrm{We}$ abstract away from recursive modification for simplicity of exposition.

[^12]:    ${ }^{11}$ We have assumed here, to simplify the presentation, that the head noun is associated with a nominal rather than an NP meaning, despite the fact that definiteness/indefiniteness is morphologically marked in Arabic.

[^13]:    ${ }^{\dagger}$ This paper has been supported in part by research grant HUM2007-61916FILO. I thank Mary Dalrymple, Helge Lødrup and participants in LFG09 for observations.

[^14]:    ${ }^{2}$ Law (2006:633) appears to also make this claim, but in fact makes the stronger claim that there are only two types of languages regarding P-stranding-those that have it and those that don't-by (in my opinion, incorrectly) analyzing Icelandic and Swedish as languages where P-passive is possible (Law 2006:655-660).

[^15]:    ${ }^{3}$ The argument that P-passive in Norwegian is essentially the same as in English is that (a) it is a passive construction (as shown by the verbal morphology and the suppression of the logical subject) and (b) the passive subject is the thematic argument of a prepositional complement of the verb. A unified analysis of P-passive in both languages is therefore preferable to a different analysis for each language.
    ${ }^{4}$ Along with control equations, c-structure annotations are eliminated altogether. This allows us to dispense with the intermediary level of the annotated c-structure in the mapping between c -structure and f -structure, thus simplifying this mapping.

[^16]:    ${ }^{5}$ The information-structural GFs used in standard LFG, TOP and FOC, are here collapsed as a single GF: OP(ERATOR). Following Bresnan (2001) (see also Falk (2001)), SUBJ and OP constitute the class of Discourse Functions (DF). The GF hierarchy assumed here places these two GFs together at the top of the hierarchy.

[^17]:    ${ }^{6}$ The closest f-binder can be defined as follows: If $\alpha$ and $\beta$ both f-bind $\gamma, \alpha$ is a closer f -binder of $\gamma$ than $\beta$ iff $\beta \mathrm{f}$-binds $\alpha$ but $\alpha$ does not f -bind $\beta$.

[^18]:    ${ }^{7}$ This proposal differs in various ways from others such as Toivonen (2003) and Farrell (2005), but I will not evaluate the differences here.

[^19]:    ${ }^{8}$ Authors like Bolinger (1975) and Ziv and Sheintuch (1981) have claimed that examples similar to (49b) are grammatical (e.g., It's never been done anything with at all, The oven hasn't been baked any cakes in yet). For Postal (1986:242) such examples "are at best jokes and must be considered ungrammatical." For dialects in which they are grammatical, an analysis like that proposed for Norwegian in section 5 would be appropriate.

[^20]:    1 I have consulted many of the available descriptive studies of Shona in the preparation of this manuscript, in particular Fortune (1995) and (1980). I am thankful to Mrs. Sandra Mavangira for the time she spent answering my queries about the language and providing me with grammaticality judgments. All errors are my own responsibility. In glossing the examples the following abbreviations are used: 3, third person; 1, first person; APPL, applicative; CAUS, causative; F, final vowel; INST, instrumental; LOC, locative; PASS, passive; PAST, past tense; Pl , plural; Sg , singular.

[^21]:    4 I am indebted to Gerlof Bouma for pointing out to me the relevance of chapter 5 of his dissertation for the analysis of word order freezing within OT. Bouma argues there that the strong version of B-OT is sufficient to account for word order freezing in Dutch. In the sections that follow I will argue that the Shona data require the weak version of B-OT. I regret that for reasons of time and space limitations I cannot explore the differences between the two approaches in more depth.

[^22]:    ${ }^{1}$ Similar cases of applicative/causative homophony are found in other languages, for example Australian Aboriginal languages (Austin 2005 [1996]) .

[^23]:    ${ }^{2} \mathrm{http}: / /$ sabda.org/sabdaweb/bible/verse/?b=19\&c=53\&v=5\&version=bis
    ${ }^{3}$ Kroeger (2007) analyses $-i$ as having the following LCS (Lexical Conceptual Structure):
    Locative object (V-i): [x ACT] CAUSE [z BECOME [FULL?] state WITH-RESPECT-TO y] [LOAD manner] $]_{\text {MANNER }}$ ] While this is the right LCS representation for many $-i$ verbs, it does not capture the broad range of the semantics of all -i verbs: in particular, those where the object is understood as a location (i.e. source) from which something (i.e. y) is taken away, as in e.g. example (15c).
    ${ }^{4}$ For psychological verbs, REACT instead of AFFECT is used (Jackendoff 1990).

[^24]:    ${ }^{5}$ Thematic roles are shorthand labels associated with positions with prominence in the Lexical Conceptual Structure.

[^25]:    a. kirim 'send' <<ARG1, ARG2> <ARG3>>
    (ag) (th) (go)

[^26]:    ${ }^{6}$ This attested example is from http://dart1202.wordpress.com/2009/04/07/review-bmw-t-shirt-bmw-citypants, accessed 27 November, 2009.
    ${ }^{7}$ There is evidence that Goal/Locative of jatuh 'fall' or datang 'come' is an oblique-like argument (i.e., associated with the conceptual unit of [PATH] of the verbs) although it is not required to be overtly present on the surface syntax. A (general) Goal/Locative adjunct cannot typically take -i in Indonesian:

[^27]:    i)

    | a. Ia | tinggal | di | Jakarta |
    | :--- | :--- | :--- | :--- |
    | 3 s | live | LOC | Jakarta |

    b. * la meninggal-i Jakarta FOR "s/he lives in Jakarta.'
    ii) a. Ali menangis di kamar Ali AV.cry LOC room 'Ali cried in the room'
    b. * Ali menangisi kamar Ali AV.cry-I room FOR: 'Ali cried in the room'
    ${ }^{8}$ We analyse that verbal bound roots carry argument structures. However, the syntactic core status of the arguments are specified when the roots are affixed, e.g. by the intransitive prefix ber- or transitive -i/-kan. Thus, in this analysis,-kunjung 'visit', for example, is a predicate with two arguments (the 'visitor' and the 'thing visited'). The second argument is understood as Goal, which then fuses with ARG2 of $-i$ in the derived $-i$ verb.

[^28]:    ${ }^{9}$ These examples were found on http://old.rumahfilm.org/esai/esai juno.htm, and an online version of an Indonesian language learners' guide `Beginning Indonesian Through Self-Instruction' by Wolff et al (1992:890), accessed 27 November, 2009.

[^29]:    ${ }^{10}$ In fact, the a-str of the type (32d) allows for double fusion if ARG1 is not filled in with an agent. Thus, the following is acceptable. The water flows because of its natural force.

    | Air | itu | mengalir-i | sawahnya |
    | :--- | :--- | :--- | :--- |
    | water | that | AV.flow-l | rice.field |

    'The water flooded his/her rice field.'

[^30]:    ${ }^{11} \mathrm{http}: / / \mathrm{www} 2$. parc.com/isl/groups/nltt/xle/

[^31]:    ${ }^{13}$ It should be noted that a di-passive verb may have its Actor realised as an enclitic, e.g. di-pukul=nya 'dihit=3s', in which case the actor behaves more like a core argument than an oblique (see Musgrave 2001; Arka and Manning 2008).

[^32]:    ${ }^{\dagger}$ I'd like to thank Mary Dalrymple, Ron Kaplan, Tibor Laczko, Chris Potts, and Ida Toivonen for helpful comments and discussion. I also thank everyone who stopped by my poster at LFG09, as well as audiences at Oxford University and BWTL 12 , for their comments. The ideas presented here grew out of a lecture in the class that I taught with Ida Toivonen at the LSA LSI in 2007; I thank the students for their insightful comments and questions. Any remaining errors are my own.
    ${ }^{1}$ In previous, unpublished work, including the poster presented at LFG09, I have used the term 'Complementizer-Adjacent Nominal Extraction'. While this makes for a nicer acronym, it is misleading, because the categorial status of the extracted subject is irrelevant:
    (i) a. Kim suspects that that Sandy burped shocked Robin.
    b. What does Kim suspect shocked Robin?
    c. *What does Kim suspect that shocked Robin?
    (ii) a. The kids claimed that under the bed was a good place to hide.
    b. Under where did the kids claim was a good place to hide?

    * Under where did the kids claim that was a good place to hide?

[^33]:    ${ }^{2}$ The categorial status of the interpolated adverbial phrase does not seem to matter, although these are both arguably PPs.

[^34]:    ${ }^{3}$ The point here is that intervention alone is not sufficient. However, (10) is also independently ungrammatical without that, which muddies things. I return to this point at the end of section 5.2.

[^35]:    ${ }^{4}$ The restriction of the metavariable to terminal nodes is stated explicitly for clarity, but it follows if we assume that (1) the $\pi$ correspondence function is an injection (one-to-one), since no two elements in a properly tokenized string map to the same c-structure node and (2) $\pi$ is not onto (there are c-structure nodes that have no string correspondents - the non-terminals). The relevant point is that we should be able to assume that $\pi^{-1}(*)$ is a string element, rather than a set of such elements. For example, we do not want to have to consider the case of a node such as a branching VP having more than one string correspondent. The $\pi$ function would likely have a more complex analysis if Lexical Sharing is assumed (Wescoat 2002, 2005, 2007, 2009).

[^36]:    ${ }^{5}$ I reject any contention that a theory can explain a phenomenon if its grammatical models cannot generate the correct pattern of data.
    ${ }^{6}$ PF stands for either Phonetic Form or Phonological Form, depending on the author (e.g., Chomsky 1995, Merchant 2001).

[^37]:    ${ }^{7}$ This amendment is necessary, because $\phi^{-1}(\uparrow$ PIV) on its own does not have the intended effect of checking for the existence of a c-structure correspondent, since $\phi^{-1}$ always returns a set, even when the set is the empty set.

[^38]:    ${ }^{8}$ Functional ID should be simpler than f-precedence, however, since it is performing an existential check on the first set, not a universal one.

[^39]:    ${ }^{9}$ The second definition of f-precedence is not the definition from Bresnan (1995:249), which has the extra clause that $\phi^{-1}(f)$ and $\phi^{-1}(g)$ must be nonempty. The nonempty clause entails that a null pronominal does not f-precede anything and is not f -preceded by anything.
    ${ }^{10}$ This definition makes the standard assumption that dominance is reflexive (Partee et al. 1990:440), which allows it to properly cover terminal nodes.

[^40]:    ${ }^{11}$ Furthermore, any complementizer maps to the same f-structure as $\mathrm{C}^{\prime}$, which dominates the c -structure correspondents of the subject and any other grammatical functions in the f-structure of the clause that the complementizer introduces. This means that $\mathrm{C}^{\prime}$ does not c-precede the c -structure correspondents of the grammatical functions inside it, so the complementizer in fact f -precedes none of the grammatical functions in the f -structure that it introduces.
    ${ }^{12}$ I assume a DP analysis of the nominal with the relative clause NP adjoining to an NP; this preserves the theory of adjunction in Toivonen $(2001,2003)$, but is not a crucial feature of the analysis.

[^41]:    ${ }^{13}$ The initial version of f－precedence in Bresnan（1995：249），which is subsequently revised to edge－based f－precedence （Bresnan 1995：250），would have the same result，because of the clause that $\phi^{-1}(f)$ and $\phi^{-1}(g)$ must be nonempty；see footnote 9.

[^42]:    ${ }^{\dagger}$ The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in the Emmy Noether project PTOLEMAIOS, on Grammar Induction from Parallel Corpora.

[^43]:    ${ }^{1}$ The statistical system uses only the disambiguated morphological analysis of the target language from the training data, not the syntactic analysis itself. In principle, it would be possible to employ a tree-to-tree translation model (Yamada and Knight, 2001; Koehn and Knight, 2003; Huang et al., 2006); however, tree-to-string translation with the capability for morphological generalizations may be a very effective middle ground.

[^44]:    ${ }^{2}$ Note that the architecture is not based on the noisy channel model, so in its purest form, the model should not be used in combination with a language model for the target language.
    ${ }^{3}$ The search strategy adopted is to (greedily) go for the most probable classification outcome in each sub-decision, although in principle it would be possible to use other strategies.

[^45]:    ${ }^{4}$ Some morphological tags are per se underspecified (since the form is identical for various feature values, e.g., starke could be nominative or accusative, hence the tag . NA for Case); here, no disambiguation is needed. We will come back to these underspecified tags in the following section.
    ${ }^{5}$ Note that the syntactic LFG grammar of the target language is not applied in application/decoding mode, since its only function was to provide a disambiguated morphological analysis of the words in the training data.

[^46]:    ${ }^{6}$ In addition, it should be noted that we are seeing the effect of a representational short-hand that was intended for a different application context.

[^47]:    ${ }^{1}$ We would like to thank Ash Asudeh, Mary Dalrymple, Aditi Lahiri, Frans Plank, and Janet Grijzenhout for detailed discussions of the issues in this paper. We would also like to gratefully aknowledge funding from the DFG (Deutsche Forschungsgemeinschaft) for the Urdu ParGram project.
    ${ }^{2}$ The constraint-based LFG architecture is neutral between generation and recognition and so is naturally compatible with processing models of both language production and language comprehension. Traditional phonological approaches are usually biased towards the generation/production direction, describing how syntactic and semantic structures can be converted into some representation of their pronunciation. It is less obvious how the traditional approaches can be incorporated into models of comprehension.

[^48]:    ${ }^{3}$ Selkirk (1995) explicitly addresses the issue of function words and their cliticization to the preceding prosodic word and takes them out of the general mapping algorithm. However, the problem is more general than just function words, as the data from Dutch below shows. Here, it is an adverb which clticizes and thus gives rise to a misalignment between prosody and syntax.
    (i) Ik $\quad\left(\left((\text { trap })_{\omega} \quad \text { te }\right)_{\omega}\right)_{\phi} \quad$ hard
    I kick too hard
    'I kick too hard.' (Wheeldon and Lahiri, 1997, 358)

[^49]:    ${ }^{4}$ We have left the bracketing of and somewhat underspecified in (4). It can represent a prosodic word on its own, but additional bracketings for (4) may occur when and is prosodically a clitic,

[^50]:    ${ }^{5}$ The analysis of Urdu ezafe is part of on-going work on building a computational grammar of Urdu within the ParGram project (Butt et al., 1999; Butt and King, 2007). The representations below reflect the output of the implemented grammar, and the Urdu examples are provided in the ASCII transliteration scheme used by the grammar.
    ${ }^{6}$ This analysis contrasts with proposals that treat items like ezafe as phrasal affixes (Anderson, 2005) that do not appear as separate syntactic elements but are instead morphologically incorporated into their hosts. This is because they share some morphological properties with inflectional affixes (Zwicky, 1987; Samvelian, 2007; Miller, 1992). Most proponents of this approach have worked within Head-Driven Phrase Structure Grammar, but this type of analysis is also compatible with GB/Minimalism in that clitics can be seen as functional items that are placed high in the tree (e.g. within IP) and can thus be thought of as postlexical inflectional items (e.g. van der Leeuw (1997)). The idea of treating a subclass of clitics in the morphology has also appeared in earlier LFG-oriented proposals (e.g. Sadler and Spencer, 2000; Luís and Otoguro, 2005). On our new account this mixture of prosodic/morphological/syntactic properties follows more straightforwardly from the interaction of separate components.

[^51]:    ${ }^{7}$ CHECK features are used by convention in the ParGram grammars to encode information that is needed to ensure syntactic well-formedness but is not theoretically interesting and is not relevant to other modules or domains of application (e.g., semantic interpretation or machine translation). Bögel et al. extend the CHECK convention to handle the cross-module interaction of syntax and prosody.

[^52]:    ${ }^{8}$ Asudeh (2009) proposes an elaboration of the LFG formalism that allows constraints associated with lexical nodes to make direct reference to the structures that correspond to preceding or following lexical nodes. He aims to account for the restrictions on Complementizer-Adjacent Extraction (e.g. that-trace and fixed-subject constraints), but his technique might offer a simpler and more intuitive account of clitic prosodic attachment within a configuration of co-described representations. Asudeh builds on a suggestion made originally by Kaplan $(1987,1989)$ to formalize the mapping between phonological tokens and the lexical nodes of the c-structure in terms of another projection function within the overall Correspondence Architecture of LFG. He observes that this function, denoted as $\pi$, is one-to-one, and its inverse is therefore a function that maps from a lexical node to the corresponding phonological token. This can be composed with functions that take phonological tokens into the tokens that precede or follow them. Asudeh defines a new designator $\succ=\phi\left(M\left(\pi\left(\operatorname{Next}\left(\pi^{-1}(*)\right)\right)\right)\right)$ to designate the f-structure of a following lexical node. Similarly, we can define $\prec_{p}$ as a designator for the p-structure corresponding to the lexical node of a preceding phonological token. Given this machinery, we can replace the CHECK assignment on the EZAFE clitic with the constraint ( $\prec_{p}$ CL-FORM) = ezafe and remove the CL-FORM and CHECK annotations from the host noun.

    This formalization avoids some of the unintuitive aspects of the Bögel et al. (2008) account, but it still requires a case-by-case distribution of constraints.

[^53]:    ${ }^{9}$ In the XLE implementation, METARULEMACRO takes three arguments: the CATegory, the BASE CATegory, and the RHS. For our purposes, the distinction between the category and the base category is unimportant.

[^54]:    ${ }^{\dagger}$ The work reported in this paper was in part supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in SFB 632 on Information Structure, project D4 (Methods for interactive linguistic corpus analysis). We would like to thank the audience at the LFG 2009 Conference in Cambridge for their comments and the fruitful discussion. Special thanks go to Farrell Ackermann, Alex Alsina, Ash Asudeh, Miriam Butt, and Anette Frank.
    ${ }^{1}$ Some of the examples in this paper are attested sentences. For these, the source is cited. For volatile sources like the WWW, we also cite the consultation date.
    ${ }^{2}$ nl. yunomi.be/artikel/uien-snijden-zonder-tranen, 28/9/2009
    ${ }^{3}$ http://www.ru.nl/csmr/disclaimer/fmpgdo_producties/, 28/9/2009
    ${ }^{4}$ http://media.europa-nu.nl/9353000/1/j9vvhjsfpncqnhs/ vi5zbeznuuqj?ctx=vh72mb14wkwh, 28/9/2009

[^55]:    ${ }^{5}$ http://www.nieuwsblad.be/Article/Detail.aspx?ref=hv\&ArticleID= GBOMB9JT

[^56]:    ${ }^{6}$ Some of these are listed in Dik (1980). Another incomplete list can be found in the grammar of the Alpino parser for Dutch (van Noord, 2006).

[^57]:    ${ }^{7}$ We are not aware of any published reference for this, although the grammar of the Alpino parser for Dutch lists this as a possible argument frame for some of the combinations with laten.

[^58]:    ${ }^{8}$ There is however a reading of weten 'manage' that takes a te-VP. There is also a very restricted AcI-like weten which can be analyzed as taking an object and a predicate in the form of a te-VP. These readings are not relevant to the LPC in (14a), however.
    ${ }^{9}$ Spoken Dutch Corpus CGN, broadcast news subcorpus, sentence fn0001946:5.
    ${ }^{10}$ http://innameoffame.web-log.nl/innameoffame/2007/03/de_eerste_ kuike.html, 7/10/2009.

[^59]:    ${ }^{11}$ As a side remark, although Dutch has the simplex verb tonen for the English to show, its use is marked. It is much more common to use laten zien.
    ${ }^{12}$ With a concrete object, the label 'verb of transmission of a message' may sound as a bit of a misnomer. However, note that the same holds for the English VTM to show, which may take a concrete theme as in He showed her the giant sequoia. A study of the precise lexical decomposition of the propositional/abstract vs concrete theme taking versions of VTM verbs is beyond the scope of this paper.

[^60]:    ${ }^{13}$ http://www.collegenet.nl/content/literatuur/tot1920lit/index005. htm, 8/10/2009
    ${ }^{14}$ http://bis.almere.nl/regelgeving/06095_00/45377965.HTML, 8/10/2009. The sentence is from a local government policy text. Realizing blijken's experiencer in a aan-PP feels highly marked and may be acceptable only in legalese.
    ${ }^{15}$ Spoken Dutch Corpus CGN, spontaneous telephone dialogue subcorpus, sentence fn006935.186

[^61]:    ${ }^{16}$ http://www.volkskrant.nl/archief_gratis/article820739.ece/Ach, _je_geeft_elkaar_even_een_knipoog, 8/10/2009
    ${ }^{17}$ http://www.primeurjagers.nl/d1896_minnares_gordon_ramsay_een_ ware_sex_tijger_in_bed_hier_alle_details_.html, 8/10/2009. Orth. corrected.
    ${ }^{18}$ www. cubra.nl/eddalderop/misdienaars1943.htm, 9/10/2009
    ${ }^{19}$ http://www.moslima.nl/nwmoslimas/linda.htm, 9/10/2009

[^62]:    ${ }^{20}$ It is a terminological question whether one wants to distinguish prototypical complex predicate constructions from the type of monoclausal phenomenon we have identified for the LPC. We use the term complex predicate to refer to the technical aspects, literally requiring the formation of a single PRED value from two items.

[^63]:    ${ }^{21}$ Note that this sketch assumes that the theory of grammar reserves a place for fully systematic, syntactic rules - they are not assumed to be just an extreme case of a construction, in that it is completely unrestricted. We believe that this is a desirable theoretical property (for which LFG has a very natural place in its rules), since otherwise, the formal place for formulating syntactic generalizations seems to be lost.

[^64]:    ${ }^{\dagger}$ We would like to thank Ruth Kempson, Ronnie Cann, Andrés Enrique-Arias and Nigel Vincent for valuable comments. Furthermore, the audiences at the 14th International LFG Conference at Cambridge and the Formal vs. Processing explanations of Syntactic Phenomena workshop at the University of York are also thanked for providing useful comments and inspiring critique. Miriam Bouzouita gratefully thanks the Ministerio de Ciencia e Innovación (Juan de la Cierva Grant) and Stergios Chatzikyriakidis the Arts and Humanities Research Council for providing partial funding to parts related to this paper. Of course, any inconsistencies remain our own.
    ${ }^{1}$ This section summarizes research reported in Bouzouita (2007, 2008a, 2008b, 2008c) and Bouzouita \& Kempson (2006).
    ${ }^{2}$ In the following examples, the constituents preceding the clitic will be underlined for ease of consultation.

[^65]:    ${ }^{3}$ It is important to note that this formal simplification does not entail a simplification in the data output. As can be seen in Table 1 , RenSp presents more variation environments than MedSp and could therefore be considered as more complex.
    ${ }^{4}$ There are varieties of the latter two languages that exhibit different variants of the constraint (Ormazabal \& Romero 2007; Bianchi 2006).

[^66]:    ${ }^{5}$ Such a fact is noted in the literature as a repair (Bonet 2007; Rezac 2008) and may involve a number of different strategies 'saving' in a way the illicit PCC combinations (object camouflage in Georgian, absolutive displacement in Basque). See Rezac (2008) for an extensive discussion of these different repair strategies.

[^67]:    ${ }^{6}$ Note that no proper treatment of negation is meant here. See Chatzikyriakidis (forthcoming) for an attempt to formalize negation in DS without the use of a [+NEG] feature.

[^68]:    ${ }^{7}$ Note that the fact that a noun exhibits some syncretism does not mean that constructive or output filter case is no longer possible. What we need to further look at is which cases in the paradigm are syncretized. For example 3rd declension nouns are always syncretized for nominative/accusative. However, this does not mean that no accusative case filter can be projected in this case, since nominative can be always parsed as a link under standard DS assumptions (Cann et al. 2005).

[^69]:    ${ }^{8}$ It is a well known fact that ethical datives escape the PCC (Ormazabal \& Romero 2007; Kempson \& Chatzikyriakidis 2009 among others). Assuming the development of an alternative parsing strategy for $1 \mathrm{st} / 2 \mathrm{nd}$ person clitics that basically parses the clitic as a linked structure, in effect matching the third strategy mentioned, such a fact is straightforwardly captured. See Kempson \& Chatzikyriakidis (2009) for the exact formulation and argumentation.

[^70]:    ${ }^{1}$ One exception are PPs headed by von which occurs very frequently with a meaning corresponding to that of a genitive, and therefore cannot be included in this generalisation.

[^71]:    ${ }^{2}$ Note that while pronouns were extracted during this process, PNV-triples containing pronouns were ignored.

[^72]:    ${ }^{3}$ http://www.collocations.de

[^73]:    ${ }^{4}$ We do not take the head noun of this DP into account at the moment.
    ${ }^{5}$ The missing results for feature combinations C and CD are due to problems with training.

[^74]:    $\dagger$ Thanks to the LFG09 audience members for many useful comments, especially Joan Bresnan, Miriam Butt, Mary Dalrymple and Yehuda Falk.
    ${ }^{1}$ The valence pattern discussed here is also found in the other Algonquian languages. See Dahlstrom (1991) for Plains Cree, Rhodes (1991) for Ojibwe, Bloomfield (1962) for Menomini, etc. Rhodes (1991) presents a Relational Grammar analysis for the Ojibwe phenomenon that treats many of the issues raised here.

[^75]:    ${ }^{2}$ Abbreviations in the examples: 3' third person obviative, AOR aorist, EP epenthetic consonant, IC Initial Change (ablaut process affecting the vowel of the first syllable of the verb, required by various verb paradigms, including participles, which are used in relative clauses), IMP imperative, IND independent indicative, OBV obviative, PART participle, REDUP reduplication, X unspecified subject. On transitive verbs " $>$ " separates the indication of SUBJ and OBJ features: e.g. " $1>3$ " for $1^{\text {st }}$ singular subject acting on a third singular object; the label of the verb's inflectional paradigm follows the subject and object agreement features.

[^76]:    ${ }^{3}$ These are the stems which are labeled "AI +O " in Algonquianist terminology.

[^77]:    ${ }^{4}$ Perhaps 'proposition' is another thematic role associated with $\mathrm{OBJ}_{\theta}$ : if the suggestion of Alsina et al. (2005) to eliminate the GF of COMP is pursued, the sentential complements of Meskwaki could be reanalyzed as propositional $\mathrm{OBJ}_{\Theta} \mathrm{S}$.

[^78]:    ${ }^{\dagger}$ I wish to thank George Aaron Broadwell for his assistance, and Ronald Kaplan and Michael Wescoat for their helpful comments. I am greatly indebted to my K'ichee' Maya consultants, in particular Felipe and Juan Barreno García of Totonìcapán, Guatemala. All the usual disclaimers apply.
    ${ }^{1}$ All K'ichee' data are from the author's field work, except (36). First, second, third person $=1,2$, 3, absolutive agreement marker $=\mathrm{ABS}$, animate pluralizer $(e e)=\mathrm{PLU}$, antipassive $=\mathrm{AP}$, completive $=\mathrm{COM}$, determiner $=\mathrm{D}(\mathrm{ET})$, distributive $($ taq $)=\mathrm{DISTR}$, distributive pluralizer $($ taq $)=\mathrm{PL}$, ergative agreement marker $=$ ERG, incompletive aspect $=I N C$, independent pronoun $=P R O$, interrogative $=$ INT, irrealis $=I R R$, negative $=$ NEG, nominalizing suffix $=$ NOM, particle $=P T$, possessive $=P O S$, transitive/intransitive phrase final marker $=T / I P F$, plural $=-P L$, preposition $=P(R E P)$, singular $=S$.
    ${ }^{2}$ The distributive taq (DISTR) is not fully addressed in this paper due to space considerations. I propose elsewhere that the distributive taq (DISTR) is a non-projecting word, that it right head-adjoins to verbal predicates only, and that its semantics is representative of distributives cross-linguistically. The paper's title reflects my hypothesis that the non-bound morpheme taq actually represents two words, that, although homophonous, differ in terms of semantics, word type, distribution, and syntax.
    ${ }^{3}$ The exception is: 'partícula que sirve para distribuir el efecto de un verbo, adjetivo, o preposición a las varias entitades de un sustantivo plural' from García Hernández and Yac Sam (1980:144).

[^79]:    ${ }^{4}$ K'ichee' Mayan is an ergative, pro-drop, head-marking language that marks agreement on the finite verb with ergative and absolutive agreement markers. Possessed nouns (POSM) agree in person and number with their possessors (POS). Complex prepositions agree in person and number with their object complements. I argue that canonical (unmarked) word order is [ $\left.\mathrm{S}^{0} \mathrm{XP}{ }^{*}\right]$.

[^80]:    ${ }^{5}$ Willson (2004) first demonstrated the interrelationship of the distributive pluralizer taq and attributive adjectives in the DP complements of prepositional phrases.
    ${ }^{6}$ To indicate plurality in complements, speakers mildly prefer the singular form of the prefixed agreement maker in conjunction with the distributive pluralizer taq, rather than the plural paradigm of agreement markers with or without the distributive pluralizer taq.

[^81]:    ${ }^{7}$ Although it is possible for the distributive pluralizer taq to follow both the preposition and the attributive adjective of the PP's DP complement at the same time, the multiple use of taq in this manner is grammatical but never used.

[^82]:    ${ }^{8}$ The distributive pluralizer taq can follow the preposition and be used in the phrasal compound at the same time but the usual warnings against multiple uses of taq apply (21c).

[^83]:    ${ }^{9}$ Note: jachina'q 'who (PL) (phrase-final)' < jachin taq 'who (PL)' (non-phrase-final)

[^84]:    ${ }^{10}$ It is possible to use the distributive pluralizer taq in both places at the same time, but repetition of the distributive pluralizer almost always never occurs.

[^85]:    a. Lee q'eq-a chi'l saq-a wakax DET black-ATT CONJ white-ATT cow 'The black and white cow'
    b. Lee q'eq-a chi'l saq-a taq wakax DET black-ATT CONJ white-ATT PL cow 'The black and white cows'
    c. *Lee q'eq-a taq chi'l saq-a wakax DET black-ATT PL CONJ white-ATT cow ('The black and white cows')

[^86]:    ${ }^{11}$ From its syntactic behaviour in nominals and at the edges of the verb complex, I suggest that the adverb $\operatorname{chi}(k)$ 'again, already' is a non-projecting word. When it is used in a nominal with the indefinite determiner jun $\sim$ jujun, the combination of the two means 'another (lit. one again).'
    ${ }^{12}$ In contrast, the distributive taq (DISTR) and the directionals, which I argue are non-projecting clitics, can together immediately follow a finite verb and can order freely with each other.

[^87]:    ${ }^{13}$ The distributive $\operatorname{taq}$ (DISTR) follows the finite verb and distributes the verb (the distributive share) over the semantically plural nominal lee chaak 'every job' (the sorting key).
    ${ }^{14}$ My translation of the K'ichee', not the Spanish.
    ${ }^{15}$ Except for the $w h$-interrogative, in which case, the distributive pluralizer taq right-adjoins to it.
    ${ }^{16}$ See Toivonen (2003) for the definitive analysis of projecting \& non-projecting clitics \& particles.

[^88]:    ${ }^{17}$ King (1995) proposes two linear precedence (LP) rules, while Falk (2001:49) proposes five.

[^89]:    ${ }^{18}$ In this paper, all phrase-structure rules are unordered.

[^90]:    ${ }^{19}$ Possessors can extract to e-topic position adjoined to CP . The binding relation remains in effect because of agreement morphology on the possessum that co-indexes the possessor's person/number.

[^91]:    ${ }^{20}$ Add to (46) the phrase structure rules shown in (43), (44), and (45).

[^92]:    ${ }^{1}$ Thanks to Alex Alsina, George Aaron Broadwell, Mary Dalrymple, and Louise Mycock for comments on a previous version. This research is supported by the Israel Science Foundation (Grant No. 207/07).

[^93]:    ${ }^{2}$ What is meant here by "subject" is "pivot" in the sense of such studies as Dixon (1994) and Falk (2006)

[^94]:    ${ }^{3}$ I will not attempt to characterize pragmatic prominence formally in terms of i-structure.
    ${ }^{4}$ In addition to the considerations here, prosody also plays a role in some island constraints. For a case in Japanese (apparent wh island constraint and the "additional wh effect"), see Mycock (2006: 188).

[^95]:    ${ }^{6}$ Mycock (2005) observes that the prosody in in-situ constructions marks the focus and the scope of interrogativity. It would be interesting to see if there are prosodic differences between the two varieties of Egyptian Arabic

[^96]:    ${ }^{`}$ This is a very tentative proposal, as there are quite a few apparent counterexamples.
    ${ }^{8}$ Erteschik-Shir and Lappin (1979) present evidence that allegedly shows that sentential subjects cannot be pragmatically prominent ("dominant" in their terminology). However, their data are problematic. For example, one of their examples is
    (i) Bill said: That Sheila knew all along is likely.
    a. which is a lie-it isn't
    b. $\quad$ *which is a lie-she didn't

    But (ib) is infelicitous here for semantic reasons: even if Sheila did not know all along, Bill's claim that she was likely to is not a lie. Their other examples suffer from similar defects.
    ${ }^{9}$ I find such examples ungrammatical.

[^97]:    ${ }^{10}$ Interestingly, these are internally-headed relative clauses. Navajo has both internallyand externally-headed relative clauses, but the internally headed ones are more common. Island constraints work the same in both. In the case of the "Sentential Subject Condition" (an infelicitous name for the Navajo implementation, since it applies to non-subjects as well), this is an example of how grammaticalization of a constraint results in the original motivation being obscured. In internally-headed relative clauses, parsing should not be an issue, since the clause is structurally complete, but the constraint has been extended to it.

[^98]:    ${ }^{11}$ I use double brackets to mark the boundaries of the wh path．

[^99]:    ${ }^{12}$ It is plausible that the bottom of a wh path is specially marked，just as the top is．I am not aware of any unequivocal evidence from wh path phenomena，but this does not negate the existence of the feature．The variety of Spanish discussed in Torrego（1984）distinguishes the lowest clause of a wh construction from others，but the data are problematic（Alex Alsina，personal communication）．If both of these features exist，they combine to create the following four possibilities：
    （i）$\quad[+\mathrm{T},-\mathrm{B}]$ ：the top of a construction in which the wh path spans more than one clause．
    －$[-\mathrm{T},+\mathrm{B}]$ ：the bottom of a construction in which the wh path spans more than one
    clause．
    －$[-\mathrm{T},-\mathrm{B}]$ ：the middle of a construction in which the $w h$ path spans more than two clauses．
    －$\quad[+\mathrm{T},+\mathrm{B}]$ ：a clause containing a one－clause $w h$ construction．

[^100]:    ${ }^{13}$ This is also true of $w h$ islands and finite-clause islands. It is not surprising that a language would choose a uniform treatment for different islands.

[^101]:    ${ }^{\dagger}$ We thank Louisa Sadler and Doug Arnold for their invaluable support and guidance, and also Mary Dalrymple for her comments.

[^102]:    ${ }^{1}$ In English, disjunctively conjoined nouns with the predisjunction and disjunction either...or show similar agreemet patterns according to Morgan (1985)

[^103]:    ${ }^{2}$ For more discussion see Eggert (2002).

[^104]:    ${ }^{1}$ VPx is a category used to allow for the attachment of extraposed constituents under VP without allowing for their attachment under each recursive expansion of VPx. Apart from this detail, the VPx category in the German ParGram LFG is equivalent to the commonly assumed VP.

[^105]:    ${ }^{2}$ We are aware that right node raising and conjunction reduction do not only affect VPs. However, these phenomena are particularly frequent with VPs. Moreover, "conjunction-reduced" constituents of other categories can often be analyzed reasonably even without the deleted material, whereas the subcategorization requirements of verbs tend to make this impossible for "conjunction-reduced" VPs. Figure 7 illustrates an analysis involving a "conjunction-reduced" VP as well as an NP that arguably lacks right-node-raised material, namely 8,4 Prozent. Note that the "conjunction-reduced" VP is analyzed by means of a special VP coordination rule, whereas the NP 8,4 is simply analyzed as a headless NP whose semantic head needs to be recovered by post-syntactic means.

[^106]:    ${ }^{3}$ For the sake of simplicity and clarity, we omit functional annotations and linear precedence constraints which regulate constituent order.
    ${ }^{4}$ In the ASCII-based XLE notation, ${ }^{\wedge}$ ' stands for ' $\uparrow$ ', '!' stands for ' $\downarrow$ ', and ' $\$$ ' stands for ' $\in$ '.

[^107]:    ${ }^{1}$ Through personal communication with John Maxwell.
    ${ }^{2}$ Note that results for our system should not be compared with results reported in Riezler and Maxwell (2006) since our transfer component is statistical while that of Riezler and Maxwell (2006) is a hybrid.
    ${ }^{3}$ There are parsing and generation resources available for LFG that are trained fully automatically (Cahill et al., 2004; Cahill and van Genabith, 2006). We plan to use these resources with our statistical transfer decoder to compare with the current hybrid system in the near future.

[^108]:    ${ }^{4}$ Through personal communication with John Maxwell.

[^109]:    ${ }^{\dagger}$ This paper builds on collaborative work with Kjell Johan Sæbø and Cathrine Fabricius-Hansen (both University of Oslo), to be published as $\mathrm{S} æ b \emptyset$ et al. (Forthcoming). The responsibility for the LFG- and Glue-based analyses presented here remains entirely my own.

[^110]:    ${ }^{1}$ For convenience I will refer to noun phrases without a determiner or a postposed article as NPs, and noun phrases with a determiner or a postposed article as DPs. Nothing hinges on this. For clarity, I sometimes speak of 'bare' NPs and 'full' DPs.
    ${ }^{2}$ The term 'small clause' is here used pre-theoretically without a specific c-structure analysis being implied; see section 2 for the analysis.
    ${ }^{3}$ The actual example continues 'but nowadays, the skis are most often worn by the mother', exploiting the control ambiguity and referring to the traditional saying that Norwegian babies come with skis.

[^111]:    ${ }^{4}$ This point is made by Aa (2006).

[^112]:    ${ }^{5}$ See for example Dalrymple et al. (2004) on the importance of a proper distinction between syntactic functions and semantic arguments.

[^113]:    ${ }^{6}$ Notice that the subject position is optional since it is unlikely that all Norwegian prepositions have subjects. However, since the subject is required by the argument structure of med, but never overtly realized, the rule always takes effect.

[^114]:    ${ }^{7}$ Notice that the (English) genitive also comes with a built-in definite article.

[^115]:    ${ }^{8}$ Notice that we did not introduce the arguments of the matrix verb here, as these will be introduced by a transitive template, see Asudeh et al. (2008).

[^116]:    ${ }^{9}$ Notice that both this meaning constructor and the one in (50) can apply not only to 'small clauses' but also to bare NPs, as we will see in section 4.5. For that reason we introduce the slight functional uncertainty $\{$ XCOMP $\mid \mathrm{OBJ}\}$.
    ${ }^{10}$ As another simplification, håndkle is introduced in the meaning constructor as the syntactic subject of over. However, this could be done in other ways. It is not clear that there is a subject position in over, but nothing really hinges on that question here.
    ${ }^{11}$ The semantic representation ignores the subject argument. Also, as shown in Asudeh et al. (2008), arguments of the verbs should be introduced by argument structure templates such as

[^117]:    $m_{\sigma}$
    $\lambda y . \lambda s . \iota x . \operatorname{tank}(x, y) \wedge \operatorname{full}(x, s):$

[^118]:    ${ }^{\dagger}$ We thank Josef van Genabith and Jennifer Foster from Dublin City University for providing the initial development data and the Natural Language Theory and Technology group at PARC for providing the XLE LFG grammar and the XFR ordered rewrite system.

[^119]:    ${ }^{1}$ The XLE f-structures generally encode standard theoretical LFG f-structure features. The one exception to this are the CHECK features which are used primarily grammar-internally to constrain the application of specific syntactic constructions or to provide information useful for debugging. By convention, the names of these features begin with an underscore. As will be seen in $\S 3.2$, for the purposes of this project, the only CHECK feature of importance is the _SUBCAT-FRAME feature which the XFR semantics uses for lexical look-up.

[^120]:    ${ }^{2}$ Most of the frames were obtained from electronic dictionaries or manually. See O'Donovan et al. (2005) for ways to bootstrap creation of such lexical resources from treebanks.

[^121]:    ${ }^{3}$ The XLE f-structures contain the nominative form of the pronoun as the PRED instead of the more commonly accepted PRED 'pro' familiar from the theoretical LFG literature and seen in the input DCU f-structure in Figure 9. This choice of PRED value for pronouns is independent of the issues in this paper, other than the fact that the XFR rewrite rules must be able to alter the PRED values correctly in order to create the structures that the XFR semantics expects and input.

[^122]:    ${ }^{4}$ Precision and recall are widely used to evaluate the output of natural language processing systems. When a set of "test" items Y (rewritten DCU f-structures) is compared to a set of "reference" items $X$ (original XLE f-structures), precision $(X \mid Y)=\frac{|X \cap Y|}{Y}$ and recall $(Y \mid X)=\frac{|X \cap Y|}{X}$ (Melamed et al., 2003) are measures to compare the output. The f -score is a weighted average of precision and recall.

[^123]:    ${ }^{5}$ The analysis of proper nouns is particularly important for the semantics because of the applications it is used in. As such, the testsuites contain examples of proper nouns in order to assure that they are being correctly processed by the syntax and the semantics.

[^124]:    ${ }^{6}$ These passage-query-answer pairs came from the regression sets (de Paiva and King, 2008) used in developing the question answering system. By using a regression set which was designed for an application that uses the XFR semantics, any changes in overall system performance could be more easily detected.

[^125]:    ${ }^{7}$ Within the XLE LFG parser, the syntax time is roughly divided as: morphology (including the textbreaker and tokenizer) ( $4 \%$ ), lexicon ( $6 \%$ ), chart ( $25 \%$ ), unifier ( $55 \%$ ), completer ( $4 \%$ ), solver (6\%).
    ${ }^{8}$ Extremely long sentences take much longer to parse, as is also the case for the XLE parser.

[^126]:    ${ }^{1}$ I gratefully acknowledge a British Academy Postdoctoral Fellowship, which has enabled me to continue this research.

[^127]:    ${ }^{2}$ I also do not know at this stage whether they follow a particular ordering within the argument structure or not. This, however, should not have a bearing on the argumentation offered in this paper.
    ${ }^{3}$ I assume the following gender values for Polish: M (masculine), F (feminine), or N (neuter) in the singular, and VIR[ILE] (masculine human) or NONVIR[ILE] (all other, i.e. non-masculine human and all non-human) in the plural. This represents a simplified view of Polish gender in its interaction with number, but it is sufficient to describe the phenomena discussed in this paper.

[^128]:    ${ }^{4}$ Dowty (2000) refers to the English variant of this construction as the 'Location-Subject Form', and notes that he adopts this term without implying a commitment to the term 'location' as a thematic role.
    ${ }^{5}$ Likewise, Dowty (2000) refers to the English variant of this construction as the 'Agent-Subject Form', also without implying a commitment to the term 'agent' as a thematic role.

[^129]:    ${ }^{6}$ Pinker (1989) suggests that the verbs involved in the locative alternation may vary with regard to which member of the alternation is the conceptual core. Interestingly, other languages may favour a particular conceptualisation of events, as reported by Schaefer and Egbokhare (2009) for Emai (a West Benue-Congo language spoken in Nigeria) which is characterised by constraints on linear argument order and the virtual absence of argument alternations, allowing only verb constructions with Figure preceding Ground (cf. Talmy 2000).

[^130]:    ${ }^{7}$ Note, however, that the first suggestion of integrating Dowty's Proto-Role proposal into LMT came from Zaenen (1993). For an overview and discussion of her approach, see Butt (2006:135-138).
    ${ }^{8}$ Subscripts in this representation are only a memory aid, helping visualise and later recall the ranking of the argument slots. It is the linear order in the representation of the argument structure that gives us the ranking information, not the subscripts.

[^131]:    ${ }^{9}$ For more detailed argumentation in support of this model, see Kibort (2007) (and earlier versions of these ideas in 2001:14-19 and 2004:349-352).

[^132]:    ${ }^{10}$ I assume that this verb is unaccusative, though nothing in the present discussion hinges on this assumption.

[^133]:    ${ }^{11}$ Note also that it is this distinction between 'referents' and 'semantic participants' that corresponds in some way to Jackendoff's functional representation of arguments at 'action tier' and the representation of their conceptual roles at 'thematic tier', respectively (1990:126ff).

[^134]:    ${ }^{12}$ For some more discussion of this construction in Polish see Kibort (2004:295-318) and (2006/2008a).

[^135]:    ${ }^{13}$ Note that the proposed variant of LMT does not need to resort to the 'Asymmetrical Object Parameter' (Alsina and Mchombo 1988) which regulates the occurrence of argument structures with two unrestricted [-r] arguments. See Kibort (2008b) for references and discussion.

[^136]:    ${ }^{1}$ In the glosses a marker like 3 SG on the noun head represents both possessivemarking and agreement-marking.

[^137]:    ${ }^{2}$ Note that the perfectivizing preverb (meg) is missing in (7c). The reason for this is that láb-on (leg-SUP) occupies a special verbal modifier position which can encode perfectivity as well.

[^138]:    ${ }^{3}$ Note that Hungarian is a pro-drop language, including the pro-drop of possessors. In (9a), for instance, the possession- and agreement-marking (3SG) is capable of encoding a $3{ }^{\text {rd }}$ person singular pronominal possessor in LFG terms. The overt pronoun version of the possessive noun in (9a) is this.
    (i) az ő közep-é-t
    the it.NOM middle-3SG-ACC
    'its middle'
    (In Hungarian ő is a gender-neutral 3SG pronoun.)

[^139]:    ${ }^{4}$ I use the following abbreviations: $\mathrm{sr}=$ social relation, $\mathrm{m}=$ member. These labels are informal here, too.

[^140]:    ${ }^{5}$ For the sake of simplicity of exposition, he use English words in these representations.

[^141]:    ${ }^{1}$ This paper was presented at the Fourteenth International Lexical Functional Grammar Conference (LFG09, Cambridge, England, July 2009). For input and discussion, I would like to thank the audience, the anonymous reviewers, and the proceedings editors. Special thanks to Tibor Laczkó. I would also like to thank my colleagues Anneliese Pitz, Jan Terje Faarlund, Eirik Welo, Dag Haug, Janne Bondi Johannessen, Hans Petter Helland, and Marianne Hobæk Haff.

[^142]:    ${ }^{2}$ Old Norse had a dative external possessor construction (Faarlund 2004:17071). Some relics can be found in fixed expressions and archaisms, such as (i).
    (i) De stakk ham en dolk i ryggen. (from Åfarli and Eide 2003:126) they stabbed him a dagger in back.DEF They stabbed a dagger into his back. [footnote continues on next page]

[^143]:    ${ }^{3}$ Assuming that the object with unergative possessor raising is non-thematic, one could imagine an analysis in which the object is a subject in a small clause, parallel to some analyses of resultatives. There seems to be no motivation for a small clause analysis, however, because there is no predication relation between the possessor and the PP. (Note that a small clause analysis could not replace possessor raising, which is necessary to account for the relation between the possessor and the body part noun.)
    ${ }^{4}$ A potential problem with analyzing these objects as non-thematic is the following: Hellan 1988:120 proposed that a non-thematic reflexive cannot be complex. Even so, a sentence with unergative possessor raising can take a complex reflexive object, as in (i).
    (i) Han tråkket seg selv på føttene.
    he stepped REFL SELF on feet.DEF
    He stepped on his own feet.
    It is not clear, however, if the generalization about non-thematic reflexives is correct. For example, the resultative expression 'sing oneself to sleep' is predicted to take the simple reflexive, but Sæbø 2009:122 points out that it is as common with the complex reflexive (synge seg selv i søvn) as with the simple reflexive (synge seg isøvn) on Norwegian web pages.

[^144]:    ${ }^{5}$ Western 1921:142 assumes that intransitive verbs take indirect objects in sentences with possessor raising. It must be noted, however, that he does not distinguish possessor raising from sentences with "free datives", and most of the examples he gives have unaccusative verbs.

[^145]:    ${ }^{7}$ In the original LFG formalism, there was a general restriction to prohibit reference to more than two attribute names on either side of an equation. This was called Functional Locality, see Kaplan and Bresnan 1982:278, note 20. This restriction cannot be considered relevant in current LFG.

[^146]:    ${ }^{8}$ The redundancy is somewhat less noticable when the noun is modified, as in (i)-(ii).
    (i) Han kysset henne på den skadede tåen hennes. he kissed her on the hurt toe.DEF her He kissed her on her hurt toe. [footnote continues on next page]

[^147]:    ${ }^{\dagger}$ The work presented in this paper stems from my doctoral dissertation, completed at the University of Cambridge in 2008. I am grateful to my Research Committee, Teresa Parodi, Jim Blevins and John Hawkins, for their valuable input.
    ${ }^{1}$ Third person si is the most widely mentioned form for Italian, partly due to the fact that it is also used to mark inchoatives, middles, and several other derived structures. The remaining reflexive and reciprocal clitics are $m i$ ( 1 st singular), $t i$ (2nd singular), $c i$ (1st plural) and vi (2nd plural).
    ${ }^{2}$ The problem in fact concerns intransitive reflexives and reciprocals in general, including the unmarked English forms (John shaved, Bill and Mary kissed) and forms marked by affixes, as in Russian, Greek, or Hungarian, or by a dedicated verbal template, as is the case in Hebrew. We will, however, limit our discussion to Italian and Serbian.

[^148]:    ${ }^{3}$ Clearly, the mapping of semantic participants in reflexive and reciprocal formation cannot be exactly the same, but if we want to assume that one of the verbal arguments is reduced, it would have to be the same argument in both cases. For the sake of simplicity, in the rest of the paper we will treat reflexive and reciprocal forms as being derived through the same argument mapping.

[^149]:    ${ }^{4}$ The sentence is grammatical with the reading 'Gianni is worried'.

[^150]:    ${ }^{5}$ Note that in Serbian the 3rd singular auxiliary form je is normally omitted in the presence of se; see e.g. Progovac $(2005,135)$.

[^151]:    ${ }^{6}$ See Alsina (1996) for a detailed explanation and a discussion of several other constructions which are best explained assuming a many-to-one mapping between arguments and grammatical functions.

[^152]:    ${ }^{7}$ See Alsina (1996, 123-134) for detailed accounts of several other phenomena.
    ${ }^{8}$ Rákosi (2008) is a notable exception to this trend, as he assumes reflexives and reciprocals in Hungarian to have different derivations.
    ${ }^{9}$ Here we refer only to forms relevant for either reflexive or reciprocal meaning, not to impersonals or middles.

[^153]:    ${ }^{10}$ Dowty's approach to thematic roles is in fact adopted by Alsina as well, but for the sake of simplicity it was not introduced earlier.

[^154]:    ${ }^{1}$ The typology covers the default case, in which the reflexive receives no special prosodic prominence and the (local) licensing predicate is other-oriented.

    I assume that the local domain relevant for binding theory is defined by the notion of Minimal Complete Nucleus (cf. Dalrymple 1993 and Bresnan 2001): the antecedent of the anaphor must be in the smallest f-structure that contains the fstructure of the anaphor and a SUBJ function. This will suffice for the purposes of this paper.
    ${ }^{3}$ Seg also has nuclear uses; see Lødrup (2007) for details.
    ${ }^{4}$ I thank the participants of the Cambridge LFG conference for calling my attention to this article.

[^155]:    5 Given that Hungarian is a pro-drop language, pronominal possessors are normally not pronounced. Note also that Hungarian does not have grammatical gender, so third person pronouns do not manifest gender-related variation in form.

[^156]:    ${ }^{6}$ This is not in contradiction with what I claimed in 4.2., namely that maga is the basic intensifier in Hungarian. Önmaga is best as an intensifier on the 'alone' reading, when it still needs to be separated from its associate or its associate needs to be pro-dropped. Maga is subject to no such restrictions on its intensifier use.

[^157]:    7 In (31) below, I assume a simplified f-structure analysis of the Hungarian possessive construction. See É. Kiss (2002) for a more detailed discussion of the data. Laczkó (2007) is a recent LFG-theoretic analysis of Hungarian possessives.

[^158]:    ${ }^{1}$ TiGer provides head annotation for all categorial nodes except NPs, PPs and PNs. Due to the flat annotation in TiGer, partly resulting from the decision not to annotate unary nodes, the problem of identifying the correct head for those nodes is more severe than for the $\mathrm{TuBa}-\mathrm{D} / \mathrm{Z}$, where the more hierarchical structure results in smaller constituents which, in addition, are all head-marked. When annotating original treebank trees, the head-finding rules are applied to NP, PP and PN nodes; when

[^159]:    ${ }^{2}$ The TiGer DB distinguishes 52 different grammatical features. We use a slightly modified version without the distinction between different prepositional objects, and without morphological features or compound analysis.
    ${ }^{3}$ The TüBa-D/Z gold standard was semi-automatically created by Heike Zinsmeister and Yannick Versley, using the conversion method of Versley [2005] on 100 randomly selected trees from the TüBa-D/Z. The feature set is similar to the TiGer DB.

[^160]:    ${ }^{4}$ Results report constituent-based evalb labelled F-scores on syntactic nodes and grammatical function labels when using gold POS tags with gold GF labels as parser input
    ${ }^{5}$ We split the gold standards into development and test set, with 500 test set trees for the TiGer DB and 125 test trees for the TIGER250. Due to its limited size, we did not split the TUBA100.

[^161]:    ${ }^{6}$ OA-MODK (conjunct of modifier of accusative object), ON-MODK (conjunct of modifier of nominative object) and OADVPK (conjunct of modifier of ADVP object) occur only once in 27,125 sentences in TüBa-D/Z Release 3, OG-MOD (modifier of genitive object) 7 times, OADJP-MO (modifier of ADJP object) 8 times, OADVP-MO (modifier of ADVP object) 10 times, and FOPPK (facultative object of PP object) 17 times.

[^162]:    ${ }^{7}$ Unfortunately, Forst [2007] does not report results for precision and recall.

[^163]:    ${ }^{\dagger}$ Many thanks go to Rachel Nordlinger for pointing me towards the Ngan'gityemerri data and to my supervisor Miriam Butt for help with the analysis and making it financially possible for me to attend the conference.

[^164]:    ${ }^{1}$ I thank Rachel Nordlinger (p.c.) for bringing up this question.

[^165]:    ${ }^{2}$ These inflecting verbs have very complex paradigms, i.e. each inflecting verb has at least 33 different forms, some have as many as 44 different forms (Reid and McTaggart, 2008)

[^166]:    ${ }^{1}$ This paper was presented at the LFG09 conference and is based on my MA thesis. I wish to thank my supervisor, Prof. Yehuda Falk and participants in the LFG09 for comments.

[^167]:    ${ }^{2}$ I adopt Shlonsky's (1991) proposal to analyze the Q with its DP/NP as a functional projection QP, with Q being its functional head in the sense of Abney (1987), for the following reasons: (cf. also Spector 2008, Fassi Fehri 1988, Shlonsky 1991)

    - The Q in Hebrew can host clitics (or incorporated pronouns) (2b) and (7b). Only lexical and functional heads can do this in Hebrew.
    - $\quad$ Q selects its DP/NP, e.g. kol subcategorizes for a definite plural or collective NP.
    - Q selects for partitive PPs (some Qs allow it and some do not).
    - Q and its DP/NP or its incorporated pronoun form a constituent.
    ${ }^{3}$ I follow Falk (2006) in his analyzing the definite noun phrases in Hebrew as NPs.
    ${ }^{4}$ In the sense of Bresnan (2001).

[^168]:    ${ }^{5}$ This section was adopted from Spector and Moldovano (2007).
    ${ }^{6}$ It can be translated into English all, any, every, each, entire(ly) and whole.

[^169]:    ${ }^{7}$ The notation $1 . .$. indicates individual (after Rullmann 2003).

[^170]:    ${ }^{8}$ There are no purple cows in this world. We are not discussing possible worlds. If possible is added, sentence (13b) becomes grammatical:

    Itaxen še ha-parot ha-sgulot notnot kulan xalav
    possible that the-cows.FEM.PL the-purple.FEM.PL ${ }^{\text {give }}$ all $_{3 . \text { FEM.PL }}$ milk
    'It is possible that the purple cows all lactate'/ 'Possibly, the purple cows all lactate.'

[^171]:    ${ }^{9}$ Although Asudeh (p.c.) believes the genericity effect follows from coordination, citing Lions are all scary as not having a generic reading, one might still argue that the floated 'all' can quantify over kinds. In Lions are all scary, it has no choice but to quantify over individuals, because there is only one kind involved (Kagan, p.c.). However, further elaboration on this matter is out of the scope of this paper.

[^172]:    ${ }^{10}$ Since one is derived from the other (cf. Sportiche 1988)

[^173]:    ${ }^{11}$ Cf. Falk (2006a): "...Quantifier Float is not a uniform syntactic construction crosslinguistically".
    ${ }^{12}$ The present paper only examines FQs in subject position.

[^174]:    ${ }^{13}$ The formal mechanism of anaphoric binding and functional uncertainty is thoroughly discussed in Dalrymple (1993). For our purposes, a simple co-indexation of the topic's f-structure and subject's f-structure is sufficient.

[^175]:    ${ }^{1}$ I would like to thank Rachel Nordlinger, Ash Asudeh and Peter K. Austin for ideas and helpful comments on drafts of this paper. I would particularly like to thank Mary Dalrymple for the lengthy discussion of how to properly implement these ideas within LFG. Given the time constraints to write up this paper, I have not implemented all of these ideas here, but I will attempt to do so later! Finally, I would like to thank the editors for their very constructive comments, particularly on the formalisms used here.
    ${ }^{2}$ The following abbreviations are used in this paper: SBJN 'subjunctive mood', IND 'indicative mood', PS 'person', 3 'third person', 2 'second person', 3sg 'third person singular', NB 'number', sg 'singular', GD 'gender', f 'feminine', nom 'nominative', acc 'accusative', dat 'dative', PRON 'pronoun', refl 'reflexive', R 'reflexive', GF 'grammatical function', SUBJ 'subject', subj 'subject', SBJ 'subject antecedent required', OBJ 'object', obj 'object', ADJ 'adjunct', COMP 'complement, LOG 'logophoric(ity)', LOG 'logophoric', PRO 'non-logophoric pronoun', ANTE 'antecedent', DEF 'definite', (single capital letter) name used in example.

[^176]:    ${ }^{3}$ Note that this heirarchy does not appear to apply to Norwegian finite LDR (Strahan, 2003).

[^177]:    ${ }^{4}$ This is a simplification of the data, since not all Faroese speakers allow LDR out of an adjunct clause (Strahan, in press).
    ${ }^{5}$ This off-path constraint is intended to include any instance of a first or second person feature anywhere, be it in the subject, object, other GF, verb, or in a non-GF.

[^178]:    ${ }^{6}$ Rachel Nordlinger (p.c.) made the intriguing suggestion that, for the Icelandic data at least, the inside-out binding constraint could work if the path was stated as 'make the last thing you go out through a COMP', as shown in (i).
    (i) $\quad\left(\left(\mathrm{COMP} \mathrm{GF}^{+} \uparrow\right) \mathrm{SUBJ}_{\sigma}=\uparrow_{\sigma}\right.$

    In fact, I can see no direct problems with this rule, as it neatly sidesteps the problematic Icelandic data in (4) and (7). My only objections are that it does not

[^179]:    ${ }^{\dagger}$ I thank my advisor, Miriam Butt, for many valuable comments on different versions of this paper, and Ingo Mittendorf and Louise Mycock for comments during the LFG09 Conference. Also, I thank the anonymous reviewers of the paper abstract for their constructive criticism.

[^180]:    ${ }^{1}$ In Irish, there are two verbs for to be: bí (often called substantive verb in the literature) and

[^181]:    is (the copula). The copula verb is generally takes essential and inherent qualities as predicatives, hence the copula predicates are most commonly noun phrases, such as occupation, nationality, group membership and the like. The substantive verb bí, on the other hand, takes as predicatives less inherent qualities, such as temporal specifications, location, (temporal) possession etc. (Ramchand, 1997; Stenson, 1981). This difference in predication is called stage vs. individual level predication in Ramchand (1997).

[^182]:    ${ }^{2}$ Note that the agreement marker $e$ is dropped together with the copula, which is noted by both Stenson (1981) and Ó Siadhail (1989). Further research relating to the exact role of this marker, and why it is dropped here, might turn out to be interesting.

[^183]:    ${ }^{3}(22)$ is ambiguous between the syntactic reading in Figure 9 and another, admittedly improbable reading where leabhar 'book' is assigned the SUBJ function and múinteoir 'teacher' is assigned the OBJ function - both the object and the subject NP of the relative clause may appear in predicate position.
    ${ }^{4}$ The f-structure in Figure 9, produced by the XLE software, has been simplified for space reasons.

[^184]:    ${ }^{1}$ With $85 \%$ token coverage.

[^185]:    ${ }^{2}$ Left/right annotation matrices play a smaller role than for English because Arabic is a lot less configurational and has a richer morphology.

[^186]:    ${ }^{1}$ I am grateful to the workshop participants for agreeing to take part, in particular to Ruth Kempson for extensive discussion beforehand about the theme and organization of the workshop, to the Mont Follick fund of The University of Manchester for financial support, to the local organizer, Anna Kibort, for agreeing to include the workshop in the conference programme and for her heroic organizational efforts, to all those who participated in the general discussion which followed the presentation of the papers, and to the following who generously found time to comment on an earlier draft of this paper at ludicrously short notice: Ash Asudeh, Joan Bresnan, Miriam Butt, Mary Dalrymple, Tracy Holloway King and Ruth Kempson. Nonetheless, responsibility for the views expressed in this paper remains with the author.
    Correspondence address: nigel.vincent@manchester.ac.uk

[^187]:    ${ }^{2}$ Although various distinctive features of DS will emerge in what follows, here is not the place for even a brief overview of the system. For this the reader is referred to Kempson et al. (2001), and Cann et al. (2005). Cooper \& Kempson (2008) is a wide-ranging collection of papers that relates to many of the issues covered in this workshop, particularly in regard to the nature of linguistic data and the modes of theoretical explanation.

[^188]:    ${ }^{3}$ For a seminal paper in relation to psychology and reductionism, compare Fodor (1974), revisited in Fodor (1995). We may safely assume that linguistics, like psychology, is in Fodor's terms a 'special science'. See too the contributions to McCauley (1996) and Fodor's trenchant review of Paul Churchland's The engine of reason, the seat of the soul (Fodor 1998: 83-89).

[^189]:    ${ }^{4}$ Interestingly, Ivan Sag's work within the Sign-Based Construction Grammar development of HPSG is taking an increasingly functionalist turn (Sag forthcoming, Sag \& Wasow forthcoming), though like mainstream LFG and unlike DS, he maintains a clear distinction between grammar (competence) and parsing (performance).

[^190]:    ${ }^{5}$ My thanks to Mary Dalrymple for suggesting that I expand this section to include more coverage of E\&L and the controversy the paper has aroused. In fact, so many potential respondents had to be excluded for lack of space in BBS that the journal Lingua will host a further round of responses in one of its 2010 issues.
    ${ }^{6}$ Page references in this section are to the various authors' contributions in the issue of BBS 32 (2009). I have not however listed every response separately in the bibliography to the present paper.

[^191]:    ${ }^{7}$ Tracy Holloway King reminds me of the meetings entitled 'Grammar Engineering Across Frameworks' which she and Emily Bender initiated in 2007 and which have taken place annually since then. She reports that in that forum researchers have shown an openness towards ideas from different frameworks, perhaps because of the overriding need to solve the practical problems of grammar implementation.
    ${ }^{8}$ There are, or at least have been, non-Chomskyan but nonetheless derivational frameworks; Relational Grammar is a case in point. But for the purposes of the present discussion and in the current theoretical climate derivationalism and Minimalism can be equated.

[^192]:    ${ }^{1}$ Udi belongs to the Lezgic family of North Caucasian languages. It is traditionally spoken in the villages of Nic and Oghuz in Azerbaijan. In the 1920s, émigrés from Oghuz took Udi to the village of Oktomberi in Georgia. More recent migration has spread Udi to Russia (Clifton et al., 2002).
    ${ }^{2}$ Special PMs exist for inversion, possession, and question structures (Harris, 2002, §2.5.1).

[^193]:    ${ }^{3}$ Initial $/ \mathrm{n} /$ in a PM assimilates to a preceding liquid; see Harris 2002, §2.5.3.1 for discussion.

[^194]:    ${ }^{4}$ Here the PM zu '1 SG' undergoes /u/-elision; see Harris 2002, §2.5.3.1 for discussion.

[^195]:    ${ }^{5}$ In fact, the 'words' in W are abstract elements labeled with word-forms. The word-form labels can occur more than once, associated with distinct elements of W, as in $\langle$ the, dog, chased, the, cat $\rangle$.

[^196]:    ${ }^{6}$ I do not claim that all clitic phenomena are analyzable with lexical sharing. However, it appears to me that most of what Zwicky (1977) calls simple clitics may be amenable to such treatment.

[^197]:    ${ }^{7}$ For the 3 SG PM, I list the alternants $n e$ and $r e$ where phonologically appropriate.

[^198]:    ${ }^{8}$ This excludes imperative $a$ in the singular; this form does not take a PM (Harris, 2002, p. 31).

[^199]:    ${ }^{9}$ I assume without argument that the sentences examined so far feature an exocentric S ; however, nothing hinges on this analysis of the Udi clause.

[^200]:    ${ }^{1}$ Unless otherwise noted, example sentences in this paper are taken from Jake (1985), with some differences in spelling and glossing. Abbreviations in glosses include $1 / 2=1^{\text {st }} / 2^{\text {nd }}$ person, $\mathrm{ABL}=$ ablative, $\mathrm{ACC}=$ accusative, $\mathrm{AN}=$ animate, CISLOC $=$ cislocative, CREF $=$ not disjoint reference, DAT $=$ dative, DESID $=$ desiderative, FUT $=$ future, INAN $=$ inanimate, $\mathrm{INCH}=$ inchoative, $\operatorname{INSTR}=$ instrumental, $\mathrm{NEG}=$ negation, OBJ $=$ non-subject argument, $\mathrm{NOM}=$ nominalization, $\mathrm{PASS}=$ passive participle, $\mathrm{PERF}=$ perfective, $\mathrm{PL}=$ plural, $\mathrm{PROG}=$ progressive, $\mathrm{Q}=$ question, $\mathrm{SBJ}=$ subject, $\mathrm{SWR}=$ not coreference, $\mathrm{SG}=$ singular, $\mathrm{TOP}=$ topic, $\mathrm{VAL}=$ validator, $\mathrm{WH}=$ ' wh'.

[^201]:    ${ }^{2}$ Role labels such as GOAL, SOURCE, INSTRUMENT, and THEME are thought of in this paper as generalized semantic roles that are derived from lexical conceptual structure (lc-str) in some way. That is to say, SOURCE names a semantically complex role category that includes both origin of a change of place, as in (5), and cause, as in (1c) (like the category associated with from in English). GOAL includes at least recipient, addressee, and destination (like the category associated with to in English). INSTRUMENT includes at least instrument and accompanier (like the category associated with with in English). THEME includes patient (what/who something is done to), theme stricto sensu (what/who changes location), and object of perception or cognition.

[^202]:    ${ }^{3}$ The natural class OBJ and $\mathrm{OBJ}_{\theta}$ is identified as the class of acting 2s in Relational Grammar (final 2 or 2 chômeur), which is appealed to for cases of double-accusative clauses, which occur in various languages, including Latin (Perlmutter 1982), Korean (Lee 1991), and Yaqui (Guerrero \& Van Valin 2004). LFG can specify this natural class with the feature [ +o ], which only OBJ and $\mathrm{OBJ}_{\theta}$ have. This is an improvement over the Relational Grammar approach, which requires a disjunctive definition. In Alsina's (1996:19) alternative LFG system of GF features, the class in question would simply be OBJ ([-subj], [-obl]). How to account for the differences between primary and secondary objects, both of which are accusative in IQ, would be a challenge for this approach.

[^203]:    ${ }^{4}$ Actually, a noun stem can also host the desiderative suffix, such that yaku-naya, for example, can mean 'want water' (Jake 1985:204).

[^204]:    ${ }^{5}$ Varieties of Quechua vary considerably in how they handle transfer-of-possession and related constructions (Wunderlich \& Lakämper 2001, Willgohs 2009). The variety of IQ described in Cole (1982) is said to generally require dative case on goal arguments. The variety portrayed in Jake (1985), which provides the primary foundation for our summary here, has a dative/accusative alternation for the goal argument.

[^205]:    ${ }^{6}$ As noted by Jake (1985:281), there is dialectal variation concerning examples such as (11c). Some speakers, including those on which the analysis in Hermon (1985) is based, would have the causative suffix -chi (without a causative interpretation) following -naya.
    ${ }^{7}$ Cole (1982:112) claims that sentences like (13c) instantiate resultative aspect rather than passive. We follow Jake's (1985:219) interpretation of this kind of sentence as passive. It is unclear what the semantic difference between resultative aspect and passive might be in the case of a stative verb like this (i.e., 'I was in the state of want-

[^206]:    ${ }^{8}$ According to Hermon $(1985,2001)$, the CREF markers are restricted to what is characterized as a control environment, i.e., an infinitival embedded phrase/clause in

[^207]:    ${ }^{\dagger}$ Thanks to Danny Bobrow, Lucas Champollion, Cleo Condoravdi, Martin Forst, Lauri Karttunen, Tracy Holloway King and John Maxwell for discussion .

[^208]:    ${ }^{1}$ In fact, some further specialized obliques are used: OBL-AG, OBL-COMPAR, OBL-PART but they do not concern the type of obliques we are discussing here

[^209]:    ${ }^{2}$ It is clear that this cannot be achieved absolutely, or rather that there is no advantage in achieving it absolutely, as at the limit, in every pair of non literally matching texts, the two texts mean something different. We mainly try to achieve sameness of propositional meaning.
    ${ }^{3}$ At least, we would know this if these semantic/thematic categories to which grammatical functions map to were well-defined. This is far from being the case but we will ignore that problem here.

[^210]:    ${ }^{4}$ See (Crouch and King, 2005) for details.
    ${ }^{5}$ It cannot be the entailment approach alone because for certain classes when (Levin 1993) clearly states that there is no entailment relation, VerbNet proposes a frame, e.g. class 11.5 (drive verbs).

[^211]:    ${ }^{6}$ The mapping from post-verbal NPs to OBJs is in most cases rather straightforward. There are also problems with the specific thematic role assignments that VN proposes, but that is not the topic of this paper.

[^212]:    ${ }^{7}$ VerbNet's agent, theme, patient, etc. roles are not inherently verb-specific in the way that finergrained roles like worker, employer, employee, etc. would be.
    ${ }^{8}$ However, a lower level mechanism for recording role restrictions on individual verb-frame pairs may be needed to either override generalizations in specific cases, or to include additional role restrictions specific to the verb.

[^213]:    ${ }^{1}$ The work reported in this paper was supported by the Deutsche Forschungsgemeinschaft (DFG; German Research Foundation) in the Emmy Noether project PTOLEMAIOS, on Grammar Induction from Parallel Corpora.
    ${ }^{2}$ The current affiliation of the first author is Universität Stuttgart, IMS.

[^214]:    ${ }^{3}$ A prototypical application context for this type of resource would be (phrasal) statistical machine translation, where additional data-driven components, such as a statistical (n-gram) language model, impose additional constraints on the usability of the induced paraphrases.

[^215]:    ${ }^{4}$ We used the code kindly made available by the authors on http://www.cs.jhu.edu/ ~ccb/howto-extract-paraphrases.html

[^216]:    ${ }^{5}$ http://maltparser.org

[^217]:    ${ }^{6}$ light verb construction

[^218]:    ${ }^{7}$ The general deletion of these meaning components is a simplification in the case of more or less fixed idioms, e.g. kick the bucket where bucket has to be singular.

