# Creating a Methodology for Large-Scale Correction of Treebank Annotation: The Case of the Arabic Treebank

**Mohamed Maamouri, Ann Bies and 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 LDC Arabic Treebank team has significantly revised and enhanced its annotation guidelines and annotation procedures over the last two years, with the goal of reducing inconsistency in annotation in the Treebank. We have now completed automatic and significant manual revisions to 738,845 tokens/words in total, bringing them into line as far as possible with the new annotation guidelines and greatly improving the annotation consistency. We created a methodology for large-scale correction of Treebank annotation during the course of this revision process, balancing the need for consistency with tight time constraints for correcting and updating a large amount of data annotated according to previous guidelines. The combination and interleaving of automatic and manual corrections were crucial to the success of the overall revision. We also demonstrate the success of the revision by reporting on an improvement in parsing results.

## Introduction

The LDC Arabic Treebank team has significantly revised and enhanced its annotation guidelines and annotation procedures over the last two years, with the goal of reducing inconsistency in annotation in the Treebank. We have now completed automatic and significant manual revisions to all of ATB1[1], ATB2[2] and ATB3[3] (738,845 tokens/words in total), bringing them into line as far as possible with the new annotation guidelines[4] and greatly improving the annotation consistency.

We created a methodology for large-scale correction of Treebank annotation during the course of this revision process, balancing the need for consistency with tight time constraints for correcting and updating a large amount of data annotated according to previous guidelines. The combination and interleaving of automatic and manual corrections were crucial to the success of the overall revision. This paper describes the correction process, the scope of correction that can be done in this way, and the type of correction that cannot. We also demonstrate the success of the revision by reporting on an improvement in parsing results.

## The Arabic Treebank

The Penn Arabic Treebank (ATB) began in the fall of 2001 (Maamouri and Cieri, 2002) and in five years has completed numerous full releases of morphologically and syntactically annotated data[5]. The ATB corpora are annotated for morphological information, part-of-speech, English gloss (all in the "part-of-speech" or "POS" phase of annotation), and for syntactic structure (similar to Treebank II style, Marcus et al., 1993; Marcus et al., 1994; Bies et al., 1995). 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 (Maamouri and Bies, 2004). 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 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.

A comprehensive description is given in Maamouri and Bies (2004) of 'Modern Standard Arabic' (MSA) as the language mostly targeted by Arabic NLP research. The Penn Arabic Treebank has therefore so far focused primarily on Arabic newswire text. This paper does not address the question of diacritization directly, but for a complete discussion of vocalization in the Arabic Treebank, see Maamouri, Kulick and Bies 2008. Syntactic clitics affecting the tree were separated after POS tagging and prior to Treebanking, resulting in an increase in the number of tokens in the Treebank data.

[1] LDC2008E61 - Arabic Treebank Part 1 v 4.0

[2] LDC2008E62 - Arabic Treebank Part 2 v 3.0

[3] LDC2008E22 - Arabic Treebank Part 3 v 3.1

[4] http://projects.ldc.upenn.edu/ArabicTreebank/

[5] Work on additional corpora, both MSA and dialectal, is on-going.

| Corpus | Source Tokens | Tokens after Clitic Separation |
|---|---|---|
| ATB1: AFP | 145,386 | 167,280 |
| ATB2: Umaah | 144,199 | 169,319 |
| ATB3: Annahar | 339,722 | 402,246 |
| ATB123 Total | 629,307 | 738,845 |

Table 1. Arabic Treebank newswire corpora sizes

Treebank annotation starts from real-life ‘raw’ data – Modern Standard Arabic or MSA newswire text without any diacritics. The Arabic Treebank corpora are annotated for morphological information, parts of speech (POS), case and mood marking, and English gloss, all in the POS phase of annotation, and for syntactic structure in the TB phase. Both annotation phases are based on semi-automatic outputs, namely (a) a set of alternative morphological analyses provided by the morphological analyzer and (b) a tree skeleton proposed by the Bikel parsing engine (publicly available at http://www.cis.upenn.edu/~dbikel/software.html).

### Choice of morphological annotation style

The output from the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) was used as the starting point for the morphological annotation and POS tagging of Arabic newswire text. For each input string, the Analyzer provides a fully vocalized solution (in Buckwalter Transliteration), including the word’s unique identifier or lemma ID, a breakdown of the constituent morphemes (prefixes, stem, and suffixes), and their POS values and corresponding English glosses.

### Choice of syntactic annotation style

When the Penn Arabic Treebank project began in 2001, we had to choose a style of syntactic annotation, while operating under considerable time and funding constraints. We considered both using a traditional/descriptive Arabic grammar style and at the same time, the overall structure of the Penn Treebank style. Annotating according to traditional/descriptive Arabic grammar, we thought, would have the advantage of being a familiar task for the Arabic-speaking annotators. However, the style, categories, and distinctions would be unfamiliar to most non-Arabic speaking researchers in the field, and there would be a considerable learning curve for these researchers to be able to use any traditional/descriptive-style annotated data.

In addition, we believed that well-educated and proficient Arabic speakers/readers could learn to operate within the Penn Treebank system as adapted to represent the structure of Arabic. Our syntactic annotation guidelines for Arabic are based on a firm understanding and appreciation of traditional Arabic grammar principles. The annotation our annotators produce should be as accurate and informative as any annotation that might be possible within the traditional Arabic grammar context, but it is more accessible to the research community in the Penn Treebank annotation style.

## The Revision Process

The overall guidelines revision process was initiated in 2006 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.

The revision process began with a complete revision of the annotation guidelines and specifications for both morphological/POS annotation and syntactic/Treebank annotation. More complete and detailed annotation guidelines overall were developed, and a period of intensive annotator training focusing on the new guidelines and on specific inconsistently annotated constructions followed. The revised 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 scores to 94.3 f-measure (Maamouri, Bies and Kulick, to appear; Maamouri, Bies and Kulick, 2008; *Arabic Treebank Morphological and Syntactic Annotation Guidelines,* 2008). With guidelines and training complete, we began the process of revising the annotation in the three existing ATB1, ATB2 and ATB3 corpora to bring existing data into compliance with revised guidelines specifications.

As noted above, Penn Arabic Treebank annotation consists of two phases: (a) Morphological/Part-of-Speech (=POS) tagging which divides the text into lexical tokens and includes morphological, morphosyntactic and gloss information, 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. (similar to the Penn English Treebank II style) (Marcus, et al., 1994; Marcus, et al., 1993; Bies, et al., 1995).

The tokens used for analysis are different for the two levels of annotation. For the morphological level, the tokens are the whitespace- and punctuation-delimited words from the source text, which receive a morphological analysis. These tokens may then be split up for the treebanking level of analysis, in order to provide access to the clitics that receive analysis in the tree. For example, the token لكتابة “lktAbp” from the text might receive the morphological analysis "li/PREP+kitAbap/NOUN", which would then be split up into two separate tokens ("li" and "kitAbap") at the treebanking level, in order to analyze the syntactic role of the preposition and noun separately.

Our revision process involved significant changes to the trees, both at the word level and in the syntactic structure, and an important aspect of the process was implementing the changes at the Treebank level while maintaining the logical connection to the level of full morphological analysis.

The five major steps in the correction process are shown in Table 2.

| Stage | Type |
|---|---|
| 1. Complete manual revision of trees according to new guidelines | Human only |
| 2. Limited manual correction of targeted POS tags | Human, based on automatic identification |
| 3. Revision of targeted tokenization and POS tags according to new guidelines, based on purely lexical information | Automatic only |
| 4. Revision of targeted tokenization and POS tags according to new guidelines, based on tree structure information | Automatic, based on human trees |
| 5. Corrections based on targeted error searches | Human, based on automatic identification |

Table 2. Correction Stages

Stage 1 focused on a human revision of all of the trees. Stages 2 through 4 focused on revising lexical information, based in part on the new tree structures, using a combination of automatic and manual changes. Stage 5 focused on error searches targeting both lexical information and tree structures.

## Stage 1: Manual revision of trees

Since so many of the subsequent correction processes depend on syntactic trees correctly annotated according to the newly revised guidelines, the necessary first step in the revision process was a complete manual revision of every tree in Stage 1.

In order to address concerns such as the inconsistent annotation of quantifiers, the decision was made to subordinate semantic needs to syntactic needs in certain constructions, for example, iDAfa with quantifiers (Maamouri, Bies and Kulick, 2008). As the iDAfa structure is a particularly frequent noun phrase structure, this decision affects the annotation of a significant portion of the corpus. In iDAfa structures syntactically headed by common nouns, the semantic and syntactic head of the noun phrase will be the same noun (as in the "grammar book" example below, where "book" is both the semantic and the syntactic head of the noun phrase).

```
(NP كتاب kitaAbu book
    (NP نحو naHowK grammar))
```

كتاب نحوٍ

*(a) grammar book*

vs.

```
(NP every –kul~u –كُلُّ
    (NP collection majomuwEapK مَجْمُوعَةٍ))
```

كُلّ مَجْمُوعَةٍ

*every collection*

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

## Stage 2: Manual correction of targeted POS tags

In Stage 2, specific tokens that were particularly ambiguous with respect either to multiple POS tags or to tokenization were revised by hand.

A careful review of the POS values of particular Arabic words in our guidelines revision process led to a change in the possible POS values for these words in the morphological analyzer and in our POS annotation. A human annotation pass was necessary to bring the POS tags into alignment with the revised guidelines when the choice of one POS value or tokenization over another could be determined only through context. The words which were targeted for this human correction at Stage 2 are

- wa و *and,*
- fa ف *and/so,*
- Hat~aY حَتَّى *until/up to; and/and even; including; so that*
- mA مَا *what, which, how, that, not, some, not be*
- layosa لَيْسَ *be not*
- <il~A إلّا *except, except for*
- <i* إذ *then; and then, as, when, while, (and) suddenly, (and) all of a sudden*
- turaY تُرَى *I wonder; think/see/believe*
- EadaY عَدَى *except, except for*
- >akovar أكْثَر *more, most, majority*
- >agolab أغْلَب *most (of), the majority (of), the greater portion of part (of)*

Every instance of these words in the full ATB123 corpus was targeted.

A listing of the possible POS values of each of the above words was determined and information provided to annotators, including a complete description of specific occurrences, Arabic reference terms whenever available, and syntactic and semantic contexts along with tests and appropriate POS and Treebank annotation specifications. These short annotation documents were given to annotators before the start of each correction cycle and used in mini-training sessions aimed at reinforcing targeted correct annotation by use of examples in text and trees and discussions of each of the above individual POS.

For instance, the revised morphological annotation guidelines and the SAMA morphological analyzer (Maamouri, et al., 2009) provide eight different POS values for mA مَا which are distinguished by semantic function and syntactic context.

(a) **mA values in SAMA**

1. mA/REL_PRON *what/which*
2. mA/NEG_PART *not*
3. mA/INTERROG_PRON *what/which*
4. mA/SUB_CONJ *that/if/unless/whether*
5. mA/EXCLAM_PRON *what/how*
6. mA/NOUN *some*
7. mA/VERB *not be*
8. mA/PART [*discourse particle*]

Two of the possibilities are mA as a negative particle (NEG_PART) and mA as a relative pronoun (REL_PRON). These two values occur in clearly different syntactic environments, with very different meanings.

(b) **mA=REL_PRON**
لِيَحْصُلَ على ما يَسُدُّ رَمَقَهُ
li+yaHoSula ElaY **mA** yasud~u ramaqa+hu
for+gets (he) **what** fill breath of life+his
*in order for him to get what he really craves*

(c) **mA=NEG_PART**
ما زالَ حَيّاً إلى الآنَ
**mA** zAla Hay~AF <ilaY Al|na
**not** finished (he) alive until the+now
*He doesn't cease to be alive now*

Since the POS tag of mA in cases like the above example is context dependent, human annotation was necessary to choose the correct POS tag.

## Stage 3: Automatic revision of targeted tokenization and POS tags based on lexical information

In Stages 3 and 4, POS tags and tokenization were revised automatically, based on either lexical or a combination of lexical and syntactic information. The guideline revisions specified changes at various different levels, including tokenization and POS tags as well as the trees. As a result, there were cases in which the existing tokenization and POS annotation were inconsistent with the revised morphological guidelines.

Implementing these changes required a reorganization of the corpus. The reason for this is that it was not possible to make automatic lexical revisions only by examining individual tokens in the Treebank, since such tokens may themselves be part of a larger original token. For example, while "limA*A" لِمَاذا formerly existed in the ATB3-v2.0 corpus both as a single token and also split into two tokens ("li" and "mA*A"), in the revised morphological guidelines it is now treated as one token only. However, the annotation as it existed in the ATB3-v2.0 corpus for the two-token analysis had already split up the word, and the individual Treebank tokens "li" and "mA*A" were both acceptable tokens unto themselves. It was only in the context of being part of a larger original word that it could be recognized that they needed to be merged back together for this revised release.

Therefore, we created a version of the corpus which associated each original token from the source text file with the one or more Treebank tokens that together make up that original token. We further characterized all original tokens for, roughly speaking, the "function words" that were the focus of the POS revisions. That is, all such tokens were automatically identified in terms of potential component morphemes and possible POS tags for each morpheme.

We then used this characterization of all the original tokens in order to modify the tokenizations to match the new guidelines. For example, all cases of the original token limA*A, whether previously split or not as Treebank tokens, were now given the same tokenization (a single Treebank token, unsplit).

This corpus reorganization allowed us to automatically implement the vast majority of tokenization decisions based only on the lexical information, without needing to refer to the syntactic tree. In many cases it was also possible to automatically modify the POS tag as well, without reference to the tree. For example, the POS tag for naHow نحـو *towards* has now been automatically revised to be NOUN instead of PREP, and this change is based only on the lexical item itself, not the tree structure.

## Stage 4: Automatic revision of targeted tokenization and POS tags based on lexical and tree information

As just discussed, to a certain extent the categorization of all the original tokens for both tokenization and possible POS values allowed us to automatically implement tokenization and POS changes based only on the lexical information. This was not always the case however, and some changes required taking into account the manually revised trees from Stage 1.

For example, both the tokenization and POS decision for the token fymA فيما were dependent on the tree annotation. fymA has two possibilities: [fiy,PREP] + [mA,REL_PRON] and [fiymA,SUB_CONJ].

In other cases the tokenization is unambiguous but the POS is ambiguous. For example, if the original token is bmA بما, the tokenization is unambiguously bi+mA, and the POS tag for bi is PREP, the POS tag for mA may be REL_PRON, INTERROG_PRON, or SUB_CONJ. The automatic assignment of a revised POS tag for the mA in the bmA cases was resolved by examining the local tree structure.

In Table 3 we show some of the most frequent cases of original tokens with alternative solutions that have the same vocalization but differ in tokenization and/or POS tags. bmA is ambiguous only for the POS tags, while the others are ambiguous both for the tokenization and the POS tags.

| **Original unvocalized token** | **Possible vocalization/POS alternatives** | **Count in ATB123** |
|---|---|---|
| <nmA or AnmA<br>إنما انما | <in~amA/RESTRIC_PART | 138 |
| | <in~a/PSEUDO_VERB+mA/REL_PRON | 2 |
| fymA<br>فيما | fiy/PREP+mA/REL_PRON | 14 |
| | fiymA/SUB_CONJ | 256 |
| kmA<br>كما | ka/PREP+mA/REL_PRON | 233 |
| | ka/PREP+mA/SUB_CONJ | 125 |
| | kamA/CONJ | 398 |
| bmA<br>بما | bi/PREP+mA/REL_PRON | 232 |
| | bi/PREP+mA/SUB_CONJ | 15 |

Table 3. Examples of tokens ambiguous for tokenization or POS tags

## Stage 5: Manual corrections of automatic search results

In Stage 5 of our revision process, we significantly improved the post-annotation quality control (QC) process for the ATB. The QC process consists of a series of specific searches targeting several types of potential inconsistency and annotation error, and we increased the number of error searches threefold during the revision process. These error searches are run after annotation is complete, and any errors found via these searches are hand corrected.

A certain residual type of correction is not possible in this context, however: corrections that require too much human decision to be made automatically, but that are too frequent or otherwise too time-consuming to be made manually. The highly complex and very frequent noun (NOUN) vs. adjective (ADJ) distinction in Arabic is an example of just this issue. Time and funding allowing, a manual revision of these cases in the Arabic Treebank will be undertaken in the future, using an appropriate combination of automatic and manual means.

# Parsing Results

An important goal is to evaluate the increase in parser accuracy as a result of the revisions described in this paper, and to compare the current accuracy to that of parsing on a more established source, namely the Wall Street Journal portion of the English Penn Treebank (PTB). Using a previously proposed data split[6] we trained and tested on each of the revised ATB1, ATB2 and ATB3 individually, as well as the combined ATB123 (738, 845 tokens/words in total), for both the newly revised versions and the older pre-revision releases. In addition, we have trained and tested on a comparable amount of PTB data for ATB3, since it is the largest of the three revised ATB corpora, as well as for the combined ATB123.

The parser used was the Bikel adaptation of the Collins parser[7]. We ran the parser in two modes. In both, the parser input contains the gold Part-of-Speech tags. The dev section results in Table 4 are for the mode in which the parser used the given tags only for words with which it was unfamiliar from training, and otherwise was free to choose its own tags. In the second mode, shown in Table 5, the parser was forced to use the given tag for each word.

| | **Old** | **New** | **PTB** |
|---|---|---|---|
| ATB1 | 78.0 | 83.5 | n/a |
| ATB2 | 79.5 | 81.8 | n/a |
| ATB3 | 77.5 | 81.0 | 87.6 |
| ATB123 | 78.8 | 82.7 | 88.6 |

Table 4. Parser results, with parser choosing its own tags

| | **Old** | **New** | **PTB** |
|---|---|---|---|
| ATB1 | 78.2 | 84.5 | n/a |
| ATB2 | 78.6 | 83.2 | n/a |
| ATB3 | 78.5 | 83.2 | 87.2 |
| ATB123 | 79.1 | 84.1 | 88.8 |

Table 5. Parser results, with parser forced to use given tags

[6] http://nlp.stanford.edu/software/parser-arabic-data-splits.shtml

[7] http://www.cis.upenn.edu/~dbikel/software.html#stat-parser

As can be seen, for both Table 4 and 5 there is significant improvement in the score for the parser with the revised data, roughly halfway bridging the gap to the PTB score. It is perhaps of note as well that there is a greater distinction in the two ways of running the parser for ATB as compared to PTB. This is perhaps indicative of greater tree/tag consistency in the ATB, or perhaps of a greater share of the burden put on the POS tags. This is a matter for further study, but in both parser modes there is noteworthy improvement for the new scores for the revised data compared to the old scores.

In order to better understand the source of the parser improvement, we performed a dependency analysis, as was also done in Kulick, Gabbard, Marcus (2006). Each parser output tree and corresponding gold tree is broken down into a collection of relations, which is a one-level slice of the context-free tree. We have selected some of the most frequent relations for ATB123 and categorized them into two groups, shown in Tables 6 and 7. In both tables the columns are (1) the relation, (2) the frequency of that relation in the new ATB123, and (3) the scores for ATB123-old, ATB123-new, and PTB.

| **Relation** | **% of all relations in ATB123-new** | **ATB123-old f-measure** | **ATB123-new f-measure** | **PTB f-measure** |
|---|---|---|---|---|
| NP → NOUN NP | 16.75 | 90.4 | 97.4 | n/a |
| PP → PREP NP | 13.40 | 96.5 | 99.2 | 95.2 |
| Base NP | 12.71 | 84.1 | 90.2 | 95.0 |
| VP → verb NP | 11.59 | 92.1 | 94.1 | 93.3 |
| SBAR → compl S | 2.59 | 91.1 | 92.9 | 92.0 |
| S → NP VP | 2.03 | 87.4 | 91.3 | 96.3 |

Table 6. Parser accuracy on core syntactic structure relations

| **Relation** | **% of all relations in ATB123-new** | **ATB123-old f-measure** | **ATB123-new f-measure** | **PTB f-measure** |
|---|---|---|---|---|
| VP → VERB PP | 6.44 | 82.6 | 83.4 | 83.5 |
| NP → NPB PP | 3.49 | 73.3 | 75.7 | 86.2 |
| NP → NP PP | 1.77 | 33.5 | 45.0 | n/a |

Table 7. Parser accuracy on PP attachment relations

Table 6 shows relations that make up what might be called the core syntactic structures. For example, the relation NP → NOUN NP is the right-branching structure, as in the iDAfa construction. PP → PREP NP is the relation of the object of a preposition to the PREP, and so on. A "base NP" (NPB) is an NP without another NP inside it. This table demonstrates the improvement in the parser recovery of these core relations, which seems strongly indicative of increased Treebank internal consistency. One place where there is certainly room for improvement is with the S → NP VP relation, in which it seems likely that the parser is getting confused over the optionality of the subject placement in Arabic.

Table 7 shows the three relations having to do with PP attachment. Here, while the score for attachment of a PP modifier to a VP is nearly identical to that of the PTB, the score is significantly lower for PP attachment to a NPB, and the very low scoring relation for PP attachment to a NP does not even exist in the PTB. The low scoring for the PP attachment to NP relation is no doubt because of the impact of the iDAfa annotation upon the PP attachment problem, as has been discussed in the literature (see e.g., Kulick, Gabbard, Marcus, 2006; Gabbard and Kulick, 2008).

## Conclusions

In a two-year effort to increase the consistency of the Arabic Treebank, the LDC research and annotation teams have significantly revised and enhanced ATB annotation guidelines and annotation procedures. The revised and enhanced morphological/part-of-speech and syntactic guidelines have been used in automatic and manual revisions in a large-scale correction of data that was already annotated according to previous guidelines in order to provide a resource of the full ATB123 that is consistent with the newly revised annotation specifications. Our combined automatic and manual revision procedure allowed us to bring this data into compliance with the revised annotation specifications as closely as possible, and also provided the annotation pipeline with better error checking and quality control for future annotation.

The revisions described above led to a substantially improved corpus, and the combination and interleaving of automatic and manual corrections backed by improved and clearer guidelines were crucial to the revision process. The overall success of this revision process has also been confirmed by a corresponding increase (5.1 absolute f-measure) in parsing accuracy.

Continued future work on combining automatic and manual annotation and correction methods is expected to lead to further improvements in corpus consistency, inter-annotator agreement and parsing results.

## Acknowledgements

We would like to thank our Arabic Treebank annotators, Fatma Kaddeche, Basma Bouziri, Badia Laadaoui and Luma Ateya, and also Wajdi Zaghouani for their contribution.

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

*Arabic Treebank Morphological and Syntactic Annotation Guidelines*. (2008). Mohamed Maamouri, Ann Bies, Sondos Krouna, Fatma Gaddeche, Basma Bouziri. http://projects.ldc.upenn.edu/ArabicTreebank/. Linguistic Data Consortium, University of Pennsylvania.

Ann Bies, Mark Ferguson, Karen Katz and Robert MacIntyre (Eds.). 1995. *Bracketing Guidelines for Treebank II Style*. Penn Treebank Project, University of Pennsylvania, CIS Technical Report MS-CIS-95-06.

D. Bikel. 2004. On the Parameter Space of Generative Lexicalized Statistical Parsing Models. Ph.D. Dissertation. University of Pennsylvania.

Tim Buckwalter. 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. LDC Catalog No.: LDC2004L02.

Ryan Gabbard and Seth Kulick. 2008. Construct State Modification in the Arabic Treeban*k. In Proceedings of ACL-08, Short Papers.*

Seth Kulick, Ryan Gabbard and Mitch Marcus. 2006. Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of Treebanks and Linguistic Theories 2006*.

Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In *Proceedings of COLING 2004*. Geneva, Switzerland.

Mohamed Maamouri, Ann Bies, Hubert Jin, and Tim Buckwalter. 2005. *Arabic Treebank: Part 3 v. 2.0*. LDC Catalog No.: LDC2005T20.

Mohamed Maamouri, Ann Bies and Seth Kulick. To appear. Upgrading and Enhancing the Penn Arabic Treebank: A GALE Challenge. In *GALE Project Publication*.

Mohamed Maamouri, Ann Bies and Seth Kulick. 2008. Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of LREC 2008*.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri. 2008. *Arabic Treebank part 1 - v4.0*. LDC Catalog No.: LDC2008E61. Special GALE release to be followed by a full LDC publication.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri. 2009. *Arabic Treebank part 2 - v3.0*. LDC Catalog No.: LDC2008E62. Special GALE release to be followed by a full LDC publication.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri. 2009. *Arabic Treebank part 3 - v3.1*. LDC Catalog No.: LDC2008E22. Special GALE release to be followed by a full LDC publication.

Mohamed Maamouri and Christopher Cieri. 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.

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Seth Kulick. 2009. *LDC Standard Arabic Morphological Analyzer (SAMA) v. 3.0*. LDC Catalog No.: LDC2009E44. Special GALE release to be followed by a full LDC publication.

Mohamed Maamouri, Seth Kulick and Ann Bies. 2008. Diacritic Annotation in the Arabic Treebank and its Impact on Parser Evaluation. In *Proceedings of LREC 2008*.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the Human Language Technology Workshop*, San Francisco.

Mitch Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, Vol. 19.