DCU 250 Arabic Dependency Bank:
An LFG Gold Standard Resource for the Arabic Penn Treebank

Al-Raheb, Y., Akrout, A., van Genabith, J. and Dichy, J.

NCLT, Dublin City University
University of Lyon
{yalraheb; aakrout; josef}@computing.dcu.ie
joseph.dichy@univ-lyon2.fr

This paper describes the construction of a dependency bank gold standard for Arabic, DCU 250 Arabic Dependency Bank (DCU 250), based on the Arabic Penn Treebank Corpus (ATB) (Bies and Maamouri, 2003; Maamouri and Bies, 2004) within the theoretical framework of Lexical Functional Grammar (LFG). For parsing and automatically extracting grammatical and lexical resources from treebanks, it is necessary to evaluate against established gold standard resources. Gold standards for various languages have been developed, but to our knowledge, such a resource has not yet been constructed for Arabic. The construction of the DCU 250 marks the first step towards the creation of an automatic LFG f-structure annotation algorithm for the ATB, and for the extraction of Arabic grammatical and lexical resources.

Keywords: LFG, treebank, Arabic, gold standard.

1. INTRODUCTION

This paper describes the construction of a dependency bank gold standard for Arabic, DCU 250 Arabic Dependency Bank (DCU 250), based on the Arabic Penn Treebank Corpus (ATB) (Bies and Maamouri, 2003; Maamouri and Bies, 2004) within the theoretical framework of Lexical Functional Grammar (LFG) (Kaplan and Bresnan 1982; Bresnan 2001; Dalrymple 2001). For automatically extracting grammatical and lexical resources from treebanks, it is necessary to evaluate against established gold standard resources. Dependency banks provide a benchmark for resource quality, allowing direct comparisons to be made between the output of differing grammar development paradigms. Gold standards for various languages such as the PARC 700 Dependency Bank (King et al., 2003) and the DCU 105 (Cahill et al. 2002) for English have been developed and used for evaluation. To our knowledge, such a resource has not yet been constructed for Arabic. The construction of the DCU 250 marks the first step towards the creation of an automatic LFG f-structure annotation algorithm for the ATB, for the extraction of Arabic grammatical and lexical resources.

Of the ATB's 23,611 parsed sentences, 250 sentences were randomly selected from Diab et al.'s (2004) test set of vocal sentences (including diacritics). The trees with complete part of speech (POS) tag information were selected instead of those with the collapsed POS tag set in order to incorporate as much morphological and grammatical information as possible. The DCU 250 has been constructed in three stages, (i) partial automatic annotation which provided annotations for over two thirds of the gold standard tree nodes, (ii) completion of the annotation process by manual annotating the remaining unannotated nodes and (iii) manual examination, and correction where necessary, of the automatic annotations.

The annotation of the DCU 250 has focused attention to a number of interesting and

problematic constructions in the ATB, e.g. NP coordinate structures, relative clauses, and the internal annotation of NPs. The mis-tagging of some words in the ATB, particularly verbs being mis-tagged as nouns, also proved problematic, e.g. in sentence 17 of the DCU 250 the verb 'Talaba' is mis-tagged as a noun. Following Cahill et al. (2004) on English, future work will focus on the implementation of a wide-coverage, robust automatic LFG f-structure annotation algorithm to automatically annotate the whole ATB corpus. The resulting f-structures will be evaluated against the DCU 250 gold standard f-structures presented in this paper.

Section 2 of this paper provides background information, including a brief introduction to LFG, and further motivates the construction of the dependency gold standard resource. Section 3 describes the process of constructing the DCU 250. Section 4 concludes and outlines plans for future work.

2. BACKGROUND AND MOTIVATION

2.1 Motivation

Scaling traditional, hand-crafted, deep, constraint-based grammars and lexical resources to provide wide coverage is prohibitively time-consuming and expensive. Automatic acquisition of linguistic resources from treebanks proved more cost-effective, however the first generation of these resources tended to be shallow, not capturing predicate-argument structure or long-distance dependencies. In recent years, an alternative approach has been explored to overcome both problems by efficiently acquiring wide coverage, deep, constraint-based grammatical and lexical resources automatically from treebanks. This approach has resulted in resources for HPSG (Miyao and Tsujii, 2002), CCG (Hockenmaier and Steedman, 2002) and TAG (Habash and Rambow, 2004).

Cahill et al. (2002, 2004) presents the automatic acquisition of deep, wide coverage LFG resources from the Penn-II Treebank (Marcus et al., 1994). This research has been extended to produce high quality resources for Chinese (Burke et al., 2004), German (Cahill et al., 2005) and Spanish (O'Donovan et al., 2005). The construction of a gold standard resource was an important initial step in all of these projects, as it (i) informed the process of developing an automatic f-structure annotation algorithm, a core component of grammar and lexicon development for each language and (ii) provided a valuable tool for resource evaluation. A gold standard also provides a platform for comparing resources produced by different research groups as explored and presented by Burke et al. (2004).

This paper presents the construction of the DCU 250, a gold standard of semi-automatically constructed LFG f-structures for 250 Arabic sentences from the ATB. This is the first step towards the development of deep, wide coverage LFG grammatical and lexical resources for Arabic. Section 2.2 provides a brief overview of LFG.

2.2 Lexical Functional Grammar

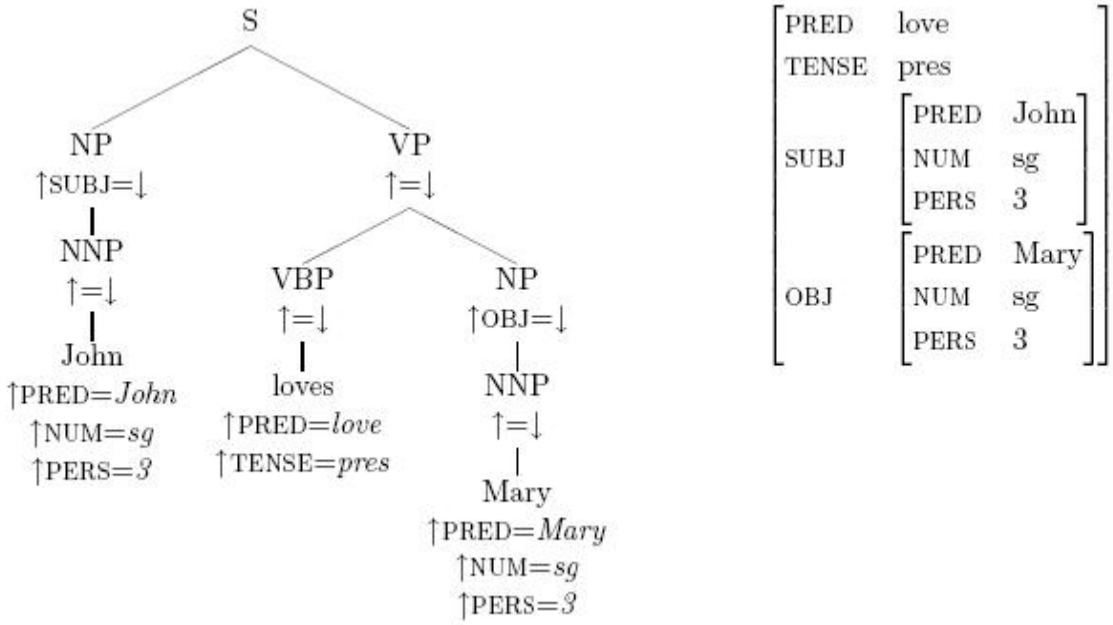
Lexical Functional Grammar (LFG) (Kaplan and Bresnan 1982; Bresnan 2001; Dalrymple 2001) is a constraint-based theory of grammar with c-structure and f-structure as the two main levels of representation, closely related in a projection architecture:

- C-structure: constituent structure is represented in terms of context-free grammar trees and captures word order and the hierarchical grouping of phrases. Nodes in c-structure trees are annotated with f-structure equations.

- F-structure: functional structure represents abstract syntactic information, such as grammatical functions (e.g. SUBJ) and morpho-syntactic properties (e.g. TENSE) which are encoded as attribute-value matrices (AVMs) approximating to predicate-argument-modifier relations or dependencies.

C-structure nodes are annotated with f-structure equations containing up (↑) and down (↓) arrows (meta-variables). Up-arrows refer to the f-structure associated with the immediately dominating tree node, while down-arrows (↓) refer to the local node. Each occurrence of meta-variables is instantiated using a unique identifier associated with the node to which the meta-variable refers, which allows a set of equations (f-descriptions)

FIGURE 1: C-structure annotated with f-structure equations and the resulting f-



structure for the sentence *John Loves Mary*.

to be created from the annotated c-structure. These equations, if satisfiable, produce an f-structure. Figure 1 provides the c-structure for the sentence "John loves Mary" annotated with f-structure equations and the resulting f-structure.

3. CONSTRUCTION OF DCU 250

3.1 Data Preparation

The 23,611 trees of the ATB are available in two forms, vowelled (with diacritics) and unvowelled (without diacritics). Diab et al. (2005) split the ATB into three parts: a training set of 18,970 trees (~80%), a development set of 2,304 trees (~10%) and a test set of 2,337 trees (~10%). The first step in constructing the DCU 250 was to randomly select trees from the ATB test set of Diab et al. (2005). Of the 250 randomly selected

trees, 14 trees consisted only of punctuation. These instances have been replaced by a set of randomly-selected sentences containing lexical instances rather that simply punctuation. The trees were chosen from the vowelled version of the treebank as these trees provide more detailed grammatical and morpho-syntactic information, e.g. case and passive, allowing a more fine-grained f-structure analysis.

Section 3.2 outlines the first step in the construction of the f-structures comprising the DCU 250, the partial automatic annotation of the ATB trees. Section 3.3 describes the processes of manually completing and inspecting these annotations.

3.2 Partial Automatic Annotation

3.2.1 Lexical Macros

In order to speed up the process of gold standard construction, the annotation of the gold standard tree nodes with f-structure information was partially automated. The first step in this process was to provide default annotations for POS nodes: lexical macros. Each word node in the ATB trees is governed by a POS node consisting of POS tags and morpho-syntactic tags containing valuable information which must be represented in the gold standard f-structures, e.g. the POS node DET+NOUN+CASE_DEF_ACC contains the POS tags DET and NOUN and the morpho-syntactic tags DEF and ACC which provide information about the definiteness of the noun and its case. Lexical macros were defined to automatically provide f-structure annotations for the morpho-syntactic tags for each POS node, as outlined in Table 1.

Arabic Treebank Tags	LFG Feature Name	LFG Values	Additional annotations
DEF INDEF	DEFINITENESS	+ -	
MASC FEM	GENDER	masc fem	
SG PL DU	NUM	sg pl dual	
FUT	TENSE	future	
PV or VERB_PERFECT IV or VERB_IMPERFECT	ASPECT	perfect imperfect	
1 2 3	PERS	1 2 3	
PASSIVE	PASSIVE	Passive	
MOOD:I MOOD:S MOOD:J MOOD:SJ	MOOD	indicative subjunctive jussive subjunctive/jussive	
ACC GEN NOM ACCGEN	CASE	accusative genitive nominative accusative/genitive	
DEM_PRON POSS_PRON REL_PRON	PRON_TYPE	demonstrative possessive relative	↑PRON_FORM=*word*
PRON	n/a	n/a	↑PRON_FORM=*word*
CONJ	n/a	n/a	↑COORD_FORM=*word*
SUB_CONJ	n/a	n/a	↑SUBORD_FORM=*word*
EMPHATIC_PART EXCEPT_PART FUT_PART INTERROG_PART NEG_PART RC_PART	PRT_TYPE	emphatic except future interrogative negative relative	↑PRT_FORM=*word*
PART	n/a	n/a	↑PRT_FORM=*word*
PREP	n/a	n/a	↑PFORM=*word*

TABLE 1: POS tags, corresponding features and values.

Applying these lexical macros to the example POS node DET+NOUN+CASE_DEF_ACC results in the annotations ↑DEFINITENESS='+' and ↑CASE=accusative, in addition to the default predicate annotation which is provided to all word nodes: ↑PRED=word. Table 1 illustrates that there are two possible values for DEFINITENESS, + and -, which are triggered by the tags DEF(inite) and INDEF(inite) respectively. Prepositions, co-ordinating and sub-ordinating conjuncts, pronouns and particles all receive additional annotations, PFORM, COORD_FORM, SUBORD_FORM, PRON_FORM and PRT_FORM respectively. All of these features have the same value as the local PRED, i.e. the uninflected word.

3.2.2 ATB Functional Tags

The second step of the partial automatic annotation process provides default annotations for phrasal nodes marked with ATB functional tags, e.g. the phrasal node NP-SBJ is marked with the functional tag -SBJ representing subject. The functional tag -SBJ triggers the partial automatic annotation process to provide the node NP-SBJ with the annotation ↑SUBJ=↓. Table 2 provides a complete set of ATB functional tags and their corresponding default annotations in the DCU 250 gold standard.

Functional tag	Description	Default Annotation
-SUBJ	subject	↑SUBJ=↓
-OBJ	object	↑OBJ=↓
-TPC	topic	↑TOPIC=↓
-TMP	temporal adjunct	↓elem↑ADJUNCT
-LOC	locative adjunct	↓elem↑ADJUNCT
-DIR	directional adjunct	↓elem↑ADJUNCT
-MNR	manner adjunct	↓elem↑ADJUNCT
-ADV	adverbial adjunct	↓elem↑ADJUNCT

TABLE 2: ATB Functional Tags and their default f-structure annotations

3.2.3 Prepositional Phrases

The most common internal structure of ATB prepositional phrases is a preposition preceding a noun phrase. There are many alternative PP structures, almost all of these have a preposition as the head node in the left-most position. The automatic partial annotation process harnesses this information to provide default annotations by (i) always annotating the left-most daughter as head if it is a preposition and (ii) for PPs with only two daughters, providing the right-most daughter with the annotation ↑OBJ=↓ if it is an NP. Figure 2 provides the annotated c-structure for the PP "dAxila Al+$ub~Ak+i" (inside the window).

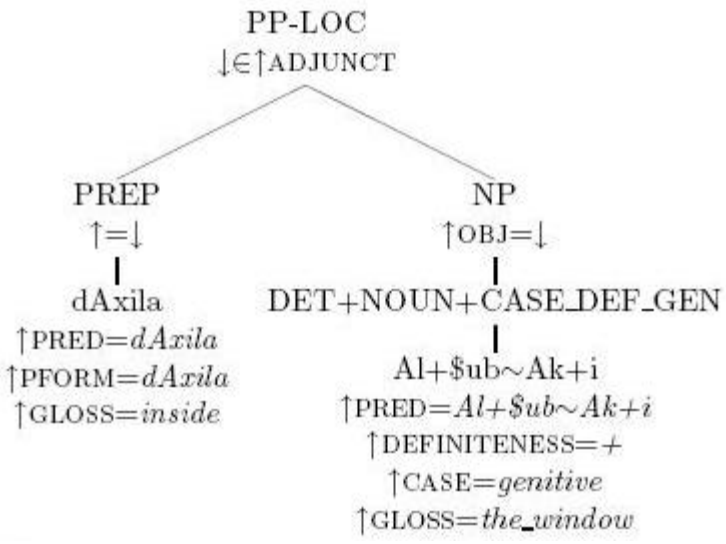


FIGURE 2: C-structure for the PP "dAxila Al+$ub~Ak+i" (inside the window) annotated with f-structure information.

The annotation processes described so far provide annotation for all but one (DET+NOUN+CASE_DEF_GEN) of the nodes in the subtree provided in Figure 2. The left-most node in the PP is a preposition and is therefore annotated as the head node as outlined above. The right-most node is an NP and is annotated as the object. The functional tag annotation process described in Section 3.2.2. provides the adjunct annotation for the parent node, triggered by the locative functional tag (-LOC). The lexical macros outlined in Section 3.2.1 provide the DEFINITENESS and CASE annotations for the word node "Al+$ub~Ak+i" and also the PFORM annotation for "dAxila".

3.2.4 Co-ordination

Annotation of co-ordinate structures was also partially automated. A conservative approach was taken, whereby only the most simple co-ordinate structures were annotated automatically, i.e. subtrees containing three daughters with the second daughter being a conjunct (CONJ). The conjunct is annotated as the head node, while both the first and third daughter nodes are annotated as elements of the co-ordination set: ↓elem↑COORD. Figure 3 shows the annotated c-structure and the resulting f-structure produced by the application of this annotation procedure (and those previously described) to the noun phrase "triyniydAd+u wa- -tuwbAguw" (Trinidad and Tobago). Example ATB phrases are transliterated using Buckwalter's (2001) morphological analyser.

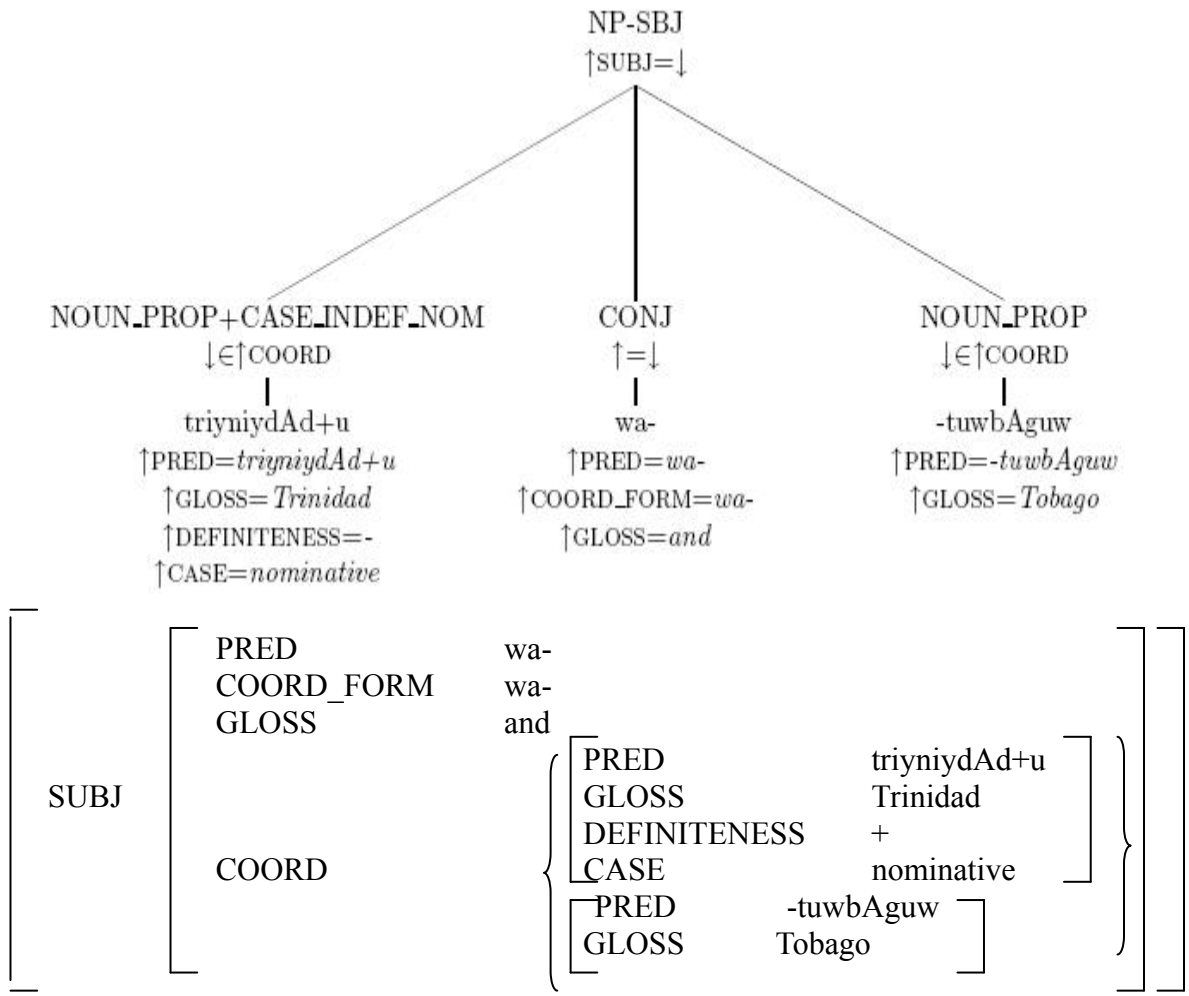


FIGURE 3: C-structure for the NP "triyniydAd+u wa- -tuwbAguw" (Trinidad and Tobago) annotated with f-structure information and the resulting f-structure.

The automatic partial annotation process outlined in this paper has been sufficient to fully annotate this c-structure. The conjunct (CONJ) is annotated as the head node (↑=↓), while both remaining nodes are added to the co-ordination set (↓elem↑COORD). The resulting f-structure shows a COORD set containing two elements, glossed as "Trinidad" and "Tobago". The subject annotation on the parent node in Figure 3 results from the functional tag annotation process outlined in Section 3.2.2. The COORD_FORM, CASE and DEFINITENESS annotations are all produced by the lexical macros introduced in Section 3.2.1.

3.2.5 Head Annotation

The final step in the partial annotation process was the identification and annotation of the head node of each local subtree in the DCU 250. Again, a conservative approach was adopted for head annotation in order to maintain a high confidence level in these annotations. This reduces the risk of introducing errors through the automatic partial annotation process, thereby minimising the later task of manual completion and

inspection. Two basic strategies were employed, which annotate a node as the head daughter if:

1. the node is the only daughter in the local subtree.
2. the node is the only remaining unannotated daughter node in the local subtree.

Applying strategy 1 to the annotated c-structure of Figure 2 annotates the one remaining unannotated node (DET+NOUN+CASE_DEF_GEN) as the head of the local subtree, producing the annotated c-strcuture and the f-structure provided in Figure 4.

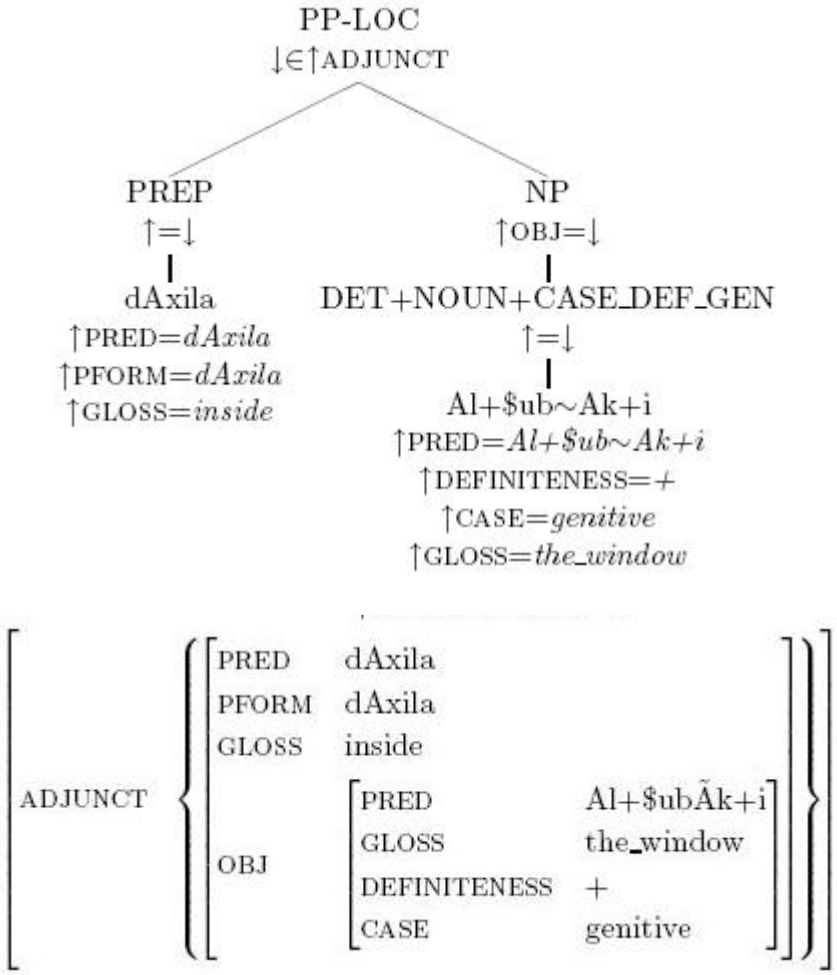


FIGURE 4: Annotated c-structure and resulting f-structure for the PP "dAxila Al+$ub~Ak+i" (inside the window) derived from the c-structure in Figure 2 and the additional head annotation from Section 3.2.5.

3.3 Manual completion and inspection

The automatic partial annotation process was applied to the trees of the gold standard and provided annotations for almost two-thirds of all nodes. The conservative approach taken to automatic annotation meant that the manual inspection process was relatively straightforward. The manual completion of the partial annotations to provide connected and covering annotations for all 250 trees was a more intensive task. However, this was a very valuable process as the annotators developed a deeper understanding of the treebank style which will help inform the process of developing an automatic f-structure annotation algorithm for the ATB.

The simplest task in the manual completion process was the annotation of the head nodes of each local subtree. A similarly straightforward task, in most cases, was the annotation of co-ordinate structures, as, in order to minimise the margin of error, some relatively simple co-ordinate structures were left unannotated by the automatic partial annotation process. However, some co-ordinate structures were quite difficult to annotate due to the flat analysis provided in the ATB trees, a problem which also occurs in the Penn-II treebank for English. Figure 5 provides the ATB subtree for the phrase ">amiyrokA Al+$amAliy~+ap+u wa- -Al-wusoTaY wa- -Al+kAriybi" (North and Central America and the Caribbean). For clarity, the unvowelled version of the ATB tree for this phrase has been provided. Rather than providing any internal structure, such as grouping "Al+$amAliy~+ap+u" (North) and "-Al-wusoTaY" (Central) as a single phrase, a flat analysis is provided. To produce the f-structure provided in Figure 5, the manual annotator must use f-structure equations to provide internal structure and ensure that the heads of any added internal phrases are not confused with the head of the overall phrase.

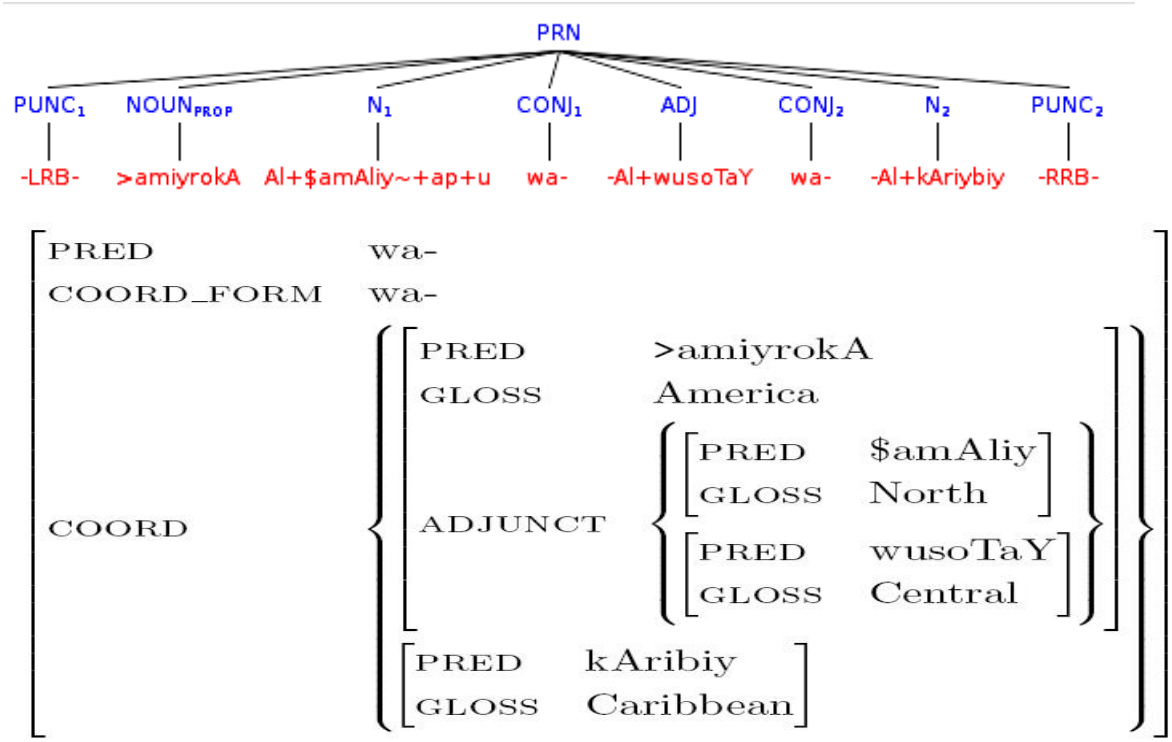


FIGURE 5: ATB tree and corresponding f-structure for ">amiyrokA Al+$amAliy~+ap+u wa- -Al-wusoTaY wa- -Al+kAriybi" (North and Central America and the Caribbean).

5. CONCLUSION AND FUTURE WORK

This paper describes the construction of the DCU 250 Arabic Dependency Bank (DCU 250), based on the Arabic Penn Treebank Corpus (ATB) within the LFG framework. Almost two-thirds of the nodes were automatically annotated, with the tasks of annotation inspection and completion performed manually. The construction of the

DCU 250 marks the first step towards the creation of an automatic LFG f-structure annotation algorithm for the ATB, for the extraction of Arabic grammatical and lexical resources. This paper contributes a valuable resource which can be used for the evaluation of Arabic grammatical and lexical resources.

REFERENCES

Beesley, K. R.. (2001). 'Finite-state morphological analysis and generation of Arabic at Xerox Research: Status and plans in 2001'. In ACL Workshop on Arabic Language Processing: Status and Perspective, pages 1--8, Toulouse, France, July.

Bies, A. and Maamouri, M. (2003). 'Penn Arabic Treebank Guidelines'. (Draft: January 28, 2003), Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA.

Bresnan, J. (2001). *Lexical Functional Syntax*, Blackwell Publishers, Oxford.

Buckwalter, T. (2001). *Buckwalter Arabic Morphological Analyzer Version 1.0*. LDC catalogue number LDC2002L49, ISBN 1-58563-257-0.

Burke, M. (2006). *Automatic Annotation of the Penn-II Treebank with F-structure Information*. School of Computing, Dublin City University, Dublin, Ireland.

Cahill, A., McCarthy, M., van Genabith, J., and Way, A. (2002). Parsing with PCFGs and Automatic F-Structure Annotation. In Miriam Butt and Tracy Holloway King, editor, Proceedings of the Seventh International Conference on LFG, pages 76-95, Stanford, CA. CSLI Publications.

Cahill, A., M. McCarthy, M. Burke, R. O'Donovan, J. van Genabith, and A. Way. (2004). 'Evaluating Automatic F-Structure Annotation for the Penn-II Treebank'. Journal of Research on Language and Computation.

Cahill, A., Forst, M., Burke, M., McCarthy, M., O'Donovan, R., Rohrer, C., van Genabith, J. and Way, A. (2005). 'Treebank-Based Acquisition of Multilingual Unification Grammar Resources'. Journal of Research on Language and Computation, Springer, Volume 3, Number 2, July 2005, ISSN: 1570-7075, pp. 247-279.

Dalrymple, M.(2001). *Lexical-Functional Grammar*. San Diego, Calif.; London: Academic Press.

Diab, Mona, Kadri Hacioglu and Daniel Jurafsky. (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. Proceedings of HLT-NAACL.

Habash, N. and O. Rambow (2004). 'Extracting a Tree Adjoining Grammar from the Penn Arabic Treebank'. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco.

Habash, N. (2004). 'Large Scale Lexeme Based Arabic Morphological Generation'. In Proceedings of Traitement Automatique du Langage Naturel (TALN-04). Fez, Morocco.

Hockenmaier, J., and M. Steedman. (2002). 'Acquiring Compact Lexicalized Grammars from a Cleaner Treebank'. Third LREC, Las Palmas, Spain.

Kaplan, R. and J. Bresnan (1982). 'Lexical-Functional Grammar: A Formal System for Grammatical Representation'. In J. Bresnan (ed.), The Mental Representation of Grammatical Relations, MIT Press, Cambridge, MA, pp. 173-281.

King, T. H., Crouch, R., Riezler, S., Dalrymple, M., and KaplanR. M. (2003). 'The PARC 700 Dependency Bank'. In Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora, held at the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest.

Maamouri, M., and Bies, A.(2004). 'Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools'. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, August 28, 2004.

Manning, D. C. (1993). 'Analyzing the verbal noun: Internal and external constraints'. In Soonja Choi (ed), Japanese/Korean Linguistics 3, Stanford, CA: Stanford Linguistics Association, pp. 236-253.

Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., Schasberger, B.. (1994). 'The Penn Treebank: annotating predicate argument structure'. In Proceedings of the workshop on Human Language Technology, March 08-11, 1994, Plainsboro, NJ.

Miyao, Y. and Tsujii, J. (2002). 'Maximum Entropy Estimation for Feature Forests'. In the Proceedings of Human Language Technology Conference (HLT 2002).

O'Donovan, R., Bodomo, A., van Genabith, J. and Way, A. (2004) 'Treebank-Based Acquisition of a Chinese Lexical-Functional Grammar'. In Proceedings of the PACLIC-18 Conference, Waseda University, Tokyo, Japan, pages 161-172.

O'Donovan, R., A. Cahill, J. van Genabith, and A. Way. (2005). 'Automatic Acquisition of Spanish LFG Resources from the CAST3LB Treebank'. In Proceedings of the Tenth International Conference on LFG, Bergen, Norway.