Enhanced Annotation and Parsing of the Arabic Treebank

Mohamed Maamouri, Ann Bies, Seth Kulick
Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104, USA
{maamouri,bies,skulick}@ldc.upenn.edu

Abstract

The Arabic Treebank at the Linguistic Data Consortium has significantly revised and enhanced its annotation guidelines and annotation procedure over the past year. The revised syntactic guidelines are now being applied in annotation production, and the combination of the revised guidelines and a period of intensive annotator training has raised inter-annotator agreement f-measure scores already. Revised morphological/part-of-speech (POS) guidelines are nearly complete as well, and will be applied in annotation production in the near future. This paper reports on an experiment in automatically enhancing the old morphological/POS tags in the right direction and the resulting parsing improvement. Finally, a new division of the POS analysis marking both morphological form and POS function is proposed.

1. Introduction

The Arabic Treebank (ATB) team at the Linguistic Data Consortium [13] has significantly revised and enhanced its annotation guidelines and annotation procedure over the past year. The revised syntactic guidelines are now being applied in annotation production, and the combination of the revised guidelines and a period of intensive annotator training has raised inter-annotator agreement f-measure scores already. Revised morphological/part-of-speech (POS) guidelines are nearly complete as well and will be applied in annotation production in the near future. However, the question of improving parser results before that production is complete has been raised. This paper reports on an experiment in automatically enhancing the old morphological/POS tags in the right direction and the resulting parsing improvement.

The overall guidelines revision process was initiated based on lower than expected initial parsing scores and on an examination of inconsistencies in the annotation. Parser scores for a statistical parser trained on ATB data were well below that of the Penn Treebank and the Chinese Treebank, roughly 14 and 9 points in absolute f-measure below, respectively. Inconsistencies within the Treebank annotation regarding the relationship between Part-of-Speech (POS) tags and the syntactic annotation as well as inconsistencies in the annotation of certain syntactic constructions were shown to contribute to the parser performance. Those inconsistencies were therefore the initial targets for improvement in both the guidelines and in annotator training. [15] reports on some of the syntactic considerations with respect to parsing results. In this paper, we will focus on morphological/POS considerations.

Automatically enhancing the existing POS tags in the ATB3a-v2.6 release subcorpus (Catalog ID: LDC2007E65) as described below results in an overall improvement in parsing performance from 74.1 to 76.2 when the parser is allowed to choose its own tags, and from 74.4 to 78.1 when the parser is forced to use the given tags.

Finally, a new division of the POS analysis marking both morphological form and POS function is proposed.

2. Overview of the Arabic Treebank

Over the past decade there has been some important progress in the computational processing of Arabic. However, because of its socio-political characteristics, highly complex morphology and significant dialectal differences, Arabic continues to challenge the NLP community. In spite of recent

progress, Arabic is still lacking in tools and annotated resources. While there has been recent progress in creating such NLP tools as Base Phrase Chunkers ([8], [9]), there remains a demand for high quality Arabic language resources and a need for greater volumes of rich and sophisticated annotated text in Arabic.

Treebanks are language resources that provide annotations of natural languages at various levels of structure: at the word level, the phrase level, and the sentence level. NLP and Human Language Technology (HLT) researchers in the academic and industrial communities seem to agree that treebanks, proposition banks, bilingual lexicons, and parallel texts are the most frequently used and needed linguistic resources in multiple areas of HLT research and development, including natural language processing, human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research. Treebanks and PropBanks, collectively called X-Banks, are at the center of activities, techniques, technologies and methodologies which automate the process of extracting and understanding information from text.

The Penn Arabic Treebank Project (ATB): This complex and difficult annotation project began at the Linguistic Data Consortium (LDC) at the University of Pennsylvania in the fall of 2001, but the set-up for Arabic – as against English or Chinese – did not include significant time for fundamental research, which as a result has become entangled with the on-going development of the annotation guidelines and process [14]. The annotation project consists of two phases: (a) Morphological/Part-of-Speech (=POS) tagging which divides the text into lexical tokens, and gives relevant information about each token such as lexical category, inflectional features, and a gloss (referred to as POS for convenience, although it includes morphological, morphosyntactic and gloss information not traditionally included with part-of-speech annotation), and (b) Syntactic analysis referred to as Arabic Treebanking (=Arabic TB) which characterizes the constituent structures of word sequences, provides function categories for each non-terminal node, and identifies null elements, co-reference, traces, etc.

The LDC Arabic Treebank team (=ATB) has now completed nearly a million words of morphologically and syntactically annotated data: (1) Arabic Treebank: Part 1 v 2.0, LDC Catalog No. LDC2003T06, roughly 166K words of written Modern Standard Arabic newswire from the *Agence France Presse* corpus (AFP); (2) Arabic Treebank: Part 2 v 2.0, LDC Catalog No. LDC2004T02, roughly 144K words from Al-Hayat distributed by Ummah Arabic News Text (UMAAH) (the annotation of this corpus includes new features of complete short vowel marking, case and mood endings, lemma IDs, and more specific part-of-speech (POS) tags for verbs and particles.), and (3) Arabic Treebank: Part 3 v. 2.0, LDC Catalog No.: LDC2005T20, roughly 350K words of newswire text from Annahar morphologically and syntactically annotated (ANNAHAR).

The ATB corpora are annotated for morphological information, Part-of-Speech, English gloss (all in the "POS" phase of annotation), and for syntactic structure (similar to the Penn English Treebank II style) ([16], [17], [6]). In addition to the usual issues involved with the complex annotation of data, we have come to terms with a number of issues that are specific to a highly inflected language with a rich history of traditional grammar.

In designing our annotation system for Arabic, we relied on traditional Arabic grammar, previous grammatical theories of Modern Standard Arabic and modern approaches, and especially the Penn Treebank approach to syntactic annotation, which we believe can be generalized to the development of annotation systems for other languages [12]. We also benefited from the existence at LDC of a rich experience in linguistic annotation. We were innovative with respect to traditional grammar when necessary and when we were sure that other syntactic approaches accounted for the data. Our goal is for the Arabic Treebank to be of high quality, to have a high level of descriptive consistency and a long shelf-life expectancy, and to have credibility with regard to the attitudes and respect for correctness known to be present in the Arab region as well as with respect to the NLP and wider linguistic communities.

3. Enhanced annotation

When the decision was made to revise the annotation guidelines and subsequently to revise the Treebank annotation of one of the Arabic Treebank corpora, ATB3-v2.0 was chosen as the first corpus to receive this revision. It is the largest single-source Arabic Treebank corpus to have been completed so far, and the 2.0 version contained the most up-to-date annotation so far, as it included all of the annotation and lexicon revisions that had been made up to that point. It therefore made sense to begin a more drastic annotation revision process with this corpus. The subcorpus of Arabic Treebank 3(a) – v 2.6 represents

the first segment of ATB part 3 (ANNAHAR) that has been revised according to the new Arabic Treebank syntactic annotation guidelines. This first segment (a) of the revised and updated Arabic Treebank ATB part 3 consists of 152 newswire stories from the An Nahar News Agency, roughly the first third of the ATB part 3 (previously released as Arabic Treebank: Part 3 (full corpus) v 2.0 (MPG + Syntactic Analysis), LDC Catalog No.: LDC2005T20). In this segment (a), there are a total of 85,497 words/tokens before clitcs are split and 100,847 words/tokens after clitcs are separated for the Treebank annotation.

3.1. Revised morphological/POS guidelines

The POS tags for nouns and adjectives in particular were revised to be more fine-grained. The core POS tag of NOUN is now further distinguished as NOUN (common noun), NOUN_NUM (number), and NOUN_QUANT (quantifier). The core POS tag of ADJ is also further distinguished as ADJ (common adjective), ADJ_NUM (ordinal number), and ADJ_COMP (comparative adjective). We also intend to introduce three additional new tags in the near future: MAS or MAS_NOUN (MASdar/ردصم or gerund), AP or AP_ADJ (Active Participle/مسا لعاف/<ism fAEil) and PAP or PAP_ADJ (PAssive Participle/لوعفم مسا). These tags will address the need for a more thorough treatment of participles and gerunds (which can have nominal or verbal readings) in Arabic. It is worth noting that 5% of VPs in the Arabic Treebank (ATB3-2.0) have a nonverbal head as against 0.5 % only in the English Penn Treebank [11].

The above greater distinctions among nouns and adjectives also follow traditional Arabic grammar categories. Additional POS changes were also made to more closely follow traditional Arabic grammar categories – for example, the number of prepositions was drastically reduced (most prepositional lexical items now being categorized as nouns, NOUN rather than PREP; PREP is now reserved for the extremely limited list of traditional Arabic prepositions), and particles are now given several POS alternatives, again closely aligned with traditional categories. For example, the particle *fa* had one POS value only in previous Treebank annotation: CONJ. It now has three different POS tags available: (a) CONJ for *fa Al-EaTf*/عاف فطعلا used for the coordination of words and sentences (b) CONNEC_PART for *fa Al-jazA'*/عاف ءازجلا used in conditional constructions before the apodosis or main clause and for *fa Al-rabT*/عاف طبرلا when used to introduce the comment after the focus particle *>am~A*/ أمّا and (c) SUB_CONJ for *fa Al-sababiy~ap*/عاف ا لسببية when it is used to introduce the result or cause of the main clause.

A new category of PSEUDOVERB has been added, to account for the verbal behavior of certain Arabic particles. These are "the sisters of إنَّ <inna" (with the exception of أنَّ ">anna," the complementizer "that"), a category regarded by Arabic grammarians as having verbal properties, such as subcategorizing for a subject and a predicate or clausal complement. Since these words display verbal behavior although they are not technically verbs, they will now be given the POS tag "PSEUDOVERB" and head a VP in the tree.[1]

3.2. Experimental automatic POS enhancements

As mentioned above, the process of enhancing the Arabic Treebank focused primarily on making the guidelines more comprehensive, more consistent and clearer at both the morphological and syntactic levels. The number of changes was important and significant, and the decision was to start implementing them at the TB annotation level first, targeting the most important volume in the Arabic Treebank segments (the Annahar Corpus ATB3). It was also decided to only make those POS changes that could be dealt with automatically in a quick engineered enhancement pass. Consequently, the December 2007 release of the subcorpus of Arabic Treebank 3 (a) – v 2.6 includes a Treebank annotation that has been revised in accordance with the new and updated Arabic Treebank Annotation Guidelines. Certain automatic changes have been made to the POS tags. These changes are described below. However, the Part-of-Speech/morphology/gloss annotation has not yet been fully and manually revised – a revision of this phase of annotation is planned for future releases.

There were 3781 automatic Part of Speech (POS) tag changes. These tag changes were an approximation to what the "correct" tags should be. Counts for the automatic enhancement changes to the ATB3-v2.6a subcorpus are as follows:

[1] For a more complete description of the new annotation policies, see [3] the *Arabic Treebank Morphological and Syntactic Annotation Guidelines.* (2008). http://projects.ldc.upenn.edu/ArabicTreebank/

Automatic POS change	Description	Instances in ATB3-v2.6
PREP to NOUN	Preposition to noun	1583
ADJ to NOUN	Adjective to noun	543
ADJ to DV	Adjective to verbal	498
NOUN to DV	Noun to verbal	343
SUBCONJ to PSEUDOVERB	Subordinating conjunction to pseudo-verb	213
ADV to NOUN	Adverb to noun	177
NEG_PART to GAYOR	Negative particle to lexically specific tag	136
NUM to SCORE	Number to sports score	108
PART to PSEUDOVERB	Particle to pseudo-verb	70
NEG_PART to PSEUDOVERB	Negative particle to pseudo-verb	58
NEG_PART to PV	Negative particle to perfect verb	46
NOUN to ADJ	Noun to adjective	4
CONJ to PREP	Conjunction to preposition	2

Table 1. Distribution of 3781 total automatic POS changes

These automatic changes fall into two distinct categories:

1. **Lexically-based changes:** These are POS tags which are mostly invariant for certain words, and account for approximately half of the automatic enhancement changes. These enhancements can also be considered as nothing more or less than quality control checks, searching out annotation errors on certain lexical items.

2. **Syntactically-based changes:** These are POS tags which are determined solely by syntactic function and context, and also account for approximately half of the automatic enhancement changes. This has been one of the most problematic areas of the Arabic annotation, since there is a significant difference in the POS tag/syntactic tree relation in Arabic as compared with English. A proposed solution to this problem is to more cleanly separate the output of the morphological analyzer and their syntactic function, indicating the syntactic function as function tags on the current POS tag.

3.2.1. Lexically determined automatic changes. The largest number of tag changes, PREP to NOUN, fall into the first category. The reason for this change is that the current release of the Arabic Treebank makes a more careful distinction than before between "true" prepositions and "prepositionals," which are all nominal in Arabic ([4], pp. 174ff). As a consequence, many of the items that used to have the tag PREP, heading a PP, are now treebanked (syntactically analyzed) as heading an NP, although the POS tag itself was not changed. For example, a typical case from the 2.6 release is:

```
(NP-ADV (PREP maEa)  with
        (NP (DEM_PRON_MS *`lika))) that
```

for which the PREP has been changed to NOUN in these experiments, to

```
(NP-ADV (NOUN maEa)  with
        (NP (DEM_PRON_MS *`lika))) that
```

The recently revised version of the POS guidelines [3], section 2.3.9 (Morphological Analysis, http://projects.ldc.upenn.edu/ArabicTreebank/), lists the 19 words that are still considered to be prepositions, which should have the PREP tag, and words (more than 40) that used to be considered prepositions, but are now considered nouns, and so their POS tags have been changed in these experiments from PREP to NOUN.

3.2.2. Syntactically determined automatic changes. About half of the changes relate to the syntactic function of the tag. This has been one of the most problematic areas of the Arabic annotation, since there is a significant difference in the POS tag/syntactic tree relation in Arabic from the way in which they relate to each other in English.

Perhaps the best way to explain this is with an analogy to English. One of the problematic areas of POS tagging for English is the ambiguity of present participles/gerunds. Present participles/gerunds can be used in English as adjectives, verbal participles, or nouns. Distributional tests for these distinct uses often yield clear results in English, as in examples 1-3 below.

1. The disturbing news gave everyone the chills.
   "Disturbing" = Adjective (JJ): modifies news, gradable (the very disturbing news); etc.

2. Disturbing people for no good reason is rude.

"Disturbing" = Verbal participle (VBG): can take direct object and assign object case (disturbing them); singular verb agreement; no determiner, etc.

3. The unlawful disturbing of the peace is punishable by a fine.

"Disturbing" = Noun (NN): has PP "of" object (not direct object); has determiner; modified by adjectives rather adverbs, etc.

The difference between the nominal reading of 'The intentional disturbing of the peace is unforgivable' vs. the verbal reading of 'Intentionally disturbing the peace is unforgivable' is explicitly marked in both the POS tags and the syntactic annotation in the English Penn Treebank [6]. Not all instances are so clearly or easily distinguishable, however, even in English, and such distinctions are considerably less easily testable in Arabic.

In addition, this sort of problematic area is much more widespread in Arabic than in English. For example, the "maSdar" (roughly similar to a verbal noun/gerund) has been tagged in two ways in the ATB: both as a verb (heading a VP) or as a noun (heading an NP). Likewise, "The active participle can function syntactically as a noun, verb or attributive adjective" ([10], cited in [18], p. 102). And even beyond these categories, there is further ambiguity between noun and adjective in Arabic.

This problem has been recognized both in the descriptive and computational literature. For example, [18] (p. 255) cites [5]: "One cannot establish for Arabic a word class of adjectives, syntactic considerations being the only identificatory criterion of an adjective." Similarly, [8] writes, with regard to her POS tagger, a comment that could hardly be more relevant: "The overall performance on the nouns and adjectives is relatively high. However, confusing these two categories is almost always present due to the inherent ambiguity. In fact, almost all Arabic adjectives could be used as nouns in Arabic" (her footnote: "This inherent ambiguity leads to inconsistency in the ATB gold[2] annotation").

The automatic POS changes that were done to change NOUN to DV, or ADJ to DV, or ADJ to NOUN, were simply a way to get around this problem by forcing the tag to be the same as its syntactic function. For example, the 343 cases of the "NOUN to DV" have to do with the maSdar functioning as a verb. The full POS tag assigned to it should therefore be something like MASDAR-DV (with the DV as the function tag), indicating that it is a maSdar functioning in a verbal context (and heading a VP). Similarly, there could also be (for the same word), a MASDAR-NOUN tag, indicating that it heads a NP, or AP-ADJ (active participle heading an ADJP), and so on. So this group of automatic pos was essentially just deriving the function tag for the POS.

Our plan is to solve this entire problem of such high ambiguity by adding function tags to the "POS tags" in future annotation. These function tags will indicate the syntactic function of the word. The morphological tag itself (that is, without the new function tag) will eventually be closer to the more traditional Arabic terms for categories such as the maSdar.

4. Parsing results

There are 152 files in the ATB3a-v2.6 release. In order to do a comparison with the ATB3-v2.0 release, we also used the older (2.0) version of those same 152 files. The files were broken up into three sections, as typically done, with 80% for training, and 10% each for development and testing. All of the results are for the development section. For comparison with earlier work[3], and for reasons of speed, we report parsing results for sentences of length <=40 words. A summary of the data used for the parsing experiments is shown in Table 2. The first number is the number of trees; the second is the number of tokens. We explain below what ATB2.6-modified is.

[2] Here, Diab's use of the term "gold" annotation refers simply to the (human) annotation in the publicly released version of the corpus, which received the usual amount of quality control. However, internally, the project refers to "gold" annotation as adjudicated and multiply corrected annotation with a higher than usual degree of consistency.

[3] The parser is [7] the Bikel Statistical Parsing Engine, available at http://www.cis.upenn.edu/~dbikel/software.html#stat-parser. For details on how it was adapted for Arabic, see [11].

Data	Total *#trees(#tokens)*	Train	Dev	Dev(<=40)	Test
ATB2.0	3154(99103)	2556(79222)	291(10211)	208(5058)	307(9670)
ATB2.6	3348(100815)	2712(80641)	307(10353)	220(5017)	329(9821)
ATB2.6-modified	3096(90335)	2505(72275)	287(9323)	209(4746)	304(8737)

Table 2. Size of subcorpora: trees (tokens)

Run#	Recall/Precision/F-measure	#nulls	data	Note
#1	73.6/76.7/75.1	1	2.0	
#3	72.0/76.3/74.1	3	2.6	
#5	74.9/77.4/76.2	1	2.6-modified	
#4	75.6/76.4/74.4	2	2.6	parser forced to use given tags
#6	76.6/79.7/78.1	3	2.6-modified	parser forced to use given tags

Table 3. Parsing results on experimental runs

The parsing results for several experimental runs are shown in Table 3:

Run#1 - The parser was trained and tested on the ATB-v2.0 corpus, using just the 152 files that correspond to the files in the ATB3-v2.6 corpus. The parsing setup was identical to that described in [11]. The 79.2 score reported in that paper was for all of the ATB3-2.0, roughly three times the size of the ATB3-2.6 corpus.

Run#3 - The corresponding results on the 152 files in the ATB3-v2.6 release.

Run#5 - The corresponding results on the 152 files in the ATB3-v2.6 after their POS tags have been modified, as described below.

Runs#4 and #6 correpond to Runs #3 and #5, respectively, except that the parser is forced to use the given tags as the only possible tags for the word, instead of being allowed to explore parsing possibilities with all of the tags that it had seen during training for a word.

The immediately striking thing of course is that there is actually an initial decrease between the 2.0 and 2.6 data. This is almost certainly because, as mentioned above, the POS tags were not modified in the 2.6 release, resulting in an increase in a large number of "mismatches" between the POS tags and the tree structure. For example, it was previously the case that prepositional nouns were annotated as prepositions (PREP), heading a prepositional phrase (PP). Under the revised guidelines, these prepositionals are now considered as nouns (NOUN) heading noun phrases (NP). While the tree annotation was changed to an NP in the intermediate 2.6 release to reflect this, the POS tag has not yet likewise been changed. Since the POS tags are used by the parser as a way to bootstrap the parsing process, this "mismatch" between the POS tag and the syntactic node category causes a decrease in the parsing accuracy.

Therefore we implemented a procedure to automatically modify the POS tags appropriately, resulting in the changes listed earlier in Table 1. We used a set of head rules, as commonly used to break down the structure of a tree. However, we modified the head rules to include in them the option to change a POS tag if an appropriate head had not been found. We also departed from the usual sort of head rules for parsing in that we did not choose a default item (e.g., the leftmost or rightmost word in a phrase) if an appropriate head was not found. This is because we wanted to determine how many instances there were of parent for which a reasonable head could not be found, even with the potential POS changes, as a way to locate annotation errors in the Treebank.

The head rules were used 111,812 times during the processing of the 3348 trees, and there were 297 instances in which a head could not be found. While these might be errors in the head rules, a spot-check indicated errors in the trees, and so these trees were thrown away for further processing, which resulted in

the elimination of 252 trees. The 2.6-modified data is therefore the 2.6 corpus with the 3781 POS tag changes, and with the 252 trees with invalid trees deleted. Of course, in the future these trees will be modified to be free of errors.

Two examples of the head rules are:

Head rule for adjective phrases (ADJP):

{{ADJP},{LEFT_TO_RIGHT, ADJ, ADJ_NUM, ADJ_COMP, CHANGE_NOUN_TO_ADJ, EMPTY_TAG, ADJP}}

Head rule for verb phrases (VP):

{{VP},{LEFT_TO_RIGHT, IV, PV, CV, IV_PASS, PV_PASS, CHANGE_NOUN_TO_DV, CHANGE_ADJ_TO_DV,CHANGE_SUB_CONJ_TO_PSEUDOVERB, CHANGE_PART_TO_PSEUDOVERB, CHANGE_NEG_PART_TO_PV_OR_PSEUDOVERB,VP}}

The idea is that given a Parent and Children in the tree, if the parent matches the first leftmost, then the children will be searched from left to right for a match with one of the rightmost categories. So for a tree fragment (ADJP child1 child2), child1 and then child2 will be checked to see if they match ADJ, and if so then that child is taken to be the head. If not, then child1 and then child2 are checked to see if they are ADJ_NUM, etc. The "heads" such as CHANGE_NOUN_TO_ADJ are special tags that do what one would guess from the name. So, e.g., for the tree fragment (ADJP NOUN), since NOUN does not match ADJ, ADJ_NUM, or ADJ_COMP, then it "matches" CHANGE_NOUN_TO_ADJ, and it gets changed to an ADJ, now the head of the ADJP.

As noted, there is at first a decrease in the parsing score using the 2.6 data (runs 1 and 3), and this was not unexpected, due to the increase in tree/POS mismatches. However, there is indeed a decent increase in the parsing score with the new tags, from 74.1 to 76.2 (runs 3 and 5). These runs were performed using the parser mode in which the parser was free to choose its own tags for each word. Runs 4 and 6 were performed using the mode in which the parser is forced to use the given tags. There is virtually no difference between runs 3 and 4, the two runs for the original (i.e., the released) 2.6 corpus, while there is a nice difference for the modified one (runs 5 and 6). This of course makes sense if the tags are more appropriate.

5. Conclusions

We have discussed some of the issues that arise when the Arabic Treebank syntactic annotation is manually enhanced as the first step, ahead of the morphological/Part-of Speech annotation. We outlined an automatic procedure that more closely aligns the POS tags and the Treebank annotation, leading to increased parsing results and additionally providing the annotation pipeline with improved error checking and quality control. A further increase in parsing results was obtained by forcing the parser to use the given tags resulting from this procedure, thus indicating the important role that a POS tagger would play in a full Arabic NLP pipeline. In future work, we intend to investigate whether certain of the tags may be more crucial for the parser to get right. It seems reasonable that many of the "function word" particles are particularly crucial.

Finally, a new division of the POS analysis marking both morphological form and POS function is proposed. Our plan is to solve the problem of enormously high functional ambiguity among POS tags in Arabic by adding function tags to the "POS tags" in future annotation. These function tags will indicate the syntactic function of the word. The morphological tag itself (that is, without the new function tag) will eventually be closer to the more traditional Arabic terms for categories such as the maSdar.

Acknowledgement

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The content of this paper does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] *Arabic Treebank: Part 3 v. 2.0*. (2005). Mohamed Maamouri, Ann Bies, Hubert Jin, Tim Buckwalter. LDC Catalog No.: LDC2005T20.

[2] *Arabic Treebank: Part 3(a) v. 2.6*. (2007). Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki. LDC Catalog ID: LDC2007E65.

[3] *Arabic Treebank Morphological and Syntactic Annotation Guidelines*. (2008). http://projects.ldc.upenn.edu/ArabicTreebank/. Linguistic Data Consortium, University of Pennsylvania.

[4] Badawi, E; M.G Carter and A. Gully. (2004). *Modern Written Arabic: A Comprehensive Grammar*. Routledge, London.

[5] Beeston, A.F.L. (1968). *Written Arabic: An Approach to the Basic Structures*. Cambridge University Press, Cambridge, UK.

[6] Bies, A., Ferguson, M., Katz, K. and MacIntyre, R (Eds.) (1995). *Bracketing Guidelines for Treebank II Style*. Penn Treebank Project, University of Pennsylvania, CIS Technical Report MS-CIS-95-06.

[7] Bikel, D. (2004). On the Parameter Space of Generative Lexicalized Statistical Parsing Models. Ph.D. Dissertation. University of Pennsylvania.

[8] Diab, Mona. (2007). Improved Arabic Base Phrase Chunking with a new enriched POS tag set. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Association for Computational Linguistics 2007, Prague, Czech Republic.

[9] Diab, M., Hacioglu, K. and Jurafsky, D. (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. In *Proceedings of HLT-NAACL 2004.*

[10] Holes, Clive. (1994). *Modern Arabic: Structures, functions and varieties*. Longman.

[11] Kulick, S., Gabbard, R. and Marcus, M. (2006). Parsing the Arabic Treebank: Analysis and Improvements. Treebanks and Linguistic Theories 2006.

[12] Maamouri, M. and Bies, A. (2004). Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of COLING 2004*. Geneva, Switzerland.

[13] Maamouri, M. and Bies, A. (To appear). "The Penn Arabic Tree Bank." In Ali Farghaly and Karine Megerdoomian, Eds. *Computational Approaches to Arabic Script-Based Languages: Current Implementations in Arabic NLP*. CSLI NLP Series.

[14] Maamouri, M. and Cieri, C. (2002). Resources for Arabic Natural Language Processing at the Linguistic Data Consortium. In *Proceedings of the International Symposium on Processing of Arabic*. Faculté des Lettres, University of Manouba, Tunisia.

[15] Maamouri, M., Kulick, S. and Bies, A. (2008). Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of LREC 2008*.

[16] Marcus, M., Kim, G., Marcinkiewicz, M, MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*, San Francisco.

[17] Marcus, M., Santorini, B. and Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19.

[18] Ryding, Karin C. (2005). *A Reference Grammar of Modern Standard Arabic* (Reference Grammars). Cambridge University Press, Cambridge UK/ New York.