FASSBL6

THE SIXTH INTERNATIONAL CONFERENCE
FORMAL APPROACHES TO SOUTH SLAVIC AND BALKAN LANGUAGES
25-28 SEPTEMBER 2008, DUBROVNIK, CROATIA

THE SIXTH INTERNATIONAL CONFERENCE
FORMAL APPROACHES TO SOUTH SLAVIC AND BALKAN LANGUAGES
25-28 SEPTEMBER 2008, DUBROVNIK, CROATIA

Organized by

Institute of Linguistics, Faculty of Humanities and Social Sciences,
University of Zagreb

Croatian Language Technologies Society

The Department of Computational Linguistics, Institute of Bulgarian Language
"Prof. Lyubomir Andreychin", Bulgarian Academy of Sciences

The Norwegian University of Science and Technology

Supported by

CEI
CENTRAL EUROPEAN INITIATIVE

ministarstvoznanostiobrazovanjaišporta

Ministry of Science, Education and Sports
of the Republic of Croatia

Bulgarian Academy of Sciences

Croatian Language Technologies Society

A CIP catalogue record for this book is available from the National and University Library in Zagreb under 678366

ISBN 978-953-55375-0-2

THE SIXTH INTERNATIONAL CONFERENCE
FORMAL APPROACHES TO SOUTH SLAVIC AND BALKAN LANGUAGES
25-28 SEPTEMBER 2008, DUBROVNIK, CROATIA

PROCEEDINGS

Edited by
Marko Tadić, Mila Dimitrova-Vulchanova, Svetla Koeva

Croatian Language Technologies Society – Faculty of Humanities and Social Sciences
Zagreb, 2008

Organizing committee

Damir Boras (University of Zagreb)
Svetla Koeva (The Bulgarian Academy of Sciences)
Marko Tadić – chair (University of Zagreb / Croatian Language Technologies Society)
Mila Vulchanova (The Norwegian University of Science and Technology)
Valentin Vulchanov (The Norwegian University of Science and Technology)

Programme committee

Damir Boras (University of Zagreb)
Anna Cardinaletti (Cá Foscari University, Venice)
Dan Cristea (University of Iaşi)
Damir Ćavar (University of Zadar)
Tomaž Erjavec (Institute Jozef Stefan, Ljubljana)
Giuliana Giusti (Cá Foscari University, Venice)
Svetla Koeva (The Bulgarian Academy of Sciences)
Iliana Krapova (Venice University)
Milan Mihaljević (Old Church Slavonic Institute, Zagreb)
Kemal Oflazer (Sabanci University, Turkey)
Stelios Piperidis (Institute for Language and Speech Processing, Athens)
Vassil Raynov (The Bulgarian Academy of Sciences)
Kiril Ribarov (Charles University, Prague)
Melita Stavrou (Aristoteles University, Thessaloniki)
Marko Tadić (University of Zagreb / Croatian Language Technologies Society)
Dan Tufiş (The Romanian Academy)
Duško Vitas (University of Belgrade)
Mila Vulchanova – chair (The Norwegian University of Science and Technology)
Valentin Vulchanov (The Norwegian University of Science and Technology)
Chris Wilder (The Norwegian University of Science and Technology)

TABLE OF CONTENTS

INVITED SPEAKERS

ORAL PRESENTATIONS

SUBJUNCTIVE COMPLEMENTS TO VERBS IN ROMANCE AND SLAVIC BALKAN

Virginia Hill*, Olga Mišeska Tomić**

*Linguistics Program, University of New Brunswick – Saint John
Tucker Park, Saint John NB Canada
mota@unbsj.ca
**University of Novi Sad, Serbia
olgamito@eunet.yu

ABSTRACT

This paper sheds new light on the left periphery of subjunctive clauses in Balkan languages by comparing the sentential complements to verbs in two groups: Romance versus Slavic Balkan. The tests indicate a systematic contrast in the organization of the complementizer field, with consequences for deeper cross-linguistic variation. The micro-parameter discussed also attests the impact of Slavic Balkan on the Romance Balkan group, especially on Aromanian..

In this paper we analyze the left periphery of subjunctive clauses in complement position of verbs in Romance and Slavic Balkan. Within the framework of the cartographic representation of the CP field (Rizzi 1997) we find a micro-parametric variation in the organization of these constructions: Romance Balkan displays an articulated CP field, whereas Slavic Balkan displays a collapsed CP field. This cross-linguistic variation has consequences for further systematic variation between the two language groups.

1. Theoretical framework

Rizzi (1997) argued that the left periphery of the clause implements two functional tasks:

(i) It assigns grammatical typing to the clause (as declarative or interrogative) and inflectional typing (i.e. determining the licensing of a certain grammatical mood, tense and phi-feature cluster). While the grammatical typing is represented as Force, and mediates the relation between the clause and discourse factors, the inflectional typing is represented as Fin(iteness) and mediates the relation between inflectional morphology and information structure.

(ii) It contains the encoding of information structure (or "discourse pragmatics" in Lambrecht 1994 a.o.) into syntax. The features of information structure are represented in two main nodes - Topic and Focus, the latter of which dominates the operators of contrastive focus and wh interrogatives and relatives.

Thus, as shown in (1), the one-node representation of the Complementizer Phrase in Chomsky (1986) has been split and articulated over several functional heads.

(1) [ForceP Force [TopP Topic [FocusP Focus [FinP Fin [IP I]]]]]

For the I(nflectional) P(hrase) segment embedded under Fin in (1) we assume a possible articulation over several functional heads with relevance for verbal morphology, such as in (2).

(2) [MoodP Mood [TP [NegP Neg [AspP Asp [vP v]]]]

The basic hierarchy in (2) has been used, with slight variation, by a remarkable number of linguists who looked at the verb movement in various Balkan languages (e.g. Mišeska Tomić 2007, Motapanyane 1991, Rivero 1994 a.o.). It takes care of the place of the subjunctive mood marker, which is ubiquitous in the Balkan area, and has an inflectional as well as clause typing function (i.e., it is compatible with both declarative and interrogative clauses).

2. Tests

The assessment criteria for the organization of the left periphery of subjunctive clauses are the following:

(a) Compound tenses with 'have' and 'be' auxiliaries;

(b) The presence/absence and location of a lexical complementizer;

(c) Wh-movement to the left periphery of the subjunctive complement.

Compound future tenses formed from 'have' or 'be' auxiliaries plus a sujunctive clause are a Balkan Sprachbund property. The configuration of these tenses provides a reliable indication of the level at which the mood marker + verb strings move in the clause hierarchy (e.g., whether they stay within IP or move to the CP field).

With respect to the lexical complementizer, where present, we have to establish its merging location, which could be Fin or Force. In either location, the complementizer must check both sets of typing features, either by overt movement or by distance 'Agree'. In configurations where the lexical complementizer does not appear, we have to determine whether a non-lexical equivalent is available or not.

Finally, wh-movement engages Force (for its +qu feature) and Focus (for its +wh feature), but not Fin. Accordingly, wh-movement must compete with the complementizers that check Force.

3. Data

We apply these tests on constructions where the verb selects a subjunctive clause in complement position. Compound tenses show that, in all the languages concerned, the string formed by the subjunctive marker and the verb moves to the highest level of IP, but does not go further to the CP field. Tests on the organization of the CP field indicate that sentence typing and inflectional typing are checked through complementizers in Romance Balkan, whereas Slavic Balkan resort to verb morphology and paradigmatic oppositions for the same purpose.

3.1. Compound tenses

This section points out that 'have' and 'be' auxiliaries take a subjunctive clause as their complement. Since the highest hierarchical position that auxiliary verbs may occupy is Fin, it follows that their complements are MoodPs or hierarchically lower functional projections. Compound tenses with subjunctive clauses are a Balkan Sprachbund property, indicating that subjunctive clauses must count as MoodP structures in both Romance and Slavic Balkan.

3.1.1. Romance Balkans

Romanian has positive and negative future tenses, as in (3a, b), where the auxiliary 'have' is followed by a subjunctive construction; Megleno Romanian displays the same formation only in the negative, as in (3c)[1]. Both constructions rule out the lexical complementizer.

(3)	a.	*Am/ai/are*	*(*ca)*	*să*	*plec/pleci/plece.*	Romanian
		have.1/2/3Sg	Subj.Comp	Subj.Mark	go.1/2/3Sg.Subj	
		'I/you/(s)he am/are/is going to leave.'				
	b.	*N-am*	*(*ca)*	*să*	*plec/plece.*	Romanian
		not-have.1Sg	Subj.Comp	Subj.Mark	go.1/3Sg.Subj	
		'I/(s)he will not go.'				
	c.	*Nu ari*	*(*ca)*	*si*	*vin.*	Megleno-Romanian
		not have.Impers	Subj.Comp	Subj.Mark	come.1Sg	
		'I won't/am not going to come.'				

3.1.2. Slavic Balkan

The same type of compound tenses occurs in Slavic Balkan, though in Macedonian and Bulgarian the 'have' auxiliary appears only in the negative form (4a-b), while in Serbian future tense forms with 'have' are lacking, but the future tense forms with the 'be' auxiliary display the same subjunctive complementation (4c).

[1] The grammaticalization of the Aromanian 'have' did not reach the level of an auxiliary. As shown in Mišeska Tomić (2006: 566) Aromanian 'have' is strongly modal even when it serves for the formation of a future tensed sequence. Hence, it is not a structural equivalent to the Romanian and Megleno-Romanian auxiliary.

(4) a. *Nema* *da* *dojdam.* Macedonian
not+have.Impers Subj.Mark come.1Sg.Perf.Pres
'I won't/am not going to come.'

b. *Njama* *da* *dojda.* Bulgarian
not+have.Impers Subj.Mark come.1Sg.Perf.Pres
'I won't/am not going to come.'

c. *Petar će* *da* *ti* *ga* *da* Serbian
Peter will.3Sg.Cl Subj.Mark 2Sg.Dat.Cl 3Sg.Neut.Acc.Cl give.3Sg.Perf.Pres
'Peter will give it to you.'

3.1.3. Common properties

While variations exist in the distribution of features within these compounds, the general conditions on derivation are stable cross-linguistically. In particular, this is a monoclausal structure, because:

(i) As seen in (3b-c) and (4a-b), the negation precedes 'have' and not the verb to the right of the subjunctive marker, which is generally a possibility when the subjunctive clause is a complement of a full (not auxiliary) verb, as in (5a-d):[2]

(5) a. *Am* *decis* *să* *nu* *plec.* Romanian
have.1sg decided Subj.Mark not go.1Sg.Subj
'I decided not to go.'

b. *N-am* *decis* *să* *nu* *plec.* Romanian
not+have.1sg decided Subj.Mark not go.1Sg.Subj
'I haven't decided not to go.'

c. *Odlučiv* *da* *ne* *odam.* Macedonian
decide.1Sg.Past Subj.Mark not go.1Sg
'I won't/am not going to come.'

d. *Ne odlučiv* *da* *ne* *odam.* Macedonian
not have decide.1Sg.Past Subj.Mark not go.1Sg
'I didn't decide not to go.'

(ii) The phi-features and tense features are shared between 'have' and the verb to the right of the subjunctive marker. The exact distribution of these features differs cross-linguistically. Thus, in Romanian, [tense] and [person] features are clustered on 'have', whereas [number] features occur on the verb, as in (6a), whereas, as shown in (6b), in Macedonian the [tense] feature can be on either 'have' or the verb, but not on both.

(6) a. **Avem/aveți/au* *să* *plecăm/plecați/plece.* Romanian
have.1/ 2/ Pl Subj.Mark go.1/2/3Pl.Subj
'We/you/they are going to leave.'

b1. *Nemaše* *da* *dojdam/dojdeš/dojde.* Macedonian
not+have.Impers.Past Subj.Mark come.1/2/3Sg.Perf.Pres
'I wasn't going to come.'

[2] Both the 'have' auxiliary and the full verb can be negated if the 'have' auxiliary functions not as a future marker, but as a verb denoting strong determination or prohibition:

(i) *Nema* *da* *ne* *odu* Macedonian
not+have.Impers Subj.Mark not go.3Sg
'He shouldn't refrain from going.'

b2.	*Nema* not+have.Impers 'I wasn't going to come.'	*da* Subj.Mark	*dojdev/dojdeše.* come.1/2/3Sg.Perf.Past	Macedonian
b3.	**Nemaše* not+have.Impers	*da* Subj.Mark	*dojdev/dojdeše.* come.1/2/3Sg.Perf.Past	Macedonian

Unlike the clauses in (6), bi-clausal structures allow complete phi-features and tense sets on both verbs. While a detailed analysis of the compound future is beyond the scope of this paper, the illustrated properties are sufficient to indicate that the verb to the right of the subjunctive marker stays within the MoodP, so it is available for selection by functional verbs. Since 'have' is not only higher but also partially inflected for subject agreement and tense, it follows that this auxiliary is merged either very high in IP or has direct contact with the IP. In the hierarchy in (1), 'have' fulfills the requirements on the Fin head, and it is presumably merged directly in this head.

Since the properties illustrated in this section apply (with small variations) to compounds with 'have'/'be' auxiliary + subjunctive complements in all Balkan languages relevant to this paper, we can generalize the conclusion on the level of verb movement in the subjunctive clause; namely, the verb string within the subjunctive clause (constituted from the verb and clausal clitics including the mood marker) moves no further than MoodP.

3.2. Subjunctive complementizers

In constructions without obligatory control the Romance Balkan group presents complementizers that precede the subjunctive mood markers, to which we refer as "subjunctive complementizers". The Slavic Balkan group is unitary in lacking such complementizers.

While the Megleno-Romanian subjunctive complementizers are formally analogous to the indicative complementizers, Aromanian and Standard Romanian have distinct subjunctive complementizers. Table 1 presents the Romance Balkan indicative and subjunctive 'that' complementizers.

	Indicative	Subjunctive
Aromanian	*ca*	*tă*
Megleno-Romanian	*ca*	*ca*
Standard Romanian	*că*	*ca*

Table 1: 'That' complementizers in Romance Balkan

In terms of the framework in (1), the complementizer checks two sets of typing features: it determines the value for the typing features of Force (i.e., declarative versus interrogative), and the value for Fin features (i.e., indicative versus subjunctive selection). Thus, the forms in Table 1 indicate that one lexical item must check two functional heads with typing features.

The position of these complementizers raises a challenge since the word order is different in the three languages, even when the complementizer is the same (i.e. *ca*). In particular, no lexical material can precede *ca* in Romanian, whereas Megleno-Romanian *ca* and Aromanian *tă* allow for word order variation. This variation is discussed below.

3.2.1. Word order of the subjunctive complementizers in the CP

Taking into consideration the hierarchy in (1), a constituent fronted to the Top position must indicate the position of the co-occurring complementizer: if the complementizer surfaces on the right, it is in Force; if it surfaces on the left, it is in Fin. The word order in (7) indicates that Romanian *ca* is located in Force, whereas Megleno-Romanian *ca* and Aromanian *tă* are located in Fin:

(7)	a.	*Narăncio (ca)* order.3Sg.Aor that	*Maria* Maria	*(ca)* that.Subj	*si* Subj.Mark	*vină* come.3Sg.Subj	*ună* one	*shi ună* and one	Megleno-Romanian
	b.	*Deade naredba* give.3Sg.Past order	*(Maria)* Maria	*tă* that.Subj	*(Maria)* Maria	*s-* Subj.Mark-	*yină* come.3Sg.Subj	*tunoară* immediately	Aromanian

c.	*A cerut*	*(*Maria)*	*ca*	*Maria*	*să*	*vină*	*imediat*	Romanian
	has asked	Maria	that	Maria	Subj.Mark	come.3Sg.Subj	immediately	
	'(S)he asked/ordered for Maria to come immediately.'							

One may wonder if the configuration in (7c) really attests that the Romanian subjunctive complement is in Force or else it is in Fin, as in the other two languages, but, for some reason, TopP is excluded and the projection of the CP field collapses Force with Fin. To eliminate the possibility of wonderment, in (8) we test the complementizer on constructions that contain preverbal constituents with topic and focus readings.

(8)	a.	*A cerut ca*	*la lucru acum să- l*		*ajute*	*cineva,*	*nu mâine.*	Romanian
		has asked that.Subj	at work now Subj.Mark-3Sg.M.Acc.Cl		help.3Sg.Subj	someone	not tomorrow.	
		'He asked that someone help him now at work, not tomorrow.'						
	b.	*A cerut (*?ca)*	*să-l*		*ajute*	*cineva.*		Romanian
		has asked that.Subj	Subj.Mark-3Sg.M.Acc.Cl		help.3Sg.Subj	someone.		
		'(S)he asked that someone help him.'						

In (8a) constituents with Topic and contrastive Focus readings intervene between *ca* and the mood marker; according to the hierarchy in (1), these functional projections occur lower than Force, but above Fin. In (8b) the adjacency between *ca* and the mood marker is judged ungrammatical in Romanian; obligatory lexical material between these two elements signals the need to maintain unambiguous parsing of *ca* as Force, instead of lowering or collapsing it with the IP. The analysis of subjunctive complementizers as merged in Force or Fin suggests that the left periphery of these clauses have a configuration articulated as in (1). Each complementizer checks the two sets of typing features: one set is checked directly through merge, whereas the second set is checked through distance 'Agree'.

3.2.2. Indirect interrogatives

So far, the tests indicate that in Romance Balkan we have a complete Force-Fin field in the left periphery. However, the word order in the indirect interrogatives points out a micro-variation within this language group. As shown in (9), wh-constituents rule out the complementizer *ca* (9a, b) but may co-occur with *tă* (9c).

(9)	a.	*Nu ştiu*	*(*ca)*	*cui*	*(*ca)*	*să-i*	*trimit*	*scrisorile.*		Romanian
		not know.1sg	Subj.Comp	whom-DAT	Subj.Comp	Subj.Mark-3Sg.Dat.Cl	send.1Sg	letters-the.Pl		
		'I don't know who to send the letters to.'								
	b.	*Nu ştiu*	*(*ca)*	*la cari*	*(*ca)*	*s-iu*	*trimet*	*prămătia.*		Megleno-Romanian
		not know.1sg	Subj.Comp	to whom	Subj.Comp	Subj.Mark-3Sg.Dat.Cl	send.1Sg	merchandise		
		'I don't know who to send the merchandise to.'								
	c.	*Nu shciu*	*(*tă)*	*a cui*	*(tă)*	*s-ălj*	*lji*	*pitrec*	*aiste cărtsă.*	Aromanian
		not know.1sg	Subj.Comp	to whom	Subj.Comp	Subj.Mark-3Sg.M.Dat	3Pl.Acc.Cl	send.1Sg	these letters	
		'I don't know who to send these letters to.'								

In constructions as in (9) the complementizer and the wh-constituent compete for the checking of Force, so they exclude each other. Under these circumstances, the presence of *tă* in (9c) indicates that this complementizer does not check Force, its function being limited to checking the Fin features. Accordingly, in declarative sentences as in (7b), there is no ForceP level, since there is no evidence on how the features of Force would be checked. Hence, the conclusion that Aromanian subjunctive clauses project only to FinP.

3.3. Configurations without complementizers

As mentioned above, the complementizers appear on an optional basis, in alternation with constructions headed by the subjunctive marker. In these configurations, the variation of word order mentioned in 3.2.1 is maintained, as further shown in (10).

(10) a.	*Narăncio*	*(Maria)*	*si*	*vină*	*ună*	*shiună*		Megleno-Romanian
	order.Sg.Past	Maria	Subj.Mark	come.3Sg.Subj	one	and one		
b.	*Deade naredba*	*(Maria)*	*s-*	*yină*	*tunoară*			Aromanian
	give.3Sg.Past order	Maria	Subj.Mark-	come.3Sg.Subj	immediately			
c.	*A cerut*	*(*Maria)*	*să*	*vină*	*(Maria)*	*imediat*		Romanian
	has asked	Maria	Subj.Mark	come.3Sg.Subj	Maria	immediately		
	'S/he asked/ordered for Maria to come immediately.'							

Aromanian and Megleno-Romanian allow for the projection of TopP above the mood marker, whereas Standard Romanian does not, in formal register. This word order indicates that the absence of a complementizer makes no difference to the pattern of derivation in Megleno-Romanian, where the construction is still Force-FinP at the left periphery, with Top and Focus in-between. There is not much difference for the configuration in Aromanian either, where the projection of TopP and FocusP may be triggered by the information structure, even if a ForceP is absent. Note that for these configurations there is no telling test on whether the absence of the complementizer in Fin triggers movement of the mood marker+verb string to Fin; giving the adjacency of Fin and MoodP, the word order effects are the same. On the other hand, the absence of the complementizer in standard Romanian forces a reversal of word order, where the mood marker + verb string must be clause initial; if constituents with Topic or Focus readings occur, they must follow the verb. Hence, the absence of the complementizer in standard Romanian clearly triggers the movement of the mood marker+verb string to Force, to replace it as an overt checker of the sentence typing features.

3.4. Slavic Balkan

The Slavic Balkan language group display a systematic lack of subjunctive complementizers. The left periphery in constructions with subjunctive clauses may show fronting of Topic or Focus constituents, as in (11), but provides no evidence for functional heads specialized in typing features.

(11) a.	*Bi*	*sakala*	*na planina*	*so MARIJA*	*da*	*odam,*	*ne so Jovana.*		Macedonian
	would	like.F.Sg.*l*-Part	to mountain	with Marija	Subj.Mark	go.1Sg	not with Jovan.Acc		
b.	*Iskala*	*bix*	*na planinata*	*săs MARIJA*	*da*	*otida,*	*ne săs Ivan.*		Bulgarian
	like.F.Sg.*l*-Part	would.1Sg	to mountain+the	with Marija	Subj.Mark	go.3Sg.Perf.Pres	not with Ivan		
c.	*Htela*	*bih*	*u planinu*	*sa MARIJOM*	*da*	*idem,*	*ne sa Jovanom.*		Serbian
	like.F.Sg.l-Part	would.1Sg	in mountain	with Marija.Instr	Subj.Mark	go.3Sg	not with Jovan.Instr		
	'I would like to the mountains with Mary to go, not with John.'								

Functional heads specialized in discourse pragmatics (e.g., Topic and Focus) are known to be able to associate with any functional field (i.e., CP or IP), and to occur at various levels in such fields, either fronted or at the bottom of the field (Belletti 2008, Kiss 1995). In fact, there is evidence that Topic and Focus features used to be clustered with inflectional features in Early Modern Bulgarian, in constructions with the particle *ta* (Mladenova 2008). From this point of view, in the absence of any kind of marking for typing features, the word order in (11) cannot guarantee the existence of an articulated CP field with the composition in (1). The value of the typing features in (11) are established through the opposition of this lack of marking with the wh-marking in interrogative/relative clauses and with the obligatory lexical complementizer for declarative complements with indicative verbs. Thus, the typing feature system in Slavic Balkan presents the distinctions in Table 2.

	Force [-qu]	Force [+qu]
Fin – indicative complementizer (articulated CP)	Fin – subjunctive no complementizer (collapsed CP/IP)	Fin – free Wh-word (articulated CP)

Table 2: Values for typing features in Slavic Balkan

According to Table 2, the subjunctive complements in Slavic Balkan are not the structural equivalent of Romance Balkan subjunctive complements without a subjunctive complementizer. The Romance Balkan subjunctive clauses have articulated CP fields irrespective of the nature of their complementizers, whereas the Slavic Balkan subjunctive clauses lack completely the functional heads in which complementizers could merge. The collapsed CP/IP system in Balkan Slavic means that the typing information needed to close off the phase (in Chomsky's 2001 terms) is inferred, partly from the inflectional heads and partly from the paradigmatic values in Table 2.

4. Typology

The presented analysis of subjunctive complementation in Balkan Romance and Balkan Slavic leads to the typological outline in Table 3.

+Force/+Fin	-Force/-Fin
Standard Romanian	Bulgarian
Megleno Romanian	Macedonian
	Serbian
-Force/+Fin	
Aromanian	

Table 3: Typology of left peripheries in subjunctive complementation

Although all the languages concerned in Table 3 display a similar verbal morphology in the subjunctive, and a similar distribution of subjunctive clauses under V selection, the way in which these subjunctives are embedded differs in a systematic way: the embedding is conditioned by an explicit typing process in Romance Balkan, whereas in Slavic Balkan it is implicit. This parametric difference may be due to historical factors, since Romance Balkan inherited the complementizers from Latin, where CP fields are generally well articulated. In this respect, language contact may be seen in the non-lexical option for subjunctive complementizers in Romance Balkan in general, and in the actual reduction of the CP field in Aromanian in particular. Aromanian has been in intensive contact with Slavic languages for a long period of time and is, therefore, prone to alteration of inherited patterns and weakening of parametric variation.

5. Conclusions

The left periphery of subjunctive clauses has been the topic of much research in formal approaches to Balkan grammars. This paper offers a comparative analysis of a less frequently studied Romance group of languages and its relation to the languages of a group with which it has close geographical contact. The analysis is instructive in two respects:

(i) Within the formal framework it discusses a type of micro-variation that can snowball to systematic cross-linguistic variation. In particular, the CP typology proposed here may also explain cross-linguistic variation in subjunctive complementation to nouns. For example, deverbal nouns select subjunctives in Slavic Balkan, but practically only infinitives in Romanian (e.g., in 'My desire to travel is strong' the verb 'travel' would take a subjunctive form in Slavic Balkan, but an infinitive form in Romanian). The typology in Table 3 provides a promising basis for the analysis of this contrast, because the CP interferes between the noun and the embedded subject position it controls.

(ii) Outside the formal framework, this study provides an example where the grammar negotiates between inheritance and language contact. More precisely, Romance Balkan inherited, from Latin, a well-articulated CP field, which came in contact with a collapsed left periphery in Slavic Balkan. Under language contact, the following phenomena occurred:

(a) weakening of the Force complementizer, by making it optional (in the entire group);

(b) reanalysis of the complementizer as a Fin element (in Megleno and in regional varieties of Romanian);

(c) complete suppression of ForceP (in Aromanian).

The impact of language contact is, thus, attesting the influence of Slavic on Romance Balkan. There is no evidence for a mutual impact in this particular parameter.

References

Chomsky, Noam 1986. *Barriers*. Cambridge, Mass: MIT Press.

Chomsky, Noam 2001. “Derivation by phase”. In M. Kenstowicz ed., Ken Hale: *A Life in Language*. Cambridge, MIT Press, 1052.

Lambrecht, K. 1994. *Information structure and sentence form*. Cambridge University Press.

Tomić, Olga Mišeska 2006. *Balkan Sprachbund Morpho-syntactic Features*, Dordrecht: Springer.

Tomić, Olga Mišeska 2007. “Mood, Negation and Pronominal Clitics: Evidence from the Balkan Languages”, *Balkanistica* 20, 111-145.

Motapanyane, Virginia 1991. *Theoretical Implications of Complementation in Romanian*. Ph.D. thesis. University of Geneva.

Rivero, María Luisa. 1994. “Clause structure and V-movement in the languages of the Balkans”. *Natural Language and Linguistic Theory* 12. 63-120.

Rizzi, Luigi 1997. “The fine structure of the left periphery”. In L. Haegeman ed., *Elements of Grammar*. Dordrecht: Kluwer, 281-337.

THE SKETCH ENGINE AS A COMMON PLATFORM FOR SHOWCASING LANGUAGE RESOURCES

Adam Kilgarriff

Lexical Computing Ltd., Brighton, UK
adam@lexmasterclass.com

ABSTRACT

The case for building good languages resources is always that they will have many users and uses. But people other than the developers will be reluctant to use them if they cannot explore them before committing. One way to enable potential users to explore a resource is to load it into a good web tool. A suitable tool for corpora (and, indirectly, for lemmatisers, POS-taggers and some parsers) is the Sketch Engine. While this is a commercial tool and web service, this can be advantageous: it means that the costs and maintenance of the service are taken care of. Both parties stand to gain: the resource developers both have their resource showcased for no cost, and get to use the resource within the Sketch Engine themselves (often also at no cost). The Sketch Engine company stands to gain as additional customers may pay for accounts (after an initial free trial period) to use the resource. The Sketch Engine already plays this role in relation to a number of resources, and case studies are presented.

1. The Problem

If you have a language resource, how do you show it off? The usual answer is, first, give talks about it, and second, send, a sample. Talks are appropriate and useful but, for someone contemplating using the resource, or even making a comparison between it and others, they only start to tell the story.

Sending a sample is often not satisfactory. I have received samples on many occasions: it is hard work to assess the quality of a resource, or to gain a sense of whether it might be useful for a project. One starts with battling through layers of XML, including close study of the DTD to try to work out what the annotation means. Eventually one finds the key elements and their relations and tries to make an informed comparison with some other resource that one knows. The next stage is to assess, for a number of words, how the information compares: this is the real work, but the summary statistics one would like are often not available, and struggling with unfamiliar annotation at every stage is slow and painful.

Showcasing resources is central to a language resource programme. A premise of language resource development is that resources, once developed, will be used by a number of groups. If they cannot easily be assessed, they will not be.

2. The Sketch Engine

The Sketch Engine is a corpus query tool. It has been widely used for lexicography, by clients including Oxford University Press, Collins, Macmillan and FrameNet, and for linguistic and language technology teaching and research at universities. Corpora for many languages have been installed. It is fast, responding immediately for most queries for billion-word corpora, and offers all standard functions (concordancing, sorting and sampling of concordances, wordlists and collocates according to a range of parameters, full regular-expression searching, subcorpus definition and handling) and some non-standard ones (in particular word sketches - one-page summaries of a word's grammatical and collocational behaviour, see Fig 1 - and also a distributional thesaurus, and keyword lists which compare words in different subcorpora).

The basic input is a corpus, preferably lemmatised and part-of-speech tagged. For the word sketches and thesaurus, either the corpus must already be parsed, or another input is required: this is a shallow grammar, written as regular expressions over words and POS-tags, in which each grammatical relation to appear in the word sketch is defined. For a computational linguist with a knowledge of the language in question, preparing a basic grammar is not a large task.

2.1 The Sketch Engine server

Lexical Computing Ltd., the owner of the Sketch Engine, provides a web service which gives easy access to corpora for (currently) ten languages: see Fig 2. Users can start using the corpus for their question directly: the user interface is simple and there is no software to install.

A Sketch Engine account also gives access to two other services: WebBootCaT (for building instant corpora from the web: see Baroni et al 2006) and CorpusBuilder. CorpusBuilder allows users to take a corpus that they have on their machine, upload it onto the Sketch Engine server, install it in the Sketch Engine, and then use the Sketch Engine to do research on it. Thirty-day free-trial accounts are available for the Sketch Engine accounts; after that users pay an annual fee (unless they are collaborators, see below).
This is the company's main income stream.

resource

British National Corpus freq = 12658

object_of	3212	2.2
allocate	192	50.55
pool	39	39.98
exploit	64	35.3
divert	38	31.35
use	311	30.21
deploy	31	29.59
devote	43	29.45
concentrate	62	29.43
reallocate	12	29.43
provide	174	27.14
utilise	22	26.28
conserve	17	25.24

subject_of	467	0.6
devote	27	32.97
remain	12	14.52
come	16	11.49
make	24	10.42
go	15	9.77

adj_subject_of	475	3.3
available	258	55.98
scarce	13	29.63
necessary	23	23.15
likely	12	14.75

modifier	6475	1.5
scarce	163	56.64
natural	321	44.64
limited	187	42.56
non-renewable	25	40.01
financial	249	38.48
mineral	89	35.8
renewable	33	34.8
additional	107	32.91
valuable	74	32.58
human	134	29.66
extra	88	29.53
meagre	21	28.03

modifies	1906	0.5
allocation	135	48.84
management	153	35.67
centre	158	32.83
committee	132	32.25
implication	46	28.48
column	20	19.91
pack	17	19.22
base	25	18.28
constraint	14	18.07
development	46	17.41
planning	19	16.08
owner	14	13.3

Fig 1. A word sketch for the English noun *resource* (reduced to fit; data taken from British National Corpus)

2.2 What corpora, and how did they get there?

Corpus developers usually want to make it easy for people to look at and explore their corpora. This fits well with Lexical Computing's business plan, which is to provide a service which gives access to a wide range of resources.
The corpora available through the Sketch Engine are mostly provided for free by the people who have developed them, in exchange for free access for them and their colleagues to the Sketch Engine server. This is a win-win scenario. The resource developer benefits in three ways:

- access to their own corpus in the Sketch Engine, which supports them in their own research on it (including maintaining and developing it)
- an easy way to show their corpus to others, in a way that allows those others to explore it in detail
- access to
  - other corpora already in the Sketch Engine
  - WebBootCaT (for building web corpora; see above).

Lexical Computing benefits because it extends the range of resources that it can offer to customers. No money need change hands in either direction. This basic model needs adapting to different circumstances (and is only applicable on a no-fee basis where Lexical Computing judges the corpus to be an interesting addition to its portfolio, and to offer the prospect of new customers commensurate with income foregone). Below we present case studies of some of the corpora and the collaborations behind them. But first, we explain the Sketch Engine's allegiance to common standards, why it is no bad thing for this role to be played by a commercial company rather than a university, 'local' *vs* 'remote' hosting of corpora, and how the model is relevant for corpus processing tools as well as the corpora themselves.

2.3 Input format and query formalisms

The Sketch Engine uses both input format and query formalism developed at the University of Stuttgart for their corpus system in the early 1990s. Since the Stuttgart system was launched in 1994, many corpus and computational linguists have used it and others have, like us, developed other systems using its input format and query formalism. In adopting these two formalisms, the Sketch Engine makes it straightforward for corpora to be prepared, installed and queried, for a large part of the community.

More recently the XML Corpus Encoding Standards have been proposed. XCES-encoded corpora can also readily be installed in the Sketch Engine.

LEX COM Lexical Computing

SKETCH ENGINE

user: Adam Kilgarriff

Help Change passwd WebBootCaT CorpusBuilder Bug reporting News Logout

Corpora

Language	Name	Tokens [?]	
Chinese	Chinese GW, simpl	706 427 624	info
Chinese	Chinese GW, trd	706 428 333	info
English	British Academic Spoken English Corpus (BASE)	1 252 256	info
English	British Academic Written English Corpus (BAWE)	7 474 757	info
English	British National Corpus	111 244 375	info
English	ukWaC	2 035 621 120	info
French	French web corpus	126 850 281	info
German	deWaC	1 644 785 836	info
Greek	GkWaC	149 067 023	info
Italian	itWaC	1 909 535 984	info
Japanese	JpWaC	409 384 405	info
Persian	WBC-Per	6 375 735	info
Portuguese	Cetenfolha, Cetempublico	66 319 147	info
Russian	Russian Web Corpus	187 965 822	info
Slovenian	Fida PLUS 620m	738 503 185	info
Spanish	Spanish web corpus	116 900 060	info

Fig 2: User's home page for Sketch Engine showing available corpora and tools

2.4 Maintenance and motivation

The maintenance of resources has often been a bone of contention for those left in charge of them. Resource developers become the victims of their own success: the more successful the resource, the greater the level of expectation that errors

will be corrected and upgrades provided, yet research funding bodies are rarely willing to fund them, since the projects have already had their funding and maintenance is not the funders' mission. So the host organisation struggles to meet users' requests for little credit or recompense. Nor does resource maintenance provide many opportunities to publish.
Lexical Computing depends for its income on the quality of its resources, so is motivated to maintain and upgrade the hardware, software and corpora. There is an income stream to fund it, from customers. For resource management and maintenance, there is much to be said for a market model, in which the people who are maintaining a resource are motivated to do it well because their income depends on it.

2.5 The 'local *vs.* remote' issue

One of the biggest questions about software, in the age of the web, is: should it be local or remote? Should we download and install, or interact through browsers and APIs? For a growing number of applications, 'remote' is gaining ground. More and more people manage their documents and photos, and read their email, on remote servers. When I want to convert a document from .ps to .pdf, I do it at http://ps2pdf.com. Corpus research is an area where 'remote' is a very appealing answer, as:

- corpora are large objects which are often awkward to copy
- copying them to other people can be legally problematic
- there are many occasional and non-technical potential corpus users who will not use them if it involves software installation
- the software is more easily maintained and updated
- the user does not need to invest in hardware, or expertise for support and maintenance.

For all of these reasons, our preferred model for most corpus use is the remote one. The corpora are on our servers, and this gives users better and easier access to them than they would get if they were on their own servers. To support users who want robot access to the corpus we provide a web API (using either cgi or JSON (see http://json.org)).
We note that the two clearing houses for language resources, LDC and ELRA, only minimally support the remote-access model.

2.6 Lemmatisers, POS-taggers, parsers

As the Sketch Engine is a corpus query tool, it most obviously serves as a common platform for corpora. Less obviously, it also provides a good platform for showcasing and exploring the behaviour of a range of NLP tools.
A potential user of, for example, a POS-tagger will have a number of questions: how fast is it, how easy is it to use, input and output options, how much does it cost - and how accurate is it. This last question is the central one, and it is also the one that is least easily answered. Published accuracy figures are usually higher than ones encountered in actual use, since, in the standard evaluation paradigm, there is a perfect match between the type of text used for training and the type used for testing: outside the laboratory, there will not be such a match. The wise potential user wants to look closely at a substantial sample of output of the tool, and form their own opinion of whether it will perform well enough for their purposes, and whether there are troublesome quirks and oddities to what it does.
If a corpus has been processed by a tool such as a lemmatiser or POS-tagger, and the data has then been loaded into the Sketch Engine, the Sketch Engine makes it easy for users to explore patterns of words and tags and to see where the tagger is correct and where it makes mistakes. There are 'view options' which allow the user to see the lemma and/or the POS-tag next to each word in the concordance (see Fig 3). There are also functions for, for example, counting different tags associated with a word. The word sketches and other statistical functions quickly draw attention to anomalies of the output (so can be useful for debugging).

A10	*/adequacy-n/NN1* of */of-p/PRF*	**resources /resource-n/NN2**	for */for-p/PRP* implementation
A0X	*/near-p/PRP* the */the-a/AT0*	**resource /resource-n/NN1**	. */.-p/PUN* Soon */soon-a/AV0*
A0C	*/with-p/PRP* limited */limited-j/AJ0*	**resources /resource-n/NN2**	. */.-p/PUN ` /`-p/PUQ* That
A1T	*/neither-a/AV0* the */the-a/AT0*	**resources /resource-n/NN2**	nor */nor-c/CJC* the */the-a/AT0*
A10	The */the-a/AT0* Age */age-n/NN1*	**Resource /resource-n/NN1**	project */project-n/NN1* , */,-p/PUN*
A10	Services */service-n/NN2* A */a-a/AT0*	**resource /resource-n/NN1**	pack */pack-n/NN1* has */have-v/VHZ*

Fig 3: Sketch Engine concordance with viewing options set so that lemmas and POS-tags are visible.

The Sketch Engine is also suitable for displaying the output of any parser which can output dependencies. Word sketches are summaries of dependencies of the form < word1, grammatical−relation,word2 >. The triples may come from either a shallow grammar written processed within the tool, or an external parser.

3 Case Studies

3.1 Slovene

A consortium of universities and publishers in Slovenia developed the Fida and FidaPlus corpora. They wanted the corpus to be as useful as possible for lexicography and other linguistic research. They approached Lexical Computing which installed the corpus in the Sketch Engine. They prepared the shallow grammar. We have recently upgraded what is available in the Sketch Engine from the 100 million word version to the 600m word version. The process is described in Krek and Kilgarriff (2006).

3.2 Japanese

Irena Srdanovic, of Tokyo Institute of Technology, was in need of a Japanese corpus for her lexicographic work. She contacted both Lexical Computing and also her colleague Tomaz Erjavec. Erjavec crawled the web for a 400-million word corpus and processed it with the CHASEN tools for Japanese word segmentation, lemmatising and POS-tagging. Lexical Computing loaded the corpus into the Sketch Engine; Srdanovic prepared a shallow grammar, so word sketches and a thesaurus could be produced. The corpus is available to all Sketch Engine account holders, and in particular to Srdanovic and her colleagues for their research. The process and corpus are described in Srdanovic (2008).

3.3 Chinese

The Chinese in Sketch Engine was developed in a collaboration between Academica Sinica, Taiwan and Lexical Computing. The Institute of Linguistics at Academia Sinica was interested in word sketches for Chinese for their research. While they had a number of corpora at their disposal, including the high-quality and carefully structured Sinica corpus, this was (at 40 million words) a little small for word sketches, so we chose instead to use the Linguistic Data Consortium's Chinese Gigaword. This was licensed and processed by the CKIP word segmentation and POS-tagging tools developed at Academia Sinica's Institute of Information Science. The shallow grammar was joint work. The corpus, word sketches and thesaurus are now available to Academia Sinica staff and students on their own server, and to Sketch Engine customers on the Sketch Engine server. The work is described in Kilgarriff et al (2005).

3.4 German, Italian, English

Marco Baroni has been exploring the issues and potential for developing very large web-sourced corpora. He and colleagues gathered three 2-billion word corpora for German, Italian and English, and lemmatised and POS-tagged them using TreeTagger (http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger/). Having gathered them, he wanted them to be optimally usable by whoever was interested. We loaded them into the Sketch Engine. He developed a shallow grammar for Italian (Lexical Computing already had one for English, and one for German is current work with University of Stuttgartt, see Ivanova et al 2008). The three corpora are available on the Sketch Engine server. The process is described in Baroni and Kilgarriff (2006) for German and Italian and Ferraresi et al (2008) for English.

3.5 Czech

The Czech National Corpus (CNC: http://ucnk.ff.cuni.cz/english/) is a large project which has been running for a number of years. A high-quality, carefully sampled 100-million word corpus has been prepared. The Institute for the Czech National Corpus wished to benefit from word sketches, and also to provide corpus access to all over the internet. The Institute chose to host the corpus themselves rather than on the Sketch Engine server, so it now has a Sketch Engine installation with the CNC loaded, which will shortly be available at no cost (on completion of an application form) to any scholar of Czech.

4 Conclusion

Showcasing corpora and NLP tools is important to their take-up and long-term success, and this is best managed through a web application for corpus access. The Sketch Engine is well-suited to the task. A commercial organisation may well perform the task better, over the long term, than a University, since it will have the income stream and hence the motivation to

maintain the website and resources. The Sketch Engine already performs this kind of role for a number of corpora and tools, and I have presented case studies of how this works. There is great potential for resource developers to benefit both from using their corpus in the Sketch Engine themselves, and also from being able to say to people who approach them about using it, “go and look at it in the Sketch Engine”.

References

Baroni, M. & A. Kilgarriff 2006. Large linguistically-processed Web corpora for multiple languages Proc. EACL. Trento, Italy.

Baroni, M,, A. Kilgarriff, J. Pomikalek & P. Rychly 2006. WebBootCaT: a web tool for instant corpora Proc. Euralex. Torino, Italy.

Ferraresi, A., E. Zanchetta, M. Baroni & S. Bernardini 2008. Introducing and evaluating ukWaC, a very large Web-derived corpus of English. In Proceedings of the WAC4 Workshop at LREC 2008.

Ivanova, K., U. Heid, S. Schulte im Walde, A. Kilgarriff & J. Pomikalek 2008. Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. Proc LREC. Marrakech, Morocco.

Kilgarriff, A., Huang C-R., P. Rychly, S. Smith & D. Tugwell 2005. Chinese word sketches. Proc. Asialex, Singapore, June.

Krek, S. & A. Kilgarriff 2006. Slovene Word Sketches Proc. 5th Solvenian/First Intnl Lanugages Technology Conference. Ljubljana, Slovenia.

Srdanovic Erjavec, I., T. Erjavec & A. Kilgarriff 2008. A web corpus and word sketches for Japanese. Japanese Journal of NLP.

DERIVATIONAL RELATIONS IN SLAVONIC LANGUAGES

Karel Pala

Centre for NLP, Faculty of Informatics, Masaryk University
Botanická 68a, 60200 Brno, Czech Republic
pala@fi.muni.cz

ABSTRACT

In the paper we touch the main derivational processes that in highly inflectional languages such as Czech (and Slavonic languages in general) form typical derivational nests (or subnets). In our view, the main attention has to be paid to the semantic nature of the derivational relations – without its systematic examination we can hardly put their formal properties under one roof. The regularity of the derivational relations in Czech and other Slavonic languages allows us to process them semi-automatically. The necessary prerequisites for dealing with derivational relations on more general level are at least two: the relevant resources, i.e. lists of roots, stems, lemmata and affixes, and the tools for handling them, such as an appropriate morphological analyzer or better, its derivational version that is able to process the basic and most productive derivational relations in a particular language (Czech) and generate the respective nests. The second tool allowing us to explore thoroughly the formal and semantic nature of the selected noun derivational suffixes and verb derivational prefixes is a special derivational interface DERIV developed in the NLP Centre at FI MU. This tool makes it possible to find for a given suffix or prefix all relevant derivational pairs which then can be further processed. Based on Czech we propose a collection of the semantically labeled derivational relations – presently 15. It is our strong belief that these 15 relations can serve as a common denominator for most of the Slavonic languages (a comparison of Czech, Croatian, Slovak and Slovene confirms this view convincingly enough).

0. Introduction

The regularity of the D-relations and their semantics in the highly inflectional language such as Czech (and other Slavonic languages as well) asks for their more formal description than is available so far in classical grammars. Such description is necessary if we want to handle D-relations formally for various computer applications like semantic analysis, searching, ontology creation, machine translation and others.

In general, derivational relations represent a system of morpho-semantic relations that definitely reflects cognitive structures underlying what may be called a language ontology. It undoubtedly exists but according to our knowledge such ontology has not been written down explicitly yet. It is also obvious that for language users derivational affixes (morphemes) function as formal means by which they express semantic relations necessary for using language as a vehicle of communication.

The researchers in the area of NLP are gradually becoming aware of this fact and recently attempts have appeared to examine the semantic of D-relations more formally in order to integrate them into semantic networks of the Wordnet type (cf. Hlaváčková, Pala, 2007, Azarova, 2008, Koeva, Krsteva, Vitas, 2008).

First attempts to capture D-relations have appeared in EuroWordNet project where the Internal Language Relation 'derivative' was introduced (cf. Vossen, 2003) but without more systematic insight into the structure of the D-relations and their comparison between EuroWordNet languages.

Wordnets as semantic networks are based on the collection of the selected semantic relations like synonymy, antonymy, hypero/hyponymy, holo/meronymy to mention the main ones. Thus if we want to integrate D-relations into the Wordnets we have to explore the semantics of the D-relations in a more detailed way as we hinted above. We have explored Czech D-relations from this point in (Hlaváčková, Pala, 2007) where we offered the collection of 14 D-relations for Czech and characterized their semantics using the particular labels. We have integrated the D-relations into Czech Wordnet with a very good result – we were able to insert approx. 30 000 new literals into it almost automatically.

The necessary prerequisite for a more detailed examination of D-relations is an appropriate algorithmical description of the inflectional morphology which in the case of Czech is implemented in the form of the morphological analyser Ajka (cf. Sedláček, Smrž, 2003). It works with the system of approx. 1800 inflectional paradigms. If we have a formal description of the inflectional morphology at our disposal we can start looking for relations belonging to derivational morphology (Osolsobě et al, 2004). We can observe that the individual inflectional paradigms are systematically linked to the particular suffixes, thus it is possible to find mapping from the inflectional paradigms to suffixes which then can serve as derivational paradigms. For

example, Czech nouns ending with suffix *-el* (*učitel*, teacher) are inflected within 4-5 paradigms where 2 paradigms are for animate nouns with *-el* and the rest for inanimate ones. In this way we get the mapping between the inflectional and derivational paradigms. This, applies to other suffixes as well and has obvious consequence for D-relations.

If we have look at what standard Czech grammars (cf. Karlík et al, 1995) say about the semantics of the parts of speech we find the formulations such as: 'nouns denote independent entities, i.e. persons, animals and things and also properties and actions. Verbs denote states and their changes and processes (actions) and their mutations'. These descriptions certainly refer to the semantics of the nouns and verbs. Then they are followed by the explanations about morphological processes typical of nouns, verbs, ..., i. e. how some parts of speech are derived from the others. What is relevant and what is missing in the standard grammars are the more detailed and extensive semantic classifications of nouns, verbs, as well as adjectives and numerals. Semantic classes of verbs are, however appearing (Hlaváčková, Horák, 2005, Levin, 1993).

It is also clear that the classical descriptions of the D-relations have been typically based on rather limited data, i. e. the main types of D-relations have been distinguished on the ground of the selected examples whose number has always been limited. Their occurrence is usually characterized as very frequent, frequent or rare without well grounded quantitative observations. Therefore, the generalizations made cannot be considered reliable, their nature is only orientational. For more complete and reliable description of D-relations in a given language we need, however, considerably larger data for the individual Slavonic languages, particularly the almost complete lists of lemmata, stems and affixes that can be obtained from corpora and then processed automatically or semi-automatically.

In fact, we face a more general task here – if we are describing the formal structure of D-relations in Czech we may as well go one step further and try to do the same also for other Slavonic languages, i.e. to attempt to compare derivational systems of the individual Slavonic languages and look for their common denominator which undoubtedly exists.

Which Slavonic languages come into consideration? Certainly the following ones can be mentioned: Bulgarian, Croatian, Czech, Polish, Russian, Serbian, Slovak, Slovene. Presently, we have no information about Ukrainian and Belorussian.

1. Resources and Data

Building the formal model of (Czech or Slavonic) D-relations relies on the necessary resources which include lists of the lemmata, roots, stems and affixes and should be as complete as possible. Ideally, the goal is to have resources that cover the whole word stock of a language. This depends on the existing lexicons and corpora which allow us to compile the required lists. Without larger data more reliable generalizations and conclusions can hardly be drawn and the adequate rules written. Thus we are going to characterize the existing resources. For Czech the following ones have been used:

- morphological database of Czech stems contains presently approx. 400 000 items where for each stem the information about the corresponding POS is given and the respective inflectional paradigm linked with the stem (there are approx. 1800 of them) is indicated as well. From this list we can generate approx. 6,5 mil. Czech word forms. The number of possible lemmata is about 600 000 and the coverage of the database is approx. 96 % of the corpus Syn2000, which is the first official version of Czech National Corpus. The database includes:
  - list of noun stems (roots) - approx. 126 000 items
  - list of verb stems (roots) - approx. 36 000 items
  - list of adjective stems contains approx. 40 000 items (?)
  - list of adverbs (non-derived about 400?, the rest is derived from adjectives, approx. 25,000)
  - lists of remaining parts of speech, i. e. pronouns, numerals, prepositions, conjunctions and particles,
- list of basic suffixes containing approx. 140 items from which 41 have been processed so far,
- list of intersegments, i. e. stem forming morphemes and derivational infixes, their number in Czech is approx. 250,
- list of prefixes which for Czech contains approx. 240 items. Presently we have selected 14 basic ones. The basic prefixes are highly productive, for instance, prefixes *na-, po-, pro-, s-/se-, u-, v-/ve-, z-/ze-* occur with 20,716 verbs (from Ajka database),
- list of the inflectional paradigms (declension, conjugation) used in a morphological analyser Ajka, it contains 1860 items,
- list of the morphological alternations (in stems and between the morphemic segments). They are necessary for

further formulating more complicated derivational rules. Few typical examples: *r - ř* (*doktor – doktoři, doctor – doctors*), *k – c* (*vlk – vlci, wolf – wolves*), epenthetic *-e-* (*letec – letci, aviator – aviators*) and others – their number in Czech is approx. 30 and they are closely related to the concrete inflectional noun and verb paradigms).

The described data are necessary for dealing with the Czech derivational rules and relations and a generalization can be made that for handling D-relations in other Slavonic languages the similar data should be at our disposal as well. Large enough data are the guarantee of the relatively complete description of the D-relations. According to our knowledge they certainly exist for Bulgarian, Croatian, Polish, Russian, Slovak, Slovenian and Serbian. Thus there is a challenge to coordinate our effort in this respect and try to prepare collections suitable for processing the D-relations in the mentioned Slavonic languages.

2. Tools

2.1. Morphological analyser

To be able to deal with inflection and prefix and suffix derivations automatically software tools are necessary. The first of them is a standard morphological analyser (Ajka for Czech, cf. Sedláček, Smrž, 2003) which handles inflectional morphology (declension and conjugation) and as its output generates lemmata and word forms together with their respective grammatical categories (tags) for the whole list of the stems (in Czech approx. 400 000 items). We can characterize it as I(nflectional)-Ajka. For handling the derivational relations in a language we also need a derivational version of the morphological analyser which is able to work both with inflectional and derivational paradigms. It has been developed for Czech and we call it D(erivational)-Ajka.

2.2. Derivational interface

The second tool we work with is a special derivational interface Deriv (developed in NLP Centre at FI MU) processing the data that are output of the D-Ajka analyser. With this tool we can obtain word derivation pairs, i.e. it links stems and prefixes or suffixes to generate pairs 'base form – derived form'. The number of noun stems in the stem list of the D-Ajka is approx. 126 000 and the list of verb stems contains approx. 36 000 verb stems. Other parts of speech can be processed as well. The Deriv tool generates derivational pairs for nouns, verbs, adjectives and adverbs in three steps:

1. a set of words (lemmata) is defined by an appropriate morphological tag and a selected suffix or prefix;
2. a derivational rule is defined, typically as a concatenation of the stem and suffixes or prefixes). Then the rule can be applied and as a result we obtain the list of the pairs 'base form – derived form' (noun – noun, noun – adjective, noun – verb, verb – noun, etc.);
3. the obtained results having the form of the lists of pairs then can be appropriately corrected and modified manually – usually they contain cases that are derived according to a given rule but they should be marked as overgenerated.

An example: take the derivational analysis of Czech sufix *–ík*: it occurs with the nouns denoting agent or instrument (means), e.g. z*ed-n-ík (bricklayer)* or *kapes-ník (hankerchief)*.

First, we want to derive agentive nouns: so we enter the suffix *–ík* and tag k1gM (noun, masculine animate) and generate the list of all pairs containing words ending with *-ík*. The output is a list of 1210 nouns including proper names (from the original list of 126 000 Czech noun lemmata). To obtain instrument nouns we input the tag k1gI (noun, masculine inanimate). As an output result we get a list of 715 nouns including proper names. The number of all words ending with suffix *-ík* (disregarding the grammatical tag) in stem dictionary of Ajka is 1830. The difference in the given numbers includes 95 items and follows from the homonymy, for instance, some nouns can be both masculine animate and masculine inanimate (e.g. noun *náčelník* can denote – *chief* as well as *čelenka – headband)*.

In our view, the interface Deriv can be developed also for other languages if the data mentioned above are at hand for them.

2.3. Tagsets – metadata

Morphological analyzers and tools like Deriv work with grammatical (morphological) tags that express the respective grammatical categories and associate them with generated or recognized word forms. The key category is a part of speech

which is typically linked with other categories such as gender, number, case (for nouns, adjectives, pronouns and numerals), person, number, tense, mode, voice, aspect for verbs, graduation for adjectives and adverbs. The list of gramatical categories or tagset used in Ajka tool is designed for Czech but, in general, there is a need for a tagset that could be used for all Slavonic languages. A suitable candidate is a tagset used in Multext East project (cf. Erjavec et al, 2003), however, not all analyzers can use it, thus the standardization is problematic since the re-programming of the individual analysers would be labourious and costly (not all analyzers can allow for easy exchange of the tagsets). Hopefully, one acceptable solution would be to write conversion scripts that might translate between the tagsets used for the particular Slavonic languages. As our experience shows it may be difficult to unify two tagsets even for one language, for example, in Czech two tagsets exist – one for Ajka and second for Prague analyser (Hajič, 2004) and we have to use the conversion scripts to be able to exploit both of them. Even then some problems remain, for instance, with differences in lemmatization.

3. D-relations and their inventory

With reference to the mentioned papers (Hlaváčková, Pala, 2007, Azarova, 2008, Koeva, Krsteva, Vitas, 2008) we offer a set of the 15 derivational relations. They have been labelled differently in the cited papers but in our view the differences are rather terminological. We try to capture the main relations here and leave aside either the marginal ones or the ones that are not exploiting strictly suffixation and prefixation, e.g. compounds.

They can be temptatively grouped in the following way:

1. role D-relations in which a relation between two POS expresses a semantic role, e.g. Agentive, Instrument, Location, etc.
2. gender, diminutive and augmentative D-relations, they can be considered symmetrical.
3. D-relations that denote various kinds of properties, some of them can be considered deverbative (first member of the pair is always a verb) and the remaining ones exist between nouns and adjectives (also possessives) and adjectives and adverbs.
4. prefix D-relations consisting only of the pairs verb – prefixed verb, their meanings are related to the verb classes , e.g. verbs of motion, verbs of drinking and eating, or verbs of weather, etc. This area represents a challenge for further research.
5. aspect relation has to mentioned as well though it is typically treated as a grammatical category in Slavonic languages and therefore it is questionable how much it can be considered derivational, however, it is close to the derivational processes. The semantics of the perfectives and imperfectives concerns the time properties of the actions denoted by them.

The first set of 14 D-relations has been proposed for the enrichment of Czech Wordnet in (Hlaváčková, Pala, 2007). The basic and most productive D-relations in Czech have been integrated into Czech morphological analyzer Ajka and their semantic labels were added to the D-relations. To the original 14 relations based on suffixation we have added the the 15th one which exists between verbs and includes prefixation only. In fact, the 15th prefix relation represents a more complex group of 11 relations that are outlined below.

The individual groups of the D-relations can be further characterized in the following way:
a) D-relations expressing semantic roles, they exist between verb – noun pairs (suffixation), we assume that they can be found in all Slavonic languages (numbers in the brackets denote frequency in the ČNK Syn2000, http://ucnk.ff.cuni.cz/):

- der-agentive: *učit – učitel* (8639 – 9924)*, teach – teacher* (pair verb - noun)
- der-patient: *trestat – trestanec* (1636 – 301)*, punish – convict* (pair verb - noun)
- der-instrument: *ukazovat – ukazovátko* (11 831 – 61)*, point – pointer, fescue* (pair verb - noun)
- der-location: *letět – letiště* (3213 – 6636)*, fly – airport* (pair verb - noun)

b) D-relations denoting gender derivation of feminine nouns from their masculine counterparts and diminutivity and augmentation, they exist between noun - noun pairs and in Czech deminutives occur also in triples (suffixation). These relations also appear in all Slavonic languages,

- der-gender: *student – studentka* (15 608 – 1260), *student – she-student* (pair noun - noun)
- der-dimin: *dům – domek – domeček* (46 485 – 5118 – 606)*, house – small house, very small house or house one likes* (triple noun – noun – noun)
- der-augm: *dub – dubisko* (752 - 24)*, oak tree – big and strong oak tree* (pair noun - noun)

c) D-relations denoting action and property (suffixation) – their typical feature is that both members of the relation denote the same meaning and they differ only in the part of speech, for instance, in the pair *učit – učení* action is denoted by a verb and deverbative noun, respectively. In the remaining pairs both members denote property and the difference consists again only in the part of speech.

- deriv-action: *učit – učení* (8639 – 2877), *teach – teaching* (pair verb – deverbative noun)
- deriv-property: *učit – učený* (8639 – 409) *teach – teaching* (pair verb – deverbative adjective)
- deriv-property: *učený – učenost* (409 – 63) *learned – learnedness (*pair adjective – noun)
- deriv-property: *učený – učeně* (409 – 33) *learned – learnedly* (pair adjective – adverb)
- deriv-property-possessive: *učitel – učitelův* (9924 – 56 (pair noun – adjective)

d) The relation Deriv-aspect (*učit - naučit*, pair verb – verb) should be mentioned but we are not dealing with it here since the category of the aspect is considered a grammatical category in Slavonic languages. Its semantics captures time relations – completion and continuation of the actions expressed by the respective verbs. Formally, aspect is realized by prefixation and infixation.

e) D-relations exploiting prefixation represent a separate group in which we pay attention only to the pair verb – verb. Their meanings depend heavily on the meanings of the verb stems they occur with. They denote a number of different semantic relations such as various sorts of motion, time and location (see below), intensity of action, inchoativity, iterativity, additivity, distributive action, obligation, result and possibly some others. Here we offer their preliminary subclassification which obviously requires a further examination. It is based on the list of 14 Czech basic (primary) prefixes we have worked with so far:
do- (to), na- (on, at), nad- (above, up), od- (from, away), pro- (for, because), při- (by, at), pře- (over), roz- (over), s-/se- (with, by), u- (at, near), v-/ve- (in, up), vy- (out, off), z-/ze- (of, off), za- (over, behind).

Verbal D-relations based on prefixes:

1. motionI: deriv-mot-to: motion to the point or place, e. g. *jít – přijít,* (*go, walking–come*), *letět – přiletět* (*fly – arrive by plane*),
1.2 motionII: deriv-mot-to-iter: iterative, repeating motion to a point or place, e. g. *přicházet – přicházívat* (*be coming – coming repeatedly*),

2. motionI: deriv-mot-from: motion from a point or place, e. g. *jít – odejít* (*go/walk – leave by walking*),
2.1 motionI: deriv-mot-from-iter: iterative motion from a point or place, e. g. *odcházet – odcházívat* (*leave by walking – leaving repeatedly*),

3. motionI: deriv-mot-over: motion across a point or place, e. g. *brodit – přebrodit* (*wade – wade through*),
3.1 motionII: deriv-mot-under: motion under a point or place, e. g. letět – *podletět* (*fly – fly under*),

4. timeI: deriv-compl-act-: to complete an action (with regard to any verb), e.g. *letět – do-letět* (*fly – finish flying*),
4.1 timeII: deriv-t-act-iter: to complete an action iteratively, e.g. *tancovat – do-tancovat – do-tanco-vá-vat* (*finish dancing – finish dancing repeatedly*),

5. obligation: deriv-oblig: to perform an action as an obligation *pracovat – odpracovat* (*work – work off*)

6. additivity: deriv-addit: action expressing adding *koupit – přikoupit (buy – buy more)*

7. distributivity: deriv-distrib: to perform an action in a distributed way *ztratit – poztrácet (lose – lose little by little, i.e. lose successively particular objects (zabíjet – pozabíjet)?*

8. result: deriv-result: to perform an action with its result *vařit -> vyvařit (boil – boil away)*

9. high intensity: deriv-high-intens: to perform an action more intensively *vařit – navařit (cook – cook a lot of sth)*
9.1 low intensity: deriv-low-intens: to perform an action with low intensity *pracovat – popracovat (work – work a little, for a while).*

The list is preliminary and in our view includes just main and most typical meanings expressed by the prefixes *do- (to), od- (from), na- (on, to), po- (after), pod- (under), pře- (over, across), při- (to), vy- (out, from).* We give examples of all meanings mentioned above but not necessarily of all prefixes. It can be seen that meanings related to motion and time can be further subclassified but we certainly have not captured all that can be related to the verbs of motion. Here the more detailed study is necessary. We also consider here only the verbs of motion with two arguments, i. e. verbs with an Agent causing motion to a Location. Verbs with an Agent, moved Object, e.g. *nést knihu domů – přinést knihu domů (carry the book home – fetch the*

book home) and Location (*home*) still have to wait for the more detailed analysis. In the list above we have included iterativity relation as well because of its regularity in Czech though iteratives are not derived with prefixes but with alternations in the stems using infixes *-áva-, -íva-, -ova-.* On one hand including the iterativity relation may seem to complicate the description but on the other its regularity allows us to handle it almost automatically. It has to be remarked that the iterativity relation is semantically close to the aspect relation, that is why some authors speak about third aspect though iteratives are imperfective by definition.
We are well aware that the D-relations exploiting prefixation can be organized differently as it is typical for any kind of semantic classification. We are attempting to outline the main ones.

4. Examples

4.1. Suffixation

The core of the derivational nest for the Czech roots *uč/uk-* (*work*) created by suffixation looks as follows:

D-relation	pair basic form - derived form	Frequency in ČNK	POS pair
deriv-ag:	*učit - učit-el (teach – teacher)*	8634-9924	verb – noun
deriv-ag:	*učit - učen-ec (teach – scholar)*	8634-2877	verb – noun
deriv-pat:	*učit - učeň (teach – apprentice)*	8634-482	verb – noun
deriv-pat:	*učit - učedn-ík (teach – disciple)*	8634-818	verb – noun
deriv-loc:	*učit - učebna (teach – classroom)*	8634-527	verb – noun
deriv-loc:	*učit - učiliště (teach, train - training institution)*	8634-156	verb – noun
deriv-instr:	*učit - učebnice (teach – textbook)*	8634-2338	verb – noun
deriv-gender:	*učitel - učitel-ka (he-teacher – she-teacher)*	9924-2303	noun – noun
deriv-gender:	*učedník - učedn-ice (he-disciple – she-disciple)*	818-21	noun – noun
deriv-gender:	učeň - učnice (he-apprentice – she-apprentice)	724-24	noun – noun
deriv-dimin:	*učebnice - učebnička (textbook - small textbook)*	2338-0	noun – noun
deriv-dimin:	*učedník - učedníček (disciple - small disciple)*	818-2	noun – noun
deriv-dimin:	*učitel - učitýlek (small teacher or* pejorative expr.	9924-3	noun – noun
deriv-augment:	*učitel - úča (teacher – she-teacher,* pejorative expr.*)*	9924-0	noun – noun
deriv-pro:	*učit - učicí (teach -teaching)*	8634-53	verb – adj
deriv-pro:	*učit - učební (teach – didactic)*	8634-893	verb – adj
deriv-pro:	*učitel - učitelský (teacher - teaching, preceptorial)*	9924-658	noun – adj
deriv-pro:	*učeň - učňovský (apprentice* noun - *apprentice* adj*)*	724-404	noun – adj
deriv-pro:	*učenec - učenecký (scholar – scholarly)*	482-10	noun – adj
deriv-pro:	*učedník - učednický (disciple -discipular)*	818-41	noun – adj
deriv-pro:	*učit - učený (teach - scholarly, learned)*	8634-409	verb – adj
deriv-pro:	*učit - učenl-ivý (teach - teachable, docile)*	8634-74	verb – adj
deriv-pro:	*učenlivý - učenliv-ost (teachable, docile - teachableness, docility)*	74-13	noun – adj
deriv-pro:	*učebnice - učebnic-ový (textbook* noun – *typical example as in a textbook* adj*)*	2338-169	adj – noun
deriv-poss:	*učitel - učitel-ův (teacher - teacher's, of a teacher)*	9924-56	noun – adj
deriv-poss:	*učitelka - učitelč-in (she-teacher – she-teacher's)*	2303-7	noun – adj
deriv-poss:	*učenec - učenc-ův (scholar – scholar's)*	482-3	noun – adj
deriv-poss:	*učedník - učedník-ův (disciple – disciple's)*	818-1	noun – adj
deriv-poss:	*učeň - učň-ův (apprentice – apprentice's)*	724-0	noun – adj
deriv-act:	*učit - učen-í (teach - teaching* noun*)*	8634-2877	verb – noun

deriv-pat:	*učit - uč-ivo (teach – curriculum)*	8634-264	verb – noun
deriv-pro:	*učedník - učednic-tví (disciple – discipleship)*	881-17	noun – noun
deriv-pro:	*učený - učen-ost (scholarly, learned – erudition) scholarship)*	409-63	adj – noun
deriv-pro:	*učit - učitel-ství (teach - teaching profession)*	8634-64	verb – noun
deriv-pro:	*učitel - učitel-stvo (teacher - teachers, teaching staff)*	9924-64	noun – noun
deriv-augment:	*učitel - učitelák (teacher – teacher's college) (pejorative)*	9924-1	noun - noun

Table 1 Nest for the Czech roots *-uč-/-uk- (teach)* - suffixation

The number of the simple derivations in Table 1 is 32. The complete list of all derivations for the roots *-uč-/-uk-* is, however, much larger and contains approx. 1000 items if we take into consideration prefixation both for nouns and verbs and all possible compounds. This means that the full list of all D-relations for roots of the similar type (e. g. *-prác-/-prac-*) will be also richer and calls for the further investigation. At the moment it is not easy to estimate how many new D-relations should be added to the list of 15 D-relations presented above but we expect that altogether the number of the main relations would not exceed 25. This does not include subclassifications that can be structured and granulated in various ways.

4.2. Prefixation

D-relation	pair basic form - derived form	Frequency in ČNK
Deriv-pref: aspect	*učit - na-učit (teach, imperfective - perfective)*	8634-7621
Deriv-pref: time, finishing	*učit - do-učit (teach - finish teaching)*	8634-53
Deriv-pref: low intensity	*naučit - po-na-učit (sth. like give advice, instruction)*, no lexicalized English equivalent	7621-1
Deriv-pref: undo	*naučit - od-na-učit (teach - unteach, wean off)*	7621-152
Deriv-pref: prescriptive	*učit - od-učit (teach -to complete teaching)*	8634-18
Deriv-pref: low intensity	*učit - po-učit (teach - instruct, advise)*	8634-1084
Deriv-prefix: redo	*učit - pře-učit (teach - re-teach, teach again)*	8634-6
Deriv-pref: low intensity	*učit - při-učit (teach -learn a bit more)*	8634-149
Deriv-pref: result	*učit - vy-učit (teach - train)*	8634 - 522
Deriv-prefix: inchoat	*učit - za-učit (teach - give initial training)*	8634- 43

Table 2 Nest for the Czech root *uč- (teach)* – prefixation

Thanks to the cooperation with the Croatian (M. Tadić), Slovene (D. Fišer) and Slovak (M. Grác) colleagues we have been able to compare derivational nests for the roots *-uč-/-uk-* in four Slavonic languages. Though we are not able to offer the complete comparison at the moment the good news is that the 'derivational' agreement between Czech, Croatian, Slovene and Slovak can be estimated as not lower than 95 % . It can be expected that agreement between more distant Slavonic languages such as Czech and Russian or Czech and Bulgarian may be lower but we assume that its value still can be around 90 %.

5. Conclusions

In the paper we outline further results of computational analysis of the basic and most regular D-relations in Czech with respect to suffixation and prefixation. The task is to extend this analysis in the direction of other Slavonic languages. We formulate the basic prerequisites that have to be fullfilled for this purpose. First, the relevant resources are characterized, i.e. lists of roots, stems, lemmata and affixes, and second, the tools for handling them, i. e. an appropriate morphological analyzer (I-Ajka) or better, its derivational version (D-Ajka) that is able to generate the particular derivational nests. For exploring the formal and semantic nature of the selected noun derivational suffixes and verb derivational prefixes a special derivational interface Deriv has been developed in the NLP Centre at FI MU. This tool makes it possible to find for a given suffix or prefix all relevant derivational pairs which then can be further processed. Both D-Ajka and Deriv interface exploit a tagset – thus a questions arises whether a tagset can be found that could work for all Slavonic languages. We come to the conclusion that the obvious candidate – the tagset developed in the project Multext East is not probably the best solution.

The possibility is to develep conversion scripts for this purpose.

Based on Czech we have proposed a collection of the semantically labeled derivational relations whose number is presently 15. We are convinced they can serve as a common denominator for most of the Slavonic languages. Preliminary comparison of the Czech, Croatian, Slovak and Slovenian and Czech derivational nests for the roots *-uč-/-uk- (teach)* confirms this view well enough – the estimated 'derivational' agreement is about 95 %.

We are well aware that these results are far from complete and they just indicate in what direction we should continue this research. The full derivation nests appear to be larger and the task is to find out how the complete list of the semantically labelled D-relations will look like. The ambition also is to test the proposed set of the D-relations for all Slavonic languages and also for others.

It is obvious that the investigation of the D-relations brings forth to the results that are relevant for computational processing of natural languages, particularly in building the Wordnets. The ultimate goal may be the derivation automaton that will allow us to reduce the large lists of stems and lemmata to the relatively small lists of the roots (approx. 60 000) and to generate the whole lexicon of a given natural language. Last but not least, we would like to know how the derivational relations and derivational subnets reflect basic cognitive structures existing in natural language. More effort is needed for exploring them in the terms of presently so popular ontologies.

References

Azarova, I.; V.; 2008. Derivational semantic relations in RussNet, presentation at the 4th Global Wordnet Conference, Szegéd.

Český národní korpus (Czech National Corpus – Syn2000), http://ucnk.ff.cuni.cz/

Erjavec, T.; C. Krsteva; V. Petkevič; K. Simov; M. Tadić; D. Vitas. 2003. The MULTEXT-East Morphosyntactic Specifications for Slavic Languages. In *Proceedings of the EACL2003 Workshop on Morphological Processing of Slavic Languages*, Budapest: ACL. (URL: http://hnk.ffzg.hr/txts/mte-mpsl03_(erjavec_et_al).pdf)

Hajič, J. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*, Charles University Press,1st edition.

Hlaváčková, D.; Horák, A. 2005. Verbalex – new comprehensive lexicon of verb valencies for Czech. In: *Proceedings of the Slovko Conference*, Bratislava.

Karlík, P. et al. 1995. *Příruční mluvnice češtiny (Reference Grammar of Czech)*, Nakladatelství Lidové Noviny, Prague, pp. 229, 310.

Koeva, S. Krsteva, C.; Vitas, D. 2008. Morpho-semantic Relations in Wordnet – a Case Study for Two Slavic Languages, In: *Proceedings of 4th Global Wordnet Conference*, Szegéd, p. 239-253.

Levin, B.: *English Verb Classes and Alternations: a preliminary investigation*, The University of Chicago Press, (1993).

Osolsobě, K.; Pala, K.; Sedláček, R.; Veber, M. 2002. A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas.

Pala, K; Hlaváčková, D. Pala, K; Hlaváčková, D. 2007. Derivational relations in Czech WordNet, In: *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, ACL, Prague, p. 75–81.

Sedláček, R.; Smrž P. 2001. A New Czech Morphological Analyser Ajka. In: *Proceedings of the 4th International Conference on Text, Speech and Dialogue*, Springer Verlag, Berlin, p. 100-107.

Vossen, P. 2003. *EuroWordNet General Document*, Version 3, University of Amsterdam.

INTELLIGENT CONTENT PROCESSING IN THE MULTILINGUAL MEDIA WORLD

Stelios Piperidis

Language Technology Applications Department
Institute for Language and Speech Processing – Athena R.C.
spip@ilsp.gr

ABSTRACT

The convergence of technological communication platforms and the evolution of broadband networks opens up new opportunities for content generation and consumption, while enabling such content to be increasingly multilingual and multimedia in nature. This talk will provide an introduction to multimedia content processing, focusing on the role and significance of natural language in multimedia discourse. We will briefly review the state-of-the-art in single-media processing (speech, text, video, etc) and discuss the problems, challenges, fall-back solutions and necessities for further advances. The role of the different media and the potential benefit from comparative analysis and fusion of single-media processing results will be discussed in the context of different settings including monolingual and multilingual subtitling, multimedia information retrieval and related applications.

1. Introduction

The explosion of multimedia digital content and the development of technologies that go beyond traditional broadcast and TV have rendered language-aware access to and further processing of such content important for all end-users of these technologies. The development of methods and tools for semantic, content-based organization and filtering of the large amount of multimedia information that reaches the user through heterogeneous channels is a key issue for its effective consumption. Despite recent technological progress in the new media and the Internet, the key issue remains “how digital technology adds value to information channels and systems”.

Contemporary language-aware content technology addresses this issue by helping people keep up with the explosion of digital content scattered over different platforms (radio, TV, Web), different media (speech, text, image, video) and different languages. Content producers and consumers are provided with search, retrieval functionalities, while more advanced semantics-based functionalities, like categorization, summarisation and translation of multimedia content, are elaborated through the use of automatically created semantic indices and links across media.

In this talk we will be focusing on three major challenges in the design of intelligent processing of multimedia content :

- how multimedia content, potentially multilingual, is augmented with semantic information like speaker identities, speaker turns, content topic(s), events and their participating entities, keyframes and face names,
- how semantic links can be established between pieces of information presented in different media and languages within the same as well as across (multimedia) documents
- how multimedia content, especially TV and DVD content can be processed in such a way that it is rendered appropriate for users with special needs, e.g. deaf and hard of hearing (monolingual subtitling), or for viewers in foreign language speaking countries (multilingual subtitling).

2. Multimedia Content Processing Architectures

Multimedia content processing technology is deployed by either content providers who want to add value to their content, restructure and re-purpose it, or by end users who wish to gather, filter and categorize information collected from a wide variety of sources. In both scenaria of use web sources, television broadcasts and radio programmes are analysed, indexed, categorized, summarized and stored in an archive. This content can be searched by a user or filtered and pushed to him/her according to his/her interests. In the Cimwos(http://www.xanthi.ilsp.gr/cimwos/) and Reveal This (http://www.reveal-this.org/) cases that will serve as examples, novice and advanced computer users are targeted; the former use mainly a simple mobile phone interface where information is pushed to, while the latter use a web interface for searching. EU politics, news and travel data are handled by the system in English and Greek.

The Reveal This architecture comprises: (i) the Cross-media Content Analysis & Indexing (CAIS), (ii) the Cross-media Categorisation (CCS) and the Cross-media Summarisation (CSS), (iii) the Cross-lingual Translation (CLTS), and (iv) the Cross-media Content Access and Retrieval subsystem. In the next sections we will describe briefly the first two subsystems.

2.1. Cross-media Content Analysis and Indexing

The Cross-media Content Analysis and Indexing Subsystem (CCAIS) caters for processing single media and automatically generating metadata for each medium such as:

- speaker turn and identity, transcriptions and stories for speech data,
- named entities (persons, places, organizations), topics and facts for web text and transcribed speech
- keyframes, shots, faces (detected and identified) and image categories for video and images

The metadata produced by the single-media processing components are aligned, synchronized and encoded in XML. The result is an XML/MPEG-7 document containing all information gathered and linked to the corresponding points of the source material (text, audio, video). Segmentation suggested by audio processing (speaker turns) and topic detection is taken into account to produce unified segments. Several consecutive segments are aggregated to form “stories”, i.e. sections of the document that deal with the same topics. The cross-media indexing component decides on the most appropriate indexing terms per story. Crossing media (or cross-mediality) is conceived of as the process of intelinking evidence, in the form of indexical data, provided by the different media participating in the message formation process within the same multimedia document (e.g. a video file).

2.2. Cross-media categorisation and summarisation

The categorization subsystem operates along the cross-lingual and cross-media directions. In the cross-lingual dimension, a categorizer based on a pivot language category model (in English) is deployed. Documents in other languages go through a translation phase, thus enabling their categorisation. Such a strategy overcomes constraints at the level of the training set, which often are not sufficiently big for building models in all languages involved.In the cross-media dimension, documents containing not only text or images but a combination of different types of media (text, image, speech, video) are considered. The multiple-view fusion method adopted builds 'on top' of two single-media categorizers, a textual and an image categorizer, without the need to re-train them. Data annotated manually for both textual and image categories is used for training the cross-media categorizer. In that set dependencies between single-media category systems are exploited in order to refine the categorization decisions made.

The task of the Cross-media Summarization Subsystem (CSS) is to determine and present the most salient parts according to users’ profiles and interests by fusing video, audio and textual metadata. The CSS subsystem consists of three major components: the textual-based summarization component, the visual-based summarization component, and the cross-media summarization component aiming at the fusion of the two analyses and creating a self-contained object. Additionally, the Cross-media summarization subsystem provides the necessary visualization interfaces enabling the user to preview a specific multimedia object before downloading a file of interest. Cross-media summaries for politics, news and travel-related audiovisual files have been designed and implemented in the project; while the module interfaces also with a web-based translation service for creating “translated” versions of the summaries for English and Greek.

The system was tested in the two domains of EU politics and travel information, and in two languages, English and Greek. The results showed that the innovative functionalities were more than welcome, the performance was generally satisfactory, though expectations for much better performance in terms of e.g. speech transcription and translation were evident.

3. Augmenting multimedia content with monolingual and multilingual subtitles

New technological developments in mass media and communication, such as digital TV and DVD, are bound to overcome the limited physical borders of countries, leading to the creation of a world-wide media audience. In such a unified framework of mass communication, subtitling – as a means of overcoming linguistic barriers between the nations – is playing a critical role. In many countries, subtitling is the most commonly used method for conveying the content of foreign language dialogue to the audience; and a broadcaster’s audience may now include several major linguistic groups (notably in the case of a satellite-broadcaster). Broadcasters, in order to meet the needs of the significant numbers of deaf and hard-of-hearing viewers, also provide subtitling increasingly. However, subtitling is far from trivial and is considered to be one of the most expensive and time-consuming tasks an interested company needs to perform, since experts mainly carry it out manually. Typically, a 1-hour programme needs around 7-15 hours of effort by humans.

In this part of the talk, we will focus on the current technological approaches to monolingual and multilingual subtitling, using as example the achievements of the “Musa” project (http://sifnos.ilsp.gr/musa/). “Musa” combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles. The system converts audio streams into text transcriptions, condenses and/or rephrases the content to meet the spatio-temporal constraints of the subtitling process, and produces draft translations. Three European languages were supported: English as source and target as far as subtitle generation is concerned, French and Greek as subtitle translation target languages. In addition, we will be referring to alternative architectures to the translation problem using recent developments in the Interreg TM BG-EL project, a project that tries to adopt best practices in machine translation applying them in the translation between Bulgarian and Greek.

3.1. Requirements and standards for subtitle production and visual presentation

Current practices and standards followed by the big media groups and the European Broadcasting Union form the basis on which the “Musa” subtitling component converts transcripts to subtitles. They provide a unifying formula based on the different subtitling conventions currently operating within various countries. They cater for standardization along:
- *spatial parameters* (layout) including position on screen, number of lines, number of characters per line, etc.
- *temporal parameters* (duration), maximum duration of a full two-line and full single-line subtitle, minimum duration of a single-word subtitle, leading-in time, lagging-out time, time between two subtitles, camera takes/cuts
- *punctuation and letter case*, including sequence and linking dots, ordinary punctuation marks, use of upper- and lower-case
- *target text editing* including single-line vs two-line subtitle, segmentation at the highest-level linguistic nodes, segmentation and line length, relation between spoken utterances and subtitled sentences, subtitles with more than one sentence, omission or retainment of linguistic items of the original, simplification of syntactic structures, use of acronyms and other literals, use of dialects and taboo words, and use of culture-specific linguistic elements.

3.2. Subtitling architecture

The architecture of the multilingual subtitle production line includes the following functional blocks:

1. an English automatic speech recognition (ASR) subsystem for the transcription of audio streams into text, but also fall back positions such as alignment of audio and script for those cases where the latter exists
2. a subtitling subsystem producing English subtitles from English audio transcriptions aiming to provide maximum comprehension while complying with spatio-temporal constraints and linguistic parameters
3. a multilingual translation subsystem integrating machine translation, translation memories and terminological banks.

The input to the ASR is an audio file and the output a time-tagged text, i.e. the word-by-word transcript of the input audio, with segments of transcript corresponding to sentences. In case the programme's script is available, then the ASR module aligns the audio with the script and provides timecodes. The subtitling subsystem comprises the constraint formulation/calculation module, the text condensation module and the subtitle editing module. The input to the subtitling subsystem is English transcript with time codes, and the output is English subtitles. Condensation is effected in two stages: first the input is scanned against a list of automatically and/or manually constructed table of paraphrases. In case the required condensation ratio is not achieved, then the input is syntactically parsed and its least important constituents are marked as deletable constituents. Deletable constituents are ordered and deletions are performed until the condensation ratio is reached. The translation subsystem comprises a translation memory module and a machine translation engine. The input to the translation is English subtitles and the output foreign language subtitles with time codes. The resulting subtitles are linguistically processed and formatted.

The infrastructure consists of a set of multifaceted multimedia parallel data, in the sense that the same content is conveyed in different media (video, audio, text), while for component development different facets of the parallel corpus are deployed. Parallel data facets include : audio-script/transcript, transcript-English subtitles, English subtitles-Foreign language subtitles.

Automatically generated subtitles were evaluated by human professional subtitlers with overall acceptability of monolingual subtitles in the range of 85-87%, while multilingual subtitles were evaluated substantially lower following limitations of current translation technology.

CORPUS-BASED COMPARISON OF CONTEMPORARY CROATIAN, SERBIAN AND BOSNIAN

Božo Bekavac*, Sanja Seljan, Ivana Simeon***

*Department of Linguistics/ ** Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
bbekavac@ffzg.hr, sseljan@ffzg.hr, isimeon@ffzg.hr

ABSTRACT

This paper explores the differences between three Slavic languages: Bosnian, Croatian and Serbian, drawing on the Southeast European Times newspaper corpus, translated to each language from the source English text and consisting of approximately 330,000 tokens for each language. The paper is an effort intended to contribute to the establishment of the criteria and methodology for measuring similarities between these languages. The differences were explored at five levels: at the level of phonology, morphology, lexis, syntax and semantics. Empirical analysis has shown that a huge portion of differences across the three languages are systematic and regular, and as such, could be formalized for automatic translation/generation. The results of this study and of similar future corpus-based studies can be used in developing NLP tools such as annotating tools, e-dictionaries, text summarizers, machine translation systems, computer-assisted language learning etc. for the three languages, as well as further linguistic investigation of their mutual relationship.

1. Introduction

As language technologies are becoming increasingly important as a way to manage the growing volume of multilingual communication in Europe as a linguistically diverse community, resources and tools for Croatian and other Slavic languages will have to be built, as a part of preparation of these countries for the accession to the European Union. Since parallel texts for these languages are scarce in comparison to widely spoken languages, such corpora could be an important resource for research.

In parallel corpora, the same information is presented in different languages, and therefore they can be used for research in terminology, lexicography, in machine translation, in computer-assisted language learning and in cross-linguistic information retrieval.

2. Corpus

Investigating parallel texts could lead us to preliminary conclusions regarding the differences between several related languages. In this case, parallel texts consisting of newspaper articles originally written in English and translated into nine languages, among which are Croatian (CR), Serbian (SE) and Bosnian (BS) were retrieved from the daily news site *Southeast European Times*[1]. Texts cover news and developments in Southeast Europe. Each corpus consists of 1,500 news documents translated to each language from the source English text, with each corpus comprising about 330,000 tokens, collected from July 2007 to April 2008. All examples are downloaded and given in the Latin script. Although parallel texts that are aligned at sentence or word level can be of considerable importance for further research, this case study was made on texts with aligned titles and paragraphs.

3. Levels of comparison

Although there are numerous historical and socio-cultural papers on Slavic languages, in this paper differences are studied:

- at the *phonological level* (e.g. use of -ije-/-je- in Croatian vs. -e- in Serbian),
- at the *morphological level* (e.g. use of -em/-om, or ending -čić or -če, inflection of abbreviations, different declensions in Croatian and Serbian or words differing in gender, e.g. second/sekunda/sekund, different verb forms),

[1] http://www.setimes.com

- at the *lexical level* (e.g. when different lexemes are used, or if words are similar but have different meanings or with pronunciation differences),
- at the *syntactic level* (e.g. more frequent use of infinitive constructions or nouns in Croatian, while in Serbian more frequent use of da constructions),
- at the *semantic level*.

3.1. Phonological level

The most obvious difference between Croatian and Bosnian on one side and Serbian on the other appears at the phonological level and concerns the reflex of the common Slavic vowel yat, which is rendered as -ije-/je in CR and BS, and as -e- in SR.

Another typical example is the -eu- diphthong in Croatian, which appears as -ev- in both Bosnian and Serbian.

In the case of loan-words derived from Greek containing -ch-, such as chemical, Christians, etc., Croatian uses -k- (*kemijski*, *kršćani*), Serbian uses -h- (*hemijski*, *hrišćani*), while both phonemes are found in Bosnian (*hemijski* vs. *kršćani*).

Croatian	Serbian	Bosnian	English
snijeg	sneg	**snijeg**	snow
povjerenje	poverenje	**povjerenje**	confidence
svjedok	svedok	**svjedok**	witness
njemački	nemački	**njemački**	German
Njemačka	Nemačka	**Njemačka**	Germany
Snježni	snežni	**sniježni**	snow
španjolski	**španski**	**španski**	Spanish
europski	evropski	**evropski**	European
kršćani	hrišćani	**kršćani**	Christians

Table 1

3.2. Morphological level

The morphosyntactic level shows consistent differences across the three languages. As these differences are very broad-ranging, touching upon the domains of morphophonology, morphology and syntax, this paper is not intended to provide a full list or formal classification of such differences, but rather an in-depth exploration of several phenomena we found to be the most representative and informative with respect to the three languages.

Croatian	Serbian	Bosnian	English
predložit će	predložiće	će predložiti	to propose
započet će	počeće	će početi	to open
sastat će se	sastaće se	održat će	to meet
izabrat će	izabraće	će birati	to elect
posjetit će	posetiće	će posjetiti	to visit

nastavit će	nastaviće	**nastavit će**	to continue
predložit će	predložiće	će predložiti	to propose
akcijski plan	**akcioni plan**	**akcioni plan**	action plan
nacionaliziran	nacionalizovan	**nacionaliziran**	nationalised
kritiziraju	kritikuju	**kritiziraju**	criticise
vršitelj dužnosti premijera BIH	**vršilac dužnosti premijera BIH**	**vršilac dužnosti premijera BIH**	acting BiH prime minister
tužitelj	**tužilac**	**tužilac**	public prosecutor

Table 2

At the morphological level several rules could be identified:

- for future tense, Croatian and Bosnian use the analytic model (verb in the infinitive form preceded or followed by the auxilliary verb) as in *sastat će se/ će se sastati, izabrat će/ će birati*, while Serbian uses the synthetic model, merging the two words and omitting the consonant *–t*, as in *sastaće se*, *izabraće*, etc.
- while the infix *-ij/-ir* is more used in the Croatian (e.g. *akcijski*, *nacionalizirati*) the Serbian uses more –io/-o (e.g. *akcioni, nacionalizovan*)
- Serbian and Bosnian use the suffix –lac to denote the agent, while Croatian generally uses the suffix –telj

3.2.1. Names

In some text genres, names are very important because they cover up to 10 percent of all tokens in text. As we are conducting our study on informative texts, we consider them as inevitable part of language comparison.

Croatian	Serbian	Bosnian	English
Burgas-Alexandroupolis	Burgas-Alexandroupolis	**Burgas-Alexandroupolis**	**Burgas-Alexandroupolis**
Bulqiza	Buljiza	**Bulqiza**	**Bulqiza**
New York	**Njujorku**	**Njujork**	**New York**
Barroso	Barozo	**Barroso**	**Barroso**
Rehn	Ren	**Rehn**	**Rehn**
Papandreou	Papendreu	**Papandreou**	**Papandreou**
Albright	Olbrajt	**Albright**	**Albright**
Di Carlo	Dikarlo	**Di Carlo**	**Di Carlo**
Rice	Rajs	**Rice**	**Rice**
Tariceanu	Taričanu	**Tariceanu**	**Tariceanu**

Table 3

As presented in table 3, names are spelled in Croatian and Bosnian[2] as they are in the original language, while in Serbian,

[2] Except for the occurrence of the token *Njujork* (eng. New York) in Bosnian

names are transcribed to match the pronunciation. This is likely the result of the extensive use of the Cyrillic alphabet in Serbian.

3.3. Lexical level

The first level we investigated is lexical. The problem found in comparing the titles of the articles is a lack of consistent translation of corresponding lexemes, even though they are a part of the lexicon of the given language. Moreover, if the same root is used by translators in another language, it is very often used in a different POS category, e.g. CR: *poništenje* (noun) and BS: *poništi* (verb), or the same word has a different MSD (e.g. different inflectional cases). Lemmatization of all texts would make this step considerably easier, but since no lemmatizers were available for Bosnian and Serbian, we focused our efforts on the manual analysis of characteristic lexemes. The following examples are gathered from the corpora, with identical tokens marked bold:

Croatian	**Serbian**	**Bosnian**	**English**
glede	u pogledu	u vezi	on/of/about/regarding
sigurnost	bezbednost	**sigurnost**	security
izvijestio	informisao	informirao	reports
paralizirao	paralisao	**paralizirao**	paralyses
tisuće	**hiljade**	**hiljade**	thousands
vanjskih	inostranih	**vanjskih**	foreign
Cipar	**Kipar**	**Kipar**	Cyprus
kompanije	**kompanije**	firme	company
tvornica	**fabrika**	**fabrika**	plant
opovrgava	demantuje	porekla	denies
crnogorski DPS	**crnogorski DPS**	crnogorska DPS	Montenegro's DPS
izjavio	**izjavio**	**izjavio**	says
s/sa	s	**s/sa**	with
diplomacija	**diplomatija**	**diplomatija**	diplomacy
točka	**tačka**	**tačka**	point
suradnja	**saradnja**	**saradnja**	co-operation
najviše sigurnosno tijelo	**najviše** bezbednosno telo	vrhovno **sigurnosno tijelo**	constitutional Court officials
vijeće	savet	**vijeće**	council
osiguranje	obezbeđivanje	obezbjeđuje	provide
reagirati	reagovati	**reagirati**	respond
zračni	vazdušni	**zračni**	air
vanjski	inostrani	**vanjski**	foreign
usmjerava	koncentriše	koncentrira	concentrate

Table 4

We found all possible combinations of lexemes overlapping across the languages, i.e. overlapping lexeme pairs in CR-SR, BS-SR, CR-BS and CR-SR-BS. There are lexical spots with different lexical choices for all three languages, as was the case with the English word denies. In the Bosnian language, a hybrid combination of the same lexical morpheme as in Serbian and the same grammatical morpheme typical for Croatian is frequently found (e.g. in Table 4, BS *koncentrira*, HR *usmjerava*, SR *koncentriše*).

3.3.1. Acronyms

Another interesting phenomenon we investigated were acronyms. None of the three languages treats acronyms consistently when it comes to morphological properties. Thus, EU is inflected as a feminine noun in certain instances, and as a masculine noun in others. This is likely caused by the fact that the headword of the acronym, *unija* ('union') is a feminine noun in all three languages; however, the acronym itself 'sounds' more like a masculine noun. Therefore, the actual use of the acronym may vary from one translator or text to another. On the other hand, certain acronyms displayed consistent differences across the three languages. For example, SAD ('USA') is treated as a plural feminine noun in both Bosnian and Serbian, presumably motivated by the fact that the headword *države* ('states') is plural feminine, while in Croatian it is treated as a singular masculine noun (again, probably because the acronym itself has the properties of a typical singular masculine noun).

Croatian	Serbian	Bosnian	English
Tužitelji **ICTY-a**	Tužioci MKSJ	Tužioci **ICTY**	ICTY prosecutors
Žalbeno vijeće **UN-a**	Žalbeno veće UN	Apelacioni sud **UN-a**	UN appeals court
dužnosti predsjednika Glavne skupštine UN-a	funkciji predsednika Generalne skupštine **UN**	dužnosti predsjednika Generalne skupštine **UN**	UN General Assembly president priorities

Table 5

It is evident from the above examples that abbreviations can either be translated, as in Serbian (e.g. *MKSJ*), or remain the same as in the original language (e.g. *ICTY*), which is the case in Croatian and in Bosnian. In the Croatian language, abbreviations are inflected (e.g. *tužitelji ICTY-a, žalbeno vijeće UN-a*), while in Serbian, they are generally translated (e.g. *MKSJ*) and remain uninflected (e.g. žalbeno veće UN), and in Bosnian, the abbreviation appears in the same form as the original, but can be either uninflected (e.g. *tužioci ICTY*) or inflected (e.g. *apelacioni sud UN-a, dužnosti predsjednika Generalne skupštine UN*).

3.4. Syntactic level

3.4.1. Prepositions, verb phrases

The preposition 'with' is highly frequent preposition (ranked as 9th on the frequency list) and it can appear in two forms in CR and BS, namely *s* or *sa*, depending on the word which follows preposition. Although the form *s* is 3 times more frequent than *sa* in CR and BS, we found less than 2% of that form occurring in SR translation.

Croatian	Serbian	Bosnian	English
zabrinuta *zbog* neuspjele *ratifikacije* CEFTA-e	zabrinuta *zbog* neuspeha *da ratifikuje* CEFTU	zabrinuta *zato što nije ratificirao* CEFTA-u	failure *to ratify* CEFTA
pokušava izabrati	se trudi da izabere	izbor	to elect
će prestati s uporabom	će prestati da koriste	će prestati s korištenjem	to stop using
OESS priopćio *kako* nema potrebe	OEBS saopštila *da* nema potrebe	OSCE saopćio *da* nema potrebe	OSCE says no need *to* monitor

Table 6

Regarding syntactic expressions the following differences have been found:

- the Croatian language uses more infinitives (*pokušava izabrati*) and noun constructions (*ratifikacije, uporaba*), similar as in Bosnian, while in the Serbian more verb constructions are used, especially *da* + verb (*da ratifikuje, da izabere, da koriste, da nema potrebe*)
- different prepositions are translated in different ways, e.g. 'failure *to ratify* CEFTA' the preposition to is translated in Croatian and Serbian by preposition *zbog* and in Bosnian *zato što*
- different conjunctions are used for the expression '*no need to monitor*' in the Croatian *kako* and in Serbian and Bosnian *da*
- different parts of speech are used in e.g. 'failure to ratify CEFTA', where to ratify is translated by noun in Croatian (*ratifikacija*), verb construction in Serbian (*da ratifikuje*) or past verb construction in negative form in Bosnina (*nije ratificirao*)
- different positive/negative forms, e.g. failure to ratify, is translated in Croatian by adjective (*neuspjele*) and by noun in Serbian (*neuspeh*) while in Bosnian is translated by negative verb form (*nije ratificirao*)
- the abbreviation CEFTA is inflected in Croatian and Bosnian by analytic form (*CEFTA-e, CEFTA-u*) or by synthetic form (*CEFTU*)

3.4.2. Noun phrases

Croatian	Serbian	Bosnian	English
Vijeće sigurnosti UN-a	Savet bezbednosti UN	**Vijeće sigurnosti UN-a**	UN Security Council
Žalbeno vijeće **UN-a**	**Žalbeno** veće UN	Apelacioni sud **UN-a**	UN appeals court
Članovi **EP-a**	Članovi EP	Članovi **EP-a**	EP members
izvjestitelji **PACE-a**	izvestioci PSSE	izvještači **PACE-a**	PACE rapporteur
kazao OESS-u	rekao OEBS-u	**kazao** OSCE-u	tells OSCE that
zatvori **CIA-e**	zatvori CIE	zatvori **CIA-e**	CIA prisons
Šef **EUPM-a**	Šef EUPM	Šef **EUPM-a**	EUPM chief

Table 7

Examples presented in table 7 show that various differences exist between the three Slavic languages at various levels within phrases:

- at the syntactic level in the three Slavic languages noun phrases are presented in the form of nominative + genitive (*Vijeće sigurnosti UN-a/ Savet bezbednosti UN; Članovi EP-a/ Članovi EP*) contrary to the English (UN Security Council; EP members)
- at lexical level in Croatian and Bosnian mainly the same word is used (*Vijeće sigurnosti, kazao*) and in the Serbian (*Savet bezbednosti, rekao*)
- at morphological level the Croatian uses –ije/je construction (*vijeće, izvjestitelji*) contrary to the Serbian –e (*veće, izvestioci*), while the Bosnian used another lexeme (*sud*) or –č construction (*izvještači*)
- the inflection is applied to abbreviations in Croatian and Bosnian (UN-a, EP-a, CIA-e), contrary to the Serbian where it is either not applied (UN, EP) or is integrated into the abbreviation (CIE).

3.5. Semantic level

It is reasonable to assume that the differences at the semantic level would be considerably more obvious, if texts were taken from the general or from the cultural domain. Although there are common lexemes in all three Slavic languages, they can have different meanings, such as '*čas'* and '*trenutak'* meaning one moment or one second which both exist in Croatian as partial synonyms, while in the Serbian '*čas'* denotes one hour. While in the Croatian the word '*tajnica'* is used as the

equivalent for the English word secretary, in Serbian and Bosinan, the word '*sekretarka'* is used. In Croatian, the collocation '*državni sekretar*' does exist, in the sense of 'secretary of state', but the feminine form, '*sekretarka'* does not exist. The word 'persons' is translated in the Serbian by '*lica'*. In the Croatian the same word denotates face, and persons translate as '*osobe'*.

4. Conclusion

Parallel corpora are valuable resources which provide insight into similarities and differences between the three languages, thereby facilitating the development of tools customized for each language, taking into the account their distinctive characteristics. To the best of our knowledge, there are no prior works or methodologies for measuring similarities between related languages which could be numerically expressed or quantified. Although they are genetically and historically related, it is evident even from this limited case study that standards are different. As the presented examples are neutral in style and deal with international relations, the differences are considerably smaller regarding syntactic constructions and lexemes, reflecting cultural differences. Many Bosnian lexemes mostly overlap with Croatian and Serbian, but there is a small number of lexemes appearing in Bosnian only.

We consider this work as a first step in establishing the criteria and methodology for measuring similarities between languages. From the perspective of comparison of Croatian, Serbian and Bosnian, it is still hard to draw statistical results; the main reason is clarity of criteria which would be used for benchmarking. Empirical analysis has shown that a huge portion of differences across the three languages are systematic and regular, and as such, could be formalized for automatic translation/generation. Differences among languages should be presented in systematic and clear manner, reflecting identity differences; otherwise their use in machine translation, in lexicography, terminology, natural language processing, text summarization or in computer-assisted language learning may give misleading results.

Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports of the Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1002, and 130-1300646-0909.

References

Barić, E.; Lončarić, M.; Malić, D.; Pavešić, S.; Peti, M.; Zečević, V. & Znika, M. 1997. *Hrvatska gramatika*. Zagreb: Školska knjiga.

Bosanski jezik (http://hr.wikipedia.org/wiki/Bosanski_jezik), August 2008.

Hrvatski jezik – posleban slavenski jezik. (http://hjp.srce.hr/index.php?show=povijest&chapter=34-poseban_jezik), August 2008.

Izjava Hrvatske akademije znanosti i umjetnosti o položaju hrvatskoga jezika. Časopis za kulturu hrvatskoga književnog jezika. Zagreb: Hrvatsko filološko društvo 2. Jezik 52, 41-80, 2005. (http://hrcak.srce.hr/file/24183)

Razlike i sličnosti. *Vijenac* 232, 2003.
(http://www.matica.hr/Vijenac/vij232.nsf/AllWebDocs/DaliborBrozovicPRVOLICEJEDNINE), August 2008.

Resnik, Ph. & Smith, N.A. 2003. The web as a parallel corpus, *Computational Linguistics* 29 (3), pp. 349-380.

Silberztein, M. 2008. *NooJ Manual, v.2*., (http://www.nooj4nlp.net), May 2008.

Southeast European Times, (http://www.setimes.com)

Stevanović, M. 1991. *Savremeni srpskohrvatski jezik*, Beograd: Naučna knjiga.

AFFECTED ARGUMENTS CROSS-LINGUISTICALLY

Solveig Bosse, Benjamin Bruening, MaryEllen Cathcart,
Anne E. Peng, Masahiro Yamada
University of Delaware*

ABSTRACT

We argue that affected arguments, common cross-linguistically, are introduced by a syntactic head Aff(ect). Possible variation in the height of the attachment of this head as well as its (non-)assertive content explain language variation. We focus on Albanian and Japanese, with some remarks on Hebrew.

1. Introduction

Many languages allow sentences to include an NP that is not selected by the verb. This NP is optional and is interpreted as affected in some way by the verbal event. An Albanian example is given below, where the affected NP is marked with dative case.[1]

1. Agim-i **i**-a theu [vazon e Ben-it] **Dritan-it**.
   Agim-Nom **3S.Dat**-3S.Acc broke [vase.Acc AD Ben-Gen] **Dritan-Dat**
   'Agim broke Ben's vase on Dritan.'
   = 'Agim broke Ben's vase, and this matters to Dritan (negatively or positively[2]).'

In this paper, we propose a semantic and syntactic analysis of affected arguments that is able to account for cross-linguistic variation in their syntax and semantics based on two parameters of variation. First, the affected argument may be introduced at different places in the syntactic structure (attachment height), and second, languages may vary in how much of the semantics is included in the assertive content. In our analysis, the affected argument is introduced by a syntactic head, whose semantic contribution consists of part assertion, part presupposition. We contend that languages can vary in how much of the semantics is assertion and how much presupposition, with syntactic consequences following from the choice.

2. Background: Event Semantics

In the event semantics we assume, verbs are understood as properties of events, and they may take an internal argument. As an example, the denotation of the verb *hit* is as shown in (2).

2. $[\![\text{hit}]\!] = \lambda x.\lambda e.\ \text{hit}(e)\&\text{Thm}(e,x)$

Following Kratzer (1996), we assume that external arguments are not arguments of the verb, but are introduced by a higher functional head, Voice:

3. $[\![\text{Voice}]\!] = \lambda x.\lambda e.\ \text{Agt}(e,x)$

VP and Voice combine via Event Identification, as follows:

---

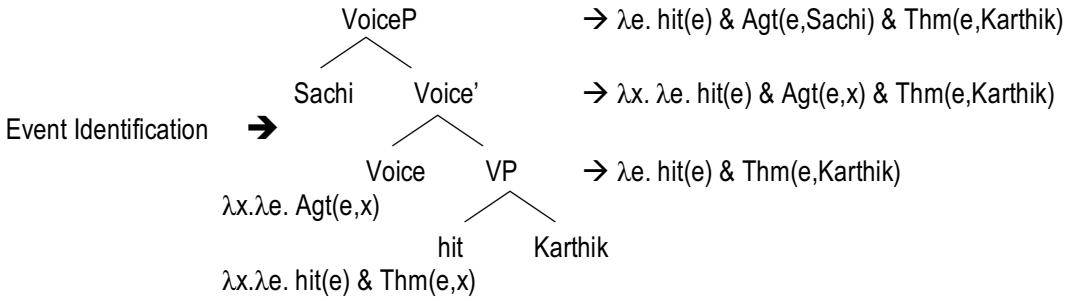
* Authors are in alphabetical order.

[1] Abbreviations: Nom=Nominative, Acc=Accusative, Dat=Dative, Gen=Genitive, AD=Adjectival Determiner, 1,2,3=1st,2nd,3rd person, S=Singular, CL=Classifier, Pass=(Adversity) Passive, Past=Past tense, Q=Question

[2] Whether the affectedness is negative or positive depends on the utterance context.

4. $f_{<s,t>}$ $g_{<e,st>}$ → $h_{<e,st>}$
   λe. f(e) λx.λe. g(x)(e) λx.λe. g(x)(e)&f(e)

5. Sachi hit Karthik.



Thus, VoiceP denotes a set of hitting events whose agent is Sachi and whose theme is Karthik. This is the desired denotation of *Sachi hit Karthik*, ignoring tense and other higher functional elements.

3. Analysis

We claim that the affected argument is introduced by a syntactic head *Aff(ect)*. This head does three things: (i) it is an identify function, passing up the predicate of events denoted by its sister; (ii) it introduces another event, namely a semantically bleached *experiencing* event, and takes an argument that is the *experiencer* of this event; (iii) and it introduces a presupposition that the *source* of the experience is the event denoted by its sister.

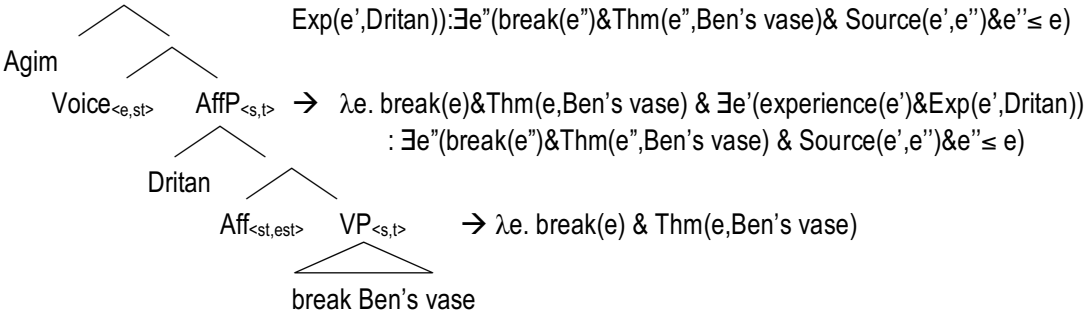
6. $[\![Aff]\!] = \lambda P_{<s,t>}.\lambda x.\lambda e.\ P(e)\&\exists e'(experience(e')\&Exp(e',x)) : \exists e''(P(e'')\&Source(e',e'')\&e''\leq e)$

The material to the right of the colon is the presupposition/conventional implicature.[3] What is presupposed is an event e'' that is identical to or a subpart of the event of the VP (*e''≤ e*), and this event is the source of the experiencing event e'. Thus, the experiencer of the event e' is in his/her cognitive state because of *e''*.

3.1. Albanian

Here, we show how Aff yields the desired interpretation of the Albanian sentence in (1). Aff is merged below the Voice head in Albanian (we motivate this below). Therefore, the derivation of (1) proceeds as shown in (7).

7. $VoiceP_{<s,t>}$ → λe.break(e)&Agt(e,Agim)&Thm(e,Ben's vase)& ∃e'(experience(e') & Exp(e',Dritan)):∃e"(break(e")&Thm(e",Ben's vase)& Source(e',e'')&e''≤ e)



[3] It is not clear whether it is a presupposition or a conventional implicature, or whether those are two different things. All that matters here is that it is not part of the assertion.

Thus, the VoiceP denotes a set of breaking events whose agent is Agim and whose theme is Ben's vase. Furthermore, there is another event, an experiencing event, of which Dritan is the experiencer. The presupposition says that the breaking of Ben's vase is the source of Dritan's experience.

Note that in this analysis, the existence of the experiencing event and the experiencer are part of the assertive content, but the source of the experience is not. This is because the presupposition of the source survives under negation and in yes-no questions (not shown):

8. Dritani-t nuk i vdiq Besa.
   Dritan-Dat Neg 3S.Dat died Besa.Nom
   'Besa didn't die on Dritan.'
   a. Besa didn't die (and if she had, it would have mattered to Dritan);
   b. *Besa died, but it didn't matter to Dritan.

But the experiencer is part of the asserted content, because it can be extracted or it can be a quantifier binding a variable elsewhere in the assertion:

9. **Kujt** i-a kafshoi qen-i mace-n e Bes-ës?
   who.Dat 3S.Dat-3S.Acc bit dog-the.Nom cat-the.Acc of Besa-Gen
   'On whom did the dog bite Besa's cat?'
10. I-a theva **çdo** **djali-t** saksinë e **tij**.
    3S.Dat-3S.Acc I.broke **every** **boy-Dat** vase.the.Acc of **his**
    'I broke his$_1$ vase on every boy$_1$.'

An additional feature of this analysis is that sentences with affected arguments are bi-eventive. The VoiceP is a predicate of events, and the Aff head existentially introduces another event, an experiencing event. We therefore make predictions regarding adverbial modification, in a particular way. We assume that VP adverbs modify predicates of events (i.e. syntactic nodes of type <s,t>), so that *Gertie attacked Peedie violently* would be *λe. attack(e) & Thm(e,P) & Ag(e,G) & violent(e)* (see Parsons 1990). In our analysis, there is no predicate of events related to the experiencing event; that event variable is introduced by an existential quantifier. We therefore predict that VP adverbs will only modify the VP or VoiceP event, and not the experiencing event, which is correct:

11. Dritan-it i vdiq i vëllai **në** **Tiranë**.
    Dritan-Dat 3S.Dat died his brother.Nom **in** **Tirana**
    'Dritan's brother died on him in Tirana.'

In (11), the PP modifier necessarily modifies the dying event. The experiencing event need not take place in Tirana, though; Dritan could hear the news in some other location and be affected.

Although it cannot be modified by VP adverbials, the experiencing event is present in the semantics in our analysis and should be able to be modified in other ways. This is also correct, but we defer a demonstration of this to the section on Japanese below.

This bi-eventivity distinguishes our analysis from other possible ones, such as that of Pylkkänen (2002). In Pylkkänen's analysis, an Applicative head introduces a malefactive argument. This head combines with the VP in essentially the same way as Voice, above, by Event Identification. In Pylkkänen's analysis of (11), then, the locative modifier *in Tirana*, the applied argument *Dritan*, and *Dritan's brother's dying* all share one event variable bound by the same lambda-operator:

12. ⟦(11)⟧= λe. die(e)&Thm(e,Dritan's brother)&InTirana(e)&Mal(e,Dritan)

Pylkkänen's analysis therefore predicts that malefactive arguments should pattern with agents in how they are treated by adverbs that modify the event. Take a transitive sentence like *Dritan killed his brother in Tirana*. It is very difficult to conceive of this situation as not having Dritan in Tirana, whereas that is very easy in (11). The malefactive should also pattern like the goal of a double-object sentence in Pylkkänen's analysis (goals are also introduced by Applicative heads in her analysis), as in *I bought Dritan a car in Tirana*. Again, the default interpretation of this sentence has Dritan in Tirana (in contrast with *I bought a car for Dritan in Tirana*, which does not; compare #*While he was in London, I bought Dritan a car in Tirana* with *While he was in London, I bought a car for Dritan in Tirana*). Our analysis therefore captures intuitions about modification better than an analysis with just one event variable.

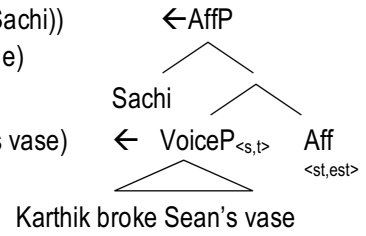
3.2. Japanese

We turn to Japanese, to further corroborate our analysis and to illustrate one of the points of variation. Japanese has a structure known as an *adversity passive* (Fukuda 2004, Kuno 1973, among others). Adversity passives also involve an affected argument, as shown below.

13. **Sachi-ga** Karthik-ni Sean-no kabin-o kowas-are-ta.
    Sachi-Nom Karthik-by Sean-Gen vase-Acc break-Pass-Past
    `Sachi had Sean's vase broken on her by Karthik.'
    = 'Karthik broke Sean's vase. This matters to Sachi, i.e. Sachi is affected by Karthik's breaking of Sean's vase.'

In Japanese, the affected argument receives nominative case (here: *Sachi-ga*). In our analysis, the affected argument is merged above VoiceP, rather than between VoiceP and VP as in Albanian. The derivation is shown below.

14. λe. broke(e)&Agt(e,K)&Thm(e, Sean's vase) & ∃e'(experience(e')&Exp(e',Sachi))
    : ∃e"(broke(e")&Agt(e",K)&Thm(e",Sean's vase) & Source(e',e") & e"≤ e)



This high attachment in Japanese versus the low attachment in Albanian leads to a difference in what is interpreted as the source of the experiencing event. In the case of low attachment, the VP is the source of the experience. That is, the external argument is not included in the source. In the case of high attachment, VoiceP, including the external argument, is the source of the experience:

15. **Low attachment:** λe.broke(e)&Agt(e,Agim)&Thm(e,Ben's vase)&∃e'(experience(e')&Exp(e',Dritan))
    : ∃e"(broke(e")&Thm(e",Ben's vase) & Source(e',e")&e"≤e)

16. **High attachment:** λe. broke(e)&Agt(e,Karthik)&Thm(e, Sean's vase) & ∃e'(experience(e')&Exp(e',Sachi))
    : ∃e"(broke(e")&**Agt(e", Karthik)**&Thm(e",Sean's vase) & Source(e',e")&e"≤e)

This is because what is interpreted as the source is the sister of Aff. In Albanian, only the VP is the sister of Aff; Voice is higher. In Japanese, the entire VoiceP is the sister of Aff.

Evidence for this difference comes from possible interpretations of the sentences. In low-attaching Albanian, it is impossible to attribute the source of the affectedness to the external argument (17). However, this is possible in high-attaching Japanese (18).

17. Bir-i m'a kafshoi Bes-ën. (Albanian)
son-the.Nom 1S.Dat-3S.Acc bit Besa-Acc
'The son bit Besa on me.'
='The son bit Besa and it matters to me because it was Besa/**#because it was the son**.'
18. Sachi-ga Karthik-ni Sean-o kam-are-ta. (Japanese)
Sachi-Nom Karthik-by Sean-Acc bite-Pass-Past
'Sachi had Sean bitten on her by Karthik.'
= 'Karthik bit Sean and it matters to Sachi because it was Sean/ **because it was Karthik**.'

This analysis also predicts different c-command relations in the two languages. In Albanian, the agent asymmetrically c-commands the affected argument, but in Japanese, it is the other way around. That this is correct can be seen from the possibility of variable binding by quantifiers. In Japanese, the affected nominative argument can bind a variable inside the agent, but the agent may not bind a variable inside the affected argument:

19. Japanese (HIGH)
a. [Go-nin-ijoo-no kodomo]$_1$-ga [sono-ko$_1$-no hahaoya]-ni odor-are-ta.
[five-CL-more.than-Gen child]-Nom [it-child-Gen mother]-by dance-Pass-Past
'More than five children$_1$ had his/her$_1$ mother dance on him/her$_1$.'
b. *[Sono-ko$_1$-no hahaoya]-ga [go-nin-ijoo-no kodomo]-ni odor-are-ta.
[it-child-Gen mother]-Nom [five-CL-more.than-Gen child]-by dance-Pass-Past
Lit. 'His/her$_1$ mother had more than five children$_1$ dance on her.
Intended: # 'More than five children$_1$ danced on his/her$_1$ mother.'

But in Albanian, our theory predicts that the agent can bind a variable inside the affected argument, but not the other way around.[4]

Just like Albanian (9,10), Japanese allows the affected argument to be questioned or to bind a variable within its scope, indicating that the experiencer is part of the assertive content:

20. **Dare**-ga Karthik-ni odor-are-ta no?
who-Nom Karthik-by dance-Pass-Past Q
'Who had Karthik dance on them?'
21. [**Go-nin-ijoo-no hito**]$_1$-ga Karthik-ni **jibun$_1$**-no heya-de odor-are-ta.
[**five-CL-more.than person**]-Nom Karthik-by **self**-Gen room-in dance-Pass-Past
'More than five people$_1$ had Karthik dance on them in self's$_1$ room.'

But, again, the source of the experience is presupposed, as it survives under negation (and in yes-no questions):

22. Sachi-wa Karthik-ni odor-are-nakat-ta.
Sachi-Top Karthik-by dance-Pass-Neg-Past
'Sachi didn't have Karthik dance on her.' (presupposition: if he had danced, it would have mattered)

---

[4] Albanian data is to be collected. Below are examples from German, which patterns with Albanian. A pronominal in the dative-marked affected argument *ihren Vorgesetzten* 'their superiors' in (i) can be bound by the quantificational agent *alle* 'everyone', while the quantificational affected argument cannot bind a variable inside the agent (ii).

i) Alle$_1$ haben ihren$_1$ Vorgesetzten den Dienst quittiert.
everyone have their.Dat superiors the service quit
'Everyone has quit the service on their superior.'

ii) *Ihre$_1$ Arbeiter haben allen$_1$ den Dienst quittiert.
their workers have everyone.Dat the service quit

Thus, Japanese and Albanian differ only in the height of the attachment of the Aff head. The semantics of the head is identical; all that differs is the input to Aff: VP in Albanian, VoiceP in Japanese.

We now return to adverbial modification. As stated above, if VP modifiers can only modify predicates of events, then we predict that any VP modifier will only modify the VP/VoiceP event, and not the experiencing event. This is true in Japanese:

23. (*Totemo) Sachi-ga Karthik-ni hageshiku odor-are-ta.
    (*very.much) Sachi-Nom Karthik-by enthusiastically dance-Pass-Past
    'Sachi had Karthik enthusiastically dance on her (*affecting her very much).'
24. (*Osaka-de) Sachi-ga Karthik-ni Tokyo-de shin-are-ta.
    (*Osaka-in) Sachi-Nom Karthik-by Tokyo-in die-Pass-Past
    'Sachi (*in Osaka) had Karthik die on her in Tokyo.'

In (24), Sachi does not have to be in Tokyo to be affected, as in Albanian, but a PP modifier cannot pick out just the experiencing event. However, a clausal modifier can:

25. Osaka-ni iru toki, Sachi-ga Karthik-ni Tokyo-de shin-are-ta.
    Osaka-in be when, Sachi-Nom Karthik-by Tokyo-in die-Pass-Past
    'When she was in Osaka, Sachi had Karthik die on her in Tokyo.'

This follows in our analysis, because there is an experiencing event in the semantics. Clausal modifiers, we assume, do not need predicates of events, unlike VP adverbials and PPs. Note the contrast with an external argument:

26. #Osaka-ni iru toki, Sachi-ga Karthik-o Tokyo-de koroshi-ta.
    Osaka-in be when, Sachi-Nom Karthik-Acc Tokyo-in kill-Past
    Literally: #'When she was in Osaka, Sachi killed Karthik in Tokyo.'

Such sentences are distinctly odd. Another verb needs to be added, as in (27).

27. Osaka-ni iru toki, Sachi-ga Karthik-o Tokyo-de koros-ootoshi-ta.
    Osaka-in be when, Sachi-Nom Karthik-Acc Tokyo-in kill-manage.to-Past
    'When she was in Osaka, Sachi managed to kill Karthik in Tokyo.'

This follows, we contend, because there is only a single event in a simple transitive.

3.3. Hebrew

We turn now to our second parameter of variation, how much of the semantics of the Aff head is asserted. We suggest that affected arguments in Hebrew *ethical datives* are entirely presupposed and are not part of the assertive content at all. The denotation of Aff in Japanese and Albanian is repeated below:

28. $[\![\text{Aff}]\!] = \lambda P_{<s,t>}.\lambda x.\lambda e.\ P(e)\&\mathbf{\exists e'(experience(e')\&Exp(e',x))} : \exists e''(P(e'')\&Source(e',e'')\&e''\leq e)$

As we showed above, the experiencer is part of the assertive content: it can be extracted or it can be a quantifier binding a variable elsewhere in the assertive content. This is not true in Hebrew. As Borer and Grodzinsky (1986) showed, ethical datives may not be questioned:

29. a. Hem kol ha-zman mitxatnim **li**. (Borer & Grodzinsky 9a)
they all the-time marry **to-me**
'They are getting married on me all the time (and it bothers me).'

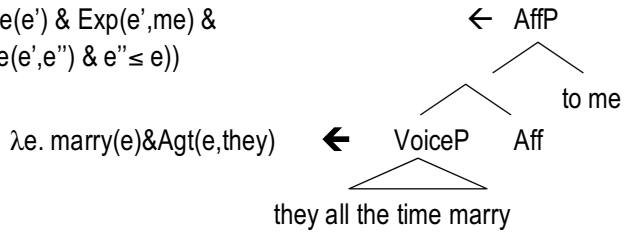
b. ***lə-mi** hem mitxatnim kol ha-zman? (Borer & Grodzinsky 11)
to-who they marry all the-time

In Hebrew, the experiencer is syntactically inert. We argue that in Hebrew the entire semantic contribution of Aff is a pressuposition, as in (30):

30. $[\![\text{Aff}_{\text{Hebrew}}]\!] = \lambda P_{<s,t>}.\lambda x.\lambda e.\ P(e) : \mathbf{\exists e'(experience(e')\&Exp(e',x)}\&\exists e''(P(e'')\&Source(e',e'')\&e''\leq e))$

We assume that Aff attaches high in Hebrew (like Japanese), although more data is needed to support this. (31) shows the derivation with $\text{Affect}_{\text{Hebrew}}$ for (29a).

31. λe. marry(e) & Agt(e,they) : ∃e'(experience(e') & Exp(e',me) &
∃e"(marry(e") & Agt(e'',they) & Source(e',e'') & e"≤ e))



In Hebrew, the experiencing event, including the experiencer, is a presupposition, along with the source of the experience. It is not part of the assertive content at all. Hence it cannot interact with anything that is part of the assertive content of the sentence, to produce a wh-question, for instance. Thus, syntactic and semantic consequences follow from how much of the denotation of Aff is part of the assertion.

Although we have limited our discussion to Albanian, Japanese, and Hebrew, we believe that this analysis can extend to numerous other languages that have affected arguments.

References

Borer H., Grodzinsky, Y. 1986. Syntactic Cliticization and Lexical Cliticization: The Case of Hebrew Dative Clitics. In: H. Borer (ed.): *Syntax and Semantics* 19, 175 -215. New York: Academic Press.

Fukuda, Shin. 2006. Japanese passives, external arguments, and structural case. In Henry Beecher, Shin Fukuda, and Hannah Rohde (eds.), San Diego Linguistics Papers 2. CA. 85-133.

Kratzer, A. 1996. Severing the External Argument from its Verb. In J. Rooryck & L. Zaring (eds.): *Phrase Strcture and the Lexicon*. Dordrecht: Kluwer Academic Publishers.

Kuno, S. 1974. *Structure of the Japanese Language*, MIT Press, Cambridge, MA.

Parsons, T. 1990. *Events in the Semantics of English. A study in Subatomic Semantics*, MIT Press, Cambridge, MA.

Pylkkänen, L. 2002. *Introducing Arguments.* Ph.D. Thesis. MIT.

LINKING A DIGITAL DICTIONARY ONTO ITS SOURCES

Dan Cristea[1, 2], Marius Răschip[1]

[1] Alexandru Ioan Cuza University of Iaşi
[2] Romanian Academy, Iaşi
E-mail: {dcristea, mraschip}@info.uaic.ro

ABSTRACT

We describe an approach to build the digital version of the Thesaurus Dictionary of the Romanian Language, an explanatory dictionary built under the auspices of the Romanian Academy since 1913. The electronic version is called eDTLR and will include, apart from the dictionary itself, the sources linked by citations to the dictionary entries, and the software to access them. We will concentrate in this paper on the structure, the technology to achieve it and the accessing capabilities over eDTLR.

1. Introduction

The Thesaurus Dictionary of the Romanian Language, built under the auspices of the Romanian Academy since 1913, contains 33 volumes, more than 15,000 pages, about 175,000 entries and more than 1,300,000 examples. The spectacular feature of this dictionary is its extremely rich collection of citations: each sense or sub-sense of each head-word is exemplified by citations that cover all periods of the written Romanian literature, more than 3,300 volumes. The electronic version is called eDTLR and will include, apart from the dictionary itself, the sources in digital form linked by citations to the dictionary entries, and the software to access them.

When browsing a dictionary for words and word senses, the common user, but also the lexicographer and the researcher, are often interested to find relevant contexts. On-line corpora are intermediated by software capable of revealing occurrences of words in contexts (without distinguishes among word senses, yet).

We will concentrate in this paper on the structure, the technology to achieve it and the accessing capabilities over eDTLR that will make possible to link the entries of the dictionary to its primary sources.

2. The structure of the dictionary

The dictionary entries are codified conforming to a subset of TEI P5[1] . Mainly, a dictionary entry resembles that of a graph in which the root is a head word, with the relevant information attached, and the nodes below it are word senses. Since a sense can have sub-senses, these – even more refined senses, a.s.o., the sense structure is that of a tree. Principal senses are placed immediately under the root and are usually paired with definitions, although is it not uncommon that definitions stay also on lower levels. Attached to nodes at all levels there could appear examples and expressions.

Examples are identified by unique labels of sources, and pages in the original editions. Examples are linked to the scanned pages that include the fragments of the examples. Linking the examples in the dictionary entries to images of the original editions is intermediated by a database and an index.

The structure of the database includes a number of tables which record: dictionary entries including examples, the original sources in scanned form and OCR-ed text, and visual pointing information identifying the exact rectangle surrounding an example. If an example is spread on two subsequent pages, the information kept in the database is capable to locate both parts. In fact, the database has a quite complex structure, capable to store and recuperate identification and content information about volumes, pages, lines and individual words.

Each example in a sub-sense of the dictionary is thus linked from the TEI representation to the scanned image in the original source.

The whole structure is also complemented by an index for approximate string matching that puts in connection each word form to all its possible occurrences in the corpus of OCR-ed texts.

[1] http://www.tei-c.org/Guidelines/P5/

3. The technology to achieve the structure

The whole production has been divided in 3 work flows:

3.1. Processing the dictionary

All volumes of the dictionary are now in electronic form either by scanning and OCR, either directly, by taking over the final, corrected, versions from the printing house. The very few volumes, produced lastly, have been received from the Academy directly in electronic form. But the vast majority has been scanned from paper versions and OCR-ed. As this process is prone to errors, a correction phase is necessary. The correction process is on-going now and will run for almost three years, partly using a collaborative approach (Cristea et al., 2008). When finished, we will have a corrected HTML form. This will be then filtered and parsed down to an XML TEI form. We have adopted a two steps parsing process: in the first run the sense tree structure is obtained, by paying attention mainly to field separators, and in the second run, the nodes on all levels are detailed, thus interpreting the morpho-syntactical information, definitions, citations, expressions, etymology, etc. This way we expect a success rate greater than parsing each entry completely guided by a grammar. Finally, the parsed version will have to be revised again by expert lexicographers.

3.2. Processing the sources

Out of the 3,300 volumes, more than 1,200 volumes are still under the copyright law (less than 70 years from their publication). We are now in the process to scan all volumes which are not under the restrictions of the copyright. A technology has been designed and implemented, which includes the following steps:

a) the volume is moved from the library to the company with which we have a contract for scanning;

b) the volume is scanned and the resulted files are stored on a local server;

c) the scanned images are downloaded on our server[2] ;

d) the scanned volumes enter a processing queue. Images are split into pages and text with visual pointing information is obtained by OCR[3]. The database is updated with the title of the book, all the identification information, and filled in with all information extracted from pages. To interpret Romanian texts written in Old Slavonic alphabet (405 volumes), we are currently training the Gamera[4] package.

e) a human operator, using a web interface, is checking page sequences, having the option to fill when missing or overwrite page numbers recognized automatically. Duplicates are removed and the missing pages or the ones requiring rescan are recorded in the database while a report is sent by email to the librarian. He checks if the volume is indeed displaying defects on the indicated places, and if yes, another copy is searched for and sent to the scanner from another library. This process only completes when the whole sequence of pages of the volume remains without errors in our database.

3.3. Linking the dictionary entries to sources

This operation aims at creating a link between each example reproduced in the dictionary entries and its image in the original scanned source. This link is intermediated by the OCR-ed copy of the volumes. It is in these documents that the example is first recuperated, by approximate string matching. Since the example identifies not only the source but also the page, the string matching should restrict the search into just one page (two, if the example is at the border of two consecutive pages). The database contains the recognized text and positions of lines and words in the scanned pages. Then, using this visual pointing information the example image location is determined and stored in the dictionary entry, together with the example.

[2] In order to cope with the vast amount of data of the described dictionary we have bought a Intel SSR212MC2R server equipped with 12 disks 1TB each, capable to store in a redundant manner up to 9 TB of useful information.

[3] http://www.irislink.com/c2-1360-189/iDRS-overview.aspx

[4] http://ldp.library.jhu.edu/projects/gamera

4. Browsing the dictionary

We have not yet inventoried the whole range of browsing capabilities that will be permitted in eDTLR. The most basic will allow the user to locate a head word in the dictionary and then to use an example in order to visualize it in the original page. In case when the volume is no more copyrighted, this display may be enlarged to include a wider context than that of the rectangle surrounding exactly the example.

But the index can be used to retrieve a lot more information. For instance, new contexts of a word, not contained in the dictionary entry set, can be located dynamically. Although the OCR-ed sources can include misspellings, these retrieval operations will certainly locate many intact occurrences. Once an occurrence is recognized, the user can have both the image of the context as well as its OCR-ed version, which he may use as a support version in the editing operation to remove misspellings.

5. Conclusions

The paper describes the organization, the process to achieve it and some of the browsing capabilities of a very large digital dictionary. The project is under development and will go on for two more years from now on. The correction phase will develop in parallel with advancing on the other phases. Tests on the parsing technology have shown an accuracy of 91% in discovering the entries' hierarchical structures. We are working to enhance this percent and to parse the nodes. Sources are scanned continuously, but we are unfortunately behind the schedule. If the present rhythm is kept we will arrive to only 70% of the original sources scanned till the end of the project, which is, fortunately, also the percent of the volumes which are no more under copyright. The web interface has been built and is now under tests on the new server. The human operator will begin to make use of it at month 8 of the project. Simultaneously, we are building the indexing solutions and enhancing the OCR results for both Latin and Old Slavonic. The alignment of scanned images with OCR results could also be used to further enhance, in a bootstrapping manner, the accuracy of the Romanian OCR, in both Latin and Old Slavonic characters, using as seeds the examples, known to be correct. When finalized, eDTLR will be one of the biggest digital dictionaries in the world, incorporating edge technologies for building and maintaining very large scale lexical and lexicographic resources.

References

Cristea, D.; Forăscu, C.; Răschip, M.; Zock, M. 2008. How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach, *LREC 2008*

PALATALIZATION AND UMLAUT

Monika Fischer

University of Szeged
10 Hattyas sor, 6725 Szeged, Hungary
fischermonika11@gmail.com

ABSTRACT

The paper is an attempt to present palatal assimilations in a unified way in a Government Phonology (GP) approach. Palatal assimilations involve fronting or raising, so they can be represented in this non-linear framework as spreading of the palatal element which contains these features. The proposed unified representation of umlaut and palatalization enables a typology of palatal assimilations. The typology clearly shows two important facts about palatal assimilations. One of them is that the key factor determining the direction of the processes is the major category of the trigger and the target of assimilation. If it coincides, that is, if both the trigger and the target are vowels or if they are both consonants, the palatal assimilation is regressive. If, however, the segments taking part in the process are of diverse major category, the process can be bidirectional. The second observation is that the former case is preferred by Germanic and the latter by Slavonic languages.

1. Introduction

The aim of the paper is to represent palatalization and umlaut in a unified way. Palatalization is used as a cover term for lexical and post-lexical, velar and coronal palatalizations. The term is also used to describe both the passive addition of secondary articulation and the active process of palatalization. Umlaut, on the other hand, signifies both the phonetically grounded diachronic phonological change as well as the morphologically conditioned alternation in Modern German.

Examples of post-lexical coronal palatalization in English be**t y**ou, hi**t y**ou [t] + [j] → [tʃ]
Examples of lexical velar palatalization in Serbian čove[**k**] – čove[**tʃ**]e 'human' NOM SG – VOC SG
Examples of umlaut in Modern German Baum – B**ä**ume 'tree' NOM SG – NOM PL

Figure 1: Examples of palatalization and umlaut

I consider palatalization and umlaut to be two sides of the same coin. They are both assimilation processes resulting in a front/palatal segment. Palatalization is a contact assimilation of consonants, umlaut is a distant assimilation of vowels. The former is by default regressive, whereas the latter can be both regressive and progressive. Palatalization is a place assimilation process and as such, one of the most frequent processes in the world's languages. It can be triggered by front high and front mid vowels, as well as the palatal glide, depending on the language.[1] Its counter-process, umlaut, is triggered by either a preceding palatal consonant (e.g. in Proto-Slavonic) or by a following front vowel (e.g. in the history of Germanic languages). Consequently, umlaut can also be progressive, as in the case of Slavonic languages, and regressive, as in the case of Germanic languages.

Palatality Harmony	**Velar palatalization**	**Umlaut**
Targets	velar Cs	velar / back vowels
Triggers	front vowels	front Cs / pal glide
Processes	fronting, raising	fronting, raising
Direction	regressive, progressive	regressive, progressive

Figure 2: Palatality Harmony in Proto-Slavonic

[1] The presence of the mid vowel in the palatalization context implies the presence of the high front vowel and the palatal glide; the presence of the high vowel implies the presence of the palatal glide in the context of palatalization.

In the theoretical framework I adopt in the paper, Government Phonology (Kaye, Jonathan, Jean Lowenstamm and Jean-Roger Vergnaud. 1985), key units are the unary elements – the building blocks of sounds. GP elements are embodiments of phonologically relevant contrasts such as palatality and roundness; in other words, they can be pronounced on their own and they do not need to combine with other elements (features) in order to form pronounceable articulations. The element inventory is the following:

I	'front' in vowels / 'palatal' in consonants
U	'round' in vowels / 'labial' in consonants
A	'low' in vowels / 'low' in uvular and pharyngeal consonants
R	'coronal' in consonants
N	'nasal' in both consonants and vowels
h	'noise', present in all released obstruents
?	'stop', present in oral and nasal stops and laterals
L	slack vocal cords
H	stiff vocal cords[2]

Figure 3: The inventory of elements in Government Phonology (Harris 1994)

Consequently, all processes can be reduced to spreading or delinking of elements. Needless to say palatalization and umlaut are cases of element-spreading, rather than delinking, although an element may also be delinked as a consequence.

In GP, the question of palatalization involves the issue of coronals and their representation with phonological primes. Namely, in velar palatalization, the coronal element (R) is present in the outcome of the process, although we do find it neither in the subject nor in the trigger of the process. Furthermore, the coronal element (R) is the only element present only in the element inventory of consonants and absent from the vocalic inventory.

2. Umlaut

Umlaut is the phonological process in which back vowels are fronted due to either a following front vowel, or a specific morpheme/morphological class. Umlaut in the Slavonic languages is similar to the one in the Germanic languages with respect to the outcome of the process. However, both the direction and the triggers of the change are different – in Proto-Slavonic, it is the preceding palatal (soft) consonant that makes the back vowel front. Similarly to Germanic languages, the phenomenon had a morphological effect: created soft (palatal) and hard (non-palatal) variants of roots (H. Tóth 1996: 79).

Old English /450/700 – 1100/[3]								
u [u:]	> y [y:] > i [i]	*m**u:**s	–	*m**y:**si	'mouse'	SG	–	PL
o [o:]	> e [e:]	*f**ō**djan	>	f**ē**dan	'food'	N	–	V
Proto-Slavonic /16thc. BC–4th c. AD/[4]								
		Proto-Sl		Old Church Slavonic				
u [u]	> y [ɨ] > i [i]	*š'**u**tei̯	>	**ši**ti	'to sew'			
ŏ [o]	> e [ɛ]	*polj**o**	>	polj**e**	'field'			

Figure 4: Umlaut data

In the abstract model of GP where only phonological contrasts are displayed, umlaut is straightforwardly represented as spreading of the palatal element. In the Germanic languages the element spreads from the vowel in the following syllable. In Figure 5 a positive level adjective (h[o]ch) fronting its back mid-high rounded vowel into a front mid-high rounded vowel (h[ø]cher). As the representation shows, only the tongue position is affected, the target sound [o] gets an additional **I** element

[2] The use of these elements is language-specific: some languages have an H–L opposition, some a nothing–L and some a nothing–H distinction. When H and L combine we witness voiced aspirates (e.g. Guajarati or Proto Indo European).
[3] Hutterer 1986, Bloomfield 1979, Bynon 1997
[4] Baleczky és Hollós 1973, Đorđić 1975, H. Tóth 1996

spread from the sound [ɛ] in the adjacent syllable. The two segments "see" each other on the level of nuclear projection, making the spreading of the element possible across a consonant.

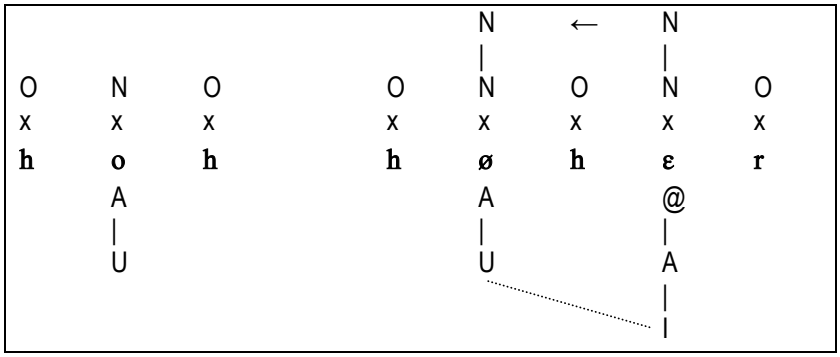


Figure 5: Representation of regressive umlaut in Germanic languages

In Slavonic languages, the element **I** spreads from the preceding palatal consonant onto the following back vowel:

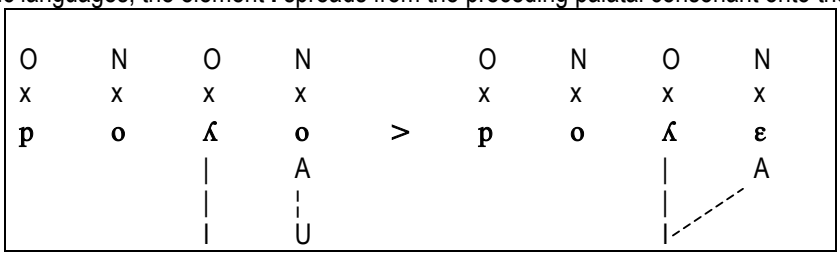


Figure 6: Representation of progressive umlaut in Proto-Slavonic

In contrast to Germanic umlaut, Slavonic umlaut is progressive. The element **I** spreads from the consonant a palatal lateral approximant[5] onto the target sound [o] present in the Proto-Slavonic, transforming it into a front (addition of **I** element) and unrounded (deletion of **U** element) vowel in the Old Church Slavonic equivalent of the word 'field'.

3. Palatalization

Palatalization is a place assimilation phenomenon in which consonants assimilate to a following, (in the case of a regressive assimilation) or to the preceding (in the case of the progressive assimilation) front vowel or palatal glide (i.e. front vocoid).

The types of palatalization are established according to several different criteria. Namely, we have a lexical–post-lexical distinction, a phonemic–allophonic distinction and one according to the segments affected by the process. The three main types of palatalization are (Bhat 1978, Lahiri–Evers 1991, Jacobs–Van de Weijer 1992):

1) addition of *secondary* articulation	p → p^{j}	all place of articulation consonants affected	bra**t'** – bra**t** 'to pick'–'brother' kro**f'** – kro**f** 'blood'–'roof'
2) shift of coronal place of articulation	t → tʃ	alveolar stops and fricatives become palato-alveolar affricates and fricatives	fac[**t**] – fac[**tʃ**]ual gra[**d**]e – gra[**dʒ**]ual
3) shift of velar place of articulation	k → tʃ	velar stops and fricative become palato-alveolar affricate and fricative	čove[**k**] – čove[**tʃ**]e 'human' NOM SG – VOC SG

Figure 7: Types of palatalization processes according to the segments affected

[5] Laterality is not indicated in the representation of the palatal lateral [ʎ], since it is not relevant to the representation of umlaut, the important factor is the palatality element.

An adequate representation of the process of palatalization depends on an adequate representation of segments taking part in the process. Velars are represented as placeless segments, and alveolars as having only the coronal element. Affricates are represented with two root nodes and one place node, since they are considered as quantitatively simple and qualitatively complex segments in the standard GP theory.

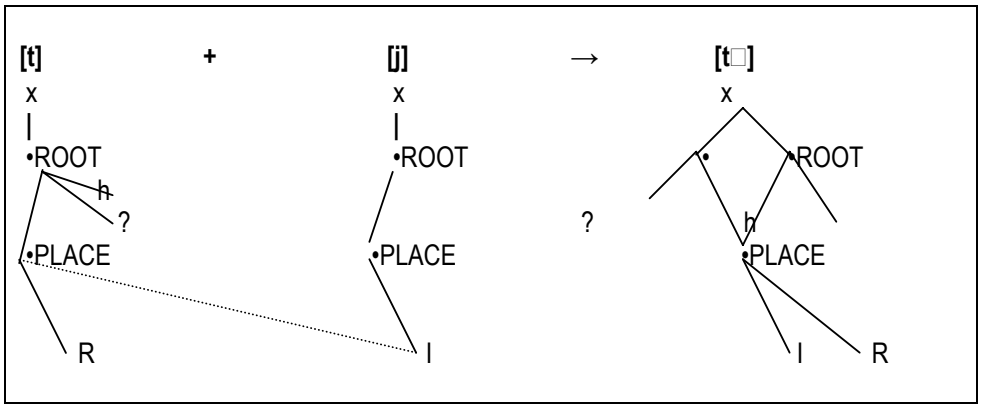


Figure 8: Representation of coronal palatalization

Secondary and Coronal Palatalization can easily be represented by spreading of the element **I** in the GP approach (Figure 9), but the representation of velar palatalizations is not straightforward because the element R has nowhere to spread from, as shown in Figure 10 (Fischer 2003):

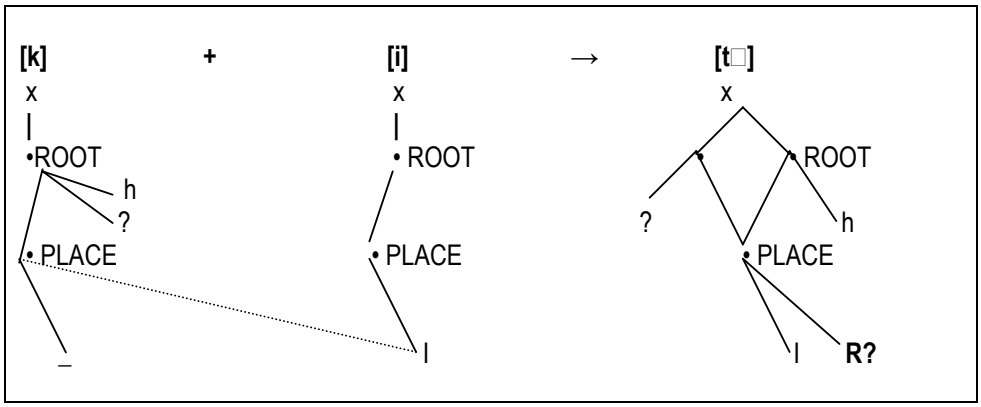


Figure 9: Representation of velar palatalization

Namely, the coronal element is not present neither in the target not in the trigger of the process. Consequently, the representation of velar palatalizations is closely connected to the issue of element coronal and the representation of coronal segments.

Phonetic research offers another possibility regarding the issue above, the representation of palatoids. According to Keating (1988), for example, X-ray data show that both the coronal and the dorsal articulator play a role in the formation of front vowels and palatals. In other words, palatals and palatoids are complex corono-dorsal segments. Consequently, their representation should reflect this fact and they should be represented with two melodic elements, i.e. two resonance elements: **I** and **A**. Plain alveolars, however, are represented solely with the element **I**. The diagram illustrates the above points.

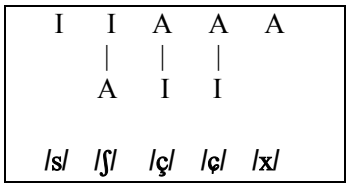


Figure 10: Representation of alveolars and palatoids

The above representations show that /ç/ and /ɕ/ are represented in the same way, since they do not usually contrast within one phonemic inventory of one language. In case they nevertheless do, Van de Weijer (1994), who works in Dependency Phonology, offers a solution combining identical elements. For example, dentals are represented as having two **I** elements.

Standard GP, however, does not allow for this kind of an extension of the theory. Consequently, I propose a different solution. If a language contrasts five types of fricatives: alveolars, dentals, palatals, palato-alveolars and alveo-palatals, and the theory recognizes two place elements, **I** and **A**, and one manner element, **h**, the only solution is to combine different head-operator relations within the representations. One variation is to represent plain palatals as headed by the element **I**, palato-alveolars headed by the element **A**, and alveo-palatals as headless. Similarly to them, dentals will have a plain element **I** in their representation without a head, whereas alveolars will have **I** as a head.

5. A unified representation of palatalization and umlaut

The paper is an attempt to show a harmonised representation of palatalization and umlaut process, in other words, a unified representation of all processes which involve a spreading of the element **I**, irrespective of the targets, trigger, outcome or even the direction of the process. This largely depends on further research in the topic of coronality and its representation. One important result of the analysis, however, is the typology of palatal assimilations:

	Type of assimilation	Direction of assimilation	Target of assimilation	Trigger of assimilation
C ← C	palatalization *English Coronal*	regressive	C	C
V ← V	umlaut *Germanic*	regressive	V	V

Figure 11: When the trigger and the target of the process belong to the same group of segments

	Type of assimilation	Direction of assimilation	Target of assimilation	Trigger of assimilation
V ← C	umlaut	regressive	V	C
C → V	umlaut, *Proto-Slavonic*	progressive	V	C
V → C	palatalization *3rd Velar*	progressive	C	V
C ← V	palatalization *1st and 2nd Velar*	regressive	C	V

Figure 12: When the trigger and the target of the process do not belong to the same group of segments

As the above tables show, the factor that determines the direction of the process is the major category of the trigger and the target. If the major category of the trigger (vowel or consonant) matches that of the target of the process, the assimilation only acts from right to left. If, however, the trigger's and the target's major category differ, both directions are possible.

Another observation that directly follows from the above tables is that the Germanic languages prefer regressive palatal assimilations happening between segments of the same type, whereas the Slavonic languages show examples of assimilatory processes happening in both directions and prefer having segments of different main category affecting each other. This latter observation is closely connected to the tendency of Palatality Harmony elaborated by Kristó (2000). Namely, according to the tendency, the vowel and the consonant mutually affect each other causing the members of a CV

sequence to be equal in palatality. There are no examples for the first row, for a regressive umlaut triggered by a consonant, but I believe that Palatality Harmony justifies this abstraction, since the tendency acknowledges that the direction of the vowel-consonant interaction in a CV syllable can be both progressive and regressive.

6. Conclusion

The paper examines the representation of palatalization and umlaut in a non-linear approach and its implications with respect to the issue of coronals in Government Phonology. This attempt sheds light on certain issues of GP and non-linear phonology in general. Further research questions include the issue of markedness. Recent work in this area has questioned the status of the traditionally unmarked segments, coronals. Certain phonologists have argued for placeless velars to be considered as the unmarked segments. Namely, markedness is a rather vague notion residing on factors such as complexity, frequency and occurrence in the world's languages, child acquisition, etc. What is more, segment and process markedness sometimes predict different (even opposite) markedness relations. For example, palatalization is considered to be unmarked since it is a simple and frequent process in the world's languages. Its outcome, however, are complex corono-dorsal segments. Consequently, do we consider palatalization to be a marked or unmarked phonological phenomenon? The answer may bring the linguists closer to the understanding of phonological changes and alternations.

7. References

Baleczky, E. – A. Hollós. (1973). *Ószláv nyelv* [Old Church Slavonic]. Budapest: Tankönyvkiadó.

Bhat, D. N. S. (1978). A General Study of Palatalization. In J. H. Greenberg (ed.), *Universals of Human Language*. Stanford, California: Stanford University Press.

Bloomfiled, L. (1933) [1979]. *Language*. London, Boston, Sydney: Allen and Unwin IX.

Bynon, T. (1979). *Historical linguistics*. Cambridge: Cambridge University Press.

Đorđić, P. (1975). *Staroslovenski jezik* [Old Church Slavonic]. Novi Sad: Matica srpska.

Fischer, M. (2003). *Representation of velar palatalizations in non-linear phonology*. MAT, Institute of English and American Studies, University of Szeged.

Harris, J. (1994). *English sound structure.* Oxford: Blackwell.

H. Tóth, I. (1996). *Bevezetés a szláv nyelvtudományba* [Introduction to Slavonic linguistics]. Szeged: JATE Press.

Hutterer, M. (1986). *Germán nyelvek* [Gemanic languages]. Budapest: Gondolat.

Jacobs, H. and Van de Weijer, J. (1992). On the formal description of palatalization. In R. Bok-Bennema and R. Van Hout. (eds.) *Linguistics in the Netherlands 1992*. Amsterdam: Benjamins. 125–135.

Kaye, J., J. Lowenstamm and J-R. Vergnaud. (1985). The internal structure of phonological elements: a theory of charm and government. *Phonology Yearbook 2:* 305-328.

Keating, Patricia A. (1988). Palatals as complex segments: X-ray evidence. *UCLA Working Papers in Phonetics 69:* 77-91.

Kristó, L. (2000). Palatality Harmony in Proto-Slavonic. In L. Varga. (ed.) *The Even Yearbook.* Budapest: ELTE.

Lahiri, A. and Evers, V. (1991). Palatalization and coronality. In C. Paradis and J-F. Prunet (eds.) *Phonetics and phonology: The special status of coronals*. San Diego: Academic Press. 79–100.

Van de Weijer, J. (1994). *Segmental Structure and Complex Segments*. The Hague: Holland Academic Graphics.

BULGARIAN FRAMENET

Svetla Koeva, Rositsa Dekova

Computational Linguistics Department, Institute for Bulgarian, Bulgarian Academy of Sciences
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
svetla@dcl.bas.bg, rosdek@dcl.bas.bg

ABSTRACT

The paper presents the conceptual model underlying the Bulgarian FrameNet. A *frame lexicon entry* is considered a unit which consists of a target verb, a unique explanatory definition, at least five annotated examples, adjacent grammar class (a set of morpho-syntactic properties) and a syntactic frame (a set of feasible syntactic structures where a constant number of arguments are individually specified for syntactic category, lexical explicitness, grammatical function and semantic features). A brief discussion on the linguistic motivation for the frame determination is also included, i.e. how the available corpora data is used when specifying the information to be encoded in the frame lexicon entry and what tests are applied to choose among the various possible classifications.

Introduction

The major goal of the Bulgarian FrameNet project[1] is to elaborate frame-semantic description of the core lexicon of Bulgarian and thus to represent the semantic (and related lexical and syntactic) knowledge in a format appropriate for multiform NLP tasks.

The aims of this paper are to present the conceptual model underlying the Bulgarian FrameNet, to outline the different types of information encoded in the frames and to discuss the linguistic motivation on which this encoding is based. In its current stage the lexicon consists of the 1 500 most frequent Bulgarian verbs (calculated over the tagged and lemmatized versions of Bulgarian Brown Corpus and Bulgarian Text Archie (Koeva et al., 2006) represented with their meanings (approx. 3 000)[2].

Related work

The Berkeley FrameNet project (Baker et al., 1998; Johnson et al., 2002; Ruppenhofer et al., 2005) is one of the most significant linguistic approaches accounting for the semantics of words and their lexical representation in relation to the syntax-semantics interface. Based on Fillmore's Frame Semantics the FrameNet aims at providing a comprehensive frame-semantic description of the core lexicon of English, while it commits to corpus evidence for semantic and syntactic generalizations underlying the representations of the semantic frames of the words described. FrameNets of languages other than English are intensively created either based on the major assumptions and generalizations of the FrameNet, or developed as independent investigations (Subirats & Petruck, 2003; Boas et al., 2006; Ohara et al., 2004; Lopatková et al., 2006; among others).

There were some attempts towards developing of Bulgarian frame description (not machine-readable one), which suffered from both inconsistency of the formal representation and incompleteness of data. Recently, such an attempt was the Model for a Syntactic dictionary of Bulgarian verbs (Penchev et al., 1998), built to the great extent in the traditional framework of valency dictionaries. The dictionary itself was compiled as a text file containing a small amount of verbs (approx. 400), their explanatory definitions, the number and types of arguments, the semantic roles, and some examples.

Conceptual model

The Bulgarian FrameNet is based on the theoretical framework and corresponding methodology described in (Koeva, 2004). We call a ***frame lexicon entry*** a unit consisting of a target word, its unique explanatory definition, annotated examples, adjacent grammatical class and syntactic frame. The ***syntactic frame*** is a set of feasible syntactic structures (masks) associated with the target word. *The* ***syntactic structure*** defines the number of arguments uniquely specified for a syntactic category (with specification of particular prepositions or complementizers, if any), lexical explicitness, grammatical function and semantic features. The main differences with the FrameNet conceptual model concern the detailed description of

[1] The Bulgarian FrameNet is under development at the Computational Linguistics Department (DCL) at the Institute for Bulgarian and was supported by the Bulgarian Ministry of Education and Research – project IO 0102.

[2] Currently five researchers at DCL are working on the Bulgarian FrameNet (the authors of the paper, Rada Vlahova, Petya Nestorova and Atanas Atanasov).

morpho-syntactic properties of Bulgarian verbs comprising the grammatical classes, on the one hand, and the approach towards the encoding of semantic relations (we do not name those relations explicitly, as we consider that the combinatory description of individual semantic features defines unambiguously the respective relation), on the other hand. The mask defines differences in the syntactic realization of arguments (exclusive of alternations since they are always related to argument transpositions).

The information is organized in linguistic modules. For example, linguistic modules could be *explanatory definition*, *semantic features*, etc. Some modules require free filling of data, while others require an option to be chosen from a predefined list of attributes and their values. Such an attribute could be *subjectivity* with values: *personal verb – impersonal verb – 3rd personal verb*, etc. Some modules are linked up with strong dependencies, while others are considered independent. For example, personal and 3rd personal verbs require a subject argument, while impersonal – do not.

Frame lexicon entry

The *frame lexicon entries* are defined within the following structure (based on early studies (Koeva, 2004)):

Target word – a verb lemma expressing a unique meaning (i.e. each target word is ascribed a single sense)[3].

All target words are assigned with the following information:

Explanatory definition – the definitions are taken from the Bulgarian wordnet – BulNet, occasionally with some modifications (Koeva et al., 2004). The frequency of the sense usage based on calculations over the Bulgarian semantically annotated corpus is also noted (Koeva et al., 2006).

Annotated examples – at least five examples (with their source cited) are provided to illustrate the definition. All the examples are annotated with the arguments.

Grammatical class is the set of the morpho-syntactic properties of the target word:

- ***subjectivity*** – each verb is specified as *personal*, *impersonal* or *third personal* according to its person paradigm;

- ***transitivity*** – each verb is classified as *transitive* or *intransitive*; both transitive and intransitive verbs may be further specified according to their lexical properties*: reflexiva tantum se* (a compound verb built with the particle *se*), *reflexiva tantum si* (a compound verb built with a particle *si*), *reciproca tantum se* (a compound verb built with a particle *se*), *reciproca tantum si* (a compound verb built with a particle *se*), *intransitiva tantum* (a 3rd person verb), *refelexiva tantum* (a 3rd person compound verb built with a particle *se*), *acusativa tantum* (with an obligatory accusative personal pronoun clitic), *dativa tantum* (with an obligatory dative personal pronoun clitic), *acusativa dativa tantum* verbs (with an obligatory dative personal pronoun clitic and a particle *se*);

- ***perfectiveness*** – each verb is specified as *imperfective* (if it expresses a process – duration, recurrence, or lack of integrity), *perfective* (if it denotes integrity and completeness), *bi-aspectual*, *imperfectiva tantum* (if a perfective correspondent of the same stem does not exist), or *perfectiva tantum* (if an imperfective correspondent of the same stem does not exist);

- ***inflectional type*** – the unambiguous formal description of the inflectional paradigm of the target word taken from the Bulgarian Grammar Dictionary (Koeva 1998).

The Syntactic frame consists of one or more syntactic structures that differ in the syntactic realization of the arguments but are identical with respect to their number, syntactic functions and their semantic relations signaled by the encoded semantic features. *The* ***syntactic structure*** defines the number of arguments, which are uniquely specified for a syntactic category, lexical explicitness, grammatical function and semantic features. Thus each argument slot in the syntactic structure is assigned with the following information concerning the possible lexical realization:

Syntactical category – the following syntactic categories are eligible: NP (noun phrase), PP (preposition phrase), AdvP (adverb phrase), AP (adjective phrase), S (clause), SC (small clause), CL/NP or CL/PP (obligatory cliticization of the respective argument), PP/ADVP (prepositional phrase that can be substituted by an adverbial phrase: these are for example PPs that denote location: He went [$_{PP}$ *in the woods*] / [$_{ADVP}$ *there*]); the subject argument slot is specified with a different set of

[3] We accept the Princeton WordNet assumption that polysemous words participate with their senses in different synonymous sets, while the explanatory definition expresses the meaning of the respective synonymous set corresponding to a unique concept.

categories than the complement argument slots); prepositional phrases are specified with respective eligible prepositions, while clauses – with respective eligible complemetisers (introductory words);

Explicitness – phrases can be obligatory explicit, or non-explicit; a PP complement can be obligatory explicit depending on the explicitness of the NP complement, an NP complement can be obligatory explicit depending on the explicitness of the PP complement;

Syntactic function – the specified functions are: subject, direct object, indirect object, adverbial, subject clause, object clause, adverbial clause, small clause); the eligible functions for the subject argument slot are subject and subject clause; for the NP complement – direct object and small clause; for the PP complement – indirect object, adverbial and small clause; for the S complement – object clause or adverbial clause; for the ADVP complement – adverbial; for the AP complement – small clause;

Semantic features – semantic features are described in three groups of complementary features: **the general semantic classification** distinguishes between abstract and concrete, animate and inanimate, person and non-person nouns, agent, experiencer, or none; **the "partitive" semantic classification** distinguishes between countable and uncountable nouns, group denoting nouns and nouns denoting a single object and between a part and a whole; **the ontology based semantic classification** mainly consists of first order entities specified in EuroWordnet (Vossen 1999).

If there is more than one possible semantic feature from the same group, each one is described separately building a new set with the selected features from the other groups. If necessary each frame lexicon entry is supplemented with some necessary comments which describe features that are not foreseen by the frame lexicon structure.

Bulgarian FrameNet software

The development of the Bulgarian FrameNet was carried out by means of a web-based system called SYNText (SYNtactic dictionary Tool) (Koeva et al. 2003).

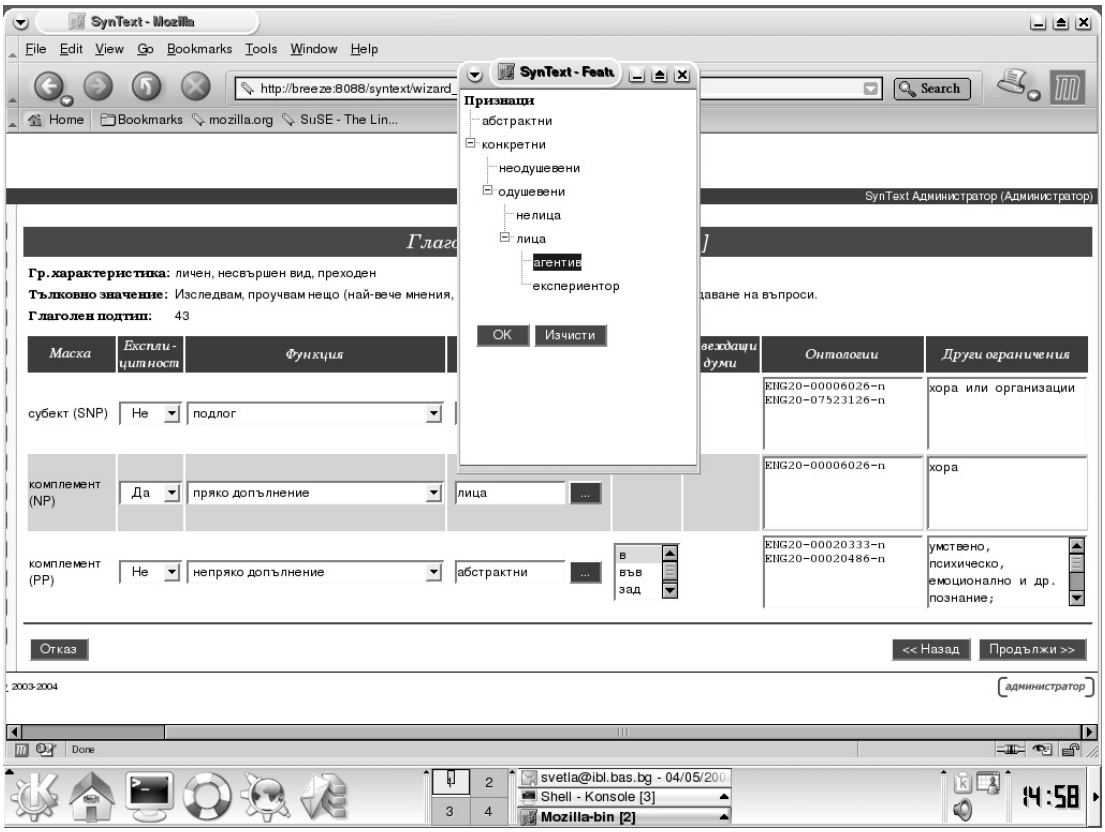
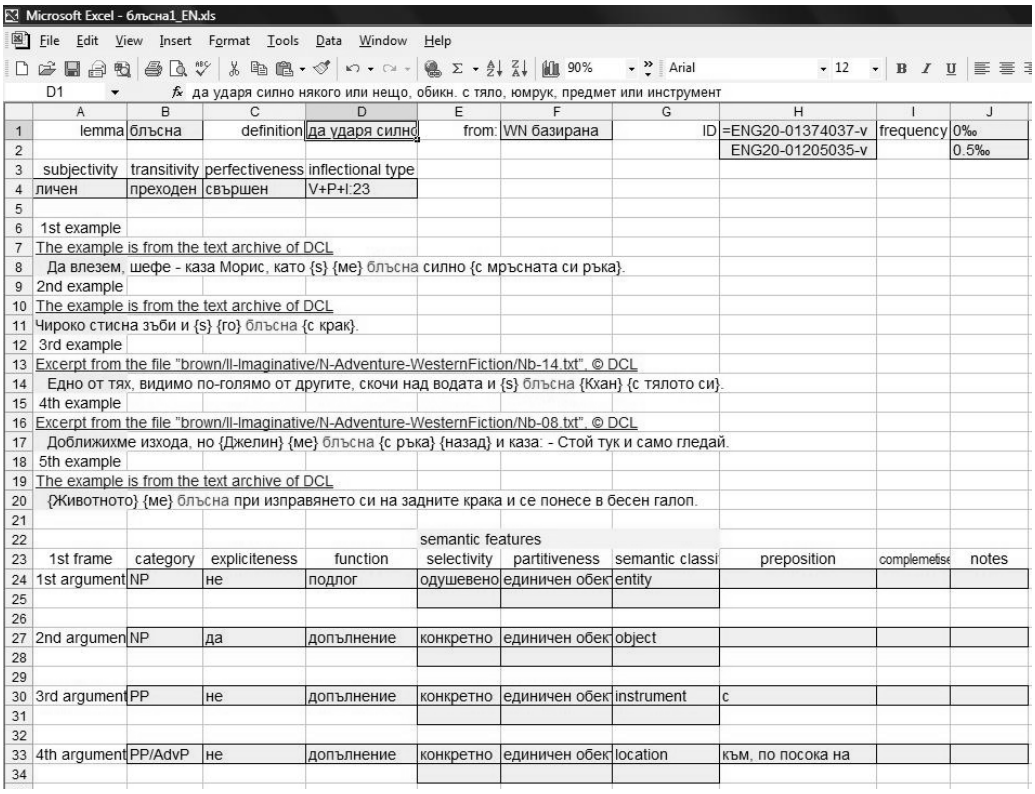


Figure 1: The SYNText system

The system (Figure 1) allows fast and easy administration of the attributes and their values inside the linguistic modules by an authorized person; supports user roles, provides authentication and special guest access; allows different checks up, i.e. to recall all units that satisfy particular criterion, to recall all units that have equal features, etc. The architecture of the SYNText system allows XML import and export of the data base and is principally organized in accordance with the Frame lexicon entry described above.

Presently the frames are being described in Excel files (conferrable to XML format) since more modules were added in the description (at least at the level of semantic features where "partitive" and ontology-based classifications were also included). The reprogramming of the SYNText system is required to support new linguistic modules added – the clarifying of all representative modules and their respective sets of attributes and values is still to be defined during the work process. Figure 2 below is an example of how data is organized within an Excel file to create the appropriate frame for each sense.



Microsoft Excel - блъсна1_EN.xls

File Edit View Insert Format Tools Data Window Help

D1: да ударя силно някого или нещо, обикн. с тяло, юмрук, предмет или инструмент

	A	B	C	D	E	F	G	H	I	J
1	lemma	блъсна	definition	да ударя силно	from:	WN базирана	ID	=ENG20-01374037-v	frequency	0‰
2								ENG20-01205035-v		0.5‰
3	subjectivity	transitivity	perfectiveness	inflectional type						
4	личен	преходен	свършен	V+P+I:23						
5										
6	1st example									
7	The example is from the text archive of DCL									
8	Да влезем, шефе - каза Морис, като {s} {ме} блъсна силно {с мръсната си ръка}.									
9	2nd example									
10	The example is from the text archive of DCL									
11	Чироко стисна зъби и {s} {го} блъсна {с крак}.									
12	3rd example									
13	Excerpt from the file "brown/ll-Imaginative/N-Adventure-WesternFiction/Nb-14.txt", © DCL									
14	Едно от тях, видимо по-голямо от другите, скочи над водата и {s} блъсна {Кхан} {с тялото си}.									
15	4th example									
16	Excerpt from the file "brown/ll-Imaginative/N-Adventure-WesternFiction/Nb-08.txt", © DCL									
17	Доближихме изхода, но {Джелин} {ме} блъсна {с ръка} {назад} и каза: - Стой тук и само гледай.									
18	5th example									
19	The example is from the text archive of DCL									
20	{Животното} {ме} блъсна при изправянето си на задните крака и се понесе в бесен галоп.									
21										
22					semantic features					
23	1st frame	category	expliciteness	function	selectivity	partitiveness	semantic classi	preposition	complemetise	notes
24	1st argument	NP	не	подлог	одушевено	единичен обек	entity			
25										
26										
27	2nd argumen	NP	да	допълнение	конкретно	единичен обек	object			
28										
29										
30	3rd argument	PP	не	допълнение	конкретно	единичен обек	instrument	с		
31										
32										
33	4th argument	PP/AdvP	не	допълнение	конкретно	единичен обек	location	към, по посока на		
34										

Figure 2: Formal representation of the Bulgarian verb *блъсна* '*to push*' in its sense "*to hit someone or something, usually by means of one's body (or body part), an object or an instrument*"

Bulgarian FrameNet corpus

The frame concomitant examples are extracted from the Bulgarian Brown corpus[4] and the Bulgarian text archive. When necessary, newspapers and fiction that are freely available on-line are used to illustrate some rare or more specific senses. Each example is annotated for the arguments specified for the particular verb sense. The lexical realizations of the arguments described in the syntactic structures are marked within each example sentence by curly brackets. The logical place of the implicit arguments is signaled respectively – subject {s}, direct object {d}, indirect object {i}, adverbial complement {a}. For example:

1) {s} Взе ли {всичко} {от колата}?

{s} *take*-2p.sg *li*-question particle {*everything*} {*from car*-the}

Did you take everything from the car?

2) {s} Взех {d} {a}.

{s} *took*-1p.sg. {d} {a}

I did.

[4] http://dcl.bas.bg/Corpus/home_en.html

Searches in the available corpora are performed with the Advanced search engine supporting regular expressions and a query language (Tinchev et al., 2007) providing elementary ordered and non-ordered conjunctive queries of words and/or feature structures (grammatical or semantic attributes and values – for example the category number and its values: singular, plural and countable form) and Boolean formulas over them. Figure 3 provides an example of the results returned from the regular expression вижда[мшть]?[емта]?е? matching the synthetic forms of the Bulgarian verb *виждам* '*to see*' in the Bulgarian Brown corpus.

A complementary corpus (approx.150 000 words) containing the most frequent verbs was extracted from the Bulgarian Brown Corpus and is autonomously annotated for syntactic tree structures.

Figure 3: A regular expression search returning the synthetic forms of the Bulgarian verb *виждам* 'to see'.

Linguistic motivation for frame determination

Much in the tradition of the current linguistic approaches, and mostly in accordance with the FrameNet project, we argue that the information that should be included in the lexical representation of verbs/words is significantly related to the encoding of different sets of arguments into lexical items and various syntactic patterns within a single language, as well as across languages (Dekova, 2006). Therefore, when providing the lexical representation of verbs we rely to a great extent on the correlation between verb sense and the possible syntactic realizations. One of the most clear-cut tests for distinguishing between verb senses is the number and lexical-syntactic realization of arguments. Thus one graphical word may be described in one or more lexicon entries, as each of the entries is given a unique explanatory definition corresponding to the particular verb sense. The meaning is based on the data found in the available corpora and it is illustrated in the entry by at least five example sentences which are chosen among all the examples found. The definitions are checked first against the Bulgarian WordNet. If the same meaning or similar definition is not included in BulNet, the definition might be taken directly or with modifications from the Bulgarian Explanatory Dictionary (Popov, 1997). For the senses not described in both sources – a new definition is created. Definitions are made in such a way that they explicitly suggest the number of arguments of the verb in the respective sense. For example, the definition of *блъсна* '*to push*' in the sense presented above is stated as follows: "*to hit someone or something, usually by means of one's body (or body part), an object or an instrument*". Sometimes, two or more definitions from BulNet are unified under one lexicon entry when no evidence was found in the linguistic data to support the split. Although the opposite – splitting one definition (synset) from BulNet into several lexicon entries – is also an option, such cases have not occurred so far.

The classification of the target verb includes also morpho-syntactical features, as subjectivity, transitivity, and perfectiveness. Different tests are used to indicate how the verb should be classified:

whether it is personal, impersonal (no subject can appear explicitly as in the Bulgarian verb *мръква се* '*it is getting dark*'), or 3rd personal (only subjects in 3rd person singular or plural can be syntactically realized as with the Bulgarian verbs *вали* '*it rains*' or *боли (ме) 'it hurts'*);

whether it is transitive or intransitive (including a test whether the reflexive particles *se* and *si* are a lexical part of the verb or they signify the presence of an argument, as for example *гордея се* '*be proud*' – **гордея себе си* '*proud oneself*' vs. *реша се* '*comb*' – *реша себе си* '*comb oneself*');

whether the verb is perfective, imperfective, perfectiva tantum, imperfectiva tantum or dual – there are certain constructions in which only verbs of particular aspect can appear; for example only imperfective verbs are allowed immediately after the explicit subject of the sentence: *Аз пиша* '*I write-imperf*' vs. **Аз напиша* '*I write-perf*', etc.

The syntactic frame consists of one or more syntactic structures (masks) depending on the possible syntactic realizations of the particular arguments. Each mask within a single syntactic frame contains the same number of arguments defined through their *syntactical category*, their e*xplicitness,* their *syntactic function*, and a number of *semantic features.*

The semantic features are divided into three complementary groups. First, the argument must be classified according to the general semantic classification, i.e. whether it is an abstract noun (if it does not relate to any physical form) or a concrete noun, and whether it can be further specified as animate (living organism) vs. non-animate, as person (humans and higher animals attributed to have human qualities or powers) or non-person, as agent (acting by one's own will) or non-agent, etc. Next, the argument is classified under the "partitive" semantic classification: whether it is countable or uncountable, a group denoting noun (such as *crew*, *institute* or *organization*, as in "*The crew numbers 30 people*.") or it denotes a single object, and whether it is a part or a whole as in "*The house consists of 2 bedrooms, a kitchen, and a bathroom*." Finally, the argument is categorized in accordance with the ontology based semantic classification: whether the possible instances of this argument can be ranked as hyponyms of some of the first order entities specified in EuroWordnet (Vossen 1999), for example: instrument, location, act, event, feeling, etc. If more sets of semantic features are applicable to one argument, each set is described separately within the same syntactic structure.

Reflecting the linguistic finds in the existing corpora *Frame lexicon entries* appear to be an appropriate representational format especially for Bulgarian. For instance, under this format different lexicon entries are created for Bulgarian aspectual verb pairs (for ex. *бутам* (*push*-impref.) – *бутна* (*push*-perf.)) because the formation of some of the diatheses depends on the verb aspect.

Passives and other syntactic alternations are not described in separate syntactic structure as they are predictable and can be generated with rules. A syntactic alternation involves a reordering of the arguments within one syntactic frame that affect grammatical functions and syntactic categories, while semantic relations and the general meaning expressed remain unchanged. Although syntactic alternations involve changes in the syntactic realization of the arguments that may be accomplished with changes at the morphological level concerning the verbal lemma, the verbal paradigm, and the verb's transitivity, these changes are fully predictable (Koeva 2008). The parameters that determine the Bulgarian syntactic transformations are: *grammatical classes*, *argumentness* (the property of the predicate to incorporate a specific number of variables that correspond to the arguments and their syntactical categories and grammatical functions in the sentence), and *selectivity* (the set of semantic features that a given phrase must satisfy in a given position).The correctness of the information encoded in the lexical entries can be easily verified. For example, if the morpho-syntactic values for a given Bulgarian verb are personal transitive (non–)perfective and the syntactic structure is NP–subject and NP–complement, and the semantic features are respectively agentive/non–animate for the NP–subject and non–animate for the NP–complement, then the syntactic diathesis *se*–passive is possible:

[NP1 *Мълнии*] *осветиха* [NP2 *небето*].	[NP2 *Небето*] *се освети* [PP *от* [NP1 *мълнии*]].
[NP1 Lightning] illuminated [NP2 the sky].	[NP2 The sky] was illuminated [PP by [NP1 lightning]].

Conclusions and future work

The proposed conceptual method for formal description of syntactic frames is (to some degree) language independent and to a great extent theory independent. However, it can cover some language specific morpho-syntactic features of Bulgarian: subjectivity, perfectivity and transitivity, as well as language specific lexical and syntactic realization. Our future work envisages enrichment of the lexicon with new verb frames, publishing the results achieved so far on-line, validation of the

Bulgarian FrameNet towards automatically extracted frames from corpora and manually annotated Bulgarian dependency Treebank. The final definition of the linguistic modules and the sets of their attributes and values is also a near future task because it predetermines the reprogramming of the Bulgarian FrameNet software system. The determination of the linguistic modules is in close relation with the definition of the necessary conditions for syntactic alternations rules. The BulNet explanatory definitions used, the morpho-syntactic information encoded in the verb grammatical classes and the third group of semantic features related to the WordNet first order entities bring together the Bulgarian FrameNet and the other extensive Bulgarian language resources. This, along with the completeness and the consistency of the encoded information, predetermines the extensive usage of the Bulgarian FrameNet in various NLP tasks.

References

Baker, C.; C. Fillmore; J. Lowe. 1998. The Berkeley FrameNet project. In "*Proceedings of the 17th International Conference on Computational Linguistics (COLING/ACL- 98)*", pages 86-90, Montreal. ACL.

Boas, H.; E. Ponvert; M. Guajardo; S. Rao. 2006. "The current status of German FrameNet," SALSA workshop at the University of the Saarland, Saarbrücken, Germany, June 2006.

Dekova, R. 2006. *Lexical Encoding of Verbs in English and Bulgarian*. Doctoral thesis. NTNU, Trondheim.

Johnson, C.; C. Fillmore, M. Petruck, C. Baker, M. Ellsworth, J. Ruppenhofer, and E. Wood. 2002. *FrameNet: Theory and Practice*. International Computer Science Institute, Technical Report-02009. Berkeley,CA.

Koeva S. 1998. Bulgarian Grammar Dictionary. Description of the linguistic data organization concept. *Bulgarian Language*, 6, 49-58.

Koeva, S.; E. Doychev and G. Cholakov. 2003. "SYNTEXT – a Web-based System Designed for Frame Lexicons", in: *Proceedings from the International Workshop Balkan Language Resources and Tools*, Thessaloniki, 2003, 41-48.

Koeva, S. 2004. "Theoretical model for a formal representation of syntactic frames" – *Scripta and e-Scripta*, Vol.2, Sofia, 31-49.

Koeva, S.; T. Tinchev; S. Mihov. 2004. Bulgarian Wordnet Structure and Validation. *Romanian Journal of Information Science and Technology*, 7 (1-2), 61-78.

Koeva, S.; S. Leseva; I. Stoyanova; E. Tarpomanova; M. Todorova. 2006. Bulgarian Tagged Corpora. In *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*. pp. 78-86.

Koeva, S. 2008. Bulgarian Syntactic Alternations, *Proceedings of the NooJ 2007 International Conference*, Publications of UAB.

Lopatková, M.; Z. Žabokrtský; K. Skwarska. 2006. Valency Lexicon of Czech Verbs: Alternation-Based Model, in Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Paris, France, pp. 1728-1733.

Ohara, K.; S. Fujii; T. Ohori; R. Suzuki; H. Saito; S. Ishizaki. 2004. The Japanese FrameNet Project: An introduction. The Fourth international conference on Language Resources and Evaluation. *Proceedings of the Satellite Workshop "Building Lexical Resources from Semantically Annotated Corpora"*, 9-11. Lisbon, Portugal. May, 2004.

Penchev, J.; S. Koeva; A. Dineva; Sv. Stoyanova. 1998. *Bulgarian Syntactic Dictionary* – delivered to Bulgarian Ministry of Education and Research: OHI 410.

Ruppenhofer, J.; M. Ellsworth; M. Petruck; C. Johnson; J. Scheffszyk. 2005. *FrameNet II. Extended Theory and Practice.* ICSI Technical Report.

Popov, D. 1997. Dimitar Popov (ed) L. Andreychin, L. Georgiev, St. Ilchev, N. Kostov, Iv. Lekov, St. Stoykov, Tsv. Todorov, *Bulgarian Explanatory Dictionary*, Publishing House "Nauka i Izkustvo", 1093 p.

Subirats C. & M. R.L. Petruck. 2003. Surprise: Spanish FrameNet! In E. Hajicova, A. Kotesovcova & J. Mirovsky (eds.), *Proceedings of CIL 17*. CD-ROM. Prague: Matfyzpress.

Tinchev, T.; S. Koeva; B. Rizov; N. Obreshkov. 2007. Sistema za razshireno tursene v korpusi, sbornik *Literaturata*, Sofia University press, 92-111.

Vossen, P. (ed.). 1999. EuroWordNet General Document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014.

ANAPHORA RESOLUTION IN CROATIAN: PSYCHOLINGUISTIC EVIDENCE FROM NATIVE SPEAKERS

Tihana Kraš

Research Centre for English and Applied Linguistics, Faculty of English, University of Cambridge
English Faculty Building, 9 West Road, Cambridge, CB3 9DP, United Kingdom
tk302@cam.ac.uk

Department of English, Faculty of Arts and Sciences, University of Rijeka
Trg Ivana Klobučarića 1, 51000 Rijeka, Croatia
tkras@ffri.hr

ABSTRACT

This paper reports the results of an experimental study on the resolution of intra-sentential anaphora in Croatian by adult native speakers. In a picture-selection task, the speakers had to identify the antecedents of third person null and overt subject pronouns in ambiguous forward and backward anaphora contexts. Similarly to adult native speakers in previous studies on Italian, the speakers tended to resolve anaphora in the subject position with the null pronoun, and in a non-subject position with the overt pronoun. In backward anaphora, they allowed the overt pronoun to co-refer with both non-subject antecedents that were available in the context of situation, and in forward anaphora, only with the one that was mentioned in the sentence. This suggests that null and overt subject pronouns have the same antecedent preferences in intra-sentential anaphora in Croatian and Italian. In both languages, the null pronoun is biased towards the subject antecedent and the overt pronoun towards a non-subject antecedent. These biases derive from a discourse-pragmatic principle which assigns the task of topic shift to the overt pronoun and that of topic maintenance to the null pronoun.

1. Introduction

Due to rich verbal inflection, Croatian allows the omission of overt pronominal subjects in tensed clauses, a phenomenon known as *pro*-drop. While the existence of two pronominal options, null and overt, in *pro*-drop languages arises as a consequence of formal grammatical properties[1], their distribution is regulated by discourse-pragmatic factors. To our knowledge, the way these forms are used and interpreted in Croatian (or other South Slavic languages) has not been closely examined so far. This paper aims to fill in this gap by addressing the interpretation of null and overt subject pronouns in sentence-internal contexts in Croatian. Given that this phenomenon, commonly referred to as intra-sentential anaphora resolution, has been extensively studied in Italian, we take previous studies on Italian as our point of reference. We present psycholinguistic evidence coming from adult native speakers that intra-sentential anaphora is resolved in a practically identical way in Croatian as in Italian, suggesting that the two null subject languages might exhibit a total overlap in this domain.

2. Interpretation of pronominal subjects in null subject languages

In null subject languages, null pronouns are felicitous only in contexts in which they are co-referential with the discourse topic (Grimshaw and Samek-Lodovici 1998, among others). Null and overt subject pronouns are, therefore, used for different purposes: the former to refer to a referent already introduced in the context (the topic), and the latter to introduce a new referent or contrast a referent with another. This division of labour between the two pronominal forms has been formally expressed by Sorace (2000, 2005) by means of the interpretable [+/-Topic Shift] feature: [+Topic Shift] contexts require an overt pronoun, while [-Topic Shift] contexts demand a null pronoun. This is illustrated in (1). In (1b), functioning as an answer to the question in (1a), only the null form is felicitous, given that the pronoun refers to the subject of the interrogative sentence, which is the discourse topic.

(1) a. *Zašto je dječak$_i$ zaspao?*
why is boy fell.asleep
'Why did the boy fall asleep?'

[1] In the generative framework, it is assumed that null subjects are licensed by the positive setting of the null-subject parameter (Rizzi 1982, 1986).

b. *Zato što je* pro$_i$/#*on bio umoran.*
because is *pro* he was tired
'Because he was tired.'

In her work on Italian, Carminati (2002) has shown that null and overt pronouns have different antecedent preferences in intra-sentential anaphora: the null pronoun prefers the subject antecedent, while the overt pronoun prefers a non-subject antecedent. This can be seen in (2). The null pronominal subject of the subordinate clause is more likely to refer to the matrix subject ('la mamma') than to the matrix complement ('la figlia'), given that the former is the default topic of the sentence, while the opposite holds for the overt subject. The overt pronoun can also refer to another referent not mentioned in the sentence (e.g. 'the grandmother').

(2) *La mamma*$_i$ *ha rimproverato la figlia*$_j$ *mentre* pro$_{i/?j}$/*lei*$_{?i/j/k}$ *cuciava.*
The mother has scolded the daughter while *pro* she was.cooking.
'The mother scolded her daughter while she was cooking.'

Psycholinguistic evidence for Carminati's generalisation has also been provided in studies testing native Italian speakers and different types of highly proficient bilingual speakers of Italian and English (Belletti et al. 2007, Serratrice 2005, Sorace and Filiaci 2006, Tsimpli et al. 2004) or Croatian (Kraš 2008). In ambiguous forward and backward anaphora[2] contexts similar to those in (2), adult native speakers preferred the matrix subject as the antecedent for the null pronoun[3] and an antecedent other than the matrix subject for the overt pronoun. In backward anaphora, they allowed the overt pronoun to co-refer not only with the matrix complement, but also with an extralinguistic referent that was present in the context of situation.

Carminati (2002:195) argues that some kind of division of labour between null and overt pronouns exists in all null subject languages, but does not predict exact correspondences across languages. According to her, possible sources of variability include defects in the verb agreement paradigm and different historical origins of overt forms. In order to determine the degree of overlap between Italian and Croatian in the domain of intra-sentential anaphora, we conducted an experimental study on Croatian, the results of which are comparable to those of the Italian studies mentioned above.

3. The study

3.1. Aims

The aim of the study was to determine whether null and overt subject pronouns in Croatian have the same antecedent preferences in intra-sentential anaphora as they do in Italian. We focused on the interpretation of third person pronouns in ambiguous bi-clausal sentences in which the pronoun was in the subordinate clause.

Due to comparable richness of the verbal paradigm in the two languages, we predicted that the participants in the study would resolve anaphora in a similar way as Italian native speakers in previous studies did. More specifically, we predicted that they would tend to resolve anaphora in the matrix subject position only with null pronouns and that they would regard both the matrix complement and the extralinguistic referent as plausible antecedents for the overt pronoun in backward anaphora.

3.2. Participants

The participants in the study were 48 undergraduate students at the University of Rijeka (Croatia), aged 20-27 (mean age: 22.02), who were native speakers of Croatian and who originated from different parts of Croatia. They were all studying towards a degree in English language and literature and another subject in the area of social sciences and humanities.

[2] Forward anaphora is the one in which the pronoun follows its referent and backward andphora the one in which the pronoun precedes its referent.
[3] This, however, applies more closely to Kraš (2008) and Tsimpli et al. (2004) than to the other studies, in which the speakers interpreted the pronoun as co-referential with either the subject or the complement of the matrix clause in forward anaphora.

3.3. Materials and design

We used a modified version of a picture-selection task that was originally used in Tsimpli et al. (2004) and subsequently in other studies on Italian. The Italian modified version was previously used in Kraš (2008). The participants had to read a sentence consisting of a main and a subordinate clause and choose a picture that corresponded to the meaning of the sentence. In this way, they were identifying the performer of the action described in the subordinate clause. The main clause contained an animate object NP matched in gender and number with the animate subject NP. The sentences were either ambiguous or unambiguous, depending on the number of possible interpretations of the subordinate clause. Unambiguous sentences corresponded to the most plausible interpretations of ambiguous sentences and served for the purposes of control. The four experimental conditions and their matching four control conditions are illustrated in (3) and (4) respectively.

(3) a. FORWARD ANAPHORA – NULL PRONOUN (FANP)
Svjedok$_{i}$ *pokazuje optuženog*$_{j}$ *dok* pro$_{i/?j}$ *ulazi u sudnicu.*
witness points accused while *pro* enters in courtroom
'The witness points to the accused as he enters the courtroom.'

b. FORWARD ANAPHORA – OVERT PRONOUN (FAOP)
Svjedok$_{i}$ *pokazuje optuženog*$_{j}$ *dok on*$_{?i/j/k}$ *ulazi u sudnicu.*
witness points accused while he enters in courtroom
'The witness points to the accused as he enters the courtroom.'

c. BACKWARD ANAPHORA – NULL PRONOUN (BANP)
Dok pro$_{i/?j}$ *ulazi u sudnicu, svjedok*$_{i}$ *pokazuje optuženog*$_{j}$.
while *pro* enters in courtroom witness points accused
'As he enters the courtroom, the witness points to the accused.'

d. BACKWARD ANAPHORA – OVERT PRONOUN (BAOP)
Dok on$_{?i/j/k}$ *ulazi u sudnicu, svjedok*$_{i}$ *pokazuje optuženog*$_{j}$.
while he enters in courtroom witness points accused
'As he enters the courtroom, the witness points to the accused.'

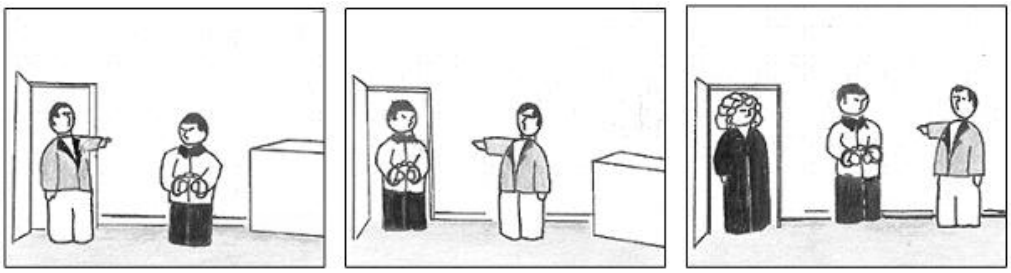
(4) a. FORWARD ANAPHORA – NULL PRONOUN – CONTROL (FANPC)
Svjedok pokazuje optuženog ulazeći u sudnicu.
witness points accused entering in courtroom
'The witness points to the accused while entering the courtroom.'

b. FORWARD ANAPHORA – OVERT PRONOUN – CONTROL (FAOPC)
Svjedok$_{i}$ *pokazuje optuženog*$_{j}$ *koji* pro$_{*i/j}$ *ulazi u sudnicu.*
witness points accused who *pro* enters in courtroom
'The witness points to the accused who enters the courtroom.'

c. BACKWARD ANAPHORA – NULL PRONOUN – CONTROL (BANPC)
Ulazeći u sudnicu, svjedok pokazuje optuženog.
entering in courtroom witness points accused
'While entering the courtroom, the witness points to the accused.'

d. BACKWARD ANAPHORA – OVERT PRONOUN – CONTROL (BAOPC)
Dok sudac ulazi u sudnicu, svjedok pokazuje optuženog.
while judge enters in courtroom witness points accused
'While the judge enters the courtroom, the witness points to the accused.'

Each sentence was accompanied by three pictures, corresponding to the choice of the matrix subject (S), the matrix complement (C), and the extralinguistic referent (ER). The thee picture types are numbered 1, 2 and 3 in Fig. 1 respectively.



1 2 3

Figure 1: Example of a picture set

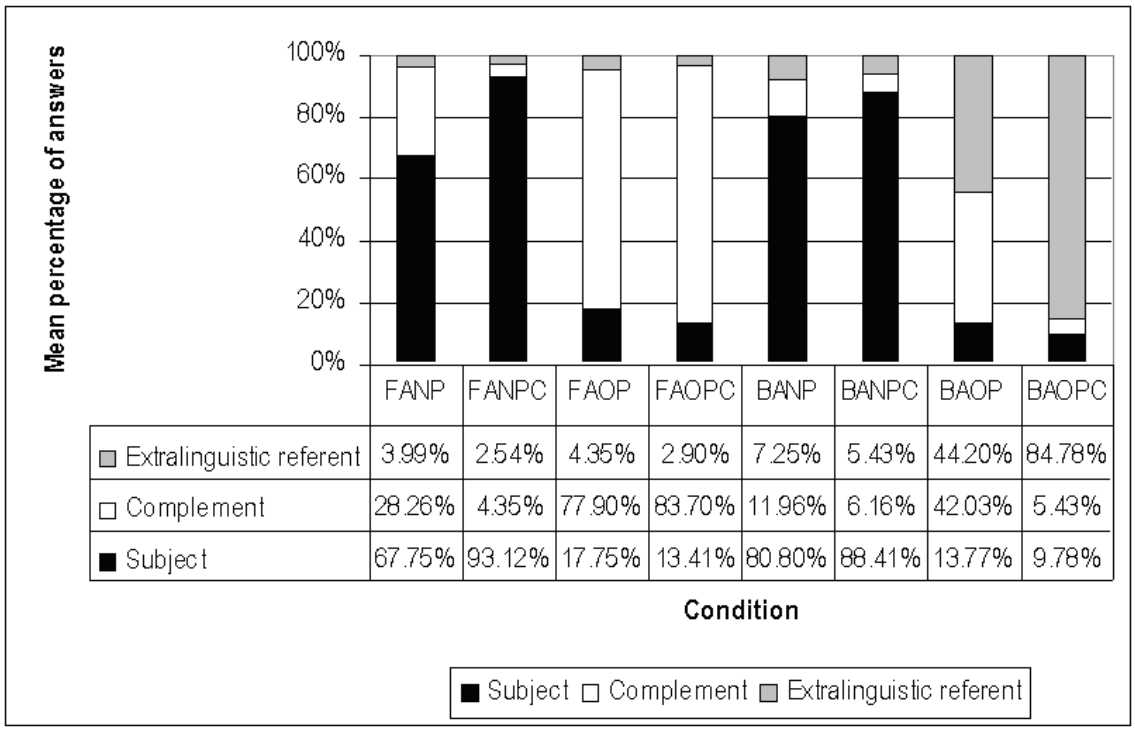
A total of 384 sentences were used in the task. They were divided into eight lists, each containing six items per condition. In a Latin Square design, forty-eight picture sets were rotated around eight conditions.

3.4. Procedure

The experiment lasted approximately ten minutes. It was implemented with SuperLab Pro 2.0 and run on an IBM ThinkPad with a 14.4" screen. Sentences were presented in a speedy word-by-word manner, in a random order for each subject. The pictures appeared on the screen immediately after the last word of the sentence had disappeared. Response time was not limited.

3.5. Results

The number of times each referent was chosen by each subject in each experimental condition was counted, and then the proportion of the three referents in each condition was calculated for each subject. The distribution of responses in all eight conditions is shown in Fig. 2.



	FANP	FANPC	FAOP	FAOPC	BANP	BANPC	BAOP	BAOPC
Extralinguistic referent	3.99%	2.54%	4.35%	2.90%	7.25%	5.43%	44.20%	84.78%
Complement	28.26%	4.35%	77.90%	83.70%	11.96%	6.16%	42.03%	5.43%
Subject	67.75%	93.12%	17.75%	13.41%	80.80%	88.41%	13.77%	9.78%

Figure 2: Choice of referent in different experimental conditions

Three ANOVAs with repeated measures with anaphora type (FANP, FAOP, BANP, BAOP) and ambiguity (ambiguous, unambiguous) as within-subject factors were performed by subject on the individual percentages of responses, one for each referent. In all three ANOVAs, there was a significant main effect of anaphora type (S: $F(3,141) = 367.343$, $p<.001$; C: $F(3,141) = 264.171$, $p<.001$; ER: $F(3,141) = 267.645$, $p<.001$) and of ambiguity (S: $F(1,47) = 15.760$, $p<.001$; C: $F(1,47) = 61.555$, $p<.001$, ER: $F(1,47) = 68.326$, $p<.001$), suggesting that the three referents were chosen to a different degree in different types of anaphora, and in ambiguous and unambiguous sentences. A significant interaction between anaphora type and ambiguity in all three ANOVAs (S: $F(3,141) = 17.367$, $p<.001$; C: $F(3,141) = 27.681$, $p<.001$; ER: $F(3,141) = 68.326$, $p<.001$) indicates that in some conditions the difference in the degree to which the three referents were chosen in ambiguous and unambiguous sentences was bigger, and in some smaller.

In Fig. 2 it can be seen that in the two types of anaphora with a null pronoun, the subjects in most cases interpreted the pronoun as co-referential with the matrix subject, especially in backward anaphora. Their second choice of antecedent for the pronoun was the matrix complement, and the third the extralinguistic referent. In the unambiguous versions of the two sentence types, preferences for the subject referent were stronger and those for the other two referents weaker in comparison to the ambiguous versions.

Contrastively, in the two types of anaphora with an overt pronoun, the subjects did not opt for the matrix subject as the antecedent for the pronoun in the majority of cases, but rather for one or either of the other two referents. More precisely, in forward anaphora they mainly chose the matrix complement, while in backward anaphora their choices were split between the matrix complement and the extralinguistic referent. Of the remaining two referents, the subjects preferred the matrix subject to the extralinguistic referent in forward anaphora. In the unambiguous version of this anaphora, the relative ratios of the three referents chosen were the same as in the ambiguous version, with the proportion of choice of the matrix complement increasing at the expense of the other two referents. In the unambiguous version of backward anaphora, the subjects had a clear preference for the extralinguistic referent, choosing the matrix complement to the lowest degree.

4. Discussion and conclusions

The study was designed to test whether null and overt subject pronouns in Croatian have the same antecedent preferences in ambiguous intra-sentential anaphora as in Italian. Considering the grammatical properties of the two languages, we expected this to be the case, and thus predicted that the participants in our study would resolve anaphora in a way similar to Italian native speakers in previous studies. More precisely, we predicted that they would prefer the subject antecedent for the null pronoun and a non-subject antecedent for the overt pronoun, and that they would allow both non-subject referents that were present in the context of situation as antecedents for the overt pronoun in backward anaphora.

All of these predictions were confirmed in the study. The fact that in unambiguous, control, conditions the subjects had a categorical preference for one of the referents indicates that the study was properly designed.

The results of our study suggest that Italian and Croatian might exhibit a total overlap in the domain of intra-sentential anaphora, i.e. that third person null and overt subject pronouns tend to establish co-reference with other sentence elements in the same way in the two languages. In both languages, anaphora resolution seems to be guided by the discourse-pragmatic principle according to which the overt pronoun signals topic shift, and the null pronoun topic maintenance.

This study represents an initial step in extending Carminati's (2002) generalisation on the antecedent preferences of null and overt subject pronouns in intra-sentential anaphora in Italian to other null subject languages. Future studies, both typological and psycholinguistic, should seek to determine to what extent this generalisation holds crosslinguistically. These studies should also consider a wider range of contexts, as it is possible that crosslinguistic differences in the co-referential properties of null and overt pronouns emerge only under very specific conditions.

References

Belletti, A.; E. Bennati; A. Sorace. 2007. Theoretical and Developmental Issues in the Syntax of Subjects: Evidence from Near-Native Italian. *Natural Language and Linguistic Theory 25*, 657–689.

Carminati, M.N. 2002. *The Processing of Italian Subject Pronouns.* Unpublished Ph.D. dissertation, University of Massachusetts at Amherst.

Grimshaw, J. & V. Samek-Lodovici. 1998. Optimal Subjects and Subject Universals. In P. Barbosa; D. Fox; P.

Hangstrom; M. McGinnis; D. Pesetsky, ed., *Is the Best Good Enough? Optimality and Competition in Syntax*, Cambridge, MA/London, England: The MIT Press, 193–219.

Kraš, T. 2008. Anaphora Resolution in Near-Native Italian Grammars: Evidence from Native Speakers of Croatian. In R. Leah; F. Myles; A. David, ed., *EUROSLA Yearbook: Volume 8*, Amsterdam/Philadelphia: John Benjamins, 107–134.

Rizzi, L. 1982. *Italian Syntax*, Dordrecht: Foris.

Rizzi, L. 1986. Null Subjects in Italian and the Theory of pro. *Linguistic Inquiry 17*, 501–555.

Serratrice, L. 2005. Anaphora Resolution in Monolingual and Bilingual Italian Acquisition. In A. Brugos; M.R. Clark-Cotton; S. Ha, ed., *Proceedings of the 29th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla Press, 504–515.

Sorace, A. 2000. Differential Effects of Attrition in the L1 Syntax of Near-Native L2 Speakers. In S.C. Howell; S.A. Fish; T. Keith-Lucas, ed., *Proceedings of the 24th Annual Boston University Conference on Language Development*, Somerville, MA: Cascadilla Press, 719–725.

Sorace, A. 2005. Selective Optionality in Language Development. In L. Cornips; K.P. Corrigan, ed., *Syntax and Variation: Reconciling the Biological and the Social*, Amsterdam/Philadelphia: John Benjamins, 55–80.

Sorace, A. & F. Filiaci. 2006. Anaphora Resolution in Near-Native Speakers of Italian. *Second Language Research 22*, 339–368.

Tsimpli, I.; A. Sorace; C. Heycock; F. Filiaci. 2004. First Language Attrition and Syntactic Subjects: A Study of Greek and Italian Near-Native Speakers of English. *International Journal of Bilingualism 8*, 257–277.

TOWARDS ENCODING EVENT STRUCTURE IN WORDNET

Svetlozara Leseva

Institute of Bulgarian Language, Bulgarian Academy of Sciences
1113 Sofia, Bulgaria, 52 Shipchenski prohod blvd., bl.17
zarka@dcl.bas.bg

ABSTRACT

The paper deals with some of the main mechanisms of verb derivation through prefixation that affect event structure. The study examines in parallel the event structure of the simplex (focused on Bulgarian activity predicates) and the derived verbs and discusses the effects induced by prefixation drawing on the lexical semantic representations laid out in the work of Levin and Rappaport Hovav (henceforth - L&RH/RH&L) and the analysis of Russian prefixation introduced by Spencer and Zaretskaya (henceforth S&Z). After an outline of the derivation mechanisms and the effects they produce in Section 2, the implications for verb encoding in wordnet are discussed in Section 3.

1. Introduction

Slavic preverbs have been found to derive accomplishment verbs from underived (activity or state) predicates by affecting the simplex (root) in such a way as to change its semantic and syntactic properties (Van Valin & LaPolla 1997, Slabakova 2005, Dimitrova-Vulchanova 2003, S&Z 1998, Svenonius 2004, among others). The changes effected thereby may be stated in terms of the number of participants involved in the eventuality and the number of subevents defined at event structure, the argument position (at event structure) in which the participants project, the relation each participant holds to the event.

2. Event structure

Event structure is understood as a structured lexical semantic representation of the verb's meaning (RH&L 1998, L&RH 1999, L&RH 2005 and similar accounts) that represents the grammatically salient aspects of verb meaning in terms of the configuration of the subevents associated with the eventuality in the form of event structure templates.

It has been established that preverbs derive accomplishments from activity verbs. For instance, Van Valin and Lapola (1997) regard the lexicalised activity-accomplishment alternation in Russian as the result of the application of a lexical rule. S&Z (1998) analyse a group of Russian prefixed resultatives (accomplishments) in parallel to the English resultative construction as derived through a process of secondary predication whereby two semantic predicates are applied to a single argument. Although the focus of their study is on verbs formed by lexical subordination[1] the claim extends to all kinds of resultatives. Both accounts recognize that preverbs effect a change in the logical structure (LS) or lexical conceptual structure (LCS), respectively. A complex predicate is formed compositionally that contains one (core) predicate corresponding to the activity and a second (or secondary) one associated with a result state (S&Z 1998). In L&RH's terms prefixed resultatives are formed by a process of template augmentation whereby a result subevent is added to the activity event structure template.

For instance the verbs *cheta* (read – 'interpret something that is written or printed'[2]) and the prefixed resultative pair ***pro-****cheta* (perfective) – ***pro****-chitam* (secondary imperfective) ('read through') involve two semantically obligatory participants (or arguments) - a human agent (reader) and the material that is read. The simplex describes the activity of reading, and although it implies that it may have a result by virtue of its nature, it does not involve such result obligatorily. Hence, its event structure (following RH&L 1998 and subsequent work) consists of a single activity subevent involving the first participant - x (the reader), while the second participant (y) is optionally expressed (1):

[1] Drawing on the concept of lexical subordination introduced by Levin and Rappaport (1986) S&Z (1998) treat certain Slavic preverbs that produce resultatives whose meaning is "incompatible" with the simplex' as "the predicator of a complex predicate, with the activity verb … as a subordinate predicator" and the process whereby such verbs are produced as lexical subordination. I assume that this mechanism underlies the interaction between the prefix and the verb's root that corresponds to non-canonical results.

[2] All definitions given below, except for Bulgarian specific prefixed verbs, are taken from Princeton WordNet 3.0. http://wordnet.princeton.edu/

(1) V *cheta*: [x ACT ($y_{(THING)}$)]

On the contrary, the prefixed verbs have a complex event structure formed by an augmentation of the simplex' through the addition of a result subevent associated with the second participant. Each subevent identified in the event structure must be associated with at least one syntactically expressed argument (Argument-per-subevent condition: L&RH 1999). Therefore the expression of the second participant with the prefixed verbs is obligatory (2):

(2) ***pro***-*V* - ***pro***-*cheta*/***pro***-*chitam*: [x CAUSE [BECOME [$y_{(THING)}$ STATE(READ)]]]

A point of note is needed here. Firstly, not all kinds of preverbal prefixation result in delimiting the event and adding a result state subevent to the event structure, for instance certain quantificational and phasal prefixes (cf. Slabakova 2005 for a basic distinction between prefixes on these grounds) do not. Secondly, perfective and secondary imperfective verb pairs share common features at the levels of event structure and argument structure. Therefore, they will be treated alike for the purposes of this analysis, and the pair will be juxtaposed to the simplex verb that has distinct semantic and syntactic properties (at least with respect to the verbs regarded herein).

2.1. Canonicity and non-canonicity

The changes in event structure and argument structure induced by prefixation will be explained here in terms of the (non-)canonicity of two components of event structure - resultative subevents, and the participants involved in these subevents.

Argument participants identified in the event structure of both the simplex verb and its prefixed counterparts will be referred to as canonical participants, whereas obligatory participants with the prefixed verbs not identified in the event structure of the simplex (although they might be present in the situation described by the root, and syntactically expressed as adjuncts) are non-canonical participants.

Canonical subevents are result subevents that are construed as the natural result of the activity denoted by the simplex, e.g. (2) above, the natural and simple result of the activity of reading being the reading through of some material. Canonical subevents involve the affected participant in the eventuality (in the sense of Jackendoff 1990 among others). Respectively, non-canonical subevents do not fulfil these requirements.

In terms of primary and secondary predication non-canonicity is involved where the prefix becomes the primary predicate and 'subordinates' the activity predicate, whereas canonicity corresponds to changes in event structure that do not alter 'the distribution of core and secondary predication' (following S&Z 1998).

2.1.1. Canonical and non-canonical participants

Example (2) above presents an event structure involving canonical participants, where the second one is realised in a canonical subevent. In comparison, the verb pair ***nad***-*byagam* – ***nad***-*byagvam* (outrun – 'run faster than') has two participants (4), while the simplex *byagam* (run - 'move fast by using one's feet, with one foot off the ground at any given time') has only one (3). Consequently, the second participant in the event structure of the derived verbs is classified as a non-canonical participant:

(3) V- *byagam*: [x ACT]

(4) ***nad***-V - ***nad***-*byag(v)am* (outrun): [x CAUSE [$y_{(PERSON)}$ BECOME [STATE (SURPASSED)]] BY-MEANS-OF(RUNNING)]

2.1.2. Non-canonical results involving canonical participants

From the above definition it follows that a non-canonical result may involve an unaffected canonical participant, a non-canonical participant, or an affected canonical participant standing in a non-canonical relation[3] to the eventuality.

Prefixed verbs that realise resultative subevents involving an unaffected canonical participant have the same set of participants and selectional restrictions as the simplex. However, the resultative subevent is associated with a participant other than the one involved in the canonical result, e.g. *lepya* (glue – 'join or attach with or as if with glue') – ***ob****-lepyam* - ***ob****-lepya* 'cover on all sides or the whole surface of by gluing)', where the affected participant with the simplex is the stuff being glued (5), while with the prefixed verbs it is the surface to which things are attached by gluing (6).

(5) V - *lepya* : x [ACT(GLUE) $y_{(STUFF)}$ ON $z_{(PLACE)}$]

(6) ***ob***-V - ***ob****-lepyam/****ob****-lepya*: [x CAUSE [BECOME [$z_{(PLACE)}$ STATE(COVERED)]] WITH $y_{(STUFF)}$ BY-MEANS-OF(GLUING)]

2.1.3. Non-canonical results involving non-canonical participants

By virtue of the above definition of non-canonicity, result states associated with non-canonical participants are by default non-canonical subevents. Non-canonical participants are ones newly introduced to the event structure of the derived verbs in a number of ways of which I will consider the following: (i) the new participant takes a specially opened position in the event structure; (ii) the new argument takes an existing position, the canonical participant being a) demoted, or moved, to a newly opened position in the resulting verb's event structure, or b) deleted - removed from the event structure.

In the first case a new argument position in the event structure is created associated with the new participant involved in the result state (7), (8):

(7) V - *kreshtya* (shout - 'utter a sudden loud cry'): [x ACT(SHOUT)]

(8) ***nad***-V – ***nad****-kreshtya(vam)* (outcry, outshout - 'shout louder than'):

[x CAUSE [BECOME [$z_{(PERSON)}$ STATE(SURPASSED)]] BY-MEANS-OF(SHOUTING)]

When a non-participant with the simplex is construed as the affected participant with the prefixed verb pair, one possibility is that the simplex' canonical participant be left out from the event structure of the derived verb (9), (10):

(9) V - *pisha* (write - 'write or name the letters that comprise the conventionally accepted form of (a word or part of a word)'): [x ACT(ПИША) $y_{(THING)}$ (WITH $z_{(INSTRUMENT)}$)] (…) - adjunct

(10) ***iz***-V- ***iz****-pisvam/izpisha* ('run out of ink[4] by writing')

[x CAUSE [BECOME [$z_{(INSTRUMENT)}$ STATE(USED UP)]] BY-MEANS-OF(WRITING)]

Demotion (movement to a lower position) takes place when the canonical affected participant is not removed from the event structure, but becomes a non-affected one and is realised in a newly opened position[5]. Consider the verbs: e.g. *pleta* (knit - 'make (textiles) by knitting') – ***v****-pleta* – ***v-****plitam* (knit in – 'intertwine a yarn in some texture') (11), (12):

---

[3] Affected canonical participants in a non-canonical relation are found in the following case : tarkam (rub - 'move over something with pressure') - pretark(v)am (meaning 'rub sore' or 'fray'), where the derived verbs are interpreted as 'injure by rubbing', although the result state involves the affected canonical participant. This kind of semantic derivation will not be considered in the paper. Suffice it to say that it is a case of non-canonical derivation and is treated alike.

[4] Example taken from S&Z (1998)

[5] Although it is theoretically possible to have a non-affected participant removed, and its place occupied by the canonical affected participant such examples have not been attested, for the time being.

(11) V: *pleta* [x ACT(KNIT) $y_{(OBJECT)}$]

(12) V: ***v**-plitam/**v**-pleta* [x CAUSE [BECOME [$z_{(OBJECT)}$ STATE (INSERTED)]] IN $y_{(OBJECT)}$ BY-MEANS-OF(KNITTING)]

3. Implications for verb encoding in wordnet

3.1. Bulgarian prefixed verbs and wordnet

Verbs in wordnet form semantic domains (Fellbaum 1990) based on their invariant conceptual meaning, and syntactic properties. Although Bulgarian wordnet (Koeva, Mihov and Tinchev 2004) follows the overall structure of Princeton wordnet, it has been enhanced with language specific synsets and features, including semantic relations within synsets (such as the lliteral relations LNOTE) and between synsets. Event structure encoding has been conformed to the general approach.

Unlike English verbs that may have an activity and accomplishment readings depending on the context they are used in, particularly the availability of a complement that delimits the event, and certain properties of this complement (such as referentiality) (L&RH 2005), in Bulgarian (and other Slavic languages) the activity-accomplishment alternation is lexicalised. For the sake of preserving the interlingual correspondences, the difference in lexical aspect is ignored. However, this is but a temporary solution since there is no relation of equivalence between the imperfectiva tantum activity verb and the prefixed pair. In fact, suggestions for the encoding of aspectual verbs in distinct synsets linked to the original one through a derivational relation have been made (Koeva 2008).

In the line of the current analysis the activity-accomplishment alternation is associated with a lexical derivation that yields a canonical subevent that encodes a natural result of an activity. The resultative verbs may be described by the metadefinition 'bring the activity of V to a result' (the treatment of prefixes' semantic contribution to verbs' semantics by means of metadefinitions for Bulgarian was suggested by Ivanova 1974).

Activity-accomplishment alternations are implicitly coded in the synset by means of the lexical and grammatical aspect of the corresponding literals (words): *imperf. t.* (imperfectiva tantum) labels activity and state verbs (unbounded events), and *nesv. v.* and *sv. v.* denote the imperfective/perfective members of the synset, whose meaning may be termed general resultativity (Ivanova 1974) (13):

(13) BG: pisha:3; LNOTE: imperf. t.; napisvam:3; LNOTE: nesv. v.; napisha:3; LNOTE: sv. v.; izpisvam:2; LNOTE: nesv. v.; izpisha:2; LNOTE:sv. v.;

EN: spell:8; write:7;

DEF: 'write or name the letters that comprise the conventionally accepted form of (a word or part of a word)'

Certain canonical resultative verbs are derived by preverbs that besides resultativity contribute an additional meaning component, referred to here as specified result. These resultative verbs denote a subordinate (more specific) concept as compared with the simplex. The metadefinition would run as follows: 'bring the activity of V to a result in a specific manner' where V is the simplex, and the manner (construed in a very general sense) is contributed by the prefix' semantics. Verbs that suffice this condition satisfy the definition of the hypernym/hyponym relation and are encoded as hyponyms of the root verb[6]. Take for example the verb synsets in (15) and (16):

(15) BG: mazha:1; LNOTE: imperf.; namazvam:1; LNOTE: nesv. v.; namazha:1; LNOTE: sv. v.

EN: spread:19;

DEF: 'distribute over a surface in a layer'

[6] Note that this is the convention for verbs that do not have correspondences in English; existing senses in PWN are accordingly encoded.

(16) BG: mazha:2; LNOTE: imperf.; namazvam:2; LNOTE: nesv. v.; namazha:2; LNOTE: sv. v.

EN: spread:20;

DEF: 'cover by spreading something over'

The first synset lexicalises a DISTRIBUTE[7] verb, while the second is a COVER verb; therefore they pertain to different hypernym trees in wordnet. The verbs ***raz****-mazvam/****raz****-mazha* ('distribute over a surface in a layer in different directions') and ***ob****-mazvam/****ob****-mazha* ('cover by spreading over the whole surface or on all sides') are derived respectively from *mazha:1* and *mazha:2*. The additional meaning components contributed by the particular prefix - ***raz***- and ***ob***-, are subsumed under the above metadefinition (a specific manner) and the resultative verbs are encoded under the relevant sense of *mazha* (17), (18):

(17) BG: razmazvam:1; LNOTE: nesv. v.; razmazha:1; LNOTE: sv. v.;

DEF: 'distribute over a surface in a layer in different directions'

HYPERNYM: mazha:1; namazvam:1; namazha:1

EN: spread:19

DEF: 'distribute over a surface in a layer'

(18) BG: obmazvam:1; LNOTE: nesv. v.; obmazha:1; LNOTE: sv. v.;

DEF: 'cover by spreading over the whole surface or on all sides'

HYPERNYM: mazha:2; namazvam:2; namazha:2

EN: spread:20

DEF: 'cover by spreading something over'

Non-canonicity involves the subordination of the simplex predicate by the secondary predicate (the preverb). Therefore, it involves a change in the lexical meaning of the prefixed verb that results in semantic class shift. For instance, the same locative prefix ***ob***- combines with Bulgarian verbs denoting various kinds of activities, e.g. DISTRIBUTE verbs (scatter, splash, etc.), ATTACH verbs (glue, nail) to produce COVER verbs; REMOVE (dig), ATTACH (sew), and PUT (thread) verbs to produce SURROUND verbs where the activity predicate's meaning is conceptualised as the manner in which the result is accomplished. The verbs resulting from the above semantic derivations have the following metadefinitions: 'cover the whole surface of/on all sides by V-ing', 'surround some place/surface (entirely) by V-ing'. The verbs that fulfil this requirement are hence encoded as hyponyms of an appropriate COVER or SURROUND synset, or possibly another more appropriate one with similar semantics (19):

(19) BG: oblepvam:1; LNOTE: nesv. v.; oblepyam:1; LNOTE: nesv. v.; oblepya:1; LNOTE: sv. v.;

DEF: 'cover the whole surface of/on all sides by gluing'

HYPERNYM: pokrivam:7; pokriya:6; zakrivam:1; zakriya:1;

EN: cover:22

DEF: 'provide with a covering or cause to be covered'

[7] The verbs spelled in capital letters denote tentative verb classes named after the hypernym that describes the common semantics of its subordinates. It may or may not be a tree root.

It should be noted that the type of result depends both on the semantics of the particular root and the prefix. As seen above ***ob***-prefixation yields a canonical result with COVER verbs such as *mazha*, and non-canonical ones with ATTACH verbs. This holds also for the rest of the preverbs illustrated here – ***raz-, v-***, as well as for many others.

For the time being the derivation from the simplex to the non-canonical verb is not marked in wordnet. A possible approach to account for it is to use the metadefinitions suggested above to formulate semantic relations and link the source and the target synsets of the derivation, e.g. 'extension over location' and 'extension over perimeter' for ***ob***-prefixation. The advantage of such way of defining relations is that they will be independent of the particular semantic classes involved and will capture the semantic component of the prefix, while at the same time keeping track of the derivation.

4. Conclusion

In the paper it has been shown on a very modest scale that mechanisms underlying the processes of prefixation affect and are describable in terms of event structure changes, and that the lexical semantic account of these processes may be applied to verb encoding in wordnet. Future work will focus on expanding the scope and size of the study as well as on elaborating and specifying the theoretical model of derivation processes and effects with a view to capturing properly event structure and argument structure alterations. Another line of investigation will be the analysis of the semantic contribution of prefixes in terms of semantic components that need to be integrated in event structure (cf. Filip 1993). Based on these analyses derivational relations will be defined that account for the type of semantic modifications on resultatives induced by prefixation, such that produce hyponyms, and the semantic derivation that results in semantic class shifts.

References

Dimitrova-Vulchanova, M. 2003. On two types of results. Resultatives revisited. In D. Beermann and Lars Hellan, eds, Proceedings of TROSS 03 – Trondheim Summer School. http://edvarda.hf.ntnu.no/ling/tross/

Fellbaum, C. 1990. English Verbs as a Semantic Net. International Journal of Lexicography, 3(4):278-301.

Filip, H. 1993. Aspect, situation type and nominal reference. Unpublished Ph.D. dissertation, UC at Berkeley.

Ivanova, K. 1974. Nacini na glagolnoto deystvie v savremenniya balgarski ezik. Sofia.

Jackendoff, R. 1990. Semantic Structures. The MIT Press, Cambridge, MA 1990.

Koeva, S., S. Mihov, T. Tinchev. 2004. Bulgarian Wordnet – Structure and Validation. In: Romanian Journal of Information Science and Technology, volume 7, numbers 1-2, 2004, pp 61-78.

Koeva, S. 2008. Derivational and Morpho-Semantic Relations in Wordnet. In: Intelligent Information Systems, 2008, pp 359-368.

Levin, B. and M. Rappaport Hovav. 1999. Two Structures for Compositionally Derived Events, Proceedings of SALT 9, 199-223.

Levin, B. and M. Rappaport Hovav. 2005. Argument Realization, Cambridge University Press 2005.

Rappaport Hovav, M. and B. Levin. 1998. Building Verb Meanings. In M. Butt and W. Geuder, eds., The Projection of Arguments: Lexical and Compositional Factors, CSLI Publications, Stanford, CA, 97-134.

Slabakova, R. 2005. Perfective Prefixes: What they are, what flavors they come in, and how they are acquired?. In S. Franks, Frank Y. Gladney and Mila Tasseva-Kurktchieva, eds., Formal Approaches to Slavic Linguistics 13: The South Carolina Meeting, 324-341. Ann Arbor, MI: Michigan Slavic Publications (to appear).

Spencer, A. and M. Zaretskaya. 1998. Verb prefixation in Russian as lexical subordination. Linguistics 36: 1–39.

Svenonius, P. 2004. Slavic prefixes inside and outside VP. Nordlyd Vol. 32, Nr. 2 – Slavic Prefixes, 205-253.

Van Valin, R. and R. LaPolla. 1997. Syntax: structure, meaning and function. Cambridge: Cambridge University Press 1997.

DOCUMENT REPRESENTATION METHODS FOR NEWS EVENT DETECTION IN CROATIAN

Nikola Ljubešić, Željko Agić, Nikola Bakarić

Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{nljubesi, zagic, nbakaric}@ffzg.hr

ABSTRACT

Constant increase in the amount of available data in the world in general demands new organizational and representational ideas and approaches. Document clustering as a method for event detection uses, supplements and upgrades existing information retrieval methods in order to improve knowledge management and representation. This article describes the research done in order to determine the impact of various methods of document representation on cluster analysis. Several statistical and linguistic NLP morphological normalization methods of document representation are tested in an event detection scenario. Event detection was conducted using online newspaper articles issued on a single day. A cluster analysis was done using the various document representation methods and a clustering algorithm. The results were then compared against a human evaluated golden standard. The results show that both statistical and linguistic methods simplify the representational complexity and minimally improve the results which lead to the conclusion that for this task statistical methods should be preferred.

1. Introduction

The ever-increasing amount of various data in the world has reached a point where standard methods of knowledge management and information retrieval are no longer adequate and need to be complemented with other methods. Contemporary processing of large amounts of data is primarily supervised because they provide better results. Due to the acquisition bottleneck problem the challenge is to create automated, unsupervised methods which can then deal with large sets of data without expensive and tedious human efforts. Therefore, we decided to investigate the applicability and effectiveness of natural language processing methods for automated document clustering.

Cluster analysis is a well-established method for creating order in large sets of data (not only textual). Document clustering as a method for organizing documents into clusters is commonly used in combination with the following information retrieval (IR) methods (Forster 2006):

- support for presentation of information retrieval results - Main-stream information retrieval methods, even plain text search (e.g. finding documents containing a certain word or words), are powerful and proven methods. However, they are often plagued by a large number of irrelevant results, especially in large data sets. Using document clustering it is possible to expand this simple query by selecting a relevant document from the list of results and use it as the query. The results of this method will be more relevant and considerably fewer in number.
- support for document retrieval - Document retrieval using document clustering is based on the same principles as classic IR but searches organized clusters instead of unorganized document collections. Therefore, the query results are not clouded by polysemy and ambiguous terms.
- direct access to documents – These methods are generally based on tracking or mediating user's actions and queries in a limited environment and do not rely directly on text retrieval.

Document clustering consists of two main components: document representation and cluster analysis. Document representation deals with 'translating' documents (articles, web pages, etc.) into structures suitable for clustering (Forster 2006). This is usually done by representing documents numerically as vectors and matrices. Cluster analysis includes methods for creating meaningful data clusters from the data structure produced by methods of document representation. Clusters are created after comparing (measuring the distance) the numerical representations of the documents.

Natural language processing is a large field with many applications. Here it is used to offer a linguistic and statistical view on document representation in order to determine if representing a document as more than a collection of random characters

has any impact on clustering. NLP methods used in this research include tokenization, stemming, lemmatization and n-grams.

Our goal is to compare the results of document clustering using several linguistic and statistical NLP methods and to determine their effectiveness against a human evaluated gold standard. The final goal of the investigated methods is to organize online newspaper articles into clusters which report on a singular event.

2. Problem

This article investigates the effects of linguistic and statistical preprocessing of data as part of document representation in the document clustering task. Our document clustering task aims at organizing online news articles into clusters where every cluster corresponds to a specific event. This task in literature is often called event detection (Yang et al. 1998, Papka et al. 1999). One of the most known online resources that organize information in such manner is Google News (Google 2008).

There are several problems concerning data clustering, the most prominent being combinatorial explosion, a common issue when dealing with large data sets. Most commonly used piece of information aimed at reducing the combinatorial explosion present in data clustering is the time of publishing of the article. Namely, to cluster data, the similarity function between every two data points has to be calculated. Almost every document clustering system uses this crucial piece of information, i.e. event detection calculates clusters on documents published in a specific time frame. In this research we calculate document clusters on articles published during a single day.

Another piece of information often used in online news event detection is the origin of the article, i.e. the publisher. The assumption is that one online publisher will not produce more than one article on a singular event. There is, of course, an objective possibility of erroneous mapping of reports from different publishers pertaining consecutive events (article A1 covering events another publisher covered in more than one article), but this problem is not of interest in this paper.

The evaluation of the efforts described earlier is most often carried out through a gold standard - a data sample organized by hand. In this research we used a small sample of online news published in a single day.

The results of this procedure are aimed at the representation of the data collected by crawlers where all articles covering one event would be presented as one entity.

3. Experiment

The data used in this experiment are obtained from the Institute for Business Intelligence (Zapi 2008) and their web crawler which collects online news on a 10-minute basis. The clustered data consists of 1028 news documents published on May 5 2008. After removing identical documents published on same domains, 1000 articles collected from 18 different domains remained in the data set.

As stated before, the emphasis in this experiment was to test the value of document representation methods as feature extraction methods by comparing their impact on the end result.

The cluster analysis was performed using the cosine similarity measure with a modified single-link hierarchical agglomerative algorithm and a defined threshold. The algorithm starts with every data point forming its own cluster. It merges clusters on the nearest-neighbor principle, merging closest clusters together. When merging clusters, the distance between them is considered as the distance between the two closest data points. The clustering threshold is the criterion that stops the clustering task when there are no clusters as close as the criterion defines. It ensures that the clustering task will not produce just one cluster. A special constraint while performing the clustering task is the fact that one cluster cannot consist of two articles published on the same domain. The clustering algorithm is implemented as follows:

1. Calculate the similarity function between documents
2. Build triples with id-s of documents and their similarity
3. Remove triples whose similarity is lower than the clustering threshold
4. Sort the remaining triples in descending order

5. Move with two nested iterations through combinations of triples
6. Form a new cluster from the first triple and add all following triples, i.e. IDs if they:
    - satisfy the threshold condition
    - are not allocated in any other cluster
    - no article from the same domain is in the cluster already
7. If an article does not meet the third condition from the previous step, do not add any following articles that have a stronger similarity with the article not meeting the condition

No emphasis is put on the weighting method of a selected feature but only the popular tf-idf measure is used (Jones 1973). In order to use the tf-idf measure, the distribution of features over documents has to be known. Therefore a corpus of 30,000 news articles and 6,985,242 tokens is constructed and the distribution of interest is calculated. Also the document space with its corresponding space complexity is defined using this data.

In this research there are six different document representation techniques investigated: TOKEN, STEM, LEMMA, 3-GRAM, 4-GRAM and 5-GRAM.

In the TOKEN representation method the lowercased corpus is tokenized by the python-like regular expression `r'['+ln+']+(?:[-.,@/]['+ln+']+)*'` where the variable `ln` contains all letters and numerals, i.e. token is defined as a sequence of letters and numerals with the characters `'-.,@/'` occurring isolated inside that sequence ('požeško-slavonska', '16.2' 'nick@127.0.0.1' etc.). The TOKEN representation method is considered the baseline of this research.

In the STEM representation method a stemming algorithm still under construction is used on the lowercased corpus since there is no other stemming algorithm for Croatian available for that purpose (the algorithm described in (Ljubešić et al. 2007) is used for normalizing basic word forms in query expansion and the algorithm described in (Šnajder 2006) has not been made publicly available yet). The stemming algorithm used deals with inflectional morphology only and the rules are primarily focused on noun and adjective paradigms.

In the LEMMA representation method the POS-tagging algorithm described in (Agić, Tadić 2006) and (Agić et al. 2008) is used.

In the remaining three representation methods a character n-gram morphological normalization approach as described in (Šilić et al. 2007) is used. In our case, character n-grams are calculated from tokens using the TOKEN representation method (e.g. the token 'imaju' is described through 4-grams '_ima', 'imaj', 'maju' and 'aju_').

These document representation methods are evaluated in the clustering process aiming at event detection. The event detection task is evaluated using a gold standard – 1,000 news articles are manually organized into clusters. Software under development is used for this task. Out of 1,000 articles, 396 of them describe events not covered by any other article, i.e. they form clusters containing just one article. The remaining 604 documents are organized into 144 clusters with an average of 4.19 elements per cluster. The median of the non-one cluster size distribution is 3 and the maximum is 18.

The basic evaluation measures used are precision and recall. They are both calculated through the best-case intersection of the gold standard and the clustering result. Additional evaluation measure is the F0.5 which favors precision twice as much as recall. Namely, the results are used for supporting information retrieval results which makes precision more important than recall.

4. Results/Discussion

The vector space complexity of different document representation methods is shown in Table 1. STEM simplifies the space complexity by 1.3 and LEMMA by almost 2. The highest space simplification is obtained by 3-GRAM, 4-GRAM is equivalent to LEMMA whilst 5-GRAM increases the space complexity.

	Number of dimensions	Simplification coefficient
TOKEN	249,136	1.0
STEM	191,676	1.3
LEMMA	125,406	1.99
3-GRAM	22,264	11.19
4-GRAM	115,693	2.15
5-GRAM	287,325	0.87

Table 1: Space complexity and simplification coefficient regarding document representation methods

In Table 2 the maximum F0.5 evaluation measures with the space simplification coefficient and corresponding clustering threshold criterion concerning the representation method is shown. The data shows no obvious difference regarding the document representation methods with STEM, LEMMA and 4-GRAM outperforming slightly the TOKEN baseline and 3-GRAM and 5-GRAM producing lower results. Taking the simplification factor into account, it would be advisable to use the 3-GRAM method for computationally intensive problems since the space simplification is large. LEMMA and 4-GRAM obtain same results and a similar simplification factor, but the 4-GRAM method is much simpler and language independent. When comparing results of language dependent methods, STEM outperforms LEMMA with a higher F0.5 measure, a lower space simplification factor, but a much simpler and faster method.

	F0.5	Simplification coefficient	Clustering threshold
TOKEN	0.858	1.0	0.25
STEM	0.868	1.3	0.35
LEMMA	0.859	1.99	0.4
3-GRAM	0.853	11.19	0.5
4-GRAM	0.860	2.15	0.45
5-GRAM	0.857	0.87	0.45

Table 2: Maximal F0.5 measure, simplification coefficient and clustering threshold regarding the document representation method

Figure 1 shows precision, recall and the F0.5 measures concerning the clustering threshold criterion (CTC) from the STEM data. As expected, with the threshold rising, precision rises and recall falls. F0.5 experiences its maximum at CTC=0.3.

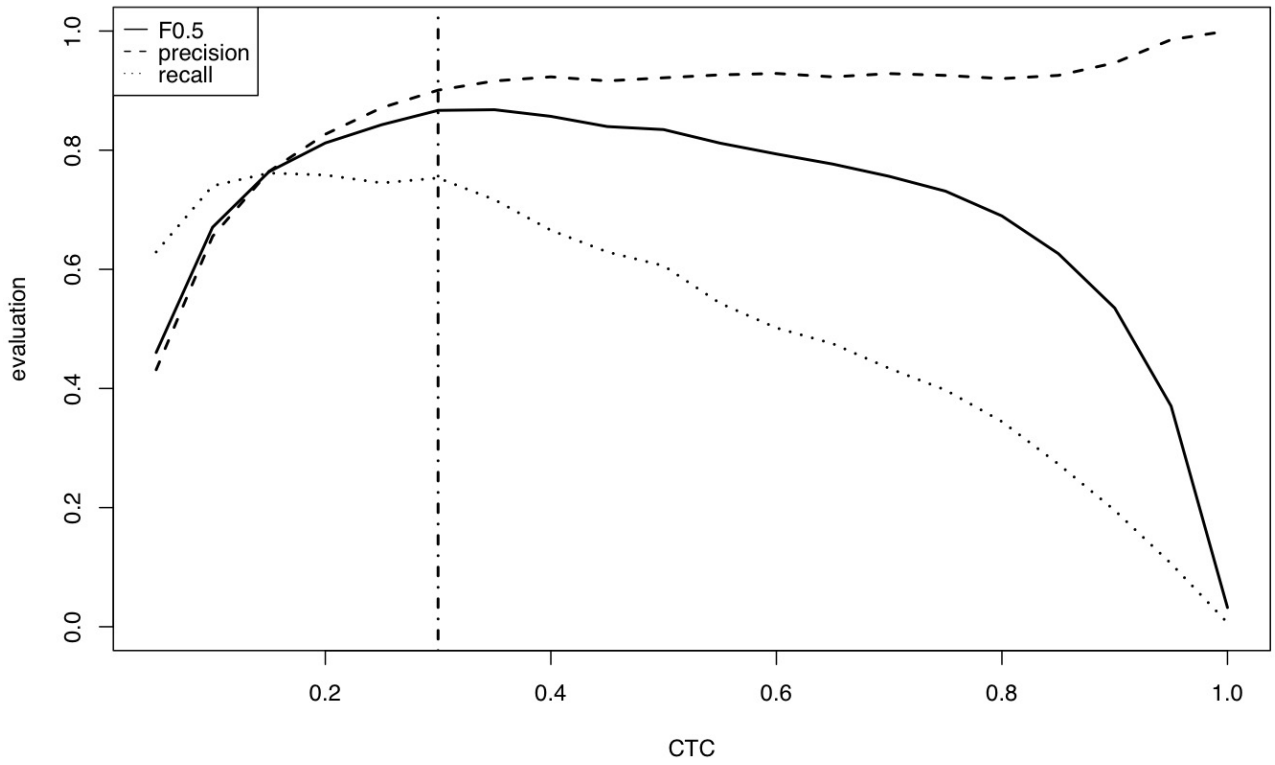


Figure 1: Precision, recall and F0.5 concerning clustering threshold criterion (CTC)

Concerning the variable of the number of elements in clusters with more than one element, it correlates strongly positively with recall (0.993) and negatively with precision (-0.531). The F0.5 measure shows its quality in measuring the overall performance of the clustering task with correlating rather highly with the described variable (0.756).

F0.5 regarding the clustering threshold criterion and ignoring the document representation method is shown in table 3. The criterion reaches its stable maximum at the value of 0.35 and could be recommended for usage regardless of the document representation method.

CTC	Average F0.5
0.05	0.416
0.1	0.580
0.15	0.687
0.2	0.758
0.25	0.806
0.3	0.833
0.35	0.845
0.4	0.851
0.45	0.846
0.5	0.841

CTC	Average F0.5
0.55	0.826
0.6	0.811
0.65	0.796
0.7	0.772
0.75	0.751
0.8	0.714
0.85	0.661
0.9	0.575
0.95	0.417
1.0	0.024

Table 3: Average F0.5 regarding clustering threshold criterion (CTC)

5. Conclusion

In this research we have analyzed the impact of different statistical and linguistic morphological normalization methods in a document clustering i.e. news event detection task. As the baseline we used pure tokenization. No significant improvement was observed when using normalization methods. Possible reason for such results could be the nature of problem because singular events tend to be described by different sources using same word forms. Highest level of the document space simplification was obtained using the character 3-grams with a slight decline in the evaluation measures and is therefore recommended when dealing with computationally demanding tasks. When comparing linguistic methods, stemming slightly outperformed lemmatization. The reason for such results could be low quality of the processed data (HTML escape sequences and such).Lemmatization yielded a higher level of document space simplification. When comparing statistical and linguistic method, they both managed to achieve similar document space simplification and evaluation measures which leads to the conclusion that the statistical methods should be preferred due to their simplicity, higher speed and language independence. Further research will include larger evaluation sets for validation and testing as well as experimenting with collocations and syntactic and semantic processing.

References

Agić, Ž.; Tadić, M.; Dovedan, Z. 2008. Combining Part-of-Speech Tagger and Inflectional Lexicon for Croatian. // Proceedings of IS-LTC (in press)

Agić, Ž.; Tadić, M. 2006. Evaluating morphosyn-tactic tagging of croatian texts. In LREC2006 Proceedings, Genoa-Paris. ELRA.

Forster, R. 2006. Document Clustering in Large German Corpora Using Natural Language Processing. Thesis presented to the Faculty of Arts of the University of Zürich for the degree of Doctor of Philosophy.

Google News. 2008. Google. http://news.google.com/.

Ljubešić, N.; Boras, D.; Kubelka, O. 2007. Retrieving Information in Croatian: Building a Simple and Efficient Rule-based Stemmer // Digital information and heritage / Seljan, Sanja ; Stančić, Hrvoje (ur.). Zagreb : Odsjek za informacijske znanosti Filozofskog fakulteta u Zagrebu. Str. 313-320.

Papka, R.;Croft, B.W.;Barto, A.G.;Danai, K.;Kurose, J.F. 1999. On-line New Event Detection, Clustering and Tracking

Sparck Jones, K. 1973. A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation

Šnajder, J. 2006. Rule-Based Automatic Acquisition of Large-Coverage Morphological Lexicons for Information Retrieval. Technical Report MZOS 2003-082, Department of Electronics, Microelectronics, Computer and Intelligent Systems, FER, University of Zagreb

Šilić, A.; Chauchat, J.; Dalbelo Bašić, B.; Morin, A. 2007. N-Grams and Morphological Normalization in Text Classification: A Comparison on a Croatian-English Parallel Corpus. // Lecture Notes in Artificial Intelligence. 4874; 671-682

ZAPI. 2008. Institute for business intelligence, http://www.zapi.hr.

Yang, Y.; Pierce, T.; Carbonell, J. 1998. A Study on Retrospective and On-line Event Detection. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM press

ON THE PRODUCTIVITY OF REFLEXIVE AND RECIPROCAL *SE* IN SERBIAN

Maja Miličević

Research Centre for English and Applied Linguistics, University of Cambridge
9 West Road, Cambridge CB3 9DP, United Kingdom
mm510@cantab.net

ABSTRACT

This paper deals with the productivity of the Serbian reflexive and reciprocal clitic *se* 'x-self/each other' and with how it relates to the alternative reflexive and reciprocal markers used by Serbian, the reflexive pronoun *sebe* and the reciprocal pronoun *jedan drugog*. The theoretical background assumed in the paper presupposes that *se*-type markers are crosslinguistically employed with verbs frequently used in reflexive/reciprocal form, while the *sebe*-type is normally employed with verbs rarely used as reflexive/reciprocal. The main question thus is whether in Serbian the division of labor between these two types of markers is grammaticalized, limiting *se* to a closed class of frequently reflexivized/reciprocated verbs (similarly to the case of the English unmarked forms), or not grammaticalized, allowing *se* with other verbs too. Since both views are represented in the theoretical literature, the issue was addressed in an extensive empirical study, involving an examination of a written corpus of Serbian and a Picture Judgment Task administered to a group of native speakers. The results obtained by both methods indicate that *se* does tend to occur prevalently with verbs often used as reflexive/reciprocal, but this tendency appears to be a matter of preference rather than a choice imposed by a grammaticalized pattern.

1. Introduction

Most of the world's languages can encode reflexive and reciprocal meaning by employing two different strategies. Since one of them is typically phonologically lighter than the other, the labels *light* and *heavy* strategies (or markers) are commonly used in the typological literature. Representatives of the former type are, for instance, the Serbian clitic *se*, the German pronoun *sich* and the English morphologically unmarked forms (as in *Peter shaved*). The latter type is exemplified by the Serbian pronouns *sebe* (reflexive) and *jedan drugog* (reciprocal), the German pronouns *sich selbst* and *einander*, and the English pronouns *x-self* (*myself*, *yourself*, etc.) and *each other/one another*. Sentences containing the relevant Serbian markers are given in (1) and (2).

(1) a. Ivana **se** oblači.
Ivana REFL.CLI dress.PRES.3SG
'Ivana is dressing herself.'

b. Ivana mrzi **sebe**.
Ivana hate.PRES.3SG herself (REFL.PRO)
'Ivana hates herself.'

(2) a. Petar i Marija **se** ljube.
Petar and Marija REC.CLI kiss.PRES.3PL
'Petar and Marija are kissing.'

b. Petar i Marija izbegavaju **jedno drugo**.
Petar and Marija avoid.PRES.3PL one other (REC.PRO)
'Petar and Marija are avoiding each other.'

The specific issue this paper is concerned with is the productivity of the Serbian light reflexive and reciprocal forms. Crosslinguistically, the distribution of light and heavy markers within languages is never random. In languages such as English, Russian or Dutch it is grammaticalized and the light strategy can only be used with a closed class of verbs such as *wash*, *shave*, *dress* (reflexives), or *kiss*, *hug*, *meet* (reciprocals), while the heavy strategy is obligatorily employed with others; compare *Peter shaved (himself)* and *Peter hates *(himself)*. In other languages, e.g. Italian or German, the contrast can at most manifest itself as a preference for the light strategy with some verbs, and the heavy strategy with others. Therefore, the question in relation to Serbian is whether it belongs to the former or the latter group. In other words, can *se* replace the heavy forms in sentences similar to those in (1b) and (2b), which contain verbs that do not fall into the mentioned closed class of verbs allowing light markers across languages?

The views suggested in the literature differ significantly, and while some authors argue that *se* is fully productive (Marelj 2004, Reinhart and Siloni 2005, Siloni 2001), others claim that it can be used only with a restricted set of predicates such as *obući* 'dress' or *obrijati* 'shave' (Perović 2003). Both views are largely based on the intuitions of a limited number of informants, and what lacks in works arguing for either approach is the support of more extensive empirical evidence. Moreover, the non-productive view deals explicitly only with the reflexive uses of *se*. Given that the findings reported in both typological (Haiman 1983, Kemmer 1993) and theoretical (Reinhart and Siloni 2005) literature indicate that, within languages, reciprocals tend to have the same productivity status as reflexives, in this paper the two will be discussed jointly.

2. Theoretical approach

The productivity of *se* is an issue that falls within a wider crosslinguistic problem of the distribution of reflexive and reciprocal markers. As noted above, this distribution is never random, and it can be grammaticalized or not, but the theories aiming to fully account for it must explain which specific properties of predicates lead to their selecting light or heavy markers.

An important group of accounts of reflexive and reciprocal marking is based on predicate meaning. According to these accounts, a light or a heavy strategy is preferred with a given verb depending directly on its semantic or pragmatic properties, for instance on whether it denotes an action typically performed on oneself or one typically performed on others (Haiman 1983), on whether the object is 'affected' by the action expressed by the verb (Hellan 1988), or on whether the participants in the action are clearly distinguishable or not (Kemmer 1993). Contrary to these views, the approach adopted in this paper, based on Haspelmath (2005), presupposes that the marker distribution might be indirectly related to predicate meaning, but is directly dependent solely on the predicate's frequency of reflexive use. Put differently, Haspelmath argues that, for reasons of language economy, if a verb has a high frequency of use with reflexive objects, it will tend to appear with light marking in reflexive use, and if it is more commonly used with non-reflexive, i.e. disjoint pronominal objects, it will normally carry a heavy reflexive marker.[1] Moreover, rather than being dichotomous, this distinction builds into a continuum of reflexive use. On this continuum, some languages have a clear cut-off point between the verbs that can use the light markers and those that have to take the heavy ones, i.e. they grammaticalize the distinction (English, Dutch), while others do not (Italian, German). Applying the same reasoning to reciprocals, we can assume that verbs often used in reciprocal form will tend to occur with light reciprocal markers, while those whose reciprocal use is rare will, obligatorily or commonly, take heavy markers.

3. The present study

The goal of the present research is to contribute to the debate about whether *se* is productive or not by examining quantitative data coming from a corpus study and from acceptability judgments expressed by native speakers of Serbian. The following subsections provide details about the sample of verbs looked at, the method used and the results obtained.

3.1. The verbs

<table>
<tr><td rowspan="2">Reflexives 1 (n=4)
[frequently reflexivized verbs]</td><td rowspan="2">obući 'dress'
oprati 'wash'
obrijati 'shave'
spremiti 'prepare'</td><td rowspan="2">Reciprocals 1 (n=4)
[frequently reciprocated verbs]</td><td rowspan="2">poljubiti 'kiss'
zagrliti 'hug'
maziti 'caress'
upoznati 'meet'</td></tr>
<tr></tr>
<tr><td rowspan="2">Reflexives 2 (n=4)
[less frequently reflexivized verbs]</td><td rowspan="2">odbraniti 'defend'
zaštititi 'protect'
maskirati 'disguise'
naoružati 'arm'</td><td>Reciprocals 2 (n=3)
[less frequently reciprocated verbs]</td><td>ignorisati 'ignore'
provocirati 'provoke'
izbegavati 'avoid'</td></tr>
<tr><td rowspan="2">Reflexives 3 (n=3)
[rarely reflexivized verbs]</td><td rowspan="2">voleti 'love'
mrzeti 'hate'
poštovati 'respect'</td><td rowspan="2">Reciprocals 3 (n=4)
[rarely reciprocated verbs]</td><td rowspan="2">ubiti 'kill'
otrovati 'poison'
raniti 'wound'
napasti 'attack'</td></tr>
<tr></tr>
</table>

Table 1: Verbs examined in the study

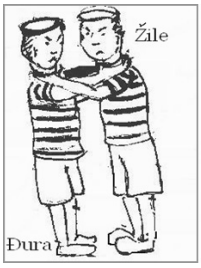
[1] Other uses, e.g. those with full NP objects, are not considered to be relevant for this problem.

Following Haspelmath's approach, the productivity of the Serbian clitic *se* was empirically tested by comparing its use with verbs placed at different portions of the reflexive and reciprocal continua, as it was assumed that such a comparison should reveal whether *se* is unacceptable with verbs less frequently or rarely occurring in reflexive/reciprocal form, or it is just not the preferred choice with them. The verbs used in the study are shown in Table 1 above. Both the verbs used reflexively and those used reciprocally were divided in three groups according to their frequency of reflexive/reciprocal use. The choice of specific verbs included in each group was based on verb lists reported in a number of typological studies (Haiman 1983, Hellan 1988, Kemmer 1993, König & Vezzosi 2004, etc.), representing verbs typically taking light vs. heavy strategies across languages.

3.2. The method

The corpus used in the study was the non-tagged part of the Corpus of Contemporary Serbian Language (*Korpus savremenog srpskog jezika*), containing 22,226,437 words. The corpus is entirely written, containing mainly narrative texts (94%) and a smaller press section (6%). All occurrences of the above 22 verbs were extracted and two separate counts were performed on them. Firstly, two types of sentences were singled out from the rest: reflexive/reciprocal occurrences of the verbs (sentences similar to (1) and (2) above), and their occurrences with disjoint pronominal objects (sentences of the type *Darko brani* ***njega*** 'Darko defends him'). Percentage ratios of these two sentence types were calculated in order to verify the assumed frequency of the verbs' reflexive/reciprocal use. Secondly, the occurrences of clitics and pronouns were counted (and their percentage ratios calculated) for the reflexive/reciprocal sentences.[2]

The same 22 verbs were used in the Picture Judgment Task administered to 20 native speakers of Serbian. All verbs were used in a sentence with the clitic *se* and in one with the pronoun *sebe/jedan drugog*. The participants were asked to mark each sentence on a scale ranging from -3 (completely unacceptable) to +3 (completely acceptable) against a picture illustrating the action expressed by the verb. Figure 1 shows a sample item representing the verb *napasti* 'attack'.



Đura i Žile su se napali. -3 -2 -1 0 +1 +2 +3

Đura i Žile su napali jedan drugog. -3 -2 -1 0 +1 +2 +3

Figure 1: Example of a PJT item

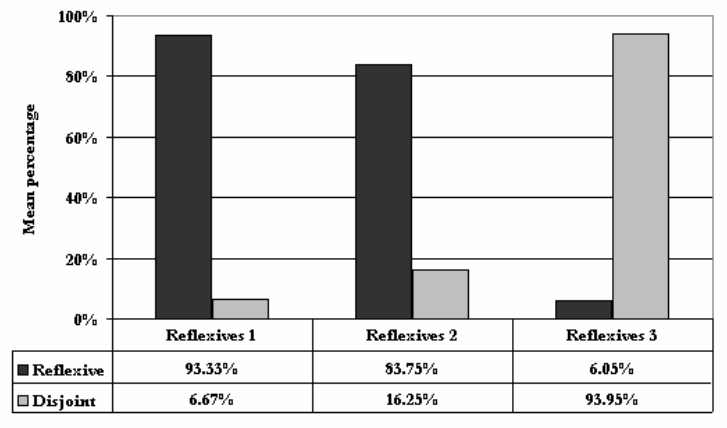
These specific methods were chosen because they enable a comparison of production and judgment data, providing thus a more comprehensive picture than either method on its own.

3.3. Results

In analyzing the results, the first issue to be looked at is the distribution of reflexive/reciprocal vs. disjoint objects in the corpus. The analysis of this distribution is important in order to ascertain whether the initial division of verbs into groups, based on non-quantified typological data, was correct.

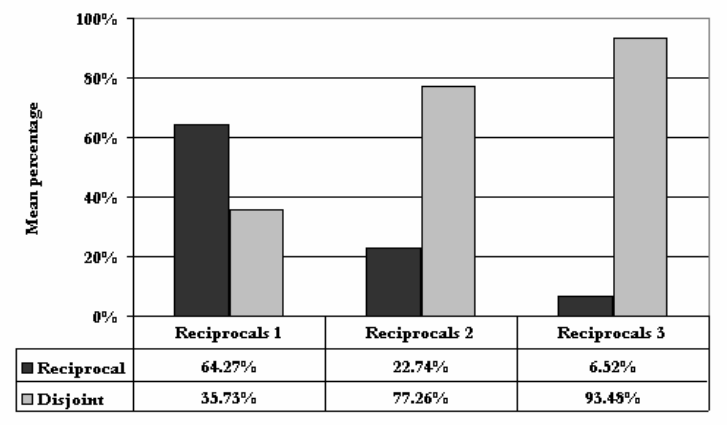
The results do confirm the assumptions about the verb groupings, as the continua of reflexive and reciprocal use are present in the corpus data trends (see Figures 2 and 3). The trends are confirmed by statistical analyses: for both reflexive vs. disjoint and reciprocal vs. disjoint (non-reciprocal) object distribution, one-way ANOVAs found a statistically significant difference between verb groups, for reflexives $F(2,8)=43.491$, $p<0.001$, and for reciprocals $F(2,8)=12.665$, $p<0.01$. The patterns are not completely gradual (for reflexives the difference is brought about by group 3, and for reciprocals by group 1), but this is due to the limited number of verbs and verb groups looked at, and it does not endanger the general assumption.

[2] Only plural verb forms were taken into account for reciprocals, as the singular ones would have created a bias toward disjoint objects.



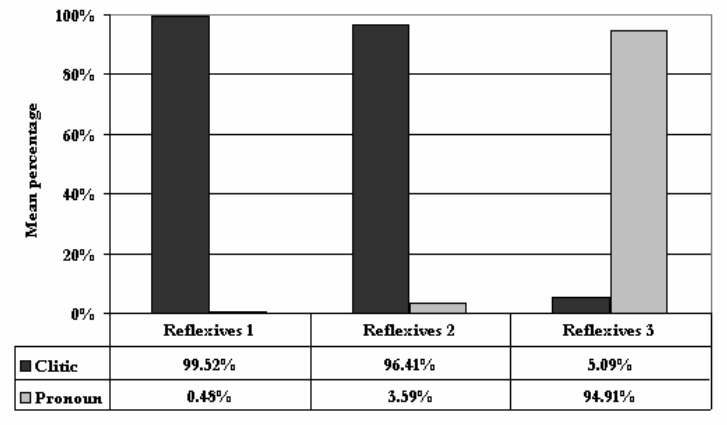
	Reflexives 1	Reflexives 2	Reflexives 3
Reflexive	93.33%	83.75%	6.05%
Disjoint	6.67%	16.25%	93.95%

Figure 2: Corpus distribution of reflexive vs. disjoint pronominal objects



	Reciprocals 1	Reciprocals 2	Reciprocals 3
Reciprocal	64.27%	22.74%	6.52%
Disjoint	35.73%	77.26%	93.48%

Figure 3: Corpus distribution of reciprocal vs. disjoint pronominal objects



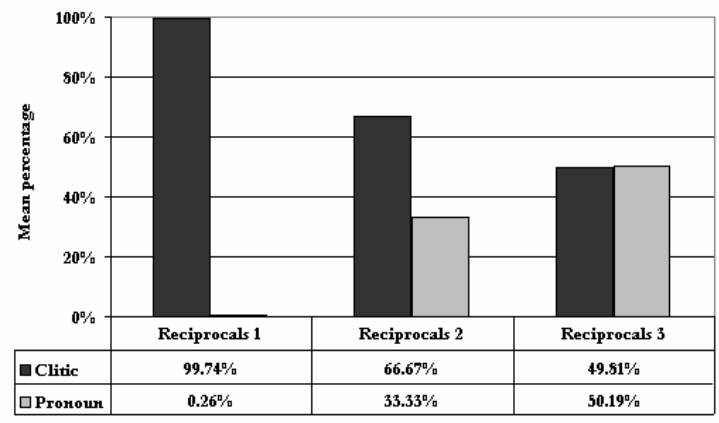
	Reflexives 1	Reflexives 2	Reflexives 3
Clitic	99.52%	96.41%	5.09%
Pronoun	0.48%	3.59%	94.91%

Figure 4: Corpus distribution of reflexive clitics vs. reflexive pronouns

For reflexives, the corpus distribution of light and heavy markers (Figure 4) is extremely similar to that of reflexive and disjoint objects. A one-way ANOVA again shows a significant difference between the three verb groups, $F(2,8)=440.665$, $p<0.001$, with reflexives 3 differing from reflexives 1 and 2. In fact, in the case of reflexives, there is a very high correlation between the object distribution and the marker distribution, r=0.97 ($R^2=0.94$). Even more importantly for the present discussion, it can be noted that, even though rare, some occurrences of the clitic *se* are found with reflexives 3, meaning that a more likely

explanation for their low percentage with respect to pronouns is preferential treatment, rather than a grammaticalized pattern.

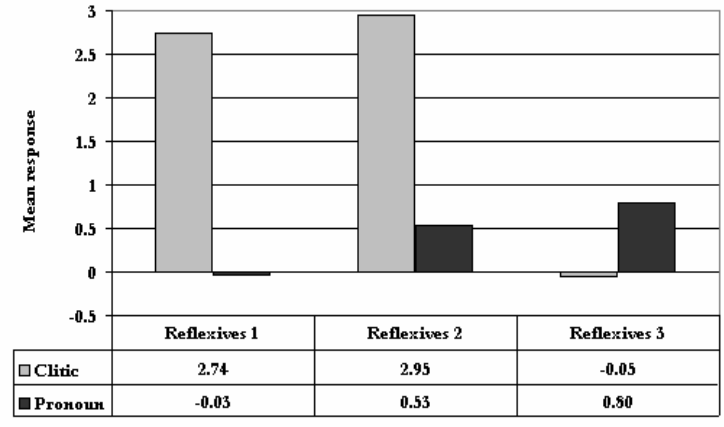
On the other hand, there are no statistically significant differences between verb groups in the marker distribution of reciprocals (Figure 5), and its statistical correlation with the object distribution is fairly low, r=0.24 (R^2=0.06). However, this finding should be interpreted in light of the fact that some of the verbs examined in this study have very low overall frequencies in the corpus and/or very few tokens in reciprocal use. It is therefore most likely the case that this was the factor that led to the lack of statistical significance. Judging from the trends, the continuum is more evident in reciprocal than in reflexive marking, and in this case it is clear that the use of *se* is not limited to verbs often used as reciprocal, as the clitic accounts for almost half the reciprocal uses even with verbs most rarely used in this form.



	Reciprocals 1	Reciprocals 2	Reciprocals 3
Clitic	99.74%	66.67%	49.81%
Pronoun	0.26%	33.33%	50.19%

Figure 5: Corpus distribution of reciprocal clitics vs. reciprocal pronouns

As for the native speakers' judgments (Figures 6 and 7), the participants made a statistically significant distinction between different verb classes in all cases (Friedman test: reflexive clitics $\chi^2(2)=35.314$, $p<0.001$, reflexive pronouns $\chi^2(2)=10.194$, $p<0.01$, reciprocal clitics $\chi^2(2)=31.194$, $p<0.001$, reciprocal pronouns $\chi^2(2)=21.536$, $p<0.001$), and their judgments correlated with the object distribution in the corpus (for reflexives r=0.94 (R^2=0.88), for reciprocals r=0.81 (R^2=0.66)). Here too we find a clear indication that the use of clitics is not grammaticalized for reciprocals. In the case of reflexives, the result for group 3, close to zero, was caused by mixed judgments, as approximately one half of the informants accepted *se* with these verbs while the other half rejected it. However, the fact that some subjects did accept *se* can be taken as indicative of its not being limited in the sense of the English or Dutch light forms, which are readily rejected with this type of verbs by all native speakers.



	Reflexives 1	Reflexives 2	Reflexives 3
Clitic	2.74	2.95	-0.05
Pronoun	-0.03	0.53	0.80

Figure 6: Judgments on reflexive clitics vs. reflexive pronouns

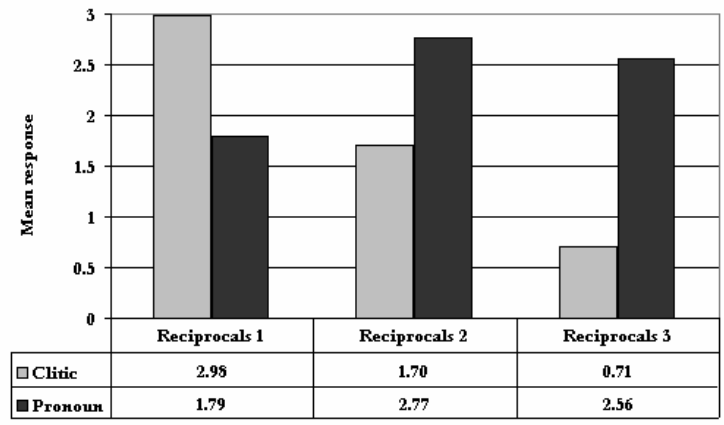


Figure 7: Judgments on reciprocal clitics vs. reciprocal pronouns

4. Conclusion and directions for future research

As shown in the previous section, the results of the present study support the productive analysis of *se*. This is so because in the corpus data there are some occurrences of *se* with verbs rarely used as reflexive (group 3) and a fairly high percentage of *se* used with verbs less frequently or only rarely occurring as reciprocal (groups 2 and 3), and also because the native speakers' judgments do not provide evidence of clear rejection of *se* with the same verb groups. In addition, the mixed judgments on reflexives 3 point to a high degree of individual and possibly dialectal variation. Similar findings are absent in the English-type restricted languages.

However, it should be highlighted that in interpreting these results, our primary goal is to argue against the fully restricted view of *se*'s distribution. Further research is needed in order to define the exact type of productivity displayed by the Serbian reflexive and reciprocal clitic, as both sources of empirical data indicate that it does not reach the same degree as, for instance, the productivity of its Italian counterpart *si* (see Miličević 2007 for Italian data).

Lastly, while the typological significance of the problem discussed in this paper is indisputable, it should be emphasized that it also has important theoretical implications, as some theories of reflexive and reciprocal binding rely heavily on the properties of predicates and markers (Reinhart & Reuland 1993, Reinhart & Siloni 2005). Future research should therefore also focus more on relating typological and theoretical accounts of reflexive and reciprocal marking.

References

Haiman, J. 1983. Iconic and economic motivation. *Language 59*, 781–819.

Haspelmath, M. 2005. A frequentist explanation of some universals of reflexive marking. Draft of a paper presented at the Workshop on Reciprocals and Reflexives, Freie Universität Berlin, 1-2 October 2004. (URL: http://email.eva.mpg.de/~haspelmt/papers.html, downloaded on 20 January 2006.)

Hellan, L. 1988. *Anaphora in Norwegian and the Theory of Grammar*, Dordrecht: Foris Publications.

Kemmer, S. 1993. *The Middle Voice*, Amsterdam and Philadelphia: John Benjamins.

König, E. & L. Vezzosi. 2004. The role of predicate meaning in the development of reflexivity. In W. Björn & W. Bisang, eds, *What Makes Grammaticalization?*, Berlin: Mouton de Gruyter.

Korpus savremenog srpskog jezika. URL: http://www.korpus.matf.bg.ac.yu/prezentacija/korpus.html. Natural Language Processing Group, Faculty of Mathematics, University of Belgrade. Last accessed June 2007.

Marelj, M. 2004. *Middles and Argument Structure across Languages*, LOT Dissertation Series 88, Utrecht: LOT Publications. (URL: http://www.lotpublications.nl/index3.html)

Miličević, M. 2007. *The Acquisition of Reflexives and Reciprocals in L2 Italian, Serbian and English*, PhD dissertation, Cambridge: Research Centre for English and Applied Linguistics.

Perović, A. 2003. *Knowledge of Binding in Down Syndrome: Evidence from English and Serbo-Croatian*, PhD dissertation, London: University College London.

Reinhart, T. & E. Reuland. 1993. Reflexivity. *Linguistic Inquiry 24* (4), 657-720.

Reinhart, T. & T. Siloni. 2005. The Lexicon-Syntax Parameter: Reflexivization and other arity operations. *Linguistic Inquiry 36* (3), 389-436.

Siloni, T. 2001. Reciprocal verbs. In Y. Falk, ed., *Proceedings of the Israel Association of Theoretical Linguistics 17*. (URL: http://atar.mscc.huji.ac.il/_english/IATL/17/. Downloaded on 24 November 2006.)

Processing WordNet with Modal Logic

Borislav Rizov

Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Science
52 Shipchenski prohod, building 17, Sofia 1113, Bulgaria
boby@dcl.bas.bg

Abstract

This paper presents an approach to WordNet processing based on modal logic. For the purposes of development and exploration of WordNet a modal language is described, which is already implemented in a tool named Hydra. WordNet database is represented as a relational database in the system and for its processing is provided an application programming interface working with the specified modal language.

1. Introduction

With the development of the Bulgarian wordnet - BulNet[1] the need for a powerful application for wordnet development enabling specific tasks related to the consistency and completeness of the wordnet database, as well as data extraction arose. Besides the above the use of wordnet database in several different applications showed the need for a powerful application programing interface (API) for wordnet processing. This API ought to use abstract language independent of the data representation. To satisfy this needs a modal logic language was designed and the application Hydra, that uses it, was developed. Besides the above features Hydra supports other user requirements including visualization of data and relations between certain portions of data, as well as editing, undo/redo functions, concurrent user access etc. Hydra supports a synchronized view of and access to wordnets for different languages through synsets encoding equivalent word senses. This provides exploitation of wordnet as a multilingual multipurpose lexicon, as well as an explanatory dictionary, a dictionary of synonyms, antonyms, hyperonyms, thematic dictionary, etc. In this respect Hydra may be regarded as a browsable and searchable lexicon user interface.
The library for wordnet processing has the following features:

- search engine working with WN modal language. It supports regular expressions;
- objects that represent entities in the wordnet structure such as synsets and literals;
- objects representing the relations between the above entities. All these objects have the appropriate interface for the modification of a wordnet structure.

Tasks in wordnet processing other than editing are reduced to retrieving a set of objects that satisfy a certain property. Provided that a property is definable by a formula in the modal language described below, the system determines all the objects in the WordNet structure validating the formula, and hence the property. Hydra's language is based on the language presented in (Koeva et al., 2004)[2].

2. WordNet

WordNet is a lexical database that organises lexical information in terms of word meanings (Miller et al., 1990) represented by sets of strong synonyms, called synsets. Thus, a synset encodes a lexicalized concept and each lexeme denoting the concept (called a literal) is a member of the synset. Each synset is supplied with a gloss (an explanatory definition), examples of the usage with real language sentences, and possibly with various notes on grammatical, semantic, pragmatic characteristics of the synset or of particular literals. The synsets and literals are interlinked through various conceptual-semantic and lexical relations having different properties. Some of the most prominent are hyperonymy (relation between a superordinate and subordinate concept), meronymy (part-whole relation), antonymy (oppositeness relation), etc. Every synset is supplied with an identification key that is identical for equivalent synsets in all languages (Miller et al., 1990).

[1] http://dcl.bas.bg/wordnet_en.html
[2] With some modifications.

3. Modal Language for WordNet

The principal purpose of the modal WordNet language introduced here is to provide a clean and uniform formalism for expressing complex queries with sufficient expressive power for the most important tasks and validation procedures required in the development and exploration of a wordnet (such as searching, validation, synchronization, etc.) to be handled. The semantic and lexical relations in wordnet are all binary. The membership of the literals in a synset can be considered as a relation between this literals and the synset that contains them. In a similar way the usage examples and the notes (we call all of them Notes) can be related to the corresponding synsets and literals. Having these in mind we can consider WordNet as a relational structure with three sorts of objects, namely Synsets, Literals and Notes, and with a set of relations (now they are over twenty). In the implementation the WordNet structure is represented as a relational database. The information retrieval and management is handled by means of SQL. Although it is more powerful query language than the one described in this paper (each formula in the language is expressed in terms of SQL) it has certain drawbacks as a front-end language. It's use is more complicated, the queries corresponding to the formulae being long and hard to write, even for short ones. The presented language is much easier to learn by common users than standard SQL. The following solution also dispenses with information retrieval procedures involving programming skills required by other types of representation.

Another advantage of the use of abstract language is that it allows unproblematic modification in the wordnet structure such as addition of new Relations or changes in the database architecture (with the corresponding translation of the formulae). It is also possible for another back-end to be used or the current one to be modified. All of the above may be performed without changing the language and the already expressed queries (formulae) and statements (e.g. ones defining wordnet consistency). It is very important that such modifications will be unnoticable for the users - including the users of hydra and the API.

First, we present the syntax of the language, and then proceed with a definition of a WordNet structure and the semantics of the language.

3.1. Syntax

Atomic formulae:

- Var – enumerable set of propositional variables.
- $\Sigma^{Literal}, \Sigma^{Note}, \Sigma^{Synset}$ – finite sets of nominals (constants).
- Sets of Boolean constants
  - $\{q^{Literal}, q^{Note}, q^{Synset}\}$
  - $B^{pos} = \left\{q_n^{pos}, q_v^{pos}, q_{adj}^{pos}, q_{adv}^{pos}, ...\right\}$
  - $B^{ili} = \left\{q^{ili}_{ENG20-06307086-n}, q^{ili}_{BUL-370295703}, ...\right\}$
  - $B^{def} = \left\{q_1^{def}, q_2^{def}, ...\right\}$
  - $B^{lang} = \left\{q_{bg}^{lang}, q_{en}^{lang}, ...\right\}$
  - $B^{bcs} = \left\{q_1^{bcs}, q_2^{bcs}, ...\right\}$
  - $B^{word} = \left\{q_{mouse}^{word}, q_{cat}^{word}, q_{person}^{word}, ...\right\}$
  - $B^{lemma} = \left\{q_{mouse}^{lemma}, q_{cat}^{lemma}, ...\right\}$
  - $B^{sense} = \{q_1^{sense}, q_2^{sense}, ...\}$
  - $B^{note} = \{q_1^{note}, q_2^{note}, ...\}$

Relational symbols:

Let $RS = \{ \equiv, R^{lnote}, R^{snote}, R^{usage}, R^{literal}\ R^{hypernym}, R^{holo\_part}, R^{holo\_member}, R^{holo\_portion}, R^{near\_antonym}, R^{be\_in\_state}, R^{category\_domain}, R^{similar\_to}, R^{also\_see}, R^{region\_domain}, R^{usage\_domain}, R^{derived}, R^{participle}, R^{eng\_drivative}, R^{subevent}, R^{verb\_group}, R^{causes}, R^{ili}, R^{bg\_derivative} \}$. $Rel = RS \bigcup \left\{R^{-1} \mid R \in RS\right\}$ is the set of relational symbols.

Formulae:

- The atomic formulae are formulae.
- If φ and ψ are formulae, then:

  $(\neg\varphi), (\varphi \wedge \psi)$ are formulae.

  We will also use the usual syntactic shortenings $(\varphi \vee \psi), (\varphi \rightarrow \psi), (\varphi \leftrightarrow \psi)$.

- If φ is formula, $R \in Rel$ is relational symbol, p and q are natural numbers then:
  $(\langle R\rangle\varphi), (\langle R\rangle_p\varphi), (\langle R\rangle_{\frac{p}{q}}\varphi)$
  are formulae.
  Then we define the syntactic shortenings $([R]\varphi), ([R]_p\varphi), ([R]_{\frac{p}{q}}\varphi)$ for
  $(\neg(\langle R\rangle(\neg\varphi)))$, $(\neg(\langle R\rangle_n(\neg\varphi)))$, $(\neg(\langle R\rangle_{\frac{p}{q}}(\neg\varphi)))$ respectively.

3.2. Semantics

The semantics of the defined modal language is based on the classical Kripke semantics. **Kripke structure** is a tuple $\langle W, I\rangle$, where:

- I is the interpretation of the nominals, boolean constants and relational symbols, where:
  - $I(c) \in W$ for any nominal c.
  - $I(q) \subseteq W$ for any boolean constant q.
  - $I(R) \subseteq W \times W$ for any relational symbol R.

A valuation over the structure is $V : Var \longrightarrow 2^W$.
A Kripke structure is called **WordNet structure** if:

- $\left\{I\left(q^{Literal}\right), I\left(q^{Synset}\right), I\left(q^{Note}\right)\right\}$ is a partition of W
- $\{I(q) \mid q \in B\}$ is a partition of $I\left(q^{Synset}\right)$, when $B \in \left\{B^{ili}, B^{bcs}, B^{lang}, B^{def}\right\}$
- $\{I(q) \mid q \in B\}$ is a partition of $I\left(q^{Literal}\right)$, when $B \in \left\{B^{word}, B^{lemma}, B^{sense}\right\}$
- $\bigcup\left\{I(q) \mid q \in B^{Note}\right\} = I(q^{Note})$
- $I\left(R^{-1}\right) = I(R)^{-1}$
- $I\left(R^{LNote}\right) \subseteq I\left(q^{Literal}\right) \times I\left(q^{Note}\right)$
- $I\left(R^{SNote}\right) \subseteq I\left(q^{Synset}\right) \times I\left(q^{Note}\right)$
- $I\left(R^{Usage}\right) \subseteq I\left(q^{Synset}\right) \times I\left(q^{Note}\right)$
- $I\left(R^{Literal}\right) \subseteq I\left(q^{Synset}\right) \times I\left(q^{Literal}\right)$
- $I\left(R^{Literal}\right)^{-1}$, $I\left(R^{LNote}\right)^{-1}$, $I\left(R^{SNote}\right)^{-1}$ and $I\left(R^{Usage}\right)^{-1}$ are functions
- $I\left(R^{ili}\right) = \left\{I(q) \times I(q) \mid q \in B^{ili}\right\} \setminus \left\{I(q) \times I(q) \mid q \in B^{lang}\right\}$
- $I(\equiv) = I\left(R^{Literal^{-1}}\right) \circ I\left(R^{Literal}\right)$

Definition: We define the truth of a formula of WN language at point $x \in W$ over WordNet structure and a valuation in it by induction on the formula construction:

- $x \Vdash c$ iff $x = I(c)$ for any nominal c
- $x \Vdash c$ iff $x \in I(c)$ for any boolean constant c
- $x \Vdash p$ iff $x \in V(p)$ for any $p \in Var$
- $x \Vdash (\neg\varphi)$ iff $x \nVdash \varphi$
- $x \Vdash (\varphi \wedge \psi)$ iff $x \Vdash \varphi$ and $x \Vdash \psi$
- $x \Vdash (\langle R\rangle\varphi)$ iff $\exists y \in W\,(xI(R)y$ and $y \Vdash \varphi)$, where $R \in Rel$
- $x \Vdash (\langle R\rangle_n\varphi)$ iff $|\{y \in W \mid xI(R)y \wedge y \Vdash \varphi\}| > n$, where $R \in Rel$[3]
  Note that for every $x \in W (x \Vdash (\langle R\rangle\varphi)$ iff $x \Vdash (\langle R\rangle_0\varphi)$
- $x \Vdash \left(\langle R\rangle_{\frac{p}{q}}\varphi\right)$ iff $|\{y \in W \mid xI(R)y \wedge y \Vdash \varphi\}|\, q > p\, |\{y \mid xI(R)\, y\}|$, where $R \in Rel$

[3] with $|A|$ we denote the cardinality of the set A.

It is easy to see that:

- $x \Vdash ([R]\,\varphi)$ iff $\forall y \in W\,(xI(R)y \Rightarrow y \Vdash \varphi)$
- $x \Vdash ([R]_n\,\varphi)$ iff $|\{y \in W \mid xI(R)y \wedge y \Vdash \neg\varphi\}| \leqslant n$
- $x \Vdash \left([R]_{\frac{p}{q}}\,\varphi\right)$ iff $|\{y \in W \mid xI(R)y \wedge y \Vdash \neg\varphi\}|\,q \leqslant p\,|\{y \mid xI\,(R)\,y\}|$

3.3. The syntax in Hydra

In our implementation of the described language we use the following syntax corresponding to the formulae construction.
The nominals we represent by unique identifiers. They are the decimal representation of integers in three disjoint sets. Each of this sets contains the identifiers for one of the types of objects in our model (WordNet) - Literal, Synset and Note.
$q^{Literal}, q^{Note}, q^{Synset}$ here are $p, $r, $q.
For the other boolean constants, let us say informally q^{type}_{value}, we write $type('value')$. For example if we want to express $q^{ili}_{ENG20-06307086-n}$ and q^{word}_{person} we write ili ('ENG20-06307086-n') and word ('person'). To use regular expression in the value the # is added before the quotes - word(#'c[au]t').
If we define a priority of the operators to be descending by their definition, then in formulae construction we can omit most of the parentheses. Let us see the formulae and their syntax in Hydra.
Assume that $\tilde{\varphi}$, $\tilde{\psi}$ are the representations of the formulae φ and ψ, respectively. Then we substitute the defined syntax using the symbols of the keyboard:

- $\neg\varphi \quad \longmapsto \quad !\tilde{\varphi}$
- $\varphi \wedge \psi \quad \longmapsto \quad \tilde{\varphi}/\backslash\tilde{\psi}$
- $\varphi \vee \psi \quad \longmapsto \quad \tilde{\varphi}\backslash/\tilde{\psi}$
- $\varphi \rightarrow \psi \quad \longmapsto \quad \tilde{\varphi} - > \tilde{\psi}$
- $\varphi \leftrightarrow \psi \quad \longmapsto \quad \tilde{\varphi} < - > \tilde{\phi}$
- $\langle R\rangle\varphi \quad \longmapsto \quad < R > \tilde{\varphi}$
- $[R]\,\varphi \quad \longmapsto \quad [R]\tilde{\varphi}$
- $\langle R\rangle_n\varphi \quad \longmapsto \quad < R, n > \tilde{\varphi}$
- $[R]_n\,\varphi \quad \longmapsto \quad [R]_n\tilde{\varphi}$
- $\langle R\rangle_{\frac{p}{q}}\varphi \quad \longmapsto \quad < R, p : q > \tilde{\varphi}$
- $[R]_{\frac{p}{q}}\,\varphi \quad \longmapsto \quad [R, p : q]\tilde{\varphi}$

  For the opposite relations we use the sign ˜ in front of the relational symbol. So:

- $\langle R^{-1}\rangle\varphi \quad \longmapsto \quad <\tilde{}R > \tilde{\varphi}$

The use of the variables is not implemented yet.

3.4. WN language in practice

The WN language defined above is used to perform simple and advanced queries in WN that are needed in wordnet exploration and validation, as well as in the work of annotators using wordnet, e.g. in word-sense disambiguation (WSD).

3.5. Example queries

- **Return all synsets which contain the word *cat***

  $$\langle R^{Literal}\rangle q^{word}_{cat}$$

  q^{word}_{cat} retrieves the Literal objects representing the word 'cat', then the modality $\langle R^{Literal}\rangle$ gives the Synsets which contain them.

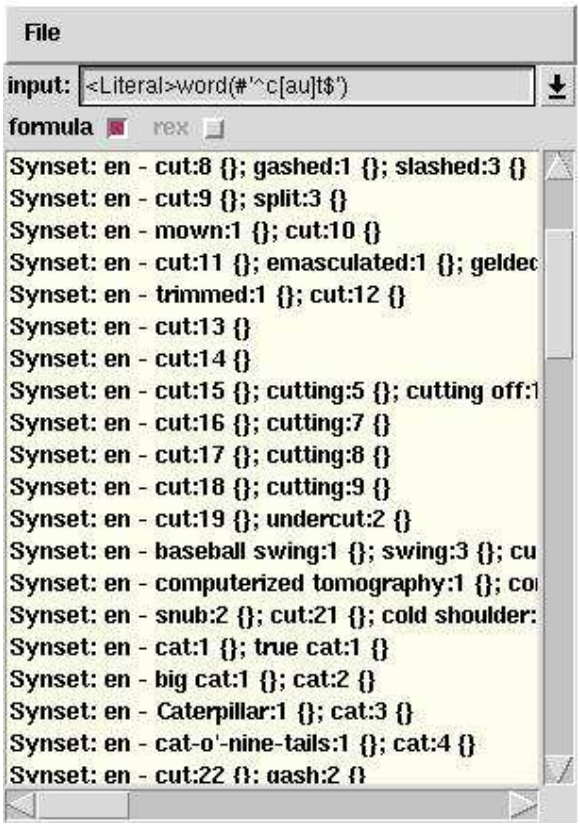


Figure 1: Search by a formula using regular expression

- **Return all hyperonyms of the synset identified by the nominal** q

$$\left\langle R^{hypernym^{-1}} \right\rangle q$$

q is a name for an object in WordNet Structure. The modality returns all the synsets which are hyperonyms of the object. In Hydra you can use $hyponym$ instead of $hypernym$ ($hypernym^{-1}$), which is the oposite relation.

- **Return all synsets which contain the word *cat* and their correspondences in Bulgarian.**

$$\left\langle R^{Literal} \right\rangle q_{cat}^{word} \vee \left(q_{bg}^{lang} \wedge \left\langle R^{ili} \right\rangle \left\langle R^{Literal} \right\rangle q_{cat}^{word} \right)$$

$\langle R^{ili} \rangle \langle R^{Literal} \rangle q_{cat}^{word}$ retrieves the synsets in other languages which are connected to the synsets containing the word 'cat'(see the previous example).

q_{bg}^{lang} gives the synsets in the Bulgarian wordnet.

$q_{bg}^{lang} \wedge \left\langle R^{ili} \right\rangle \left\langle R^{Literal} \right\rangle q_{cat}^{word})$ returns the intersection of the upper two sets.

Finally the disjunction is interpreted as union of the sets returned by its members.

- Return the set of the literals in a synset denoted by the nominal q

$$\left\langle R^{literal^{-1}} \right\rangle q$$

It is similar to our second example. This query is used in Hydra's consistency checks. If this set is empty, Hydra forbids saving of the edited synset (q).

4. Representation of WordNet structure

Although there are other alternatives, the relational nature of WordNet and the necessity of fast concurrent access to large amount of data determine the choice of RDBMS. SQL is the background of the query system that uses the defined modal language. A modal formula φ is automatically translated into an equivalent SQL

query ϕ. In other words, for any $x \in W$, $x \Vdash \varphi$ if and only if $\tilde{x}$ belongs to the results of the query ϕ evaluated in the database, where $\tilde{x}$ is the representation of x in the database.
The database is organized in such a way as to be self-explanatory. For example information about the binary relations in the wordnet representation is stored in the database and is used in every aspect of wordnet processing - visualization, editing, validation etc.
In this model we recognize three sorts of objects.

— Synset (representing the synonym sets in a WordNet structure)

— Literal (representing the graphical words)

— Note (representing some text data in a WordNet structure as usage examples and explanatory notes)

We call these objects linguistic units (LU). The literals are in the relation 'literal' with the synsets containing them. Notes are found in a number of relations with Synsets and Literals, such as Usage, LNote, SNote. Wordnet has a synset centric organization. Our model is made with respect to this concept. Every LU is associated with a single synset.A Synset is associated with itself. A Literal is associated with the synset connected with by the literal relation. A Note is related with a single LU. So the note is associated with the synset that this LU is associated with.

4.0.1. The tables of the relational database

SYNSET – the table representing the synsets.

LITERAL – the table representing the literals

NOTE – the table representing the notes

REL – the table storing the binary relations between LUs

There are several tables which make the database self explanatory.

4.1. Data retrieval

Hydra enables search in the WordNet database by means of formulae in the WN language. The search engine returns all linguistic units at which the formula is true in the WordNet structure.
We define the translation of the formulae in SQL queries by induction.
The result of each query produced by formula is a table, containing the identifiers (id) of the LU at which the formula is true. The identifiers are natural numbers, belonging to three disjoint intervals, so that the type of the LU is recognizable by its identifier. This translation remains hidden to the end user of Hydra.

- c is nominal

  select c as id;

- For the constants $q^{Literal}, q^{Synset}, q^{Note}$ the corresponding queries are:

  select id from Literal;
  select id from Synset;
  select id from Notes;

- Let $B \in \left\{B^{ili}, B^{pos}, B^{def}, B^{bcs}, B^{lang}\right\}$, $q^{type}_{value} \in B$. For q^{type}_{value} we have:

  select id from Synset where type = value;

- Let $B \in \left\{B^{word}, B^{lemma}, B^{sense}\right\}$, $q^{type}_{value} \in B$. For q^{type}_{value} we have:

  select id from Literal where type = value;

- Let $q^{type}_{value} \in B^{note}$. For q^{type}_{value} we have:

  select id from Notes where type = value;

- Let VT be the table presenting the valuation of the variable x, then:

  select id from VT;

- We retrieve the universe of the model using the following formula:

$$q^{Literal} \vee q^{Synset} \vee q^{Notes}$$

Let P, Q and R be the translations of the formulae φ, ψ and the relation R while F is the translation of the above formula (defining the universe). Then we have:

- $\neg\varphi$

```
select id from F
   where not exists
(select 1 from P where P.id = F.id)
```

- $\varphi \wedge \psi$

```
select P.id from Synset
   inner join Q
on P.id = Q.id;
```

- $\varphi \vee \psi$

```
P union Q;
```

- $\langle R \rangle \varphi$

```
select distinct Rel.id1 as id from Rel
inner join P
on Rel.id2 = P.id and rel = R;
```

- $\langle R \rangle_n \varphi$

```
select id2 as id from
(select id2, count(id2) as c from R
inner join P
on Rel.id1 = F.id and rel=R
group by id2) as D where c ¿ n
```

- $\langle R \rangle_{\frac{p}{q}} \varphi$

```
select T1.id2 as id from
(select id2, count(id2) as c from Rel
  inner join P
on Rel.id1 = P.id and rel=R
  group by id2) as T1
inner join
(select id2, count(id2) as d from Rel
group by id2) as T2
 on T1.id2 = T2.id2 where c*q ¿ d*p
```

$\varphi \rightarrow \psi, \varphi \leftrightarrow \psi, [R]\varphi, [R]_n \varphi, [R]_{\frac{p}{q}} \varphi$ are expressed by means of the already defined formulae.

5. Implementation

The tool is implemented in Python [4]. GUI is in Tkinter and Tix. As a platform-independent system, Hydra has been successfully tested under Linux and Windows. The RDBMS used in the implementation is MySQL. Almost all of the design patterns proposed by GoF (Gamma et al., 1995) are used in the implementation. The resulting system is a robust and extendable one. UTF-8 database encoding makes Hydra language-independent. The system has been tested successfully on Bulgarian, English and French.

[4] http://www.python.org

6. Interoperability with other systems

To contribute the exchange of the wordnet databases with other system Hydra provides easy to use tools for data import and export. Currently, these tools support a xml format and also is fully compatible with the VisDic and DebVisDic's[5] formats. VisDic and lately DebVisDic are very good and famous tools for wordnet development and visualization. There's wide range of functionalities is almost fully supported by our system, but Hydra has many advantages such as the modal language, robust consistency checks (expressed in the language) and the more advanced view concepts.

7. Conclusion

We presented a modal logic vision to wordnets processing. The work on the defined language and Hydra's development are still in progress. Recently, the mojarity operators were added - $\langle R \rangle, \langle R \rangle_p, \langle R \rangle_{\frac{p}{q}}$. In the future, the propositional variables will be exploited. The options for variables include they to be evaluated over the result of a query or to be initialized with a value represented by a table (this table may be retrieved outside the system). A Web Interface GUI implementation is also considered.
Hydra is currently used at the Department of computational linguistics, IBL-BAS in the development of the Bulgarian wordnet. Hydra's data retrieval engine is used in several other applications like corpora search engines and dictionaries.

8. References

E. Gamma, R. Helm, R. Johnson, and J. Vlissides. 1995. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley.

S. Koeva, S. Mihov, and T. Tinchev. 2004. Bulgarian wordnet - structure and validation. *Romanian J. Of Inf. Sci. And Technology*, 7, No. 1-2:61–78.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography*, 3, No 4:235–244.

[5] http://nlp.fi.muni.cz/projekty/visdic/

EVALUATING SENTENCE ALIGNMENT ON CROATIAN-ENGLISH PARALLEL CORPORA

Sanja Seljan*, Željko Agić*, Marko Tadić**

*Department of Information Sciences
**Department of Linguistics
Faculty of Humanities and Social Sciences
University of Zagreb
Ivana Lučića 3, HR-10000 Zagreb
{sanja.seljan,zeljko.agic,marko.tadic}@ffzg.hr

ABSTRACT

This paper describes an experiment in applying sentence alignment methods to Croatian-English parallel corpora and systematically evaluate their performance within the recall, precision and F-measure framework. It is our primary goal to provide an insight and a reference point on sentence alignment accuracy for Croatian-English language pair and also to extend the scope of (Tadić, 2000) – to our knowledge, the first experiment dealing with automatic sentence alignment of Croatian-English parallel corpora – by utilizing newly implemented tools, creating corpora subsets defined by genre and finally by expanding and formalizing its preliminary observations on alignment accuracy. Therefore, in this paper we start off by briefly describing and argumenting sentence alignment paradigms of choice and presenting available language resources, subset of Croatian-English parallel corpus described in (Tadić, 2000) being our primary asset. These descriptions are followed by a formal definition of our testing framework. Results are then discussed in detail and conclusions are stated along with a brief insight on possible future work.

1. Introduction

Parallel corpora and especially sentence-aligned bilingual corpora can be very effectively used as a resource for numerous research projects or in creation of new resources, such as sentence or word alignment projects for computer-assisted translation, machine translation, multilingual information retrieval, language learning, multilingual terminology bases and semantic networks. Translation memories used in computer-assisted translation, machine translation or for terminology extraction, created by the sentence alignment process from the parallel corpus, directly reflect two main problems: corpus size and divergences in the layout of parallel texts, which can differ regarding the expert intervention in the set up of the alignment program. For this reason, corpus aligning has caused a great interest and numerous aligners relying on different methods.

For lesser spread languages, but still having rich cultural and historic pool of texts, creation of electronic tools and resources is of considerable interest. In the situation when Croatia approaches the European Union, use of common resources has become important not only for translators, but also for researchers and everyday users. Use of common resources and translation tools has become an absolute demand in any kind of cooperation activities and international communication.

One of the most well known examples of shared resources is Europarl: the European Parliament corpus, published on the web (URL http://www.statmt.org/europarl), which has numerous applications in NLP, significant because of its size, number of languages and various linguistic phenomena included, but also because of its provenance since the sources included are mainly from the United Nations, European Union or member governments (Koehn, 2005).

The JRC-Acquis Communautaire corpus and its documentation, consisting of 20+ languages are freely available from the web (at URL http://wt.jrc.it/lt/Acquis). This most voluminous multilingual parallel corpus, consisting of 22 languages, is especially suitable for cross-information language retrieval, machine translation and machine learning. Part of it is sentence-aligned using the Vanilla aligner, applying Gale-Church algorithm (Gale & Church, 1993) and HunAlign (Varga et al., 2005), which allows benchmarking of alignment tools and algorithms.

Therefore, the first experiment dealing with evaluation metrics of sentence-aligned Croatian-English corpora has been made and is described in this paper. As sentence-aligned corpora are more efficient and generally more valuable as a resource than non-aligned parallel corpora, while planning this investigation, we considered presenting results of automatic sentence-level alignment of Croatian-English parallel corpora using several different sentence alignment tools, namely SDL Trados WinAlign, Vanilla aligner (Danielsson & Ridings, 1997), Hunalign (Varga et al., 2005) and CORAL aligner in order to derive a decision on which of these tools should be utilized in aligning corpora of language pairs Croatian-Language$_x$ in the future. We also considered Moore's aligner (Moore, 2002) as one of possible solutions but since it was built with the main purpose of compiling the language models for SMT, which feature only exact (one-to-one) alignments and avoiding other types of alignments (non-one-to-one), we shifted our interest towards other algorithms. Most of them were discussed and compared in (Och & Ney, 2003) but what we actually wanted to put in the focus of our interest were the implementations. However, being that various sentence alignment tools are in fact based on underlying paradigms provided in forms of published or proprietary algorithms, we chose the scientific method and approach and decided to evaluate sentence alignment of Croatian-English language pair on paradigms exclusively, focusing on Gale-Church algorithm implemented in CORAL aligner.

In the following section of the paper, we present in more details these tools and resources utilized in the experiment and provide argumentation on choices made. Section 3 provides insight on test environment setup, while results are discussed and future research paths indicated in section 4.

2. Resources and tools

In order to provide a measure sentence alignment accuracy for Croatian-English language pair, there are two basic assets required – the language sample, i.e. manually (or otherwise) previously aligned parallel corpora and a sentence aligner, i.e. a program implementing an algorithm that automatically aligns non-aligned corpora. Manual and automatically aligned corpora are then compared and various methods of inspecting differences provide measures of alignment accuracy. In this section, we provide insight on these two basic assets while the following section focuses on evaluation framework.

Aligned subsets Croatian-English parallel corpora used in this research consist of:

1. Croatian-English parallel corpus of legislative documents (JOC – Journal of the European Community) and
2. Subset of Croatian-English parallel corpus from the newspaper Croatia Weekly, presented in (Tadić, 2000).

The legislative Croatian-English subset consists out of 6 Croatian legislative documents of about 20 pages consisting of bylaws, regulations, and decisions of the Croatian government and their translation into English documents. The documents consist of 6791 words in Croatian and corresponding 8510 words in English, i.e. 48982 characters in Croatian and corresponding 55567 characters in English due to highly inflective nature of the Croatian language. Documents were provided in plain text format, manually aligned using CORAL aligner and used as a reference point for automatic alignment evaluation. Document stats are given in Table 1.

	Pages		Words		Characters	
document	**CRO**	**ENG**	**CRO**	**ENG**	**CRO**	**ENG**
Bylaws	6	5.5	1311	1982	9238	12301
AMI	4	4.5	1577	1805	11143	11460
e-Sig	10.5	9	3903	4723	28601	31806
Total	20.5	19	6791	8510	48982	55567

Table 1. Legislative documents subcorpus statistics

Legislative documents are included in the experiment because various statistical machine translation platforms – relying on sentence- and word-alignment preprocessing – are largely utilized exactly in tasks of translating legal documents i.e. in multilingual environment of the European Union. It was therefore important to provide results of aligning Croatian and English in this domain in order to indicate the quality of sentence alignment platform on which future research is to build machine(-aided) translation systems.

However, regardless of overall importance of achieving high alignment accuracy on legislative documents, hand-annotated subset of Croatian-English parallel corpus was the main resource for this specific investigation, being linguistically well-formed and properly annotated by XML structure. The parallel corpus itself was previously described in detail (Tadić, 2000) as being sourced from the Croatia Weekly newspaper corpus and consisting of approximately 1.6 million tokens for Croatian and 1.9 million tokens for English.

Stats given in Table 2 indicate that subset size is approximately 32% of the entire parallel corpus when comparing token counts. As expected, both Croatian and English parts of the parallel subcorpus consist of same numbers of articles, sections, main titles, subtitles and paragraphs. Minor differences in section and main title counts are caused by human annotation errors as numbers match exactly when checking for specific errors; i.e. section count for Croatian lacks one section that is easily found if `<DIV0>` is replaced by `<DIV` in search query, clearly indicating a mistyped tag. Findings are similar for main title counts and a claim can be made that counts of document structure entities are the same.

Croatian	English	
1435	1435	Articles `</BODY>`
1599	1600	Sections `</DIV0>`
1597	1600	Main titles `<HEAD type='NA'>`
493	493	Subtitles `<HEAD type='PN'>`
6327	6327	Paragraphs `</P>`
22985	25412	Sentences `</S>`
514428	618462	Words
3402532	3300409	Characters
3918959	3920825	Characters with spaces
4913825	5017745	Bytes

Table 2. Cro-Eng subcorpus statistics

When comparing occurrence counts for sentences, words and characters, Croatian and English subcorpora start to differ, English getting higher numbers in all figures. This is an expected distribution and desired behaviour as structure of newspaper articles – captured by rows 1 to 5 of Figure 1 – remains the same in Croatian and English, while translation functions for sentences are never bijective. This conclusion also applies on word selection and subsequently character counts. As with the entire corpus, manually annotated gold standard subcorpus implements higher token count on English side with a factor of 1.2 compared to 1.19 overall. Subcorpus sample given in Figure 1 contains XML-wrapped article in English and Croatian in which structural and textual similarities and differences are illustrated.

```
<BODY><DIV0 type="MAIN"><HEAD type="NA">
<S id="CW047199812241407hr.S1">Bestseler u Aucklandu</S>
</HEAD><P>
<S id="CW047199812241407hr.S2">Roman "Croatia Mine" novozelandske Hrvatice Floride Vele za kratko je vrijeme postao bestselerom u Aucklandu, a autorica se ove godine našla među 20 odabranih pisaca tamošnje utjecajne književne priredbe "World Book Day".</S>
<S id="CW047199812241407hr.S3">Roman "Croatia Mine" (u izdanju Quin Pressa iz Christchurcha) prvi je roman Floride Vele i nadahnut je stanovitim autobiografskim elementima.</S>
<S id="CW047199812241407hr.S4">Prati sudbinu jedne hrvatske obitelji iz Podogore - od iseljeništva iz Jugoslavije pa do snova i ljubavi prema dalekoj domovini Hrvatskoj.</S>
</P>
<BYLINE>(Večernji list)</BYLINE></DIV0></BODY>
```

```
<BODY><DIV0 type="MAIN"><HEAD type="NA">
<S id="CW047199812241407en.S1">BOOK BY CROATIAN AUTHORS BECOMES BESTSELLER IN AUCKLAND</S></HEAD>
<P>
<S id="CW047199812241407en.S2">The novel <I>Croatia Mine</I>, by Florida Vela, a Croat from New Zealand, quickly became the bestseller in Auckland.</S>
<S id="CW047199812241407en.S3">The author was placed on the list of twenty select writers of World Book Day, an influential local literary event. <I>Croatia Mine</I> (published by Quin Press from Christchurch) is the first novel by Florida Vela, and it was inspired by certain autobiographical elements.</S>
<S id="CW047199812241407en.S4">It follows the fate of a Croatian family from Podogora - from their emigration from Yugoslavia to their later dreams and yearning for their distant homeland of Croatia.</S>
</P>
<BYLINE>(<I>Večernji list</I>)</BYLINE></DIV0></BODY>
```

Figure 1. Croatian-English parallel corpus sample

Second choice was that of sentence aligner. As stated in first section, among many alignment tools implementing many standard and specialized algorithms, we chose to provide figures on Croatian-English pair using a less-known aligner named CORAL (CORpus ALigner), being developed in Java at the Faculty of Electrical Engineering and Computing, University of Zagreb. There are two main reasons behind this choice: one is that it implements a standard Gale-Church algorithm that we wanted to evaluate and the other is encompassed by joint programme *Computational Linguistic Models and Language Technologies for Croatian* and its goals described in (Dalbelo Bašić et al., 2007): CORAL aligner is envisioned to be a default platform for sentence alignment (automatic and human assisted) of language pairs Croatian-Language$_X$ and thorough evaluation is required in order to develop newer and better versions of the tool. CORAL was previously evaluated on English-Slovene parallel corpus extracted from MULTEXT-East v3 specification (Erjavec, 2004) but was not yet presented to the community by the time this paper was published.

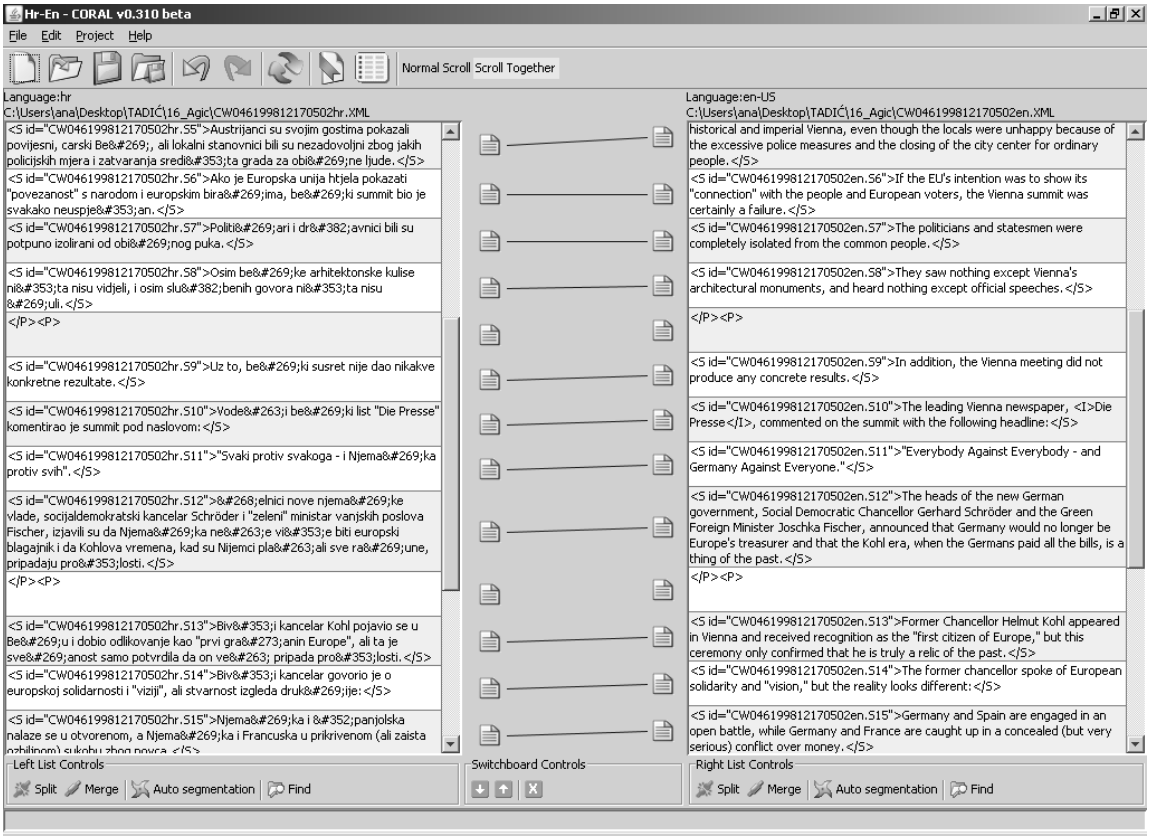


Figure 2. CORAL screenshot

3. Evaluation method

Evaluation method used in this experiment was highly influenced by the one presented in (Langlais et al., 1998), i.e. set of methods used during the ARCADE text alignment evaluation project we chose as a starting point for our own experiment. Figure 3 provides an example on which evaluation techniques are demonstrated.

A_R	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj*	t_2	*This is the second sentence.*
			t_3	*It looks like the first.*

A_S	s_1	*Ovo je prva rečenica.*	t_1	*This is the first sentence.*
			t_2	*This is the second sentence.*
	s_2	*Ovo je druga rečenica i nalik je prvoj.*	t_3	*It looks like the first.*

Figure 3. Example reference and system alignment

Formal description is as follows. We consider source text S and output text T as a sequence of alignments $\{s1, \ldots, sn\}$ and $\{t1, \ldots, tm\}$, respectively. This basic setting is also shown by Figure 3. An alignment A is then defined rather straightforward as a subset of Cartesian product of power-sets $2^S \times 2^T$. We then call the 3-tuple (S, T, A) a bitext and each of its elements is called a bisegment. Given these definitions, we set up two basic evaluation methods and consider two additional tweak or helper methods as proposed by (Langlais et al., 1998).

3.1. Basic F1-measure

Recall and precision are easily defined on a bitext:

$$recall = \frac{|A_S \cap A_R|}{|A_R|}, \quad precision = \frac{|A_S \cap A_R|}{|A_S|}$$

Being that recall basically measures coverage alone and precision deals with counting correct hits, F1-measure (Rijsbergen, 1979) is chosen for merging these two outputs:

$$F_1 measure = 2 \frac{recall \times precision}{recall + precision}$$

On example in Figure 2, the measures calculate as follows:

$A_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2, t_3\})\}$

$A_S = \{(\{s_1\}, \{t_1\}), (\{\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$

$A_R \cap A_S = \{(\{s_1\}, \{t_1\})\}$, $|A_R \cap A_S| = 1$; $|A_R| = 2$; $|A_S| = 3$

recall = 0.50; precision = 0.33; F1-measure = 0.40

Being that this framework is rather harsh – an average observer would intuitively state that the alignment presented in Figure 2 is better than F1-measure indicates – and also rather high-level-oriented, we introduced, once again according to (Langlais et al., 1998) metrics, other and more finely-grained bi-segment subdivisions and cast the F1-measure framework onto them. In the presented example some segments are only partially correct, e.g. $(\{s_2\}, \{t_3\})$, which is the reason to measure recall and precision at the sentence level, and not at the alignment level.

3.2. Sentence track F1-measure

Given alignments $A_R=\{ar_1,\ldots,ar_n\}$ and $A_S=\{a_1,\ldots,a_m\}$, with $a_i=(as_i,at_i)$ and $ar_j=(ars_j,art_j)$, sentence-to-sentence level metrics can also be defined:

Once again, on example set in Figure 2, the sets are defined and measures calculated as follows:

$A'_R = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_2\}), (\{s_2\}, \{t_3\})\}$

$A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}$

$A'_R \cap A'_S = \{(\{s_1\}, \{t_1\}), (\{s_2\}, \{t_3\})\}$, $|A'_R \cap A'_S| = 2$; $|A'_R| = 3$; $|A'_S| = 2$

recall = 2/3=0.66; precision = 2/21; F-measure = 0.80

It is now obvious that sentence granularity and measure is much more forgiving than that of an alignment granularity and that it is also somewhat closer to human evaluation. We thus chose sentence track F1-measure as a solid base for our experiment.

Method of (Langlais et al., 1998.) suggests tuning sentence track F1-measure by added granularity as A'_R and A'_S set cardinality could be expressed in terms of token count and character count. These tweaks are called word granularity and character granularity by (Langlais et al., 1998.) and we chose to waive them for purposes of this experiment. We find them somewhat useful, as they introduce reward to partial correctness of sentence alignment, but also judge them as inherent to the Gale-Church algorithm by default and therefore not to be of major effect to overall results. We proceed to results presentation for alignment track and sentence track evaluation in the following section.

4. Results and discussion

Evaluation results on alignment level and sentence level F1-measure track are provided in Table 3 for both legislative documents corpora and Croatian-English parallel subcorpus.

When considering results provided by (Gale and Church, 1993) for the core algorithm and results of (Langlais et al, 1998.) for various specific algorithms, pre- and post-processing steps encapsulating Gale-Church algorithm, these results delivered by CORAL are rather predictable and expected. Being that Gale-Church algorithm is proven to work excellent in detecting one-on-one alignments and legislative texts provided in our test case are both really small – 6791 words for Croatian, 8510 words for English overall – and straightforward in terms of manual alignment complexity, figure of 97-98% correct alignments is not surprising.

	Alignment track			Sentence track		
	Precision	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Legislative	96.80	96.15	96.47	97.82	97.47	97.64
Cro-Eng	91.40	93.65	92.51	94.20	93.97	94.08

Table 3. Alignment and sentence track F1-measure on Croatian-English parallel corpora

Alignment on Croatian-English parallel corpus could be improved, on the other hand. Lower results are seen as a direct consequence of increased complexity on newspaper texts, where the basic Gale-Church algorithm encounters larger number of non-one-to-one (one-to-n, n-to-one, one-to-zero, zero-to-one, n-to-n, n-to-m) manual alignments and resolves these with decreased accuracy, as reported in (Gale and Church, 1993) and many papers that followed.

5. Future work

The work presented in this paper certainly leaves room for improvements. Results of this research would without doubt be better with larger and/or annotated corpora, which then could be used for tasks such as word alignment, terminology extraction, creation of thesauri, online dictionaries, semantic networks, etc. If corpora were also POS/MSD tagged and/or lemmatized, it would considerably reduce information search, ambiguities and would enable significantly better exploitation of the text. This we would like to leave for further directions of investigation. Integration of the Croatian language into this kind of multilingual surrounding that exists for other European languages, would enable adding one more language and additional research on new cross-language relations and identities.

Beyond the scope of building additional corpora and enriching existing ones with linguistic annotation, technical improvements might include implementing pre- and post-processing steps around the core Gale-Church algorithm in order to handle possible non-one-to-one alignments with higher recall and precision. The algorithm itself – as a dynamic programming method – might enable additional tweaks or integration with other language preprocessing modules. Future research activities could include alignment experiments at the lower linguistic level i.e. word level or they could include building basic language models and finally experimental systems for statistical machine translation on results presented in this paper.

6. Acknowledgements

This work has been supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-0645, 130-1300646-1776, 130-1300646-0909, 036-1300646-1986.

7. References

Arcade. 2007. Evaluation of parallel text alignment system. URL http://www.up.univ-mrs.fr/veronis/arcade/arcade1/index-en.html

Ceausu, A., Stefanescu, D.; Tufiş, D. 2006. Acquis Communautaire Sentence Alignment using Support Vector Machines. In Proceedings of the LREC2006, Genoa-Paris: ELRA-ELDA.

Dalbelo Bašić, B., Dovedan, Z., Raffaelli, I., Seljan, S., Tadić, M. 2007. Computational Linguistic Models and Language Technologies for Croatian. Proceedings of the 29th ITI Conference. Zagreb : SRCE, 2007. pp. 521-528

Danielsson, P., Ridings, D. 1997. Practical presentation of a "vanilla" aligner. TELRI Workshop on Alignment and Exploitation of Texts. Institute Jožef Stefan, Ljubljana.

EC-DG-JRC. The JRC-Acquis multilingual parallel corpus and Eurovoc (v. 3.0). Italy, 2007. (http://wt.jrc.it/lt/Acquis/JRC-Acquis.3.0/doc/README_Acquis-Communautaire-corpus_JRC.html)

Erjavec, T. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. Proceedings of the LREC 2004, ELRA, Paris.

Gale, W. A., Churck, K. W. 1993. A Program for Aligning Sentences in Bilingual Corpora. Computational Linguistics, vol. 19, pp. 75 – 102. MIT Press, Cambridge, Massachusetts, SAD, 1993.

Koehn, P. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. Machine Translation Summit 2005. URL http://people.csail.mit.edu/koehn/publications/europarl

Langlais, P., Simard, M., Veronis, J. 1998. Methods and practical issues in evaluating alignment techniques. In COLING-ACL98, 1998.

Moore, R. C. 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora. In Machine Translation: From Research to Real Users. Proceedings, 5th Conference of the Association for Machine Translation in the Americas. Springer-Verlag, Heidelberg, Germany.

Och, F. J., Ney, H. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, volume 29, number 1, pp. 19-51 March 2003.

Seljan, S., Gašpar, A., Pavuna, D. 2007. Sentence Alignment as the Basis For Translation Memory Database. INFuture2007 – The Future of Information Sciences: Digital Information and Heritage. Zagreb: Odsjek za informacijske znanosti, Filozofski fakultet, 2007.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. 2006. The JRC-Acquis : A Multilingual Aligned Parallel Corpus with 20+ Languages. Proceedings of the 5th International Conference on Language Resources and Evaluation LREC2006, Genoa, Italy, 24-26 May 2006.

Tadić, M. 2000. Building the Croatian-English Parallel Corpus. In Proceedings of the Second International Conference on Language Resources and Evaluation. ELRA, Paris – Athens 2000, pp. 523-530.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. 2005. Parallel corpora for medium density languages. In Proceedings of the Recent Advances in Natural Language Processing 2005 Conference, pp. 590-596.

THE ARGUMENT STRUCTURE OF THE DATIVE DESIDERATIVE IN SLAVIC: BOSNIAN AND RUSSIAN

Tatyana Slobodchikoff
University of Arizona
slobodch@email.arizona.edu

ABSTRACT

In this paper, I argue that the argument structure of the dative desiderative predicate in Bosnian and Russian can be analyzed in terms of the lexical-semantic features residing in *v*. Specifically, a different combination of features in *v* determines the argument structure of the dative desiderative predicate in both languages. In Bosnian, *v* is specified for the features [-intent; +activity] whereas in its Russian counterpart *v* hosts the features [-intent; +affect].

0. Introduction

All Slavic languages are characterized by the 'dative reflexive' construction, which I call the dative desiderative (1-2).

(1) Jede mi se. (Bosnian)
eat.3.SG.PRES 1.SG.DAT SE
'I feel like eating.'

(2) Mne khoch-et-sja est'. (Russian)
1.SG.DAT want-3.SG.PRES-SJA eat.INF
'I feel like eating.'

The construction has three striking properties. First, it always occurs with reflexive morphology. Second, the experiencer argument is realized in the dative case. Third, the construction has an interpretation of an involuntary desire. Though the dative desiderative construction obtains the same interpretation in both Bosnian and Russian, there is one formal difference which sets the two languages apart. In Russian, the construction requires an overt desiderative verb *khochetsja* ('want') whereas in Bosnian it does not.

I propose that the Bosnian dative desiderative should be analyzed as a mono-clausal construction with a functional desiderative head whereas its Russian counterpart should be treated as a bi-clausal construction with a lexical desiderative verb. Extending Kallulli's (2006, 2007) proposal to Slavic, I provide an analysis of the argument structure of the dative desiderative predicate in terms of the lexical-semantic features present in *v*.

In Russian, the desiderative meaning is encoded by means of the matrix lexical verb, and the involuntary state of the experiencer is reflected in the lexical-semantic feature combination [-intent; +affect] hosted by its *v*. In Bosnian, the involuntary state of the experiencer is encoded by a different *v* which has the feature combination [-intent; +activity]. I claim that an additional element, such as the desiderative functional head with the feature [+des] is the source of the desiderative meaning in the Bosnian predicate.

1. The Dative Desiderative Construction in Slavic Languages

The dative desiderative construction is ubiquitous in all Slavic languages. Its basic morpho-syntax includes reflexive morphology, dative experiencer subject, and either default 3rd person singular agreement with the matrix verb (East Slavic), or singular/plural agreement with the nominative object (South Slavic). We can observe some cross-linguistic variation in the morpho-syntactic realization of the desiderative meaning. In the East and West Slavic languages, the desiderative meaning is realized by means of an overt desiderative verb (3). In the South Slavic languages, the desiderative meaning is not realized by any overt morpho-syntactic element (4).

(3) Jankowi chce się spać. (Polish)
John.DAT want.3.SG SIE sleep.INF
'John feels like sleeping.'
(Rivero and Sheppard 2003:137)

(4) Spav-a mi se. (Bosnian)
sleep-3.SG.PRES 1.SG.DAT SE
'I feel like sleeping.'

In the East and West Slavic, due to the presence of the desiderative verb, the construction obtains an interpretation of an involuntary desire. An involuntary desire reading becomes unavailable if the overt desiderative verb is omitted from the morpho-syntactic structure of the dative desiderative construction (5).

(5) *Mne est-sja. (Russian)
1.SG.DAT eat.3.SG.PRES-SJA
'I feel like eating.'

Since the South Slavic languages have a designated functional head responsible for the desiderative interpretation of the construction, the presence of an overt desiderative verb results in ungrammaticality (6).

(6) *Hoće mi se jesti jagode. (Bosnian)
want.3.SG 1.SG.DAT SE eat.INF strawberries.PL.NOM
'I feel like eating strawberries.'

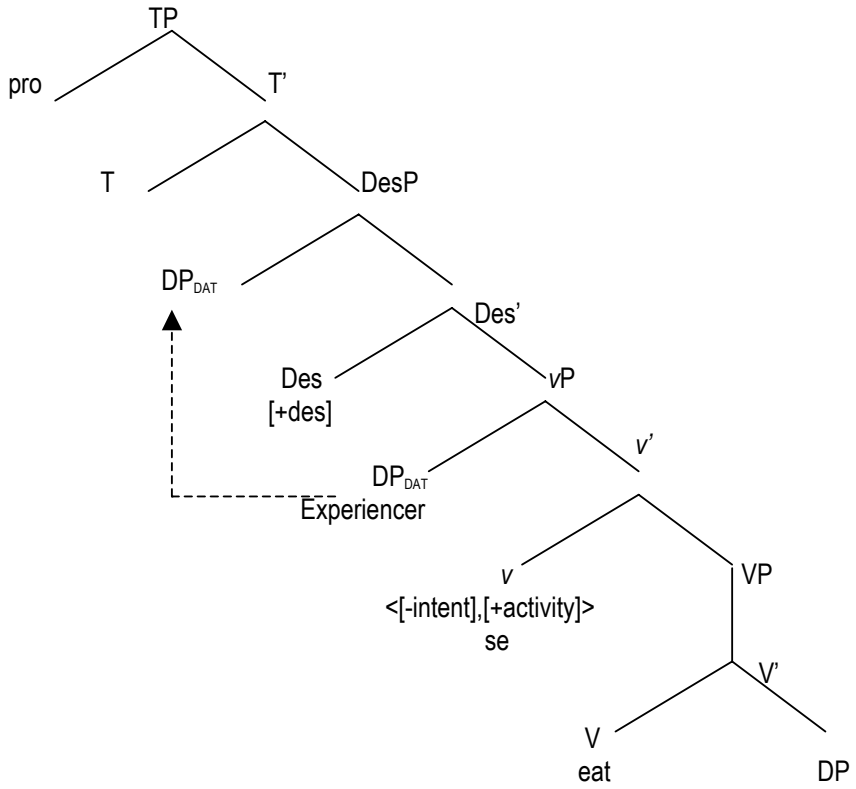
2. The Structure of the Bosnian Dative Desiderative Predicate

I propose to analyze the Bosnian dative desiderative (1) as an overtly mono-clausal structure with the dative experiencer argument introduced by the lexical-semantic features [-intent; +activity] hosted by *v.* I extend Kallulli's (2006, 2007) proposal for the dative desiderative predicate in Albanian to Bosnian. I make the following assumptions. First, I assume that there are four lexical semantic features [±intent], [±cause], [±activity], and [±affect] that define the ontological type of a predicate. Second, I assume that theta-role assignments arise due to a certain combination of lexical semantic features in *v.*

Following Kallulli (2006, 2007), I assume that the features [±cause] [±activity] are primitives since they define different ontological types of events; namely, causative predicates and activity predicates. The third primitive [±intent], which refers to intentionality/agency, can compose with both causative and activity predicates. The fourth primitive refers to the affectedness of an argument, and can combine with the [±intent] feature.

The Bosnian dative desiderative is defined by the feature bundle [-intent; +activity] hosted by *v,* i.e. it is a non-agentive activity predicate (7). In the *v*P projection of the matrix verb, the first feature of the bundle is [-intent] which prevents an agent argument from being realized in the Spec *v*P position. The only possible realization option for the dative argument is that of an experiencer. The second feature [+activity] introduces the dative argument into the syntactic structure. Base-generated in the Spec of *v*P, the dative DP receives its case inherently from *v*. The feature bundle [-intent; +activity] as a whole gives the dative DP an interpretation of an experiencer. However, this feature bundle is insufficient for the predicate to obtain an interpretation of involuntary desire. An additional functional head is necessary to introduce the desiderative interpretation.

(7) *Bosnian Dative Desiderative*

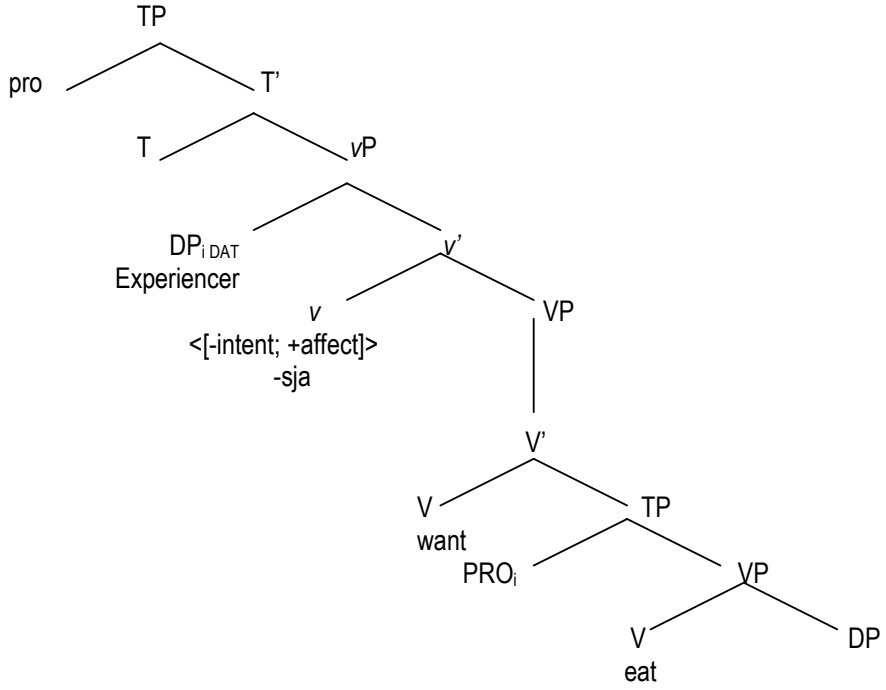


In the absence of an overt desiderative verb, I suggest (similar to Kallulli 2006b) that the desiderative functional head Des is responsible for the desiderative interpretation of the predicate. I claim that the desiderative functional head selects for the feature bundle [-intent; +activity]. It sits above the matrix *v*P and is specified for the [+desiderative] feature. The desiderative feature [+des] triggers movement of the dative DP to its Spec DesP position. Consequently, the dative argument gets interpreted as an experiencer who has an *involuntary desire* to participate in some activity.

3. The Structure of the Russian Dative Desiderative Predicate

I argue that the Russian dative desiderative predicate (2) has a bi-clausal structure with an overt desiderative verb *khochetsja* in the matrix clause and the dative experiencer. The argument structure of the predicate is determined by the lexical semantic feature combination [-intent; +affect] specified in *v* (8). The reflexive suffix *–sja* morphologically realizes this specific feature combination.

(8) *Russian Dative Desiderative*



The structure of the Russian dative desiderative construction is similar to its Bosnian counterpart in that both Russian and Bosnian share the [-intent] feature, which refers to the lack of intentionality. The presence of this feature prevents an agent argument from being realized in the Spec of *v*P making the Russian dative desiderative predicate non-agentive. The Russian dative desiderative differs from its Bosnian counterpart in that it is specified for the feature [+affect]. I introduce the feature [+affect] into the structure of *v* of the matrix desiderative verb to capture the notion of affectedness of the dative argument. I claim that the feature [+affect] introduces the dative argument into the syntactic structure. The dative argument is base-generated in the Spec of *v*P and gets its case assigned inherently by *v*. As a result of the feature bundling [-intent; +affect] in *v*, the dative DP is interpreted as an affected argument or an experiencer. Because the overt desiderative verb *khochetsja* is present in the structure of the dative desiderative predicate, it is responsible for the desiderative interpretation. Thus, the dative DP obtains an interpretation of an experiencer who has an *involuntary desire.*

4. The Russian Dative Habitual Construction

I argue that there is a crucial difference in the structure of the dative habitual and the dative desiderative constructions.

The dative habitual construction is typical of the West and East Slavic languages, and is absent in the South Slavic subgroup. It is a mono-clausal construction with the following morpho-syntactic properties: the dative experiencer, the 3rd person singular matrix verb or neuter participle, reflexive morphology, and either a negative particle or an adverb. Formally, the dative habitual construction in the East Slavic (9) looks similar to the dative desiderative construction in the South Slavic languages (10), but has a completely different interpretation, namely an interpretation of an argument's *habitual state*.

(9) Mne ne spit-sja. (Russian)
1SG.DAT NEG sleep.3.SG.PRES-SJA
'I can't sleep.'

(10) Ne spava mi se. (Bosnian)
NEG sleep.3.SG.PRES 1.SG.DAT SE
'I don't feel like sleeping.'

Compare the Russian dative habitual (9) to the Bosnian dative desiderative (10). Despite the fact that these sentences are syntactically almost identical, their interpretation is very different. The Russian sentence has an interpretation of a habitual state; namely, someone is unable to sleep. In contrast, the Bosnian sentence has an involuntary desire interpretation; specifically, someone does not want to sleep.

The Russian construction can never have an interpretation of an involuntary desire since its structure lacks a desiderative predicate; therefore, negation applies to the main verb only. The Bosnian construction contains the desiderative functional head which negation takes within its scope; therefore, it always has an interpretation of an involuntary desire.

Marušič and Žaucer (2006) propose a structural analogy between the South (Slovenian) and East Slavic (Russian). They claim that if analyzed as a bi-clausal predicate with a phonologically null lexical verb GIVE, the interpretation of involuntary desire can obtain in an overtly monoclausal Russian sentence (11).[1] As I have shown, Russian monoclausal constructions with dative subjects cannot receive a desiderative interpretation due to the absence of a desiderative predicate in their structure. Such constructions always obtain an interpretation of a habitual state.

(11) Mne ne rabota-et-sja. (Russian)
1.SG.DAT NEG work-3.SG.PRES-SJA
*'I don't feel like working.'
'I can't work.'

Analyzing Polish and Slovenian, Rivero and Sheppard (2003) propose to explain the semantic difference between the dative habitual and the dative desiderative construction in terms of the operation of the dative existential disclosure with different types of disclosers in Logical Form. While in Slovenian and other South Slavic languages an Atelic Operator performs the disclosure, in Polish the discloser could be either an adverb, or an overt modal verb, or a negative particle. Rivero and Sheppard's analysis employs three different types of formal disclosers for the dative habitual construction. As I have shown in my approach, the dative habitual construction does not have a desiderative (modal) predicate; therefore, it seems problematic to resort to an overt modal verb as its formal discloser.

5. Conclusion

I have shown that Bosnian and Russian differ in how they encode the desiderative meaning of the predicate. The Russian dative desiderative has a bi-clausal structure with an overt desiderative verb *khochetsja.* In contrast, the Bosnian dative desiderative does not allow an overt desiderative verb in its predicate structure. I proposed to analyze this difference in terms of lexical-semantic features present in *v.* I have argued that the Russian dative desiderative has a matrix desiderative verb whose *v* is specified for the features [-intent; +affect]. Its Bosnian counterpart has a different *v* with the features [-intent; +activity].

[1] Based on my own as well as other native speakers' judgments, I marked the original translation in (11) as ungrammatical and provided my own translation below.

References

Kallulli, Dalina. 2006a. Unaccusatives with dative causers and experiencers: a unified account. In D. Hole, A. Meinunger and W. Abraham (eds.), *Datives and Other Cases: Between argument structure and event structure,* 271-300. Amsterdam: John Benjamins.

Kallulli, Dalina. 2006b. A unified analysis of passives, anticausatives and reflexives. In O. Bonami and P. Cabredo Hofherr (eds.), *Empirical Issues in Syntax and Semantics* 6:201:225. http://www.cssp.cnrs.fr/eiss6

Kallulli, Dalina. 2007. Rethinking the passive/anticausative distinction. *Linguistic Inquiry* 38.4:770-780.

Marušič, Franc and Rok Žaucer. 2006. On the intensional FEEL-LIKE construction in Slovenian: A case of a phonologically null verb. *Natural Language and Linguistic Theory* 24: 1093-1159.

Rivero, Mària Luisa and Milena Milojević Sheppard. 2003. Indefinite reflexive clitics in Slavic: Polish and Slovenian. *Natural Language and Linguistic Theory* 21: 89-155.

LINGUISTIC DATABASE IN AN INTELLIGENT PORTAL FOR EDUCATION AND INFORMATION ("FOUND IN THE NET")

Stanimir Stojanov*, Radka Vlahova**

*University "Paisii Hilendarski", Faculty of Mathematics and Informatics,
Plovdiv, Bulgaria, stani@pu.acad.bg

**Sofia University, Faculty of Slavic Studies
1504 Sofia, 15 Tsar Osvoboditel Blvd., rvlahova@slav.uni-sofia.bg

ABSTRACT

21st century brings forward state-of-the-art language technologies that apply NLP methods and resources drawing largely on mathematics and logic. One of the strategic aims of the software and informatics technologies today is focused on the development and the adoption of commonly accepted and standardized computer and communication infrastructures of completely new generation. The main objective is the development of an effective intelligent distance education system. The information and education portal will incorporate three main structural information units: lecture materials, knowledge assessment tests and diverse information materials (resources in different thematic domains, announcements, news, forums). The realization of the project is ensured by mutual interaction of three different modules: ***Infrastructure for support, management and execute of intelligent electronic services, Information Structure, presentation of humanitarian knowledge in computer format*** which will allow distant tuition and evaluation in the educational sphere of the BULGARIAN LANGUAGE AND LITERATURE.

1. Introduction

21st century brings forward state-of-the-art language technologies that apply NLP methods and resources drawing largely on mathematics and logic. One of the strategic aims of the software and informatics technologies today is focused on the development and the adoption of commonly accepted and standardized computer and communication infrastructures of completely new generation. Trends in computer science indicate that such infrastructures should integrate platforms and architectures oriented to offering of intelligent eServices (eServices) that are tight integrated with telecommunications. Despite of some successes achieved in eCommerce, eLearning, eAdministration (especially eGovernment), the Internet based eServices offered today are still with unsatisfactory quality, low effectiveness and small strength.

The main objective is the development of an effective intelligent distance education system. The information and education portal will incorporate three main structural information units: lecture materials, knowledge assessment tests and diverse information materials (resources in different thematic domains, announcements, news, forums).

The team proceeds from the assumption that the main structural information units should be interrelated while at the same time remaining relatively independent of each other. Therefore, the chief approach underlying the project is modularity, in this way ensuring the portal's specification according to particular criteria, on the one hand, and its openness to incorporation of new components, on the other. Modularity also enables the portal's orientation both towards the high-school education and higher education as well as towards other specific education and information needs such as re-qualification courses.

The autonomous information units may be specified in relation to their creation – either directly in the portal's data base or on their publication on the Internet. The first type subsumes tests being worked out by lecturers, tests being filled out by students, publication of additional information, while the second – publication of lecture courses and additional information (although strict differentiation of the two should not be drawn as the creation of most of the information units will be enabled both ways.

The realization of the project is ensured by mutual interaction of three different modules: ***Infrastructure for support, management and execute of intelligent electronic services, Information Structure, presentation of humanitarian knowledge in computer format*** which will allow distant tuition and evaluation in the educational sphere of the BULGARIAN LANGUAGE AND LITERATURE. The project is intended to facilitate the transition between the various

educational degrees by using uniform methodology of presenting the educational content and testing the results. The project will provide comparable modules of information and will secure a threshold of competence.

In order to achieve these goals it is necessary to study critically the methods for testing and the modes of external evaluation. One of the goals of the project is to offer a new model for evaluating in accordance with the basic objective of the project. The team describes an exemplary research cycle of evaluation in the two degrees of the secondary school, in the initial stage of the higher education, and in the A.M. programs, and also offers a variety of research designs for external evaluation. One of the main objectives of the project is the creation of and extensive data base with tests which will be statistically evaluated for their most important parameters: facility value, discrimination index, test consistency.

As a final product the product is aimed at the creation of multimedia products which will present the content of the language competence in Bulgarian, and tests for self-evaluation and external evaluation of the knowledge and skills in Bulgarian.

The result is seen as intelligent, web-based educational and information gateway dedicated to humanities.

The results obtained throughout the project may be used not only for the purposes of distance education but will also be applicable in different types of Internet services aimed at the development of the e-society.

2. Theoretical background

2.1. New computer and communication infrastructure

For the creation of a new generation of intelligent eServices effectively-working approaches, models and methods were offered in the last few years along with the implementation of related software modules, components and entire technologies. The World Wide Web delivers a universal infrastructure for the development of client-server applications, as for data accessing there are universal client (Web browser), universal communication protocol (HTTP), universal server (Web server) and universal user interface (HTML) available. The lack of some basic features and components preventing it to be complete and total platform (because it is static and does not support states) can be compensated through the integration of Java and J2EE technological components. For the integration of the existing applications (and legacy systems) realized in different software languages and working on different platforms, the use of the middleware broker CORBA is especially successful. With the appearance of Web, the EAI obtained more significance extending beyond the borders of association of applications only within single company or organization. Application servers support already huge amount of data and process a significant part of the applications' functionality. The service provision through the Web emerges as a strategic direction in the evolution of the information technologies. The Web-based integration of applications facilitates the access to data and services. Companies must integrate the existing applications and systems in order to manage B2C (Business-To-Client) and B2B (Business-To-Business) interactions along with the Web-services. The Web-services use a model for applications integration with three XML-based protocols:

- WSDL (Web Services Description Language) – a standard for service description;
- UDDI (Universal Description, Discovery, and Integration) – a standard protocol for publication and search of Web-services;
- SOAP (Simple Object Access Protocol) - a standard protocol for Web-services binding.

Another trend gaining speed is related to the development of semantic Web, eContent systems based on agent technologies integrating Web services by means of the communication protocol DAML-S (OWL-S) developed by the DARPA consortium.

2.2. Electronic learning (eLearning)

Applying of information and communication technologies in education is one of the most substantial scientific topics today. The eLearning could be used effectively in all types of schools including secondary schools. It allows for accessible, effective, mobile and flexible education not only for regular students but also for students using remote form of education or students with specific needs.

The research in this area is focused on finding solutions for the following main problems:

- Development of suitable information and technological infrastructure for eServices used in education. Especially suitable for this is the portal technology;

- Creation of effective developing environments for eContent generation;
- Assurance for interoperability between different eLearning systems, which allows to connect them in regional and national educational networks.

2.3. Semantic networks & Text categorization

Semantic networks are among the most popular Artificial Intelligence formalisms for knowledge representation that have been widely used in the 70's and 80's to represent structured knowledge. Like other networks, they consist of nodes and links. Nodes represent concepts, i.e., abstract classes whose members are grouped together on the basis of their common features and/or properties, while arcs between these nodes represent relations between concepts and are labeled so as to indicate the relation they represent. The semantics of the concepts resides not in the name of the associated labels, but in the concepts' properties and relations to other concepts of the semantic network. The sets of words, called synsets (synonymy sets), constitute the building blocks for representing the lexical knowledge reflected in WordNet, the first implementation of lexical semantic networks. As in the semantic networks formalisms, the semantics of the lexical nodes (the synsets) is given by the properties of the nodes (implicitly, by the synonymy relation that holds between the literals of the synset and explicitly, by the gloss attached to the synset and, sometimes, by specific examples of usage) and the relations to the other nodes of the network. These relations are either of a semantic nature, similar to those to be found in the inheritance hierarchies of the semantic networks, and/or of a lexical nature, specific to lexical semantics representation domains. The convergence of the representational principles promoted both by the domain-oriented semantic networks and ontologies, and by WordNet's philosophy in representing general lexical knowledge, is nowadays an apparent trend, motivated not by fashion, but by the significant improvements in performance and by the naturalness of interaction displayed by the systems that have adopted this integration.

Text categorization has witnessed a booming interest recently, due to the availability of ever larger numbers of text documents in digital form and to the ensuing need to organize them for easier access and use. The dominant approach nowadays is one of building text classifiers automatically by learning the characteristics of the categories from a training set of pre-classified documents. State-of-the-art machine learning methods have recently been applied to the task, leading to systems of increased sophistication and effectiveness. At the same time there is a progressive adoption of automatic or semi-automatic (i.e. interactive) classification systems in applicative contexts where manual work was the rule. Bulgarian wordnet and the domain ontologies will be the basic knowledge representation and integration models which will supply the system with the backbone resources for the semantic processing of the web documents in carrying out text categorization and question answering tasks. To this end techniques will also be employed that combine in an effective way the achievements of modern information technologies in natural language processing (NLP).

3. Objectives, hypothesis, approach

This project will be natural continuation and expansion of the team's research work carried on in the previous collaborated projects. The main aim of the project is to develop a standardized infrastructure for the integration of intelligent services, testing and verification of the infrastructure for a particular application area, namely eLearning and intelligent eLearning portals.

For the realization of this aim the following ***main tasks*** under the first module are defined***:***

1. Development of entire model of an infrastructure for eServices integration including elements of semantic Web such as:
   - Intelligent agents;
   - eServices ;
   - Appropriate protocols for communication between agents and eServices;
   - Ontologies.
2. Development of architecture of eLearning educational portal as a concrete realization of the infrastructural model.
3. Development and implementation of models allowing additional extra portal intelligency and flexibility:
   - Models of educational modules/courses (domain models);

- Pedagogical model.
- User model.

The second proposed project module ***Information Structure Design*** includes the design and implementation into the Intelligent Information and Education Portal of modules for automatic correction of information units (proofing tools), automated text categorization, and automated question answering. On the creation of information units intelligent user-assisting modules will be implemented directly in the portal's data base drawing on up-to-date language technologies – a module that will ensure correct spelling in conformity with the norms of the Contemporary Bulgarian Language (Spell Checker) and a module that will enable word substitution with synonyms, sense-related words and provide other useful language information (Thesaurus).

The achievement of these objectives will require the carrying out of several individual, yet interrelated, tasks presented within the description of each module.

The objectives towards text categorization can be described as follow:

- Automatically process large volumes of Web documents so as to represent them by sequences of thematically related keywords;
- Computing a thematic category of the domain ontologies for each of the documents;
- Automatically fetch, categorize and sort documents within thematic domains.

The objectives towards development of a Question – Answering system are as follows:

- Build topical focused crawlers for the topics covered in the domain ontologies;
- Analyze large volumes of (Web) data already categorized in order to trace events, names, places, topics, etc. within their contents;
- Acquire and create training paradigms for Question-Answering;
- Implement a Question-Answering System.

The aim of ***third project module*** is to work out a presentation of humanitarian knowledge in standardized computer format allowing for distant tuition and evaluation. The goal is to help the transition between different educational degrees by applying uniform methodology of evaluation based on comparable modules of information ensuring a threshold of competence.

In order to achieve this objective it is necessary to study completely the methodology of standardization, as well as the manners of evaluations and the varieties for external evaluation. The project team undertakes the task of describing an exemplary research cycle for evaluation in the two degrees of the secondary school and in the M.A. programmes and offers variants of research design for external evaluation

4. Methodology, research technologies, data processing, analysis

4.1. Common model of eServices infrastructure

As a foundation for the development of this model will be used The DeLC infrastructure. The infrastructure is oriented to support different types of eServices. The eLearning Nodes are the main elements of this infrastructure.

The DeLC Infrastructure Model specifies the basic building blocks of the DeLC. Furthermore it characterizes all the possibilities for integration and management of eServices within the defined clusters. The DeLC model consists of DeLC Nodes (DeLCNs) that are established and supported by real administrative units offering a complete educational cycle (e.g. laboratories, departments, faculties, colleges, universities). In order to support also mobile eServices a so-called *Expanded DeLC Network Model* is developed, which has a 3-tier structure. The integration of eServices is provided within individual clusters (i.e. controlled cluster-oriented integration is supported).

The model of the infrastructure specifies also a 3-tier architecture supporting mobile eServices, called Expanded DeLC Network Architecture.

4.2. Creation of the infrastructure of the intelligent educational portal

The corporate portal is an application of a broad set of technologies following customized information design. There are many challenges in providing users with a personalized, single-point-of-access desktop that integrates both existing corporate information systems and external information sources.

We intend to implement an additional layer, called ***Learning Loop***. The portal Learning Loop differs from the other architectural components in that is not concerned with a specific aspect of information management, but in the ongoing effectiveness of the portal itself. This component enables the portal to adjust heuristically to changes in the organization's and information environment.

4.3. Text categorization & Question-Answering

We will collect a large volume of Web documents and process them in a two-steps approach. Lexical chaining concerns representing the documents context into a sequence of thematic keywords, which will be used for indexing the Web data into an appropriate thematic domain. For lexical chains' generation, the monolingual wordnets will be extensively used. The lexical elements contained in each documents' lexical chain will be mapped against the domain concepts of the specialized ontologies. Mapping will be performed semi-automatically keeping human effort overheads to a minimum. Afterwards, well-known semantic similarity techniques will be employed together with a categorization algorithm in order to define which of the ontology domains match the semantics of each of the documents. Several algorithmic and scoring techniques will be applied thereafter in order to compute the best matching categories into which to assign the documents.

4.4. Presentation of humanities knowledge

The numbers of competing theories and evaluation models have made some authors use the term "Balkanization", by which they mean fragmentation and lack of consent and consistency in the description of various theoretical tendencies in the field of evaluation studies. We share the view that both approaches should receive considerable attention in the field of evaluation studies, but with regard to the nature of the proposed strategy as a strategy for external programme evaluation, we think that the leading role should be given to the quantity approach.

References

Andreev, R., Ivan Ganchev, Martin O'Droma, Innovative SCORM-Based Approach for the Creation of Reusable VUIS eLearning Content, 2004

Ganchev I., S. Stojanov, M. O'Droma, I. Popchev. 2004. "Enhancement of DeLC for the Provision of Intelligent Mobile Services". In Proc. of the 2nd International IEEE Conference on Intelligent Systems (IS'2004), vol. 1, Pp. 359-364, 22-24 June, Varna, Bulgaria. IEEE Cat. No. 04EX791. ISBN 0-7803-8278-1. Library of Congress 2003115853.

Glushkova T., Framework for eLearning in Secondary School by application of DeLC system, 2nd International Workshop on eServices and eLearning, September 28-29, 2004, Smolian, Bulgaria

Glushkova T., Electronic cluster "ECL-School Brezovo" of the DeLC system, Scientific-practical conference "New technologies in education and vocational training", Sofia 2003.

Glushkova T., S. Stojanov, DeLC architectural framework adaptation to a second level education, Scientific-practical conference "New technologies in education and vocational training", Plovdiv, 2004.

Glushkova T., M. Sarafov, A model for the application of eLearning in second level education, in "Good pedagogical practices", Sofia, 2004, p. 19.

Fellbaum 1998: Fellbaum, C. (ed.). WordNet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press, 1998.

Koeva 1998: Koeva, S., Grammar dictionary of Bulgarian. Description of the conception of the language data organization. – *Bulgarian Language*, 6, 49-58

Koeva 2001: Koeva S., S. Mihov "INTEX 4.0 for Bulgarian (Error checking as an INTEX application)" – in: Revue Informatique et Statistique dans les Sciences humaines, Anne Dister, ed., Universite de Liege, 2001, 231-241.

Koeva 2004a: Koeva S., G. Totkov, A. Genov, Towards Bulgarian WordNet, *Romanian Joutnal on Information Science and Technology*, 2004, Vol. 7, 45-61.

Koeva 2004b: Koeva S., T. Tinchev, S. Mihov, Bulgarian WordNet – Structure and Validation, *Romanian Journal on Information Science and Technology*, 2004, Vol. 7, 61-79.

Koeva 2005: Koeva, Sv., Rizov, B., Lesseva Sv. Flexible Framework for Development of Annotated Corpora в International Journal "*Information Theories & Applications"*, под печат.

Koper, R. Use of the Semantic WEB to Solve Some Basic Problem in Education, Journal of Interactive Media in Education, 2004

Learning System Architecture Lab (LSAL) at Carnegie Mellon University, SCORM Best Practices Guide for Content Developers 1st edition, 2003, www.lsal.cmu.edu/lsal/expertise/projects/developersguide/index.html

Miller 1990: Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J. (1990). Introduction to WordNet: An On-Line Lexical Database. In *International Journal of Lexicography*, vol. 3, no. 4 (winter 1990), 235-244.

SCORM 2004 Overview, http://www.adlnet.org

SCORM, ADL Initiative, SCORM 2004 specification, http:// www.adlnrt.org

Schultz 2004: Klaus Schulz, Stoyan Mihov and Svetla Koeva "Precise and Efficient Text Correction Using Levenshtein Automata, Dynamic Dictionaries, Web Directories and Optimal Correction Model" in: *Proceedings from 1st International Workshop on proofing Tools and Language Technologies*, Patras, Greece, 2004.

Silberzstein 1993: Silberztein, Max. Dictionnaires électroniques et analyse automatique de textes. Le système INTEX. Masson: Paris. 1993. (240 p.).

Silberzstein 1999: Silberztein, M. INTEX: a Finite State Transducer toolbox, in *Theoretical Computer Science,* 1999. 231:1.

Stamou 2002: Stamou S., Oflazer K., Pala K., Christodoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit

D., Grigoriadou M. (2002b). BALKANET: A Multilingual Semantic network for the Balkan Languages. In *Proceedings of the st International Global Wordnet Conference*, Mysore, India, 2002.

Stojanovic L., eLearning based on the Semantic Web, http://www.aifb.uni-karlsruhe.de

S. Stojanov, I. Ganchev, M. O'Droma, A Model for Integration of Services in a Distributed eLearning Center. In Proceedings of 14th Annual EAEEIE Conference, Gdansk, Poland , June 2003.

Stoyanov S., A. Stoyanova-Doycheva, M. Trendafilova, Functional model of services offered by a Virtual University, National scientific conference "Informatics in the science knowledge", 13-15 June 2002, Varna.

Vossen 1998: Vossen P. (Ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic publishers, Dordrecht, 1998.

http://www.lsal.cmu.edu/lsal/expertise/projects/compio/compio2001finalreport/virtual_univ/portaltaxonomy.html

http://uk.teachers.net

http://www.web-learning.org/osservation/Enti

http://www.educoas.org/portal/en/tema/editorial2005

http://www.reload.ac.uk

HIGHER-ORDER FUNCTIONAL REPRESENTATION OF CROATIAN INFLECTIONAL MORPHOLOGY

Jan Šnajder and Bojana Dalbelo Bašić

Department of Electronics, Microelectronics, Computer and Intelligent Systems,
Faculty of Electrical Engineering and Computing, University of Zagreb
Unska 3, Zagreb, Croatia
{jan.snajder, bojana.dalbelo}@fer.hr

ABSTRACT

Many language processing tasks rely on some sort of morphological processing, such as word-form generation, lemmatization, or morphological tagging. These are inherently difficult tasks for morphologically complex Slavic languages. For these languages morphology representation becomes an issue as well. This paper describes higher-order functional morphology (HOFM), a convenient formalism for representing inflectional concatenative morphologies, and its application to morphologically complex Croatian language. In HOFM, inflectional paradigms are represented by inflectional rules and morphological transformations are represented by higher-order functions. HOFM representations resemble the highly expressive morphology descriptions often found in traditional grammar-books, allowing for compact and comprehensible morphology models. We describe a HOFM implementation in Haskell, a modern functional programming language, and demonstrate its use on the task of generating and lemmatizing Croatian words.

1. Introduction

Many natural language processing tasks rely on some sort of morphological processing, such as word-form generation, lemmatization, or morphological tagging. For morphologically complex languages, this is an inherently difficult task. Because of this complexity, morphology representation becomes an issue as well. A morphology representation formalism should strike a good balance between expressiveness and complexity. This presents a challenge for morphologically complex Slavic languages such as Croatian.

Previous work on modeling of Croatian morphology uses the two-level formalism Koskenniemi (1983) to describe only a subset of Croatian inflectional and derivational morphology (Lopina 1992). Of more practical importance is the work by Tadić (1994), who models Croatian inflection with over 600 patterns and uses this model to generate an inflectional lexicon (Tadić and Fulgosi 2003). It could be argued that in both cases morphology modeling is extremely labor intensive and requires a substantial amount of linguistic expertise.

This paper describes *higher-order functional morphology* (HOFM) – a convenient formalism for representing inflectional concatenative morphologies – and describes its application to Croatian language. The HOFM formalism utilizes the word-and-paradigm approach first proposed by Hockett (1954). In HOFM, inflectional paradigms are represented by *inflectional rules*. An inflectional rule defines the various word-forms obtainable from the word's stem. Because defining the various morphological transformations directly would be tedious, HOFM uses the notion of a *higher-order function* – a function that returns a function or takes functions as arguments – to define these transformations indirectly. The idea is to use one higher-order function for each type of morphological transformation, such as suffixing, prefixing, phonological alternation, or some combinations of those. This way, HOFM representations resemble the highly expressive morphology descriptions often found in traditional grammar-books. This allows for compact and comprehensible morphology models. The feasibility of this approach has been proven in practice: in (Šnajder et al. 2008) HOFM description of Croatian morphology has been used successfully to acquire automatically an inflectional lexicon for morphological normalization.

In this paper we describe a HOFM implementation in Haskell (Jones 2003), a functional programming language that has gained widespread acceptance. The potential of functional programming for natural

language processing has only recently been recognized (Frost 2006). Related to our work is the functional morphology formalism developed by Forsberg and Ranta (2003). In functional morphology, an inflectional rule is encoded as a Haskell function over finite algebraic types representing the grammatical categories. Given an input string (the lemma) and the grammatical features, the function generates the corresponding word-form. However, it is not possible to compute the inverse of such functions, hence functional morphology cannot be used to lemmatize a given word-form. This is in contrast to HOFM, which is more abstract, but may be implemented so that both directions of computation are possible.

The paper is organized as follows. Section 2. describes the HOFM formalism and Section 3. describes its Haskell implementation. In Section 4. we describe the HOFM representation of Croatian inflectional morphology. Section 5. concludes the paper and discusses future work.

2. HOFM formalism

2.1. Transformation and Condition Functions

HOFM describes the inflection of a word as morphological transformation of the word's stem (the part of the word common to all inflected forms). Let $\mathcal{S}$ be the set of all strings. Morphological transformation is defined by a partial injective *transformation function* $t : \mathcal{S} \to \mathcal{S}$. Let $\mathcal{T}$ be the set of all morphological transformations. Let nul denote the identity transformation, i.e., $\forall s \in \mathcal{S}.nul(s) = s$.

Defining the transformation functions directly would be a tedious job considering all the possible morphological transformations stems undergo. Instead, HOFM uses higher-order functions to define these transformations indirectly. The idea is to use one higher-order function of type $f : X \to \mathcal{T}$ for each *type* of morphological transformation, where values from X act as transformation parameters. For concatenative morphologies, such as Croatian morphology, common transformation types are suffixing, prefixing, various phonological alternations, or some combinations of those. Thus, to define suffixation, we can use a higher-order function $sfx : \mathcal{S} \to \mathcal{T}$, defined so that $sfx(r)$ returns a transformation function of suffixing r to stem s:

$$sfx(r) = \lambda s.(s +\!\!+ r),$$

where '++' denotes string concatenation. With this higher-order function, morphological transformation corresponding to suffixation of '*-a*' can simply be represented as $sfx(\text{`}a\text{'})$, with suffix '$a$' acting as transformation parameter. This transformation can then be applied to distinct stems to yield their corresponding word-form, e.g.: $sfx(\text{`}a\text{'})(\text{`}slik\text{'}) = \text{`}slika\text{'}$, $sfx(\text{`}a\text{'})(\text{`}rib\text{'}) = \text{`}riba\text{'}$, etc. To define transformations featuring phonological alternations, we can use a higher-order function $alt : \wp(\mathcal{S} \times \mathcal{S}) \to \mathcal{T}$, where $\wp$ is the power-set operator. Here the phonological alternation is defined by a set of string pairs, each defining the stem ending and its replacement. Thus, sibilarization (the replacement of velar consonants *k*, *g*, and *h* with *c*, *z*, and *s*, respectively) may be defined as follows:

$$sbl = alt(\{(\text{`}k\text{'}, \text{`}c\text{'}), (\text{`}g\text{'}, \text{`}z\text{'}), (\text{`}h\text{'}, \text{`}s\text{'})\}). \tag{1}$$

To combine phonological alternation with suffixation, we make use of functional composition, denoted '$\circ$'. For instance, morphological transformation $t = sfx(\text{`}i\text{'}) \circ sbl$ will first change the stem's ending, and afterwards concatenate to it the suffix '*-i*'. For example, $t(\text{`}slik\text{'}) = \text{`}slici\text{'}$ and $t(\text{`}duh\text{'}) = \text{`}dusi\text{'}$.

Another kind of functions that HOFM uses are the *condition functions* of type $c : \mathcal{S} \to \{\top, \bot\}$, where $\top$ denotes the truth and $\bot$ denotes falsity. Condition functions will be used to define the applicability of an inflectional rule to a particular stem. As in the case of transformation functions, we can use higher-order functions to define indirectly the various condition functions. A useful higher-order function would be $ends : \wp(\mathcal{S}) \to \mathcal{C}$, used to test whether the stem ends in a particular suffix. For example, $ends(\{\text{`}i\text{'}\})(\text{`}slici\text{'}) = \top$. More complex conditions can be formulated as compound logical statements in which higher-order functions are used as terms. To this end, we extend the usual logical operators to operate on functions:

$$c_1 \wedge c_2 = \lambda s.(c_1(s) \wedge c_2(s)), \quad c_1 \vee c_2 = \lambda s.(c_1(s) \vee c_2(s)), \quad \neg c = \lambda s.(\neg c(s)).$$

For example, expression $ends(velars) \vee ends(\{\text{`}st\text{'}\})$ yields a condition function that tests whether a stem ends in a velar (set $velars$) or the suffix '*-st*'.

2.2. Inflectional Rules

A HOFM *inflectional rule* defines the distinct word-forms of a single word as transformations of the word's stem, as well as the condition which the stem must satisfy in order for the rule to be applicable. Formally, an inflectional rule r is a pair:

$$r = (c, \{(t_0, d_0), \ldots, (t_n, d_n)\}) \ \in \mathcal{C} \times \wp(\mathcal{T} \times \mathcal{D}). \tag{2}$$

The condition function c defines the condition that the stem s must satisfy in order for the rule r to be applicable. Set of pairs $\{(t_0, d_0), \ldots, (t_n, d_0)\}$ defines the transformations and the corresponding (arbitrarily formatted) morphosyntactic descriptions. In other words, functions $t_0, \ldots, t_n$ define the transformation of stem s into n distinct word-forms, $t_0(s), t_1(s), \ldots, t_n(s)$, morphosyntactically described by $d_0, d_1, \ldots, d_n$, respectively. By convention, string transformation t_0 transforms the stem s into the lemma l, i.e., $l = t_0(s)$. As an example, consider the inflectional paradigm of noun *slika* (picture). The corresponding inflectional rules is defined as follows:

$$\begin{aligned} r_{slika} = \Big(ends(velars), \Big\{ & (sfx(\text{`}a\text{'}), d_0), (sfx(\text{`}e\text{'}), d_1), (sfx(\text{`}i\text{'}) \circ sbl, d_2), (sfx(\text{`}i\text{'}), d_3), (sfx(\text{`}u\text{'}), d_4) \\ & (sfx(\text{`}o\text{'}) \circ sbl, d_5), (sfx(\text{`}om\text{'}), d_6), (sfx(\text{`}ama\text{'}), d_7) \Big\}\Big), \end{aligned} \tag{3}$$

where $d_0, \ldots, d_7$ are the morphosyntactic descriptors (we will fill in the details of this definition in the next section).

An inflectional rule defines the word-forms obtainable from a particular lemma, provided the rule is applicable to it. An inflectional rule r defined by (2) is regarded *applicable* to a lemma l, denoted $r \vDash l$, if (i) each transformation function from r is defined for the string it is applied to, and (ii) the condition of the inflectional rule r is satisfied on the stem s obtainable from lemma l. Formally:

$$(c, \{(t_i, d_i)\}) \vDash l \qquad \text{iff} \qquad \forall t_i.\ (t_i \circ t_0^{-1})(l) \downarrow \ \wedge (c \circ t_0^{-1})(l), \tag{4}$$

where $t_0^{-1}(l)$ transforms the lemma into the stem by applying the inverse of transformation t_0, and $t(s) \downarrow$ denotes that transformation t is defined for string s. Because an inflectional rule must define all word-forms of a given lemma, no transformation is allowed to be undefined for the stem obtain from l. We assume that if $x \uparrow$, then $t(x) \uparrow$, i.e., a transformation of an undefined value is also undefined.

2.3. Word-form Generation and Lemmatization

Based on the notion of an inflectional rule, we now can define functions for word-form generation and lemmatization. The word-forms (and their corresponding morphosyntactic descriptors) of lemma l can be obtained by the word-forms function $wf : \mathcal{S} \times \mathcal{R} \to \wp(\mathcal{S} \times \mathcal{D})$, a one-to-many mapping defined as follows:

$$wf(l, r = (c, \{t_i, d_i\})) = \begin{cases} \left\{ \left((t_i \circ t_0^{-1})(s), d_i \right) \right\} & \text{if } r \vDash l, \\ \emptyset & \text{otherwise}, \end{cases} \tag{5}$$

where '$\vDash$' is the applicability relation defined by (4). Computationally speaking, each word-form is obtained by first applying the inverse of t_0 to transform the lemma l into the stem s, and then by applying t_i to transform the stem s into the word-form. If rule r is not applicable to lemma l, an empty set is returned.

The lemmatization function $lm : \mathcal{S} \times \mathcal{R} \rightarrow \wp(\mathcal{S} \times \mathcal{D})$ is defined as follows:

$$lm(w, r = (c, \{t_i, d_i\})) = \left\{ \left((t_0 \circ t_i^{-1})(w), d_i \right) : \ r \vDash (t_0 \circ t_i^{-1})(w) \right\}. \tag{6}$$

This function basically works the other way around: it first transforms each word-form into the stem, and then attempts to transform each stem into the lemma. From the lemmas obtained, only those to which the rule is applicable are retained. Note that, because an inflectional rule may be ambiguous in general, this function too is a one-to-many mapping.

3. Haskell implementation

Functional programming languages are based on the lambda calculus, a theory of computation evolved around the notion of a mathematical function. This makes functional programming languages a natural choice for the implementation of HOFM. Our language of choice is Haskell (Jones 2003), a modern functional programming language.

3.1. Condition and Transformation Functions

Condition functions can be implemented in Haskell in a straightforward manner. A condition function maps strings to Boolean values, thus its type in Haskell is:

```
type Cnd = String -> Bool
```

Unfortunately, the same approach will not work for transformation functions. The problem is that (4), (5), and (6) require us to be able to compute at run-time the inverse of a transformation function. Because the computation of inverses is intractable in general, we cannot use Haskell functions directly to implement the transformation functions. Instead, our approach will be to implement the transformation functions indirectly using a set of *string operation primitives*, each corresponding to a simple bijection. Three such operations will be used: suffix replacement (`RS`), prefix replacement (`RP`), and infix replacement (`RI`).[1] These will be represented in Haskell as values of so-called algebraic type, named `StringOp`:

```
data StringOp = RS String String | RP String String | RI String String
```

The inverse of a string operation can simply be obtained by switching the order of the strings to be replaced, i.e., the inverse of `RS x y` is `RS y x`. A transformation function t can now be implemented as a list comprised of a list of string operations. Thus, transformation functions will in fact be of the following type:

```
type Tr = [[StringOp]]
```

where inner lists represent the composition of string operations, while the outer list represents the (mutually exclusive) transformation options. For example, transformation sbl defined by (1) will be represented as:

```
[[RS "k" "c"],[RS "g" "z"],[RS "h" "s"]]
```

In list of options, exclusivity is enforced by giving higher priority to operations at the beginning of the list.

For convenience, we will define function wrappers that resemble the original transformation functions. For example (in the following Haskell operator '`::`' is used to declare a type of a value or function):

[1] Note that infix replacement is by itself not injective unless additional assumptions are made (e.g., that only the leftmost infix is replaced.)

Table 1: Basic HOFM condition and transformation (higher-order) functions implemented in Haskell.

Name and type	Description	Example of application
`always :: Cnd`	always true	`always "slik"` ⇒ `True`
`ends :: [String] -> Cnd`	suffix testing	`ends ["k","g","h"] "slik"` ⇒ `True`
`starts :: [String] -> Cnd`	prefix testing	`starts ["naj"] "najmanji"` ⇒ `True`
`land :: Cnd -> Cnd -> Cnd`	logical AND	`starts ["naj"] ‘land‘ ends ["k"] "najmanji"` ⇒ `False`
`lor :: Cnd -> Cnd -> Cnd`	logical OR	`starts ["naj"] ‘lor‘ ends ["k"] "najmanji"` ⇒ `True`
`neg :: Cnd -> Cnd`	logical NOT	`neg(ends ["k"]) "slika"` ⇒ `True`
`nul :: Tr`	nul transformation	`nul $$ "slik"` ⇒ `"slik"`
`sfx :: String -> Tr`	suffixation	`sfx "a" $$ "slik"` ⇒ `"slika"`
`pfx :: String -> Tr`	prefixation	`pfx "naj" & sfx "iji" $$ "crn"` ⇒ `"najcrniji"`
`alt :: [(String, String)] -> Tr`	alternation	`sfx "i" & alt [("k","c")] $$ "slik"` ⇒ `"slici"`
`rsfx :: String -> String -> Tr`	suffix replacement	`sfx "ovi" & rsfx "ao" "l" $$ "ugao"` ⇒ `"uglovi"`
`rpfx :: String -> String -> Tr`	prefix replacement	`rpfx "naj" "" $$ "najmanji"` ⇒ `"manji"`
`rifx :: String -> String -> Tr`	infix replacement	`rifx "ije" "e" & sfx "na" $$ "vrijeme"` ⇒ `"vremena"`

```
alt :: [(String,String)] -> Tr
alt ss = [[RS s1 s2] | (s1,s2) <- ss]
```

Composition of transformation functions amounts to inner lists concatenation. For this purpose we define in Haskell the infix operator ‘`&`’. For instance, computation of $sfx(\text{`}i\text{'}) \circ sbl$ results in (‘ ⇒’ denotes the result of a computation returned by the Haskell interpreter):

```
sfx "i" & sbl ⇒ [[RS "k" "c",RS "" "i"],[RS "g" "z",RS "" "i"],[RS "h" "s",RS "" "i"]]
```

Since $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$, computing the inverse of a transformation amounts to reversing the inner lists and inverting each individual string operation. This will be accomplished by function `inv`. For example, the computation of $(sfx(\text{`}i\text{'}) \circ alt(jot))^{-1}$ is as follows:

```
inv(sfx "i" & sbl) ⇒ [[RS "i" "",RS "c" "k"],[RS "i" "",RS "z" "g"],[RS "i" "",RS "s" "h"]]
```

To perform the actual transformation, i.e. the application of a transformation function, all we need is a function that takes a list of type `Tr` and a string to be transformed as inputs, and then executes the string operations in left-to-right order. For this we define in Haskell the ‘`$$`’ function. For example:

```
sfx "i" & sbl $$ "slik" ⇒ Just "slici"
inv(sfx "i" & sbl) $$ "slici" ⇒ Just "slik"
```

Table 1 summarizes the basic HOFM condition and transformation functions implemented in Haskell, most of which are higher-order functions. It should be noted that this set is not minimally expressive (some functions can be defined in terms of others).

3.2. Inflectional rules

To implement HOFM inflectional rules defined by (2) in Haskell, we define a Haskell data-type `IRule` as follows:

```
type Label = String
type Msd = String
data IRule = IRule Label Cnd [(Tr, [Msd])]
```

The data-type tells us that a rule consists of a label (a unique string), a condition function, and a list of pairs each consisting of a transformation function and a list of morphosyntactic descriptors (strings). In our implementation the morphosyntactic descriptors are based on the MULTEXT-East Specification (Erjavec et al. 2003). MULTEXT-East compactly encodes the values of various morphosyntactic attributes in a single string. The value of each attribute is represented by a single character at a predefined and fixed position within the string. Non-applicable attributes are marked by a hyphen ('-'), with the omission of trailing hyphens. For example, for word-form *slika* (picture) one would use the MULTEXT-East description 'Ncfsn-n': the first five values indicate a common (c) feminine (f) noun (N) in singular nominative (sn) case, then follow two non-applicable attributes (definiteness and clitic), while the last value (n) indicates noun's non-animateness. We extend this specification with an additional marker, '=', to indicate that a value of an attribute is not deducible at the morphological level. Thus, in case of noun *slika*, we will use the descriptor 'N=fsn', because noun type and animateness cannot actually be determined by an inflectional rule. As revealed by the above data-type declaration, each morphological transformation will be associated with a list of MULTEXT-East descriptors, instead of a single descriptor. This allows for a more compact representation and faster computation. As an example, consider Haskell implementation of rule (3):

```
r_slika :: IRule
r_slika = irule "r_slika" "N=f"
  (ends velars)
  [sfx "a"         # ["sn","pg"],
   sfx "e"         # ["sg","pn","pa","pv"],
   sfx "i" & sbl   # ["sd","sl"],
   sfx "i"         # ["sd","sl"],
   sfx "u"         # ["sa"],
   sfx "o"         # ["sv"],
   sfx "om"        # ["si"],
   sfx "ama"       # ["pd","pl","pi"]]
```

Instead of directly defining a value of type IRule, we use function irule as a wrapper. This function takes one additional argument: a partial morphosyntactic descriptor common to all word-forms producible by the rule ('N=f' in this case). The partial descriptor will be combined with the more specific descriptors associated with the various morphological transformations. This way, we allow for incremental specification of MULTEXT-East descriptors. The infix operator '#' is merely syntactic sugar: it constructs a pair comprised of a given transformation and a list of partial descriptors.

The above example illustrates how conveniently one can encode HOFM inflectional rules in Haskell, but the rule itself is rather simple. For more complex paradigms, such as the Croatian adjectival paradigm discussed in the next section, writing comprehensive inflectional rules becomes more difficult. To ease the definition of more complex inflectional rules, we define two additional Haskell operators: '<&' and '<#'. The former distributes a function composition, while the latter distributes combining of partial morphosyntactic descriptors. In other words, we can use the former to add a transformation function and the latter to add morphosyntactic descriptors to a list of transformation-descriptor pairs. For example:

```
[sfx "i" # ["msn","msa"], sfx "og" # ["msg","nsg"]] <& sfx "ij" <# ["c..."]
```

is equivalent to:

```
[sfx "i" & sfix "ij" ["cmsn","cmsa"], sfx "og" & sfix "ij" ["cmsg","cnsg"]]
```

Note the use of dots ('.') as placeholders in the partial descriptor to ensure proper positioning of the attributes in the combined descriptor string.

3.3. Word-form Generation and Lemmatization

Haskell functions for word-form generation and lemmatization are straightforward implementations of (5) and (6), respectively:

```
wf :: String -> IRule -> [(String, Msd)]
lm :: String -> IRule -> [(String, Msd)]
```

We can use this function to generatand lemmatize word-forms. For example:

```
wf "slika" r_slika ⇒
  [("slika","N=fsn"),("slika","N=fpg"),("slike","N=fsg"),("slike","N=fpn"),
  ("slike","N=fpa"),("slike","N=fpv"),("slici","N=fsd"),("slici","N=fsl"),
  ("sliki","N=fsd"),("sliki","N=fsl"),("sliku","N=fsa"),("sliko","N=fsv"),
  ("slikom","N=fsi"),("slikama","N=fpd"),("slikama","N=fpl"),("slikama","N=fpi")]

lm "slici" r_slika ⇒ [("slika","N=fsd"),("slika","N=fsl")]
```

Because word-form *slici* is (internally) homographic (*slici* may be both dative and locative singular), function `lm` returns two distinct morphosyntactic descriptors.

4. Croatian HOFM

4.1. Inflectional Rules

HOFM implementation described above is general and language-independent. Before we can use it to the inflectional rules for Croatian language, it needs to be tailored to the peculiarities of Croatian morphology. Based upon the functions listed in Table 1, we make a number of language-specific definitions:

- Specific letter groupings to ease the definition of condition functions: list of consonants (`cons`), palatals (`pals`), non-palatals (`nonpals`), velars (`velars`), and the list of all sequences of two consonants, so-called consonant groups (`cgr`). The latter is needed because the applicability of many rules of Croatian morphology depends on whether the stem ends in a consonant group or not.
- Transformation functions `sbl`, `plt`, and `jot`, as a shorthand for common phonological alternations:

  ```
  sbl = alt [("k","c"),("h","s"),("g","z")]
  plt = alt [("k","č"),("g","ž"),("h","š")]
  jot = alt [("sp","šplj"),("sb","šblj"),("sv","švlj"),("sm","šmlj"),...,("p","plj"),
             ("b","blj")("v","vlj"),("m","mlj"),("t","ć"),...,("s","š"),("c","č")]]
  ```

  Note that more specific replacements are listed first to gain precedence over more general ones.
- A transformation function `ins` for extending the stem by inserting an '*a*' at penultimate position, so that, for example:

  ```
  ins $$ "napredk" ⇒ "napredak"
  ```

To define the HOFM inflectional rules, we used a traditional grammar-book (Diklić 1979) as reference. The implementation consists of 98 inflectional rules: 49 for nouns, 32 for verbs, and 17 for adjectives. Our implementation covers most of Croatian inflectional morphology, leaving aside some rare exceptional cases. The rules for verbs are generally the most simple ones and will not be discussed here. The rules for nouns are more diverse and thus in principle they are more complex. An example of a noun rule has been given in the previous section. A slightly more complicated example is the rule `r_n42` defining, for instance, the inflection of the noun *napredak* (progress):

```
r_n42 :: IRule
r_n42 = irule "N42" "N=m"
        ends cgr
        [t               # ["sn","sa"],
         sfx "a"         # ["sg"],
         sfx "u"         # ["sd","sv","sl"],
         sfx "om"        # ["si"],
         sfx "i" & sbl   # ["pn","pv"],
         sfx "a" & t     # ["pg"],
         sfx "e"         # ["pa"],
         sfx "ima" & sbl # ["pd","pl","pi"]]
       where t = ins & alt [("tk","dk"),("sk","zk"),("šk","žk")]
```

This rule, besides using sibilarization, uses a locally defined transformation function `t` to extend the stem and alternate its ending in the nominative and accusative singular, as well as the genitive plural case. The rule's condition, `ends cgr`, could have been omitted in this case: because transformation `t` is defined only for stems ending in specific consonant groups, rule as a whole will not be applicable to any other stems. In most cases it is in fact not necessary to explicitly define a condition function.

The above examples illustrate how compact Croatian inflectional paradigm can be represented in HOFM. This is even more evident in case of more complex paradigms, such as the adjectival paradigm. The adjectival paradigm is more complex then the noun or verb paradigm in that it gives three degrees of comparison, each using a different stem. The comparative degree typically uses a suffixed or alternated stem of the positive degree, and the superlative degree additionally prefixes '*naj-*' to the stem. Moreover, most adjectives have separate inflectional patterns for indefinite and definite word-forms of the positive degree. To represent this compactly by an inflectional rule in HOFM, we first divide the paradigm into four separate inflectional patterns, defining separately the indefinite positive, the definite positive, the comparative, and the superlative degree. We then make the first two patterns (the indefinite and definite positive degree) operate on the original stem, and the other two patterns (the comparative and the superlative degree) on a modified stem. As an example, consider the rule `r_a06` defining, for instance, the inflection of adjective *star* (old):

```
r_a06 :: IRule
r_a06 = irule "A06" "Af....."
        (ends (nonpals ++ ["š"]) `land`
          (neg(ends cgr) `lor` ends ["st", "št", "rt","rn","rm"]))
        (ai01 <# ["p...n"] ++
         ad01 <# ["p...y"] ++
         ad02 <& tc <# ["c..."] ++
         ad02 <& tc <& ts <# ["s..."])
      where tc = sfx "ij"
            ts = pfx "naj"
```

This rule is applicable to all stems that ending in non-palatals or the letter '*š*', provided the stem does not end in a consonant group, except for the five groups explicitly listed. In the transformations section of the rule, we use pattern `ai01` for the indefinite positive, pattern `ad01` for the definite positive, and pattern `ad02` for both the comparative and the superlative degree. In order to modify the stem for the comparative and superlative degree, we use operator '`<&`' discussed earlier to distribute into the pattern `ad02` transformations `tc` and `ts`. For the comparative degree, we suffix '*-ij*' to the stem, while for the superlative degree we additionally prefix the stem with '*naj-*'. We use the operator '`<#`' to define the full morphosyntactic descriptor for each pattern separately. The patterns `ai01`, `ad01`, and `ad02` are defined as follows (full definitions are omitted):

```
ai01 = [nul       # ["msn","msa","msv"],
        sfx "a"   # ["msg","nsg","npn","npa","npv","fsn","fsv"],
```

```
        sfx "u"   # ["msd","msl","nsd","nsl","fsa"],...]
ad01 = [sfx "i"   # ["msn","msa","msv","mpn","mpv"],
        sfx "og"  # ["msg","nsg"],
        sfx "om"  # ["msd","msl","nsd","nsl","fsi"],...]
ad02 = [sfx "i"   # ["msn","msa","msv","mpn","mpv","nsn","nsa","nsv"],
        sfx "eg"  # ["msg","nsg"],
        sfx "em"  # ["msd","msl","nsd","nsl"],...]
```

This patterns are shared among other adjectival rules, thus allowing for a very compact representation of the Croatian adjectival paradigm. The rather short inflectional rule defined above compactly encodes as many as 229 word-forms and their morphosyntactic descriptors, all of which can be generated using the `wf` function described earlier.

4.2. Practical Difficulties

The most common difficulty that we have encountered when implementing the rules of Croatian morphology were the ones arising from informal or incomplete descriptions in the grammar-book (Diklić 1979). There are a number of cases in which the description of rule applicability refers to concepts outside the scope of morphology, such as animateness, abstractness, definiteness, etc., making impossible a precise definition of rule's condition function. A further problem is that the grammar-book turned out to be incomplete, completely omitting a number of arguably common inflectional paradigms.

A potential solution to this problem is an iterative approach to morphology modeling conducted as follows. In the first step, the grammar rules are encoded as described in a grammar-book. In the second step, the rules are verified against a corpus by attempting to lemmatize each word-form from corpus. Word-forms for which lemmatization fails are not covered by any of the defined rules and therefore indicate rule omissions. In the third step, the obtained lemmas are used to generate the corresponding word-forms. A rule that has generated word-forms that are not attested in the corpus should be considered potentially erroneous. In the fourth step, the rules would then be added and corrected, after which the process would be repeated until reaching a satisfactory level of quality.

4.3. Lemmatizing Croatian Words

The described implementation of Croatian HOFM can be used with function `lm` described earlier to lemmatize any given word-form. An example was given in Section 3.3. Another example, one that illustrates another aspect of grammar ambiguity, considers the lemmatization of word-form *starijom* (the feminine singular instrumental case of the comparative degree):

```
lm "starijom" r_a06 ⇒
  [("starijom","Afpmsnn"),("starijom","Afpmsan"),("starijom","Afpmsvn"),("star","Afcfsi")]
```

In this case, four solutions are possible, of which only the last one is correct. The ambiguity is even larger in practice because we do not know which inflectional rule to use with a given word-form, hence we have to consider each of the 98 implemented rules, in which case we end up with as many as 27 possible solutions. This illustrates the high level of ambiguity of Croatian inflectional morphology, which is exactly why automatic lemmatization of inflectionally complex languages is such a difficult task.

5. Conclusion

We have described higher-order functional morphology (HOFM), a formalism for representing inflectional concatenative morphologies, and its application to the morphologically complex Croatian language. HOFM utilizes the word-and-paradigm approach and uses inflectional rules to represent inflectional paradigms. An inflectional rule defines the word-forms obtainable from the word's stem by means of different higher-order functions, one for each type of morphological transformation. Higher-order conditional functions are used to define the applicability of each inflectional rule. This way, HOFM representations resemble the highly expressive morphology descriptions often found in traditional

grammar-books. This allows for compact and comprehensible morphology models, hopefully making the task of morphology modeling less difficult. The convenience of HOFM is perhaps best evident in case of Croatian adjectival paradigm, which, despite being rather complex and potentially generating over two hundred distinct word-forms, could be represented very compactly in HOFM.

To implement HOFM, we used Haskell, a modern functional programming language. Implementation is relatively straightforward, the only exception being the indirect treatment of transformation functions to make possible the computation of their inverses. Our implementation of Croatian HOFM can be used to generate and lemmatize Croatian words.

As part of future work, we intend to extend HOFM formalism to support optional morphological transformations, allowing for even more compact representation. We also intend to investigate how HOFM descriptions can be mapped into finite-state transducers to meet the demand for fast morphological processing. Finally, we hope to gain new insights by applying HOFM to other morphologically complex languages.

Acknowledgments

This work has been jointly supported by the Ministry of Science, Education and Sports, Republic of Croatia and Government of Flanders under the grants 036-1300646-1986 and KRO/009/06 (CADIAL).

References

Diklić, Z. 1979. *Priručna gramatika hrvatskoga književnog jezika*. Školska knjiga.

Erjavec, T.; Krstev, C.; Petkevič, V.; Simov, K.; Tadić, M. and Vitas, D. 2003. The MULTEXT-East morphosyntactic specifications for Slavic languages.

Forsberg, M. and Ranta, A. 2003. Functional morphology. In *Proceedings of the Ninth ACM SIGPLAN International Conference of Functional Programming ICFP'04*, 213–223.

Frost, R. A. 2006. Realization of natural language interfaces using lazy functional programming. *ACM Computing Surveys*, 38(4).

Hockett, C. F. 1954. Two models of grammatical description. *Word*, 10:210–234.

Jones, S. P. 2003. Haskell 98 language and libraries: The revised report. `http://www.haskell.org/definition`.

Koskenniemi, K. 1983. Two-level morphology: A general computational model for word-form recognition and production. Publications of the Department of General Linguistics, University of Helsinki.

Lopina, V. 1992. Dvorazinski opis morfonoloških smjena u pisanome hrvatskom jeziku. *Suvremena lingvistika*, 34:185–194.

Tadić, M. 1994. Računalna obrada morfologije hrvatskoga književnoga jezika, PhD dissertation.

Tadić, M. and Fulgosi, S. 2003. Building the Croatian morphological lexicon. In *Proceedings of EACL'2003*, 41–46.

Šnajder, J.; Dalbelo Bašić, B. and Tadić, M. 2008. Automatic acquisition of inflectional lexica for morphological normalisation. *Information Processing and Management*, 44(5):1720–1731.

COMPILATION OF INFLECTIONAL DICTIONARIES USING WORDEDITOR

Maria Todorova, Nikola Obreshkov

Department of Computational Linguistics, Institute of Bulgarian Language – BAS
e-mail: maria@dcl.bas.bg, nikola@dcl.bas.bg

ABSTRACT

The paper presents a tool for manual editing and generation of grammatical and inflectional types for single words and multi word expressions - WordEditor (WE) - developed at the Department of Computational Linguistics, IBL-BAS. WordEditor is designed to facilitate lexicographers' work in dictionary construction. Morphological relationships have been taken as a starting point in the design of the program. Some of its key features are language and theory independence, extendibility and user-friendliness. The resulting lexical resource allows validation of the lexical description and is accessible to external applications. The data are reusable in different NLP tasks as POS annotation, search according to lemma or grammatical characteristics, as a sub function in sense annotation, text summarization, text generation, question answering, etc.

1. Introduction

Computational morphological dictionaries are a key component in morphological processing needed in various projects involving natural language processing. The compilation of such dictionaries poses various specific problems as the boundaries of the described linguistic units and the type of grammatical information relevant for the dictionary entries may vary considerably. The experience gained in the work on a number of large-scale Bulgarian language resources at the Department of Computational linguistics (DCL)[1] such as SemCor (Koeva et al. 2006) and BulNet[2] have proved the necessity of a tool for editing, construction and exploration of morphological lexicons that handles both simple and multi-word lexical units. The formalization of the morphological patterns as attested in rich collections of linguistic data will allow the unification and enlargement of DCL's grammatical resources[3] along a clear methodological framework. In the paper we give special attention to WE's functions for incorporation and formalization of multi-word lexical units which present special interest due to their specific morphological properties. Comprehensive formal description is useful for tasks as automatic corpus processing, linguistic annotation and enriching wordnet synsets with grammatical information.

WordEditor is constructed to support lexicographers' analyses and the manual construction of inflectional paradigms. Based on a combination of rules and statistical models the program is platform-independent and works both under Linux and Windows. The tool's business logic and the data mode modules are written in Perl.

2. Dictionary format

WordEditor uses input lists of lexical units, which after manual supervision and correction, are pre-processed and two outputs are generated: an inflectional type data and a lexicon. The lexical entries in the resulting lexicon have the general structure, used in the most NLP lexical approaches: *lexical unit, formal description.*

All entries in the dictionary, separated by a new line are considered to be lexical units, respectively single or graphical words are continuous sequence of characters between blanks and MWE are "contiguous sequence of graphical words separated by blanks" (Koeva 2005), graphical words, which compose MWE, are termed MWE simple words, or MWE components (MWEC).

The information, preprocessed by means of **WE**, is attached to different input entries and encoded in the output morphological lexicon following the hierarchy of grammatical features represented in Koeva (1998, 2005): *Lexical unit – lemma; grammar type; grammar subtype; inflectional type.* Grammar types represent information about the lexical category of lemmas (noun, verb, adjective, numeral, pronoun, adverb, preposition, conjunction, particle, interjection). Grammar subtypes represent the paradigmatic characteristics corresponding to lemmas' POS (number and gender for nouns and adjectives; number, person, tense for verbs, etc.) Inflectional types represent particular classes of lemmas grouped

[1] http://dcl.bas.bg
[2] http://dcl.bas.bg/BulNet/general_en.html
[3] http://dcl.bas.bg/dictionaries_en.html

according to identical word form generation patterns. The inflected forms are produced after the inflectional set of rules (flextype) is either generated manually by the user, or the lemma is associated with an already existing set of rules. Rules generally follow the widely used formats for inflectional alternations such as DELA (Courtois and Silberztein, 1990; Silberztein 2005), and are similar to Xerox FST approaches (Beesly 2001). They are composed of a set of basic operations over characters (deletion, insertion, copy, etc.) which are applied to the lemma in the generation of the inflected forms.

3. WordEditor – functions and interface

With a view to flexibility, different modules were implemented each displayed in a different window in the user interface. The interface offers three main functions: edit, search and validate, organized in a main and validation modules. The main module is used for the processing of new words, whereas the validate module is used for editing lexical units already encoded in the dictionary.

3.1. Edit functions:

The main module has two types of edit functions since single words and MWE require different features to be edited.

3.1.1. Single words editing allows the user to add single words to the dictionary. It is automatically activated when some graphical word, including a component of a MWE is not presented in the dictionary as single word entry. A list of grammatical types defined in advance (POS and the corresponding subtype according to the morphological features associated with the lexical category), is offered to the lexicographer. On selection of a grammatical type a list of the corresponding paradigm features is displayed and the user has to enter the relevant inflected forms. The inflectional type consists of a set of rules that generate the inflected forms of lexemes belonging to the corresponding "flextype". The user is allowed to create new inflectional types or to choose among already existing flextypes. To facilitate the user WE suggests grammatical types and inflectional paradigms for single words using frequency statistics based on words already processed in the lexicon

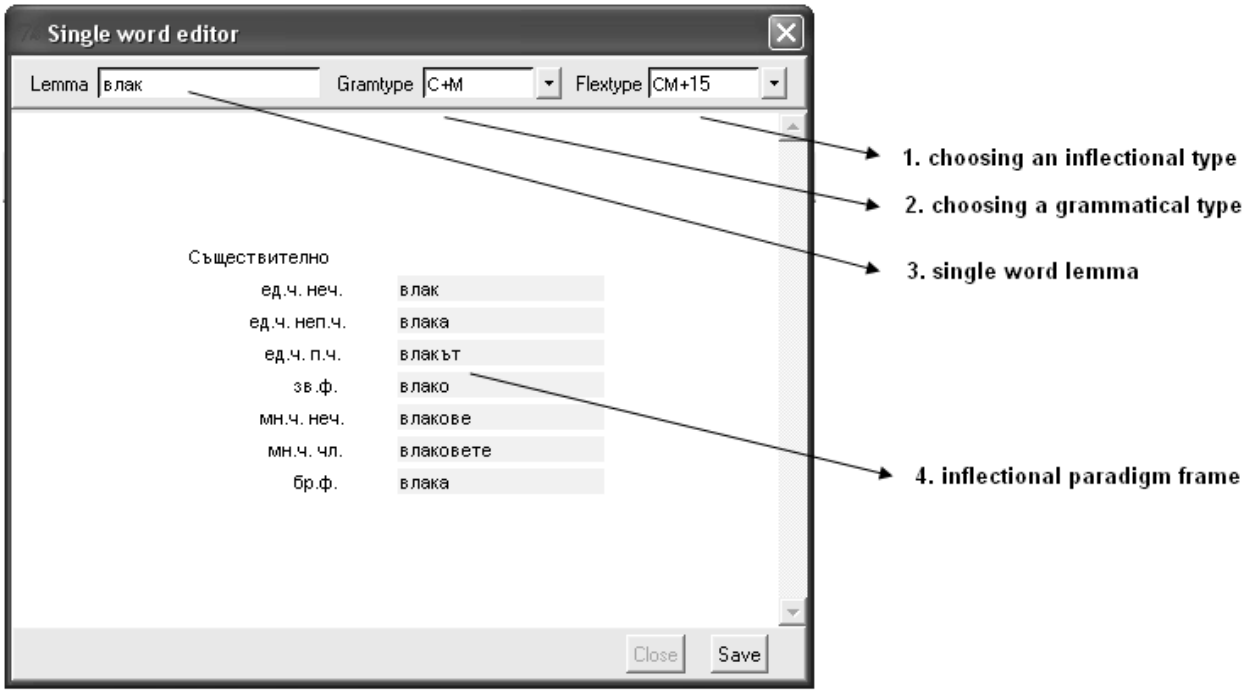


Figure 1. Single word editor

3.1.2. MWE editing allows adding multiword lexical entries to the dictionary. They are elaborated only after their components are represented as single word entries in the dictionary. The clustering of MWE into grammatical classes is based on information about MWEC and the head word. MWE look-up and editing are initiated upon selecting the head

component of a MWE and the type of combination mode among the components. The window displays a list of all possible combinations of word forms of the component words. By means of checkboxes and a hide button the relevant constraints on the paradigm of the MWE are encoded and the resulting paradigm - displayed. Space type function is defined by the individual users according to the described features. It allows the definition of some morpho-syntactic properties such as word order and modifiability.

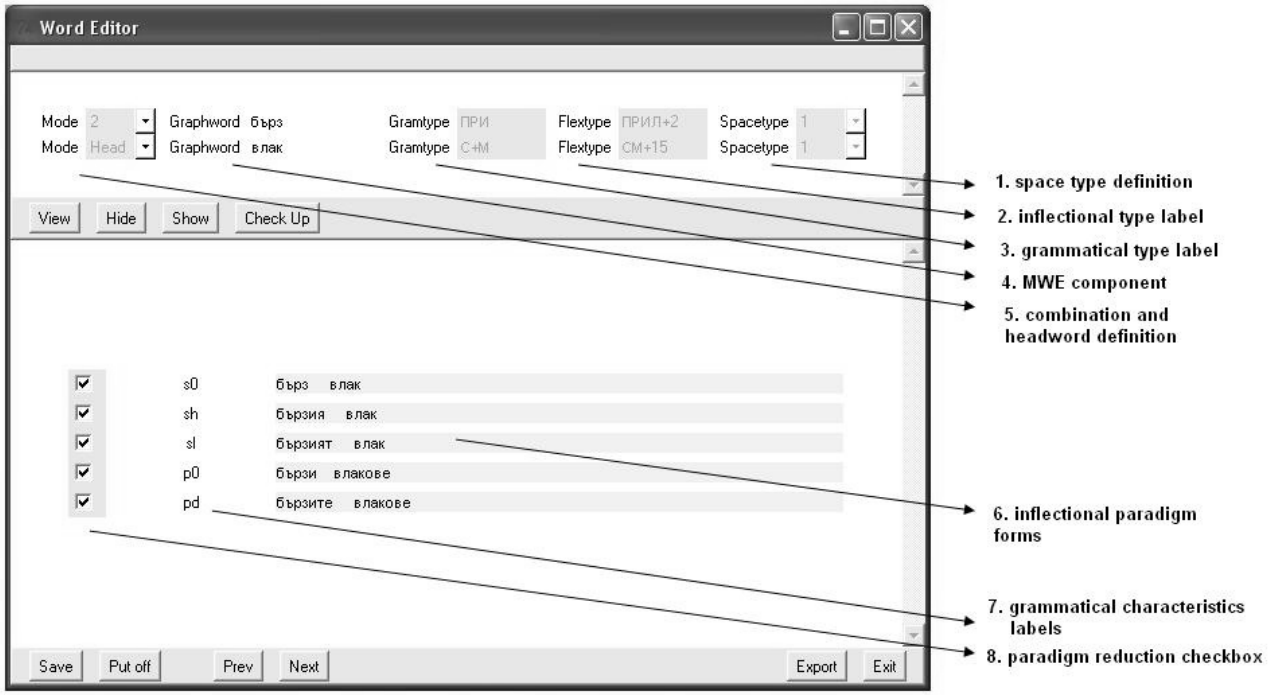


Figure 2. MWE editor

3.2. The validation module has a single edit function for both single and multi words. This module operates only over lexical entries already present in the dictionary. The editing of lexical entries is performed after activating the module's search function. Editing allows updating or adding new instances of the selected word to the dictionary, when a single lemma is associated with more than one inflectional paradigms.

3.3. Search functions, incorporated in the main and validate modules both allow the user to search entries in the constructed dictionary according to different criteria - such as grammatical type, inflectional type and the graphical form of the lemma.

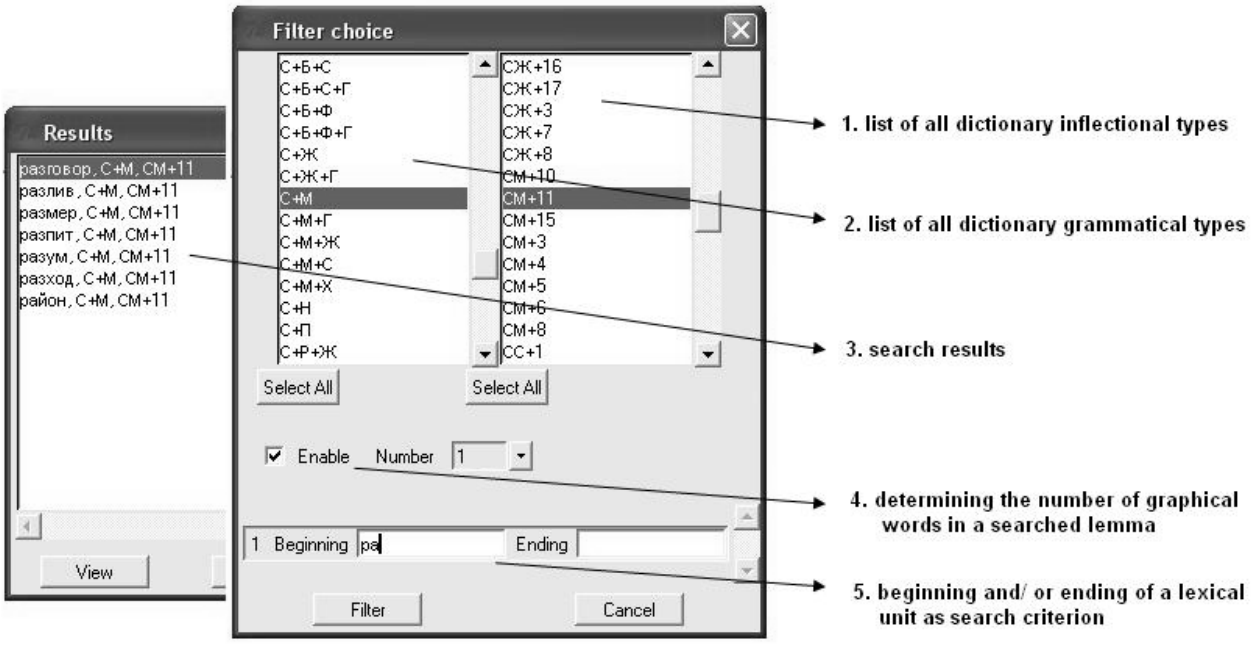


Figure 3. WE Search

Grammatical and inflectional types are listed in select boxes and any combination of them can be selected as a search criterion. Search results are displayed as lists of all lemmas present in the dictionary that satisfy the selected morphological characteristics in a separate window as shown on Figure 3. The graphical properties of lemmas are also used as a criterion for searching – in the search box a user can write a single or MWE lemma or only part of a lemma – its beginning and / or ending. If the lemma consists of several graphical words it can be search by any of them.

4. Application of WE in the treatment of morpho-syntactic anomalies in MWEs

The general class of multi-word expressions, also known as non-free expressions (Melchuk 1995), are usually defined as words with spaces or lexical units that cross word boundaries (Sag et al. 2002). MWEs present great variability ranging from closed sets which may be listed to open sets that are formed according to specific morphological, syntactic, word order, etc. properties. MWEs are usually built from the inventory of simple words, but as constituents of a particular MWE, single words are subject to changes in their paradigm and constraints with respect to some of their forms. WordEditor makes it possible to integrate a wide variety of MWE types in the morphological output database – from grammaticalized constructions such as compositional prepositions and conjunctions (*vupreki che - though)*, pronouns' full forms (*nego go – to him*), reflexive verbs (*smeya se - laugh)* through support verb constructions (vzemam uchastie - take part), noun - prepositional constructions *(s ogled na – with a view to*), to lexicalized constructions such as compounds (*kiselo mlyako - yogurt)*, compositional named entities (*Stara planina*), idioms (*hvurlyam topa – kick the bucket*) etc. The greatest advantages of the WE formalism lie in the detailed linguistic description that allows the organization of the classes of multi-word units and the provision of additional information about MWEs' internal structure including the number and word classes of constituents, paradigmatic constraints, subordination dependencies determined by the headword, etc. Following the formal grammatical features described in Koeva (2005) the MWE description in the output dictionary contains: MWE paradigmatic information - number, order and POS of components and MWE grammatical information – the type of components' paradigmatic combinations; the type of components' paradigm constraints; MWE morpho–syntactic information – the type of space. Except for the description of inflectionally compositional MWE (Savary 2008), whose morphological properties can be deduced from the respective properties of their constituents and from the head word definition function, WE's functionality allows the description of morphological anomalies (morphological irregularity (Melchuk 1995)) in paradigm combinations attested at semi-fixed and fixed MWE. Through the mode selection, paradigm reduction and space type definition functions every MWE group is further subdivided according to the particular inflectional type of its variable components and the restrictions on its paradigm and modifications if any.

4.1. Morpho-syntactic non-compositionality described by WE's MODE selection function.

The Mode selection function accounts for different types of combinations among MWE components including combinations of invariable POS and fixed components. It allows combinations among all inflected forms of a given component with all inflected forms of another, or combinations of all inflected forms of a particular MWEC with only one form of another component, or combinations of fixed components. This function resolves the formal description of such cases of morpho-syntactic non-compositionality as MWEs in which both head and non-head constituents vary or MWEs with fixed head components, where the paradigmatic information is carried by non-head constituents. Such MWEs are verb idioms with a fixed 3rd person verb head, where the paradigmatic information for person is represented by the form of a personal pronoun component, for example *broyat mi se rebrata (*literally *-ribs can be counted,* meaning *' be too slim').*

4.2. The constraints on MWE inflection paradigm forms, described by reduction paradigm design.

The possibility of reduction of some of the MWEC combinations allows the definition of formal inflectional paradigms of MWE that display formally unpredictable features as defectiveness of paradigm formation, paradigmatic constraints on the form of the head word and fixedness of a non-head constituent in a particular form. Separate MWE inflectional subtypes for example are defined when:

- a particular form of the head word does not occur in the idiomatic meaning. Such are constraints on the tense or number paradigm of verbs, etc., for example the reduction of the singular forms of both the verb and the noun in the expression *broim se na prasti (*literally *- be counted on the fingers,* meanin*g 'there are surprisingly few of').*
- a particular form in the paradigm of the head word or in a non-head constituent is formed in an irregular way. In such a case all forms except for the particular category are considered to be impossible. For example in the expression *pishi go na cheloto si* literally "*write it on your forehead"*, meaning '*demonstrate the exaggerated value of something done*', the verb "*pishi*" occurs only in the imperative form on the idiomatic reading.

4.3. Encoding of different degrees of internal morpho-syntactic variability and modifiability of MWEs by means of the space definition function.

This function allows the user to specify separator constituents (see Savary 2008) and distinct blank types between constituent words (see detailed classification in Koeva 2005) and in this way to describe different interrelations among MWECs – word order fixedness, context words insertion, insertion of modifiers (adjectives, adverbials, prepositional attributes etc.), etc. Different space types are defined for every possible type of modification where types are specified in terms of the possible number of modifiers and their lexical categories. For instance different formal MWE subtypes are defined for MWE constructions where both an adjective and an adverb modification is possible, e.g. *navlicham si (burzo/ golyama) belya na glavata (*literally *bring (fast/ a great) trouble to one's head",* meaning *'inflict (great) trouble upon oneself')* and for MWE constructions where only adverbial modification is licensed, e.g. *davam (ponyakoga) uho (*literally *"give ear (sometimes)",* meaning *'eavesdrop (sometimes)'*, etc. Another group of formal MWE subtypes is defined for MWE that disallow the occurrence of concrete words like the interrogative particle "li" or of the negative particle "ne".

5. Architecture and implementation

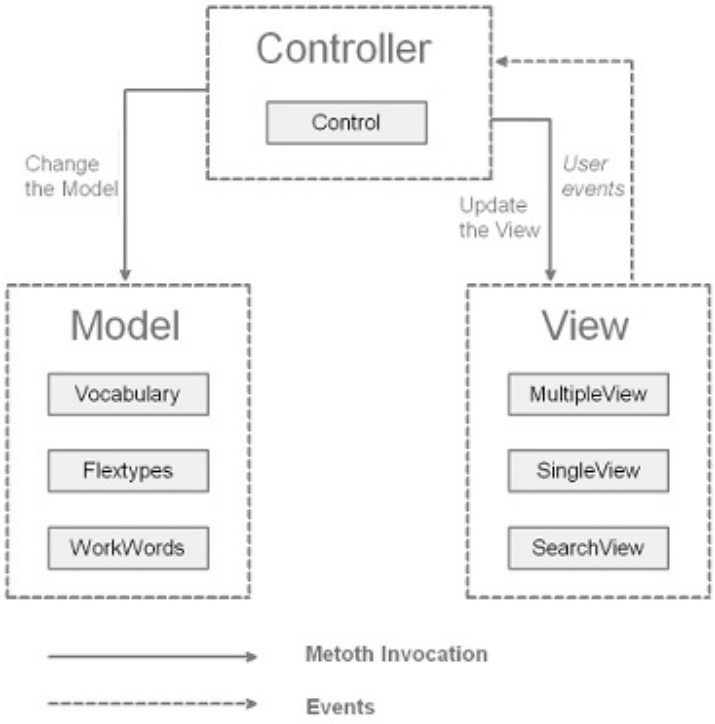
The program is underlain by the Model-View-Controller architecture pattern (first described by Trygve Reenskaug 1979)[4] which provides a flexible and easy-to-maintain system, because changes in the data model do not affect the user interface and vice versa. It is very appropriate for specific tasks and requirements as the creation of dictionaries covering wide polystratal lexical phenomena.

The user interface (UI) of the program is intended to speed up the process of the dictionary compilation and provides several visual components for that purpose. The main view of the program is created using the MultipleView object and provides

[4] http://heim.ifi.uio.no/~trygver/themes/mvc/mvc-index.html

functionalities for processing of MWE. If some of the MWECs are not present in the dictionary as single entries, the system creates a SingleView object, which provides the controls needed for adding single words to the dictionary. A Guess module suggests grammtype (grammatical paradigm subtypes) and flextype (inflectional forms) of the currently processed single word, using frequency statistics based on the endings of already processed words in the lexicon. The search function and the definition of searching criteria are provided by a SearchView object.

- **Model:** the input language resources are managed by Flextypes, Vocabulary and WorkWords modules. The Flextypes module is a collection of flextype objects where every Flextype object encodes a flextype. The Vocabulary module is a collection of Lexeme objects where every lexeme consists of a lemma, grammatical characteristics and a Flextype object. The rules defining a flextype are applied to the lemma and generate the inflectional paradigm of the lexeme. The WorkWord module contains the list of words which will be processed.
- **View:** the visualization layer consists of a SingleView module, representing single words and a MultipleView module, responsible for the visualization of MWEs. Each of them consists of a LemmaView, a ParadigmView and a FormView module.
- **Controller:** The business logic of the application is implemented in a single Control module. Events invoked by user actions are sent to the Control module by implementing



the Observer design pattern that provides the communication between the view and the model. According to the event the Control object may initiate changes in the Vocabulary and Flextypes objects and update the UI.

The business logic and the data model modules are written in Perl. The implementation of the user interface uses the Perl/Tk[5] module, which is an extension of the popular GUI toolkit Tcl. The used programming language provides platform independence of the program, successfully tested on Linux (Ubuntu, Debian, Suse) and Windows (Windows XP).

6. Conclusion

WordEditor is a lexicon-building software useful both in the creation of various large-scale morphological lexicons and in lexical analysis. It assists the manual construction of elaborate dictionaries according to a unified formal description of grammatical features. The resulting lexical resource efficiently combines detailed grammatical information for single and multiword lexical units and can be imported into different formalisms and automatically applied to texts for the purposes of identification, recognition and annotation of lexical units. It has diverse applications in various NLP tasks such as question answering, text summarization and generation, information retrieval and extraction, machine translation etc.

[5] www.perltk.org

References:

Beesly (2001).Kenneth R. Beesley and Lauri Karttunen. Finite-State Morphology: Xerox Tools and Techniques. Cambridge University Press, Cambridge.

Burbeck, S. (1987). Applications Programming in Smalltalk-80: How to Use Model-View-Controller (http://st-www.cs.uiuc.edu/users/smarch/st-docs/mvc.html).

Copestake Ann et al. Ann Copestake Villavicencio Aline, Benjamin Waldron, Fabre Lambeau. Lexical Encoding of MWEs. (http://acl.ldc.upenn.edu/acl2004/MWU/pdf/villavicencio.pdf)

Courtois, Blandine and Max Silberztein, eds. 1990. Les dictionnaires électroniques du français. In: Larousse, Langue française, vol. 87.

Gross, M. (1986) Lexicon-grammar. The representation of compound words. In: Proceedings of COLING '86 (Bonn: University of Bonn).

Koeva (1998). Grammar Dictionary of Bulgarian. Representation of Linguistic Data. In: Bulgarian Language, vol. 6: 49-58.

Koeva Sv. (2005). Infelction Morphology of Bulgarian Multiword Expressions. In: Computer Applications in Slavic Studies - Proceedings of Azbuki@net, International Conference and Workshop, Sofia, pp. 201-216.

Koeva et al. (2006). Koeva, S., S. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova. Bulgarian Tagged Corpora. In: Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages. pp. 78-86.

Melchuk (1995) Phrasemes in language and phraseology in linguistics. In: Idioms: Structural and Psychological Perspectives, chapter 8. Lawrence Erlbaum Associates

Calzolari Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Antonio Zampolli, Catherine MacLeod. Towards Best Practice for multi-word Expressions in Computational Lexicons (http://gandalf.aksis.uib.no/lrec2002/pdf/259.pdf)

Paumier S. (2006) UNITEX 1.2

Sag (2002) Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. Multiword expressions: A pain in the neck for NLP. In: Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002), Mexico City, Mexico, 2002. pp. 1-15

Savary, 2005 A. Savary, Towards a Formalism for the Computational Morphology of Multi-Word Units in Proceedings of 2nd Language & Technology Conference, ed. Zygmunt Vetulani, pp. 305-309

Silberztein M. (2005). NooJ's Dictionaries. In: Proceedings of LTC 2005, Poznan University.

Silberztein M. (2004). NooJ: A Cooperative, Object Oriented Architecture for NLP. In: Intex pour la Linguistique et le traitment automatique des langues. Cachiers de la MSH Ledoux, Presses Universitaires de Franche-Comte

Silberztein M. (1999). INTEX: a Finate State Transducer Toolbook. In: Theoretical Computer Science. 33 – 46.

CONNOTATION ANALYSIS

Dan Tufiş

Research Institute for Artificial Intelligence, Romanian Academy
13, 13 Septembrie, 050711, Bucharest
tufis@racai.ro

ABSTRACT

Language ambiguity as produced by humans, is often unnoticed and as such, is most of the times involuntary. In an original context, a sentence might be very clear with respect to the producer's intentions, but if it contains some unnoticed ambiguities (obliterated by the context), when put in another context, might convey a very different meaning, sometimes funny, sometimes embarrassing.

1. Introduction

There are various ways to model the processes of opinion mining and opinion classifications and different granularities at which these models are defined (e.g. documents vs. sentences). For instance, in reviews classification one would try to assess the overall sentiment of an opinion holder with respect to a product (positive, negative or possibly neutral). However, the document level sentiment classification is too coarse for most applications and therefore the most advanced opinion miners are considering the sentence level. At the sentence level, the typical tasks include identifying the opinionated sentences and the opinion holder, deciding whether these opinions are related or not to a topic of interest and classifications according to their polarity (positive, negative, undecided) and force.

Irrespective of the methods and algorithms (which are still in their infancy) used in subjectivity analysis, they exploit the pre-classified words and phrases as opinion or sentiment bearing lexical units. Such lexical units (also called senti-words, polar-words) are manually specified, extracted from corpora or marked-up in the lexicons such as General Inquirer or SentiWordNet (Esuli &Sebastiani, 2006) etc.

While opinionated status of a sentence is less controversial, its polarity might be rather problematic. The issue is generated by the fact that words are polysemous and the polarity of many senti-words depend on context (some times on local context, some times on global context). Apparently, bringing into discussion the notion of word sense (as SentiWordNet does) solves the problem but this is not so. We argued elsewhere (Tufiş,2008) that it is necessary to make a distinction between words intrinsically bearing a specific subjectivity/polarity and the words the polarity of which should be relationally considered. The latter case refers to the head-modifier relations; compare the different polarities of the modifier *long* in the two contexts: "the response time is long" vs. "the engine life is long".

Research in this area has been for some time monolingual, focused on English, which is "mainly explained by the availability of resources for subjectivity analysis, such as lexicons and manually labeled corpora" (Mihalcea et al., 2007). Yet, in the recent years, there are more and more languages for which required resources are developed, essentially by exploiting parallel data and multilingual lexicons. With more than 40 monolingual wordnets (see http://www.globalwordnet.org/), most of them aligned to the Princeton WordNet (Fellbaum, 1998), the recent release of SentiWordNet, and several public domain language independent tools for opinion mining and sentiment analysis, the multilingual research in opinion mining and sentiment analysis has been boosted and more and more sophisticated multilingual applications are expected in the immediate future. Such an application is briefly introduced in the next section.

2. Analysis of potential unwanted connotations

Many commercials make clever use of the language ambiguity (e.g. puns, surprising word associations, images pushing-up a desired interpretation context etc.) in promoting various products or services. Many of these short sentences, when used in regular texts, might have their connotations obliterated by the context and unnoticed by the standard reader. This observation is also valid the other way around: specific sentences, conceived in the context of a given text, when taken out of their initial context and placed in a conveniently chosen new context may convey a completely new (potentially unwanted)

message/attitude. It is relatively easy to find, especially in argumentative texts, examples of sentences which taken out of their context and maliciously used could have an adverse interpretation compared to the intended one.

The subjectivity and sentiment analysis methods are usually concerned with detecting whether a sentence is subjective and in case it is, establishing its polarity. Our approach takes a different position: we are interested in whether a given sentence, taken out of context, may have different subjective interpretations. We estimate, on a [0,1] scale, the potential of a sentence being objective (O), positively subjective (P) or negatively subjective (N), based on the senti-words in the respective sentence. Usually, these scores are uneven with one of them prevailing. We found that sentences which may have comparable subjective (positive or negative) scores are easier to use in a denotation/connotation shift game.

The SentiWordnet has been initially developed for English but the subjectivity information can be imported into any other language's wordnet which is aligned with Princeton WordNet or use it as an interlingual index. Such a wordnet, with subjectivity annotation imported from English (via the synset translation equivalence relations) will be referred to as a sentiwordnet.

3. CONAN (CONotation ANalyzer)

The CONAN system has been developed in a language independent way, and it should work for various languages, provided the analyzed texts are appropriately pre-processed and there are sentiwordnets available for the considered languages.

The necessary text preprocessing, required by CONAN includes: tokenization, tagging, lemmatization and chunking and, optionally, dependency linking. These fundamental operations for any meaningful natural language processing application have been largely described in previous publications and recently have been turned into public web-services (Tufiş et al., 2008) on our web server (https://nlp.racai.ro). Currently our linguistic web-services platform (which is based on standard web technology: SOAP/WSDL/UDDI) ensures processing for Romanian and English.

After the text is preprocessed as required, the second phase identifies all senti-words, i.e. those words which in the associated sentiwordnet (in our case the Romanian one) have at least one possible subjective interpretation (that is, their objectivity score is less than 1). There has been mentioned by various authors that the bag-of-words (BoW) approaches to subjectivity analysis is not appropriate since the subjectivity priors (the lexicon mark-up subjectivity) may be changed in context by the so-called valence shifters (Polanyi & Zaenen, 2006): intensifiers, diminishers and negations. The first two operators increase and respectively decrease the subjectivity scores (both the negative and the positive ones) while the latter complements the subjective values. As the valence shifters do not necessary act on the senti-word in their immediate proximity, the chunking pre-processing step mentioned earlier is necessary for taking care of delimiting the scope of the operators action. For instance in the sentence "He is NOT VERY *smart*", the word in italics (smart) is a (positive) senti-word, while the upper case words are valence shifters: NOT is a negation and VERY is an intensifier. The intensifier acts on the senti-word, while the negation act on the result of the intensifier: NOT(VERY(smart)). As a consequence, the sentence above has a negative subjectivity score. In (Tufiş, 2008) we showed that most wrong subjectivity mark-up existing in SentiWordNet can be explained due to a BoW approach to sense definitions analysis. The majority of synsets with wrong computed subjectivity markup have in their definitions valence shifters which apparently were ignored.

CONAN takes input either from a file or from the keyboard. In case of input from a file, CONAN expects the file to be already preprocessed and encoded the same way our linguistic web services platform provides the output (XCES format). In Figure 1 we exemplify the encoding of a sentence from the SEMCOR corpus processed by the RACAI web service platform..

```
<s id="br-a01.5.5.en">
 <w lemma="the" ana="2+,Dd" chunk="Np#1">The</w>
 <w lemma="jury" ana="1+,Ncns" chunk="Np#1" wns="ili:ENG20-07903245-n"> jury</w>
 <w lemma="say" ana="1+,Vmis" chunk="Vp#1" wns="ili:ENG20-00983145-v"> said</w>
 <w lemma="it" ana="13+,Pp3ns" chunk="Vp#2">it</w>
 <w lemma="do" ana="3+,Vais" chunk="Vp#2">did</w>
 <w lemma="find" ana="1+,Vmn" chunk="Vp#2" wns="ili:ENG20-00939971-v"> find</w>
 <w lemma="that" ana="31+,Cs">that</w> <w lemma="many" ana="22+,Pi3-p"> many</w>
 <w lemma="of" ana="5+,Sp" chunk="Pp#1">of</w>
 <w lemma="Georgia" ana="8+,Np" chunk="Pp#1,Np#2" wns="ili:ENG20-08512235-n"> Georgia</w>
 <w lemma="'s" ana="21+,St" chunk="Pp#1,Np#2">'s</w>
 <w lemma="registration" ana="1+,Ncns" chunk="Pp#1,Np#2" wns="ili:ENG20-00045146-n"> registration</w>
 <w lemma="and" ana="31+,Cc-n">and</w> <w lemma="election" ana="1+,Ncns" chunk="Np#3" wns="ili:ENG20-
    00171672-n">election</w>
```

```
 <w lemma="law" ana="1+,Ncnp" chunk="Np#3" wns="ili:ENG20-06129345-n"> laws</w>
 <c>"</c>
 <w lemma="be" ana="1+,Vmip-p" chunk="Vp#3" wns="ili:ENG20-02526983-v"> are</w>
 <w lemma="outmoded" ana="1+,Afp" chunk="Vp#3,Ap#1" wns="ili:ENG20-00931211-a"> outmoded</w>
 <w lemma="or" ana="31+,Cc-n">or</w>
 <w lemma="inadequate" ana="1+,Afp" chunk="Ap#2" wns="ili:ENG20-00054916-a"> inadequate</w>
 <w lemma="and" ana="31+,Cc-n">and</w>
 <w lemma="often" ana="14+,Rmp" chunk="Ap#3" wns="ili:ENG20-00035649-b"> often</w>
 <w lemma="ambiguous" ana="1+,Afp" chunk="Ap#3" wns="ili:ENG20-00107395-a"> ambiguous</w>
 <c>"</c>
  <c>.</c>
</s>
```

Figure 1: XCES encoding of a sentence contained in an input file for CONAN

In Figure 2 we show screenshots with input taken from a file (exemplifying the sentence in Figure 1). The left window of the lower panel displays the analysis of the sentences in the input file. The sentences are displayed in the order of the *interpretability scores* (see below). The mid window of the lower panel displays the interpretability scores of the sentences shown in the lower left window. The right window of the lower panel displays the wordnet sense IDs and definitions of the word clicked by the user in the analysis window (the left window of the lower panel). By selecting the Analysis tab in the CONAN panel the user has the options to see the various interpretations of the selected sentence: positive, negative or objective or to *force an desired interpretation: force non-negative, force non-positive, force non-objective, force most-negative, force most-positive, force most-objective.* (see Figure 2).

In case of keyboard input, the raw text (one or more sentences) is preprocessed via our linguistic web services platform and the rest of the interaction is the same as described above.

The algorithm computes the subjectivity scores of the grammatical chunks (taking into account the subjectivity operators and their scopes) summing their values to get the score for each sentence.

The user can ask for the most *objective* interpretation (the considered senses for the senti-words are the ones with the highest objective scores), the most *positive subjective* interpretation (the considered senses for the senti-words are the ones with the highest subjective scores), the most *negative subjective* interpretation (the considered senses for the senti-words are the ones with the highest negative scores) or all the three interpretations.

When the user is asking for *forced* interpretations (be it negative, positive or objective) CONAN replaces the words in the current analyzed sentence with synonyms which have less interpretative scores then the current ones. The rationale is to help the user in avoiding words which could be interpreted in different ways, by synonyms which have less (ideally no) connotation variability.

We define the *interpretability score* (IS) as a quantitative measure of the potential for connotation shift of a sentence. It is computed as a normalized sum of the *interpretability scores* of the senti-words (sw) of the considered sentence as described in the equations below:

IS (sentence) = $\sum_{k=1}^{|senti-words|} IS(sw_k)$ with |senti-words| representing the number of the senti-words in the current sentence

IS(sw$_k$) = $\frac{0.5 * (\max P(sw_k) + \max N(sw_k))}{1 + |\max P(sw_k) - \max N(sw_k)|}$ with *max P(sw$_k$)* and *max N(sw$_k$)* representing the highest positive and negative scores among the senses of *sw$_k$* senti-word.

The rationale for this empirical formula is that when a senti-word has one sense highly positive and another one highly negative and these values are comparable, the respective word is a major constituent for a possible connotation shift of the sentence in which it appeared. The interpretability score of a senti-word is maximum (1) when it has one exclusively positive sense (P(sw$_k$)=1) and another sense which is exclusively negative (N(sw$_k$)=1). For the current SentiWordNet annotations, the senti-words with the highest interpretability score (IS= 0,875) are *pretty, immoral* and *gross*.

Figure 2: CONAN in action. Lower panel**:** main screen with the connotation analysis; Upper left panel**:** the most positive paraphrasing; Upper right panel**:** the most negative paraphrasing

The valence shifters (intensifiers, diminishers and negations) are specified in three external text files (user editable) which are read-in each time CONAN is launched. Currently all the valence shifters are uniformly dealt with, irrespective of the arguments they take: the intensifiers and diminishers increase or decrease with 20% the score of their argument (senti-word or *senti-group-phrase*) while the negations switch the P/N scores of their arguments. Therefore, the valence shifter files are simple lists of words. A more elaborated approach (under development) will specify for each valence shifter, its grammar category its sense number (if necessary) and preferred argument-type as well as an argument-sensitive valence shifting function .

Concerning the valence shifters, it is interesting to note that, in general, translation equivalence preserves their type distinctions. However this is not always true. For instance, in Romanian *destul* (either adjective or adverb), when followed by the preposition *de*, is arguably a diminisher. In English, its translation equivalent *enough* acts more like an intensifier than as a diminisher[1].

References

Esuli A.& F. Sebastiani. 2006. SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. Proceedings of the 5th Conference on Language Resources and Evaluation LREC-06, Genoa, Italy, pp. 417-422

Fellbaum C., editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Mihalcea R.; Banea C.; Wiebe J. 2007. Learning Multilingual Subjective Language via Cross-Lingual Projections. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics,.Prague, Czech Republic, June, pp. 976-983

Polanyi L. & Zaenen A. 2006. Contextual Valence Shifters. In James G. Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing Attitude and Affect in Text: Theory and Application.* Springer Verlag.

Tufiş D.; Ion R.; Ceauşu A.; Ştefănescu D. 2008. RACAI's Linguistic Web Services. In Proceedings of 6th Conference on Language Resources and Evaluation LREC-08, Marrakech, Marocco

Tufiş D. 2008. Subjectivity mark-up in WordNet: does it work cross-lingually? A case study on Romanian Wordnet. Invited talk on the Panel "Wordnet Relations" at the Global WordNet Conference, January 22-25, 2008

[1] This comment was made by an anonymous reviewer, gratefully acknowledged here.

BUILDING LANGUAGE RESOURCES AND TRANSLATION MODELS FOR MACHINE TRANSLATION FOCUSED ON SOUTH SLAVIC AND BALKAN LANGUAGES

Dan Tufiş[1], Svetla Koeva[2], Tomaž Erjavec[3], Maria Gavrilidou[4], Cvetana Krstev[5]

[1]Research Institute for Artificial Intelligence, Romanian Academy, 13, Calea 13 Septembrie, 050711, Bucharest, Romania, tufis@racai.ro
[2]Institute for Bulgarian Language, Bulgarian Academy, 52, Shipchenski prohod, 1113, Sofia Bulgaria, svetla@idcl.bas.bg
[3]Jožef Stefan Institute, Jamova cesta 39, SI-1000, Ljubljana, Slovenia, tomaz.erjavec@ijs.si
[4]Institute for Language and Speech Processing, 6, Artemidos, GR15125, Marousi, Greece, maria@ilsp.gr
[5]University of Belgrade, 16, Studentski trg, 11000, Belgrade, Serbia, cvetana@poincare.matf.bg.ac.yu

ABSTRACT

The paper presents the results of a small and short-term SEE-ERA.net project the purpose of which was to investigate the feasibility of machine translation (MT) research and development for several South Slavic and Balkan languages. For these languages MT systems are scarce and for some of them even non-existent. We argue that by investing efforts in building appropriate language resources, the current technology can be successfully used for a quick development of acceptable MT prototypes, easy to further extend to working systems. The paper describes the parallel corpus compiled in the scope of the project, concentrating on its composition, format, and linguistic analysis. Word-alignments automatically derived from the annotated parallel corpus are also discussed. The paper concludes with direction for further work.

Introduction

Since the seminal work of the IBM group in statistical word-based translation (Brown et al., 1993), new methodologies (memory-based, phrased-based, syntax-based etc.) and techniques (reification, factorization) emerged in multilingual data-driven approaches to machine translation. Yet, several studies underlined the idea that the quality of data to be fed into any machine learning system is of a crucial importance and cannot be compensated by using mass raw multilingual data. In spite of numerous attempts to construct MT systems entirely based on raw parallel data, the evaluations showed that although useful and encouraging results can be obtained in a short period of time, the translation quality can hardly be further improved by increasing the volume of data. The ongoing EuroMatrix project[1] started from this finding and adopted a very promising hybrid approach, combining the strength of rule-based and statistical machine translation and exploiting more and more linguistic knowledge[2]. The Factored Translation Models (Koehn & Hoang, 2007) allow for exploiting, where available, different levels of linguistic pre-processing: lemmatization, part-of-speech tagging, chunking, parsing, word-sense disambiguation, etc. For most of European languages there exist already tools for ensuring the basic pre-processing steps required for a factored translation approach. In fact, with current MT technologies (Och & Ney 2000; 2003; Koehn et al. 2007) which, to a large extent, are language independent, the development of large enough and high quality training data became the critical part of an MT development project.

In this paper we present some results of a small and short-term SEE-ERA.net project[3], the main objective of which was to provide necessary linguistic and technological resources that will foster machine translation RTD for South Slavic and Balkan languages. The partners in the project were from Bulgaria, Greece, Romania, Serbia and Slovenia. Some partners harmonized the objectives of this project with the objectives of other local or bilateral running projects and the project thus includes Czech, French and German as additional languages. Although the project officially ended in July 2008, we hope that this preparatory phase will be followed by another concerted action for further enhancing and exploiting the multilingual resources that have been created.

[1] http://www.euromatrix.net/
[2] For a nice demo with Moses MT, which is the basis for the EuroMatrix MT development, see http://demo.statmt.org/webtrans/.
[3] http://dcl.bas.bg/ssbc/home.html

The Multilingual Data

The Acquis Communautaire is the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of EU Member States. Thus, the Acquis Communautaire is a collection of parallel texts in the following 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene and Swedish. A significant part of these parallel texts have been compiled by the Language Technology group of the European Commission's Joint Research Centre at Ispra into an aligned parallel corpus, called JRC-Acquis (Steinberger et al. 2006), publicly released in May 2006. In November 2007, the European Commission's Directorate General for Translation (DGT) and the Joint Research Centre (JRC) have made available a multilingual Translation Memory (DGT-TM) of the Acquis Communautaire in the above mentioned official European Union languages. These unique language resources[4] are among the few available parallel corpora containing the languages we were interested in: Bulgarian, Greek, Romanian, Slovene plus Czech, English, French, and German (called further SEE-ERA.net Administrative Corpus - SEnAC). This resource does not yet exist for Serbian, and for that reason an additional resource, based on Verne's novel "Around the world in 80 days"[5] (called further SEE-ERA.net Literary Corpus - SEnLC), has been compiled**.**

SEnAC Corpus Construction and Encoding

From the entire JRC-Acquis, which uses the same identifiers (Celex numbers) for the same documents (trailed with the language code); we selected all the documents available in all our target languages. This resulted in a list of 1204 files per language. Since we have noticed several errors in the sentence alignments of the original JRC-Acquis corpus, we re-aligned the 1204 files for Bulgarian, Czech, French, Greek, German, Romanian, Slovenian against the corresponding files in English, using RACAI's SVM sentence aligner (Ceauşu et al. 2006). From the XX-EN aligned sentences, we retained only the 1-1 alignment pairs (more than 99% on average of the total alignments) and each partner had the responsibility to check and correct, if necessary, the sentence alignment. We are not aware of any alignment error for the retained 1-1 XX-EN sentences[6]. Finally we merged the alignments into one XML document, containing 60,389 translation units, each containing one sentence translated in 9 languages, as exemplified in Figure 1.

```
<tu id="3936">
  <seg lang="bg">
     <s id="31985L0337.n.83.1">
        Резултатите от консултациите и информацията , събрана съгласно членове 5,6 и 7 , трябва да
        се вземат предвид при процедурата по издаването на разрешението.</s></seg>
  <seg lang="cs">
     <s id="31985L0337.n.83.1">
        Informace shromážděné podle článků 5 , 6 a 7 musí být brány v úvahu v povolovacím řízení.</s> </seg>
  <seg lang="de">
     <s id="31985L0337.n.85.1">
        Die gemäß den Artikeln 5 , 6 und 7 eingeholten Angaben sind im Rahmen des
        Genehmigungsverfahrens zu berücksichtigen.</s></seg>
  <seg lang="el">
     <s id="31985L0337.n.85.1">
        Οι πληροφορίες που συγκεντρώνονται δυνάμει των άρθρων 5 , 6 και 7 πρέπει να λαμβάνονται υπόψη
        στα πλαίσια της διαδικασίας για τη χορήγηση αδείας.</s></seg>
  <seg lang="en">
     <s id="31985L0337.n.84.1">
        Information gathered pursuant to Articles 5 , 6 and 7 must be taken into consideration in the
        development consent procedure.</s></seg>
  <seg lang="fr">
```

[4] http://langtech.jrc.it/JRC-Acquis.html

[5] Since this novel has been translated world-wide, we think it is a good candidate for building a parallel corpus for languages which are not yet included into JRC-Acquis corpus.

[6] The sentence aligner took advantage of the specific structure of the corpus, and besides the usual sentence delimiters (period, semi-colon, exclamation mark, etc.) we took into account the hard line breaks. This is why, besides proper sentences, the alignments contain pairs of section titles (e.g. ("Article 1" ; Articolul 1), or pairs of dates or locations.

```
    <s id="31985L0337.n.83.1">
        Les informations recueillies conformément aux articles 5 , 6 et 7 doivent être prises en considération
        dans le cadre de la procédure d'autorisation.</s></seg>
  <seg lang="ro">
    <s id="31985L0337.n.83.1">
        Informațiile culese conform art. 5 , 6 şi 7 trebuie să fie luate în considerare în cadrul procedurii de
        autorizare.</s></seg>
  <seg lang="sl">
    <s id="31985L0337.n.83.1">
        Informacije , zbrane skladno s členi 5 , 6 in 7 , se morajo upoštevati v postopku za pridobitev soglasja
        za izvedbo.</s></seg>
</tu>
```

Figure 1: A translation unit from the 9-language parallel corpus

This corpus was then tokenized, tagged and lemmatized by each partner. The tagsets used for all languages (except Bulgarian and German) were compliant with the MULTEXT specifications, for the most part with the MULTEXT-East specifications Version 3[7] (Erjavec 2004) (see http://nl.ijs.si/ME/V3/msd/). The Table 1 shows some statistics concerning the result of the pre-processed corpus:

Language	No. of tokens	Avg no. of tokens/sentence
BG	1436925	23.79
CS	1238981	20.51
DE	1314441	21.76
EL	1469642	24.33
EN	1466912	24.29
FR	1527241	25.29
RO	1422995	23.56
SL	1271011	21.04

Table 1: Statistical data on the compiled parallel corpus

After tokenization, tagging and lemmatization, this annotation was added to the XML encoding of the parallel corpus. Depending on the available processing tools for different languages, additional information could be added to each language-specific segment of a translation unit. Figure 2 shows the representation of the Romanian segment of the translation unit displayed in Figure 1.

```
<tu id="3936">
...
        <seg lang="ro">
                <s id="31985L0337.n.83.1">
                        <w lemma="informație" ana="Ncfpry">Informațiile</w>
                        <w lemma="culege" ana="Vmp--pf">culese</w>
                        <w lemma="conform" ana="Spsd">conform</w>
                        <w lemma="art." ana="Yn">art.</w>
                        <w lemma="5" ana="Mc">5</w>
                        <c>,</c><w lemma="6" ana="Mc">6</w>
                        <w lemma="şi" ana="Crssp">şi</w>
                        <w lemma="7" ana="Mc">7</w>
                        <w lemma="trebui" ana="Vmip3s">trebuie</w>
                        <w lemma="să" ana="Qs">să</w>
```

[7] http://nl.ijs.si/ME/V3/msd/

```
                    <w lemma="fi" ana="Vasp3">fie</w>
                    <w lemma="lua" ana="Vmp--pf">luate</w>
                    <w lemma="în" ana="Spsa">în</w>
                    <w lemma="considerare" ana="Ncfsrn">considerare</w>
                    <w lemma="în_cadrul" ana="Spcg">în_cadrul</w>
                    <w lemma="procedură" ana="Ncfsoy">procedurii</w>
                    <w lemma="de" ana="Spsa">de</w>
                    <w lemma="autorizare" ana="Ncfsrn">autorizare</w>
                    <c>.</c>
                </s></seg>
...
</tu>
```

Figure 2: A linguistically analysed sentence in a language-specific segment of a translation unit

SEnLC Corpus Construction and Encoding

One reason that we have chosen Jules Verne's novel is that this text is available in digital form for many of the languages that we were interested in. Moreover, for the majority of these languages lexical resources exist in the same format, which enables comparable processing of the text in different languages. Translation of the novel in sixteen languages have been acquired, namely: French, English, German, Spanish, Portuguese, Italian, Romanian, Russian, Serbian, Croatian, Bulgarian, Macedonian, Polish, Slovenian, Hungarian and Greek. Not all of these texts have yet been aligned; alignment was done for the five Balkan languages, French original and English.

In the preparatory phase each translation was marked in accordance with the TEI-standard in XML, and the title (<head>), paragraph (<p>) and "sentence" (<seg>) were included as units of text logical layout. Before alignment, each text was transformed to the TEI-conformant format[8]. The XAlign system[9] was used for the alignment process. Starting from the French version, the goal of the alignment was to establish 1:1 relations on the segment level (<seg> tag) with all other languages. In order to achieve this goal segments had to be further divided. So, the total number of segments in all texts is 4409. This type of text alignment of bitexts required an intensive manual control of the output of the XAlign system. In this way, the missing segments or the inconsistencies between the source text and its translations were also identified.

```
<tu id="n569">
<seg lang="fr">
    <s id="Verne80days.n569">
        Vous savez que cette formalité du visa est inutile, et que nous n'exigeons plus la présentation du passeport?</s></seg>
<seg lang="sr">
    <s id="Verne80days.n569">
        Vi znate da je ova formalnost viziranja izlišna i da se više ne traži pokazivanje isprava?</s> </seg>
<seg lang="bg">
    <s id="Verne80days.n569">
        Знаете ли, че тази формалност с паспортите е безполезна и че ние вече не изискваме да представяте паспортите
        си?</s></seg>
<seg lang="en">
    <s id=" Verne80days.n569">
        You know that a visa is useless, and that no passport is required?</s></seg>
<seg lang="el">
    <s id="Verne80days.n569">
        Ξέρετε ότι αυτή η τυπική διαδικασία της βίζας δεν είναι αναγκαία και δεν απαιτείται πλέον η εμφάνιση του
        διαβατηρίου;</s></seg>
<seg lang="sl">
```

8 http://www.tei-c.org/index.xml

9 http://led.loria.fr/download/source/Xalign.zip

```
<s id="Verne80days.n569">
        Ali vam je znano, da je ta formalnost vidiranja nepotrebna in da ne zahtevamo več predložitve potnega lista ?</s></seg>
<seg lang="ro">
    <s id=" Verne80days.n569">
        tiți că formalitatea vizei e inutilă şi că noi nu mai cerem prezentarea paşaportului.</s>
```

Figure 3: One sentence from a 7-language corpus of Verne's novel: French original, English as a hub language, and five South Slavic and Balkan languages

The total number tokens in this text in French is 71,793, while the total number of unique tokens (types) is 9,433 (ratio 7.6). The figures for other languages are different, e.g. for Serbian the total number of tokens is 58,722, while the total number of types is 12,733 (ratio 4.6), for Bulgarian the total number of tokens is 58,678, while the total number of types is 11,217 (ratio 5.2), while for Greek the total number tokens is 68,615, and the total number of types is 11,809 (ratio 5.8).

For the present, all the language versions of this corpus for which DELA type lexical resources exist were tagged, but disambiguation has been done for Serbian and Bulgarian. The initial tagsets were those used in the corresponding lexical resources, but they were latter mapped into MULTEXT-East specifications (Krstev et al. 2004). After tagging and lemmatization, this annotation information was added to the XML encoding of the parallel corpus. Figure 4 shows the representation of the Serbian segment of the translation unit displayed in Figure 3:

```
<tu id="n569">
    <seg lang="sr">
        <s id="Verne80days.n569">
            <w lemma="vi" ana="Pp2-pn">Vi</w>
            <w lemma="znati" ana="Vm-p2p-an-n---p">znate</w>
            <w lemma="da" ana="C-s">da</w>
            <w lemma="jesam" ana="Va-p3s-an-y---p">je</w>
            <w lemma="ovaj" ana="Pd-fsn">ova</w>
            <w lemma="formalnost" ana="Ncfsn--n">formalnost</w>
            <w lemma="viziranje" ana="Ncnsg--n">viziranja</w>
            <w lemma="izlišan" ana="Afpfs1">izlišna</w>
            <w lemma="i" ana="C-s">i</w>
            <w lemma="da" ana="C-s">da</w>
            <w lemma="se" ana="Q-">se</w>
            <w lemma="više" ana="Rgp">više</w>
            <w lemma="ne" ana="Q-">ne</w>
            <w lemma="tražiti" ana="Vm-p3s-an-n---p">traži</w>
            <w lemma="pokazivanje" ana="Ncnsn--n">pokazivanje</w>
            <w lemma="isprava" ana="Ncfpg--n">isprava</w>
            <c>?</c>
        </s></seg>
</tu>
```

Figure 4: A tagged and a lemmatized sentence from the Serbian version of Verne's novel

Word-Alignment of the SEnC

Based on the pre-processing discussed in the previous section, we built, using GIZA++ (Och & Ney 2003) 8 unidirectional translation models (EN-RO, RO-EN, EN-BG, BG-EN, EN-SL, SL-EN, EN-EL, EL-EN). The processing unit considered in each language was not the wordform but the string formed by its lemma and the first two characters of the associated morphosyntactic tag (e.g. for the wordform "informațiile" we took the item "informație/Nc"). We used for each language 20 iterations (5 for Model 1, 5 for HMM, 1 for THTo3, 4 for Model3, 1 for T2To4 and 4 for Model4). We did not include Model 5 nor Model 6 as we noticed a degradation of the perplexities. Given the formulaic language used by the Acquis-Communautaire documents, the perplexities of the resulting language models were encouraging, and range from 13.07 (RO-EN) to 19.88 (EN-BG). Based on these models we word-aligned the bitexts using the iterative high precision COWAL aligner (Tufiş et al. 2006). As described in (Tufiş et al. 2006), translation pairs prescribed by each unidirectional translation model were unconditionally included in the alignment skeleton. The rest of the links were established in the subsequent iterations of

the aligner. The training corpora SEnC, the alignments and the perplexities for each translation model are available on the project's site. Additionally, an alignment viewer and editor (see Figure 5) was implemented for allowing the visualization and correction of the alignments with the purpose of further fine-tuning the translation models.

At the time of this writing, the translation model for the RO and EN has been used for some preliminary experiments in translation. The results are very encouraging and in Table 2 we provide a translation example (from Romanian into English), as produced by our translation system prototype and by Google's MT. The input text is the following:

articolul 1
utilizând vehicule înmatriculate pe teritoriul unei părți contractante , care datorită construcției şi echipamentelor lor sunt adecvate pentru a transporta mai mult de nouă persoane , inclusiv conducătorul auto , şi sunt destinate acelui scop ; în înțelesul prezentului acord , serviciile internaționale înseamnă serviciile care tranzitează teritoriul a cel puțin două părți contractante . în înțelesul prezentului acord , termenul teritoriul unei părți contractante , acoperă , din punctul de vedere al comunității economice europene , acele teritorii unde se aplică tratatul de instituire a comunității respective şi în condițiile stabilite în tratatul respectiv .

Our translation	Google translation
article 1 using vehicles registered in the territory of a contracting party , which because of the construction and their equipment are appropriate to carry more than nine persons , including the driver , and are intended for that purpose ; within the meaning of this agreement , the services of the international means services in transit through the territory of at least two contracting parties . within the meaning of this agreement , the term the territory of a contracting party , shall cover , from the point of view of the european economic community , those territories which shall apply to the treaty establishing the community in question and under the conditions laid down in the treaty in question .	Article 1 using vehicles registered in the territory of a contracting party, which due to their construction and equipment are adequate to carry more than nine persons including the driver, and are intended for that purpose; the meaning of this Agreement, international services means services transiting the territory of at least two contracting parties. the meaning of this Agreement, the term territory of a contracting party, cover, in terms of European economic community, those areas where applicable treaty establishing the community and the conditions laid down in the treaty.

Table 2: An example of our translation using the SEnC RO-EN translation model vs. Google's MT

Figure 5: The alignment of a translation unit (no. 69) in four bitexts

Future work

The presented work is still in progress. We plan to conduct experiments with the other models for translating from the XX language into English. The other direction (EN-XX) will be also considered. We plan to extend the experiments for other language pairs present in SEnC corpus It is obvious that due to the same level of annotation we could experiment with any of the XX-YY language pair in SEnC corpus, but we are equally interested in studying the effect on translation models building by using word and phrase alignments derived from a pivot/hub language alignment. We described in (Tufiş & Koeva 2007) a system for automatically deriving from pivot alignments PIVOT-X1 and PIVOT-X2 word alignments the X1-X2 alignment.

Future work will address the more challenging task on building translation models from the SEnLC literary corpus and provided adequate data will be available, experiments with other South East and Balkan languages.

References

Brown et al. 1993. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, Robert J. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2): 263–311.

Ceauşu et al. 2006. Alexandru Ceauşu, Dan Ştefănescu, Dan Tufiş. 2006. Acquis Communautaire sentence alignment using Support Vector Machines. In *Proceedings of the 5th LREC Conference*, Genoa, Italy.

Erjavec 2004. Erjavec, T. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation*, LREC'04, pp. 1535 - 1538, ELRA, Paris.

Krstev et al. 2004. Cvetana Krstev, Duško Vitas, Tomaž Erjavec. Morpho-Syntactic Descriptions in MULTEXT-East - the Case of Serbian, in Informatica, No. 28, pp. 431-436, The Slovene Society Informatika, Ljubljana.

Koehn & Hoang 2007. Koehn Philipp, and Hieu Hoang. *Factored Translation Models*. EMNLP.

Koehn et al. 2007. Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.

Och & Ney 2000 2000. Franz J. Och, Herman Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Conference of ACL*, Hong Kong: 440-447.

Och & Ney 2003. Franz J. Och, Herman Ney. A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, 29(1):19-51

Steinberger et al. 2006. Ralph Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiş.. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147

Tufiş et al. 2006. Dan Tufiş, Radu Ion, Alexandru Ceauşu, Dan Ştefănescu. Improved Lexical Alignment by Combining Multiple Reified Alignments. In Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006), Trento, Italy, 3-7 April, 2006, pp. 153-160.

Tufiş & Koeva 2007. Dan Tufiş, Svetla Koeva: "Ontology-supported Text Classification based on Cross-lingual Word Sense Disambiguation". In Francesco Masulli, Sushmita Mitra and Gabriella Pasi (eds.). Applications of Fuzzy Sets Theory. 7th International Workshop on Fuzzy Logic and Applications, WILF 2007, Camogli, Italy", LNAI 4578, Springer-Verlag Berlin Heidelberg, 2007, pp. 447-455

BARE PLURALS IN ALBANIAN

Etleva Vocaj
Université du Québec à Montréal
evi_vocaj@yahoo.fr

ABSTRACT

Unlike English Bare Plurals, Albanian Bare Plurals (BPs) only have an existential reading. To create a Kind or Generic reading, Albanian uses definite nominal expressions. We present a new analysis of little known Albanian data that shows that the behaviour of BPs and their variation observed in languages such as English and French are not due to the fact that they are NP constituents when interpreted existentially and DP constituents when interpreted generically (Kallulli, 1999), which constrains their clausal distribution and interpretation. We show that the particularities of Albanian and the observed variation are due to the Number that BPs express. The plurality expressed by a BP in Albanian is not realized as a 'sum', but as a 'cluster'—a unique plural individual, an indivisible whole with an opaque structure that disallows any access to its parts. The BP may not refer to the totality of individuals that have the property denoted by the noun (the 'nominalization' of the property (Chierchia, 1998)), reference that is required in order to create a Kind or Generic reading. It may only indicate that its referent contains more than one element, which allows them to have a weak indefinite interpretation.

0. Introduction

In both singular and plural form, the Albanian noun exhibits a three-way formal opposition. It may appear bare, or appear with either a definite article or an element like *një 'one', ca/disa* 'some'.

(1)

libër	*libr -i*	*një libër*
book	book-the-sg.	a book
libr -a	*libr -a t*	*ca libr-a*
book-pl.	book-pl the-pl'	some book-pl

The presence of *ca/disa* is considered optional (2) and traditional grammars state that Albanian has no grammatically-realized form for the plural indefinite article, despite there being a singular indefinite determiner *një.*

(2) Në pritje ishin të pranishëm edhe (ca) aktorë.
in receiving were present too some actor
'In the receiving were present some actors too.'

A number of tests, including those employed by Carlson (1980), allow us to conclude that BPs in Albanian are not perfectly synonymous with plural noun phrases containing *ca/disa* 'some' (NIP).

-BPs are not ambiguous in contexts containing propositional attitudes. They may only be interpreted opaquely ('*de dicto*'), as in (4a). However, propositional attitude contexts create ambiguities concerning the interpretation of NIPs (4b). In addition to the 'de dicto' reading, they may also be interpreted 'de re' ('there are many orthopedists that Anna wants to consult').

(4) a. *Ana dëshiron të konsultojë ortopedist -ë.* (opaque reading) / *(transparent reading)
Ana want consult-subj.3.sg. orthopedist-m. pl.
'Anne wants to consult orthopedists'
b. *Ana dëshiron të konsultojë disa ortopedist -ë.* (opaque reading) / (transparent reading)
Ana want consult-subj.3.sg. some orthopedist-m. pl.
'Anne wants to consult some orthopedists'

- The differences between BPs and NIPs are clearly visible in anaphoric contexts. BPs may not serve as the antecedent to *të tjerë(-t)* anaphors (5a). In sentences with resumptive pronouns, coreference is possible with an NIP, but not with a BP (5b vs 5b').

(5) a. *Ana më rekomandoi *(disa) libr - a dhe Savi me rekomandoi të tjerë -(t).*
Anne me recomended some book pl. and Savi me recommended others the-pl.
'Anne recommended me some books and Savi others / the others.'
b. *Eri pranon në provim* ***disa student -a***$_{i}$ *pa librezë, kurse Adri* ***i***$_{i}$ *përzë.*
Eri accept in exam some student pl. without notebook while Adri cl.-3.pl. expel
'Eri accept in the exam some students without notebook while Adri expel them.'b. *Eri pranon në provim* ***student -a***$_{i}$ *pa librezë, kurse Adri* ***i***$_{*i}$ *përzë.*
Eri accept in exam student pl. without notebook while Adri cl.-3.pl. expel
'Eri accept in the exam some students without notebook while Adri expel them.'

- Another difference between these two types of nominals is the aspectual effect that BPs have on the verb. With a BP argument, predicates denoting accomplishments and achievements acquire the characteristics of activities (6a). Also, non-durative predicates acquire a durative interpretation, when used with a BP.

(6) a. *Bruna shkroi artikuj *brenda një viti / gjatë një viti.*
Bruna wrote paper-pl in one wear during one wear
'Bruna wrote papers during one wear.'
b. *Kaja vazhdoi të therë pul -a gjithë fundjavën.*
Kaja continued to butcher-3.sg.subj. hen pl. all weak-end
'Kaja continued to butcher hens all over the weak-end.'

Thus, in many respects, Albanian BPs resemble English BPs. However, the distinction between generic BPs and existential BPs does not hold in Albanian. Albanian BPs may only be interpreted existentially (7a). The use of the definite article is necessary for a generic interpretation. (7b,c).

(7) a. *Adi bleu çokollat -a për Silvin.*
Adi bought chocolate pl. for Silvi
'Adi bought chocolates for Silvi.'
b. *Arinj -*(të) janë në rrezik zhdukje.*
bear-pl. the-pl. are on the verge of extinction
'Bears are on the verge of extinction.'
c. *Arinj -*(të) janë gjitarë.*
bear-pl. the-pl. are mammals
'Bears are mammals.'

The goal of our talk is to show that the particularities of BPs in Albanian, as well as the variation observed with English, are due to the role played by the type of referent that grammatical number allows BPs to pick out in each of these languages. To this end, we will show that number marking on the noun is not simply a morphological marker, but rather that it plays a semantic role. We will then examine the type of plurality that this marking identifies, and we will conclude with the implications that the contribution of Number has to the behaviour of BPs. However, before presenting our analysis, we will briefly present an analysis of bare nouns by Kallulli (1999).

1. Bare Nouns: DP or NP?

Kallulli (1999) proposes that the behaviour of Albanian BPs, and the interpretative and distributional differences that they display with respect to English BPs, are due to the fact that there exist two differents kinds of BPs: Generic BPs and Existential BPs, which are syntactically, and thus semantically, different. Generic BPs are DP constituents with a null D head that denote variables. Unlike English, Albanian does not allow empty D heads; therefore, generic BPs are not permitted in this language. Existential BPs, being the plural counterpart to *Bare Singulars*, are NPs that completely lack the D projection. They are therefore neither constants nor variables, but actually predicates. The authors claims that «their existential force comes from a source external to the BPs themeselves, namely from the verb» (Kallulli, 1999, p. 151). Their distribution is constrained:

- semantically, by the fact that they are unsaturated structures. They may be neither generic nor specific.

- syntactically, by the fact that plural existentials may only be generated as the sister to V (194). They may not appear in specifier position, where, according to Kallulli (1999), arguments are generated. Bare, plural nouns may be neither subjects nor complements of individual-level or kind-selecting predicates. They function as predicates or form part of the Focus domain, where they do not serve to identify or refer to a particular object.

Data from Albanian is inconsistent with Kallulli's analysis. Contrary to its predictions, Albanian bare nouns can refer to a particular *haecceitas*[1]; they introduce a referent into the discourse that may be referred to by a later pronoun. As shown in (8), the bare noun may serve as an antecedent for the pronoun just as well as the definite description that also occurs in the sentence.

(8) *Kamionçin-a*$_i$ *e Pjeros ka rimorkjo*$_j$. *Ajo*$_{i/j}$ *përdoret shpesh.*
van the of Piero-gen.sg. has trailer she to find use-non-act. often
'Peter has a trailer for his van. He often uses it.'

Furthermore, it is not clear how Kallulli can disallow the predicative use of bare singulars in English (9).

(9) *John is doctor.

Given that predicates that are syntactically NPs are possible in this language (9'), it must be the case that a non-syntactic property differentiates bare singulars from their plural counterparts.

(9') John and Mary are doctors.

We propose that what distinguishes there two forms of the noun is Number.

2. Number

The role of grammatical number in the behaviour of nominal expressions has been frequently noticed in many languages, and a number of different authors have proposed to account for this category and its realization in the analysis of nominal expressions (Delfitto et Schroten, 1991; Deprez, 2004 ; Munn et Schmitt, 2005 ; Bouchard, 2002 ; 2005). According to Bouchard (2002), Number plays an important role because it is one of the features that can restrict the reference of a common noun to a particular subset that includes the individuals or objects designated by that noun. But how is this subset represented in Albanian? As the list of atoms that make it up, or as a singular or plural individual? Which part of the NP encodes the semantic contribution of Number?

2.1. The Realisation of Number in Albanian

It has been proposed (Dimitrova-Vulchanova, 2002) that, like in Romanian and in Bulgarian, the number marking that is semantically pertinent is realized on the determiner, and thus Albanian resembles a language like French. The following tests also suggest that the number marking on the noun also has a semantic contribution.

-The encoding of Number is obligatory on nouns, not just on determiners (10a);

-There are no [V+N] compounds headed by V (10b).

(10) a. *breg-u* vs. *brigj -e -t*
side the-sg. side pl. the-pl.
'the side vs. the sides'
b. *hekur-punu -*(es) - a*
iron work -er pl.

The existence of two semantically pertinent number markers is unsurprising because each of these markers contributes to the semantics of the noun phrase.

- The Number marking on the determiner allows for the atomization of the set. It indicates that the original set consisting of individuals or objects to which the common noun is applicable has a cardinality, i.e. that it contains a certain number of elements that are the participants of the event.

- The Number marking (plural marking) on the noun indicates that the members of the set are pluralities.

2.2 Number and BPs in Albanian

In studies of plurality, regular count nouns such as 'table', 'cat' etc., which are generally considered to denote sets of discrete elements are often juxtaposed with collective nouns like 'orchestra', 'committee', 'mafia' etc., which are said to denote 'groups' (Landman, 1989, Simons 1987, Moltmann 1997, Mari, 2005). These set-theoretic objects are principally distinguished from sets by the following property: unlike the members of sets, members of groups cannot be individuated. We propose that the differences between plural indefinites and BPs in Albanian are due to the fact that they denote different types of objects. We propose that, while plural indefinites denote transparent plural individuals (formed from applying a *sum/union* operation to the discrete atoms of a set), BPs denote fluid groups, i.e. unique plural individuals that form indivisible wholes (Mari, 2005). These groups have an opaque structure that disallows any quantitative or qualitative reference to their parts. To avoid terminological confusion, we will call this second type of plural individual a *cluster*. The following tests show the opaque structure of the denotation of a BP.

-BPs may only recieve a collective interpretation; they may never create distributative dependencies with indefinites that appear within their scope. A sentence such as (11) may only be interpreted as meaning that the first grade students sang a song together, never that they each sang a different song.

(11) *Nxënës të klasave të para kënduan një këngë.*
Student-m.pl. class-f.gen.pl. first sang a song
' First grade students sang a song'

Albanian BPs may not be used in 'Dependent plural' contexts.

(12) *Luan-ët meshkuj kanë *krifk -a /krifkë.*
Lion –pl.déf. male have mane-pl. mane
'Lions have manes'

The ungrammaticality of BPs in these constructions is due to the fact that the elements that make up the denotation of a BP are not accessible, making the establishments of a one-to-one relation between lions and manes impossible. The use of a bare plural creates only one interpretation: 'Each of the lions has more than one mane', which is rejected by our encyclopedic knowledge.

- BPs may not be the antecedent to neither a reflexive nor a reciprocal pronoun. These anaphors impose a distributive interpretation on their plural antecedents.

(13) * *vajz -a admironin vetveten në pasqyrë.*
girl pl. admire-3.pl.pres.non-act. themselves in mirror

(14) **vajz -a admironin njëra-tjetrën.*
girl pl. admire-3.pl.pres.non-act. each other

-BPs may not be interpreted as "the range of a binomial distributive construction' (Safir & Stowell, 1989; Gil 2005).

(15) **Vajz -a flenë secila në një shtrat (të ndryshëm/vet).*
girl pl. sleep every one in one bed different own

In the absence of an indefinite determiner in (15), it is impossible to access each of the girls to predicate *sleeping* of them.

2.3 Distribution of BPs in Albanian

In the preceding section, we showed that Albanian BPs are clusters that disallow access to their internal structure. The only information that they convey is that this structure contains more than one element. Unlike English BPs, they do not have the type of Number that allows them to define and individuate a class (cf Bouchard 2002 for an analysis of BPs in English). Thus, they may not be interpreted as generic, kind-denoting, as strong indefinites. Furthermore, since they cannot be intentional individuals (Bouchard 2002), they may not be an argument of a psychological verb.

These plural nouns may appear in subject position, but only when they are interpreted collectively.

(16) *Studentë ishin grumbulluar në oborr.*
student-pl. were gathered in yard
'Students were gathered in the yard.'

Modifying identifiers, adjectives such as *i njohur* 'known', *i famshëm* 'famous', *i caktuar* 'specific', etc., and specifying relative clauses, have a specifying effect that enables the speaker, and even sometimes the listener, to access individuals that are part of the cluster, thus permitting plural nouns to receive a distributive reading.

(17) *Gjuhëtarë me famë janë të fortë në abstraksione.*
linguist-pl. with renown are good in abstractions
'Renowned linguists are good with abstractions.'

Conclusions

In this paper, we showed that the facts about Albanian follow from the way in which the language encodes number through set-theoretic objects. Since the Albanian plural noun carries a semantically pertinent plural feature, the Albanian plural has a greater distribution than the French plural, where the semantically interpretable number is marked only on the determiner. However, since the Albanian number marking on the noun only indicates that it refers to a plurality, all the while prohibiting access to the internal structure of that plurality, the distribution of Albanian BPs will be more limited than that of their English counterparts. English plurals individuate the set denoted by the noun and indicate plurality at the same time; therefore, there are no restrictions on the distribution of English BPs: they may have generic, kind-denoting, and weak-indefinite readings, and may also combine with stage-level predicates and those that apply to kinds.

References

Bouchard, D. 2002. Adjectives, number, and interfaces: why languages vary: North-Holland linguistic series ; vol. 61. Amsterdam: North-Holland.

Bouchard, D. 2005. «Exaption and linguistic explanation». *Lingua* 115, p. 1685-1696.

Carlson, G.N.1980. Reference to kinds in english: Out standing dissertations in linguistics. New York : Garland.

Chierchia, G. 1998. Reference to kinds across languages. Natural Language Semantics 6:339-405.

Delfitto, D. andJ. Schroten. 1991. «Bare plurals and the number affix in DP». *Probus* 3, p.155-185.

Deprez, V. 2004. «Morphological Number, Semantic Number and bare Nouns». *Lingua* 6, p. 857-883.

Dimitrova-Vulchanova, M. 2002. «The Realization of Number in the Balkan Languages». *Papers from the Third Conference on Formal Approaches to South Slavic and Balkan Languages*, Dimitrova-Vulchanova, D.L. Dyer, I. Krapova et C. Rudin (éds.), *Balkanistica* 15, p. 171-192. Mississipi : The University of Mississipi Printing Services.

Gill, D. 2005. «Distributive numerals». Dans *World Atlas of Language Structures*, M. Dryer, M. Haspelmath, D. Gil et B. Comrie (éds.). Oxford: Oxford University Press.

Kallulli, D. 1999. **«**The Comparative Syntax of Albanian. On the Contribution of Syntactic Types to Propositional Interpretation». Thèse de doctorat, University of Durham.

Landman, F. 1989. Groups I and II. Linguistic and Philosophy 12 (5,6).

Link, G. 1983. The logical Analysis of Plurals and Mass Terms : A Lattice Theoretic Approach. In R. Bauerle, C. Schwartze et A. von Stechow (eds.) : 302-323.

Mari, A. 2005.«Intensional and epistemic wholes». Dans *The compositionality of Meaning and Content. Vol I Foundational Issues,* E. Machery et M. Werning. (éds). Ontos Verlag, p. 189-212.

Moltmann, F. 1997. *Parts and Wholes in Semantics*. Oxford: Oxford University Press.

Munn, A. and C. Schmitt (2005). «Number and indefinites». *Lingua 115*, p. 821-855.

Safir, K. and T. Stowell. 1989. «Binominal each». *Proceedings of NELS*, 18., p. 426-450. GLSA Publications. Amherst : University of Mass.

Simons, P. 1987. *Parts: A Study in Ontology*. Oxford: Oxford University Press.

[1] The notion *haecceita* is used like in medieval philosophical texts, meaning: *that which makes an object what it uniquely is.*

Supported by

Ministry of Science, Education and Sports
of the Republic of Croatia

Bulgarian Academy of Sciences

Croatian Language Technologies Society

ISBN 978-953-55375-0-2