In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010).

From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News

Mohamed Maamouri, Ann Bies, Seth Kulick, Wajdi Zaghouani, David Graff, Michael Ciul
Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
E-mail: {maamouri,bies,skulick,wajdiz,graff,mciul}@ldc.upenn.edu

Abstract

The Arabic Treebank (ATB) Project at the Linguistic Data Consortium (LDC) has embarked on a large corpus of Broadcast News (BN) transcriptions, and this has led to a number of new challenges for the data processing and annotation procedures that were originally developed for Arabic newswire text (ATB1, ATB2 and ATB3). The corpus requirements currently posed by the DARPA GALE Program, including English translation of Arabic BN transcripts, word-level alignment of Arabic and English data, and creation of a corresponding English Treebank, place significant new constraints on ATB corpus creation, and require careful coordination among a wide assortment of concurrent activities and participants. Nonetheless, in spite of the new challenges posed by BN data, the ATB's newly improved pipeline and revised annotation guidelines for newswire have proven to be robust enough that very few changes were necessary to account for the new genre of data. This paper presents the points where some adaptation has been necessary, and the overall pipeline as used in the production of BN ATB data.

1. Introduction

The Arabic Treebank (ATB) Project (Maamouri and Bies, 2004) at the Linguistic Data Consortium (LDC) has embarked on a large corpus of Broadcast News (BN) transcriptions, and this has led to a number of new challenges for the data processing and annotation procedures that were originally developed for Arabic newswire text (ATB1[1], ATB2[2] and ATB3[3]). The corpus requirements currently posed by the DARPA GALE Program, including English translation of Arabic BN transcripts, word-level alignment of Arabic and English data, and creation of a corresponding English Treebank[4], place significant new constraints on ATB corpus creation, and require careful coordination among a wide assortment of concurrent activities and participants.

Nonetheless, in spite of the new challenges posed by BN data, the ATB's newly improved pipeline and revised annotation guidelines for newswire (Kulick, Bies and Maamouri, 2010; Maamouri, Bies and Kulick, 2009; Maamouri, Bies and Kulick, 2008) have proven to be robust enough that very few changes were necessary to account for the new genre of data. This paper presents the points where some adaptation has been necessary, and the overall pipeline as used in the production of BN ATB data.

2. Issues of Broadcast News

2.1 Metadata, Speech Effects

Unlike newswire data, BN transcripts include metadata in several forms to convey several kinds of information in addition to the text of what each speaker is saying. Some forms of metadata have no relevance to treebank annotation and must be ignored, such as indications of coughs, laughter, background noise or music. Some forms may have relevance or impact for treebanking, despite being unrelated to the grammar of the spoken message, such as indications of discourse markers, hesitation sounds, word fragments, mispronunciations and other disfluencies: because these are part of what is spoken, their presence must be acknowledged in treebank annotation, in such a such a way that every verbalized token in the transcript has a coherent and appropriate annotation label, identifying how the token functions within the utterance as a whole. Even when tokens carry no semantic or syntactic value, their distribution needs to be known in order for machine learning algorithms to build higher-level models from speech data. Then there are types of metadata that determine which portions of a BN recording can be addressed using the MSA-based ATB annotation conventions: notations indicating that the speech in a given region is in a language other than Arabic, or that the speaker is using a colloquial dialect of Arabic rather than MSA.

2.2 Indistinct Audio Signal

Another problem is that the audio signal is sometimes indistinct: yet another form of metadata is the use of double-parentheses to allow the transcriber to indicate that speech could be heard but not understood, or could

[1] LDC2008E61 – Arabic Treebank Part 1 v 4.0
[2] LDC2008E62 – Arabic Treebank Part 2 v3.0
[3] LDC2008E22 – Arabic Treebank Part 3 v3.1. As of this writing, ATB3-v3.2 is scheduled for publication in April 2010, LDC Catalog Number: LDC2010T08.
[4] LDC2009E55 – English Translation Treebank Part 3 v2.0, for example, is the pre-existing English-Arabic Treebank BN data release that the ATB5 data was selected to parallel.

only be understood or guessed at from context rather than from the audio signal.

When some portion of an utterance is not recoverable from an audio recording, this will tend to have a cascading impact on higher-level annotations. Even when the loss is relatively small, affecting only a few words that are inferable from context, the annotation must somehow convey the fact that it is not the audio signal that accounts for the linguistic information in that region.

3. Tool Development for ATB BN Data

The tools for processing and annotation in the ATB data pipeline had to be adapted to filter out the metadata that ATB would ignore, while preserving the ability to align the annotation results to the initial transcripts. The other metadata that would be useful or required in ATB annotation had to be retained in a manner that would inform but not obstruct or overly complicate the annotation tasks, and would support verifiable alignment and quality control. In addition, while using the annotation tool for the initial stage, selecting the correct vocalization of the undiacritized transcripts and assigning part-of-speech labels to disambiguate the text, annotators also had access to the original audio files when necessary, which is to say, when the POS annotators needed to listen to the audio in order to disambiguate doubtful words in the transcript or to recognize and confirm that a token, which could be otherwise fine, is in fact a typo.

For example:

- Transcribed typo "zbr"[5] زبر 'to prune' in place of "brz" برز 'to appear,' or
- Transcribed typo "lmE" لمع 'to shine' in place of "Elm" علم 'to learn'

4. Guidelines Development for BN Data

Aside from the extra challenges posed by the nature of BN transcripts, the ATB team has adapted the Penn English Treebank Switchboard annotation guidelines (Taylor, 1996; Bies et al., 1995) for use with Arabic BN data. As the Switchboard Bracketing Guidelines focus on the treatment of speech effects, disfluencies and metadata, which is not language-specific, that methodology could be adopted fairly straightforwardly. In addition, specific dialect-related structures were addressed, so that the occasional dialect speech (in field interviews, or other less highly monitored speech that occurs within the BN) could be consistently annotated as well (Maamouri et al., 2009c). For the annotation of the syntactic structures in general, the revised and enhanced Arabic Treebank Syntactic Guidelines[6] were followed (Maamouri et al., 2008). This has led to a treebank annotation procedure that improves the overall consistency of annotation.

5. ATB Annotation Pipeline

The ATB annotation and processing pipeline has been improved overall, and has also been adapted to support the production of treebanked broadcast news corpora such as the Arabic Treebank part 5 - v1.0 (LDC Catalog No. LDC2009E72), roughly 100K words of Broadcast News from Aljazeera, Dubai and Alhurra News (Maamouri et al., 2009a), and for all BN corpora following this.

Several components of the pipeline are devoted to the handling of word forms that fall outside the vocabulary and grammatical repertoire of SAMA (Kulick, Bies and Maamouri, 2010), including feedback to upgrade its lexicon and morphotactic tables (Maamouri et al., 2009b), and careful vetting of POS labels and glosses assigned to novel terms.

5.1 Speech Transcription and SU Annotation

The current pipeline shown in Figure 1 begins with the transcription process, which uses the LDC's "XTrans" transcription tool and creates one tab-delimited-format (tdf) file for each BN recording, with one phrasal "semantic unit" (SU) per time-stamped region of audio.

The transcription guidelines[7] describe how the audio should be segmented into time-stamped regions to identify "sentence units" (SUs), how these units should be labeled, what punctuation to use, and what sorts of additional metadata need to be included in the Arabic orthographic transcription (for things like noises, foreign words and phrases, mispronunciations, etc.).

Considerable attention in the guidelines was given to identifying the SUs, segmenting them coherently, and assigning final punctuation to indicate their type (statement, question, or incomplete). The SU decisions made by transcribers needed to be held firm throughout all subsequent stages of annotation, because two or more independent downstream annotations needed to be done in parallel, rather than serially. In particular, translation of the Arabic transcripts into English (and treebanking of the English[8]) was done in a separate pipeline, which ran independently from (and concurrent with or prior to) ATB morpho-syntactic annotation. In order to maintain a consistent SU segmentation across annotation projects, Arabic and English Treebank annotators did not alter the

[5] Throughout this paper we use the Buckwalter transliteration http://www.qamus.org/transliteration.htm

[6] For a more complete description of the revised annotation policies, see *Arabic Treebank Morphological and Syntactic Annotation Guidelines*. http://projects.ldc.upenn.edu/ArabicTreebank/.

[7] http://projects.ldc.upenn.edu/gale/Transcription/Arabic-XTransQRTR.V3.pdf

[8] Such as LDC2009E55 – English Translation Treebank Part 3 v2.0, for example.

pre-existing SU annotation.

Of course, despite best intentions, transcribers would sometimes make mistakes in SU segmentation, through either fatigue/inattention, or being unaware of subtle factors affecting treebank annotation. This, like obscured speech in the audio signal, has a cascading effect on the final result.

5.2 Morphological Analyzer and Morphological/ Part-of-Speech Annotation

The completed but undiacritized transcripts are then processed through the Standard Arabic Morphological Analyzer SAMA (Maamouri et al., 2009b), an expansion of the Buckwalter Arabic Morphological Analyzer used in previous ATB corpora, to list, for each Arabic word token, all known/possible annotation solutions, with assignment of all diacritic marks, morpheme boundaries (separating clitics and inflectional morphemes from stems), and all Part-of-Speech (POS) labels and glosses for each morpheme segment.

The novel properties of BN transcripts (in contrast to newswire data) involved a couple of issues: (a) watching out for "out-of-band" characters that would never occur in newswire, such as the Persian character "keheh" being used mistakenly for the MSA letter "kaf" (because the two have the same shape in some contexts); and (b) making sure that the AG-based stand-off annotation skips over the metadata annotations (foreign words, tags that mark regions of colloquial Arabic, etc.). These needed to be resolved in a manner that would not risk disrupting the integrity of the source transcript, and thereby jeopardizing the ability to sustain cross-references between ATB and other, parallel annotations.

After an AG XML file has been created with possible solutions for each word included from SAMA, it is given to an annotator using the SelectPOS tool for selecting morphological/part-of-speech analysis (referred to together as POS for ATB).

The input to SelectPOS is a set of solutions generated by SAMA, the Standard Arabic Morphological Analyzer. The SAMA tool makes use of very high quality data about Modern Standard Arabic, which has been verified multiple times for correctness. SelectPOS aims to relate this data to the text, and improve on it where the correct analysis for a word is not available. Everywhere possible, SelectPOS attempts to limit data entry to values that could possibly be correct. This means to avoid requiring the annotator to type in new data, and to force elements of solutions such as number of segments to be consistent with each other.

The annotator selects a solution for each word, making note of problems along the way. The output is then prepared for the parsing step.

In this morphological stage of annotation, if the correct solution for a word is missing from SAMA, the annotator has no choice but to mark the word as a "NO_MATCH," indicating that no solution is available. After SelectPOS annotation is completed, a separate "NO_MATCH" tool is used to fill in annotations of words for which there was no correct SAMA solution. This process allows for a limited or pending annotation to be entered for words without a SAMA solution, and these annotations are carefully tracked and flagged for possible later integration into SAMA (see Kulick, Bies and Maamouri (2010) for details). Tokens having a DIALECT tag are by definition not in SAMA (since SAMA includes Modern Standard Arabic only), and in the current pipeline, these tokens are not further analyzed unless they include a clitic that must be separated for syntactic annotation (see section 5.3 below). However, DIALECT tokens will be analyzed in the future when the project begins to prioritize Broadcast Conversation data, in which a higher rate of dialectal Arabic occurs (with an expected rate of approximately 50% of the tokens).

A new version of the SelectPOS annotation tool is currently in development that will allow for proposed solutions to be entered on the first POS annotation pass for NO_MATCH tokens, and a second pass will be possible within the same tool.

5.3 Clitic Separation, Parsing, and Syntactic Annotation

Once the POS annotation is done, the clitics are separated automatically according to the tags provided by the POS annotation, in order to prepare the segmentation necessary for the treebanking phase. Next, the data is parsed using Dan Bikel's parsing engine[9] (Bikel, 2004), and presented to Treebank annotators using the LDC TreeEditor Annotation tool to correct the parse output and add function tags and empty categories.

The clitics are separated based on a simple algorithm that selects the various "core" POS tags from the morphological analysis resulting from the POS annotation. For example, a token that received the analysis

> kutub/NOUN/books + i/CASE_DEF_GEN/def.gen
> + hi/POSS_PRON_3MS/its-his

is broken up into two tokens for treebanking:

> kutub/NOUN/books + i/CASE_DEF_GEN/def.gen

and

[9] The Bikel Statistical Parsing Engine, available at: <http://www.cis.upenn.edu/~dbikel/software.html#stat-parser>

hi/POSS_PRON_3MS/its-his

See Kulick, Bies and Maamouri (2010) for detail related to this splitting of tokens.

A dialect token in the current pipeline that includes a clitic will also be split, so that the syntactic annotation can be completed fully. The clitic receives the necessary POS tag (and vocalization), but the remaining dialect token has the POS tag DIALECT. For example, the dialectal token "wrAH" is analyzed as

wa/CONJ/and +rAH/DIALECT/(he) went, started

and is split into two tree tokens for treebanking:

wa/CONJ/and

and

rAH/DIALECT/(he) went, started

Once the tokens are separated into the tokens for treebanking, the Bikel parser is used to automatically create syntactic trees for treebanking. See Kulick, Gabbard, Marcus (2006) for a description of the modifications of the parser as used for parsing Arabic in this pipeline. The "gold" POS tags resulting from the POS annotation, as split for the treebank tokens, are used as input to the parser along with the "unvocalized" form of the token, which is simply the vocalization with the diacritics stripped out. (See Kulick, Bies and Maamouri (2010) for more information about this distinction.)

In the next step of the annotation process, treebank annotators correct the parser output in accordance with the syntactic annotation guidelines for the project. This annotation step includes:

1. The correction when necessary of the constituents and attachment structure provided by the parser.
2. The insertion of function tags not included by the parser, and the correction when necessary of function tags included in the parser output. The parser currently includes a subset of all the possible function tags, including SBJ, CLR, TPC, and OBJ.
3. The insertion of empty categories with appropriate co-indexing. The parser does not currently include empty categories in its output.

While the parser gets the "unvocalized" tokens as input, as mentioned above, the resulting trees are simply overlaid on top of the complete morphological analysis for each token. Therefore, the treebank annotators have access to the full morphological analysis of each token, together with the parse tree output.

The Treebank annotation tool itself (LDC's TreeEditor) is a simple graphically-based tree annotation tool, which displays the tree using the "vocalized" transliterated tree tokens and allows the annotators to manipulate the tree in the necessary ways. The tool also displays the full morphological analysis, the Arabic script source tokens and the English gloss for each token as separate listings for the annotators' convenience.

It is also occasionally the case that treebank annotators will wish to modify an earlier morphological analysis, in order to be consistent with the desired syntactic annotation. This may be a simple change in the POS tag, or a more substantial change which may therefore require adjustment of the tokenization. The TreeEditor tool allows the annotators to make these modifications in a limited format.

Annotators also mark speech disfluencies (repetitions and restarts, etc.) as they appear in the trees, according to the BN syntactic annotation guidelines.

5.4 Quality Control Searches and Corrections

Finally, quality control (QC) passes are performed to check and correct any error of annotation in the trees. The Corpus Search tool[10] is used with a set of 93 error-search queries to locate and index a range of known problems involving improper patterns of tree structures and node labels. Once this indexing is done, each of the affected files goes through a manual pass using LDC's TreeDiag annotation tool to seek and repair the problems. TreeDiag is a version of the TreeEditor tool with a "diagnostic mode" that displays the search results and allows the annotators to click through directly to the affected portion of each tree.

Throughout the pipeline, there are numerous stages and methods of sanity checks and content validation, to assure that annotations are coherent, correctly formatted, and consistent within and across annotation files, and to confirm that the resulting annotated text remains fully concordant with the original transcripts, so that cross-referential integrity with the original speech data and with English translations is maintained.

[10] CorpusSearch is freely available at: <http://corpussearch.sourceforge.net/>

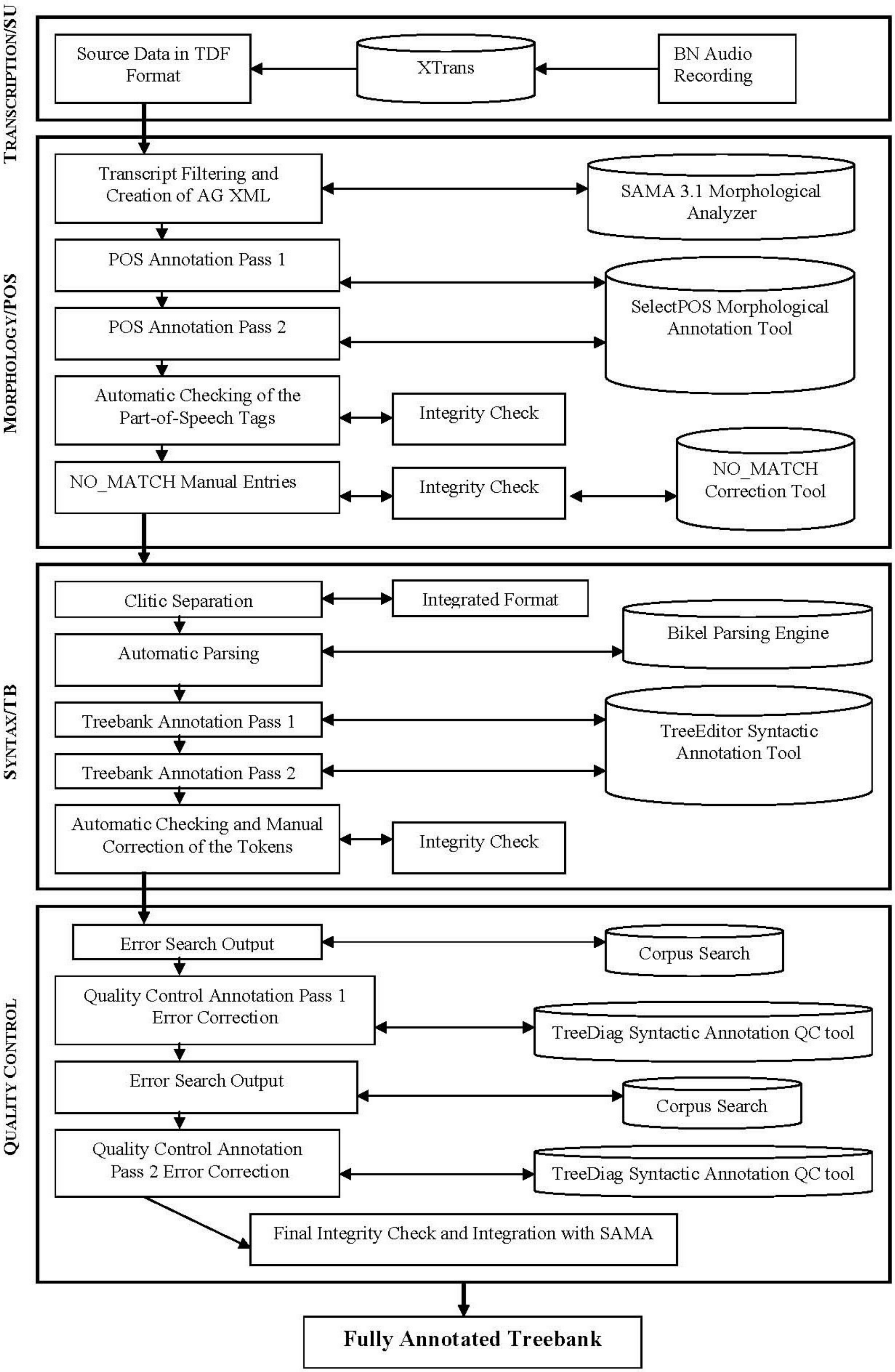
TRANSCRIPTION/SU
Source Data in TDF Format
XTrans
BN Audio Recording
MORPHOLOGY/POS
Transcript Filtering and Creation of AG XML
SAMA 3.1 Morphological Analyzer
POS Annotation Pass 1
POS Annotation Pass 2
SelectPOS Morphological Annotation Tool
Automatic Checking of the Part-of-Speech Tags
Integrity Check
NO_MATCH Manual Entries
Integrity Check
NO_MATCH Correction Tool
SYNTAX/TB
Clitic Separation
Integrated Format
Automatic Parsing
Bikel Parsing Engine
Treebank Annotation Pass 1
Treebank Annotation Pass 2
TreeEditor Syntactic Annotation Tool
Automatic Checking and Manual Correction of the Tokens
Integrity Check
QUALITY CONTROL
Error Search Output
Corpus Search
Quality Control Annotation Pass 1 Error Correction
TreeDiag Syntactic Annotation QC tool
Error Search Output
Corpus Search
Quality Control Annotation Pass 2 Error Correction
TreeDiag Syntactic Annotation QC tool
Final Integrity Check and Integration with SAMA
Fully Annotated Treebank

Figure 1. The Arabic Treebank Annotation Pipeline

6. Conclusion

In spite of the new challenges posed by Broadcast News data, the ATB's newly improved pipeline and revised annotation guidelines have proven to be robust enough that very few changes were necessary to account for the new genre of BN data. We have presented the ATB annotation pipeline and addressed the points where adaptation was necessary to accommodate BN data. Similar adaptations will be made in the future to account for additional new data genres (such as webtext and dialectal speech), and it is hoped that the current pipeline will continue to prove flexible and robust enough to accommodate the morphological and syntactic annotation of the necessary data.

7. Acknowledgements

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.

We would also like to thank the Arabic Treebank annotators for their many contributions.

8. References

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.

Ann Bies, Justin Mott, Colin Warner. (2009). English Translation Treebank, Part 3 v2.0 (EATB BN). LDC Catalog ID: LDC2009E55.

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

Seth Kulick, Ann Bies and Mohamed Maamouri. (2010). Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010).

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

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, Fatma Gaddeche, Sondos Krouna, and Dalila Tabessi Toub. (2009c). *Guidelines for Treebank Annotation of Speech Effects and Disfluency for the Penn Arabic Treebank, v1.0.* http://projects.ldc.upenn.edu/ArabicTreebank/. Linguistic Data Consortium, University of Pennsylvania.

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

Mohamed Maamouri, Ann Bies, and Seth Kulick. (2009). Upgrading and enhancing the Penn Arabic Treebank: A GALE challenge. In Joseph Olive (Ed.), *In progress for publication (book describing work in GALE program).*

Mohamed Maamouri, Ann Bies and Seth Kulick. (2008). Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (LREC 2008), Marrakech, Morocco, May 28-30, 2008.

Mohamed Maamouri, Ann Bies, Seth Kulick and Fatma Gaddeche. (2009a). Arabic Treebank part 5 - v1.0. Linguistic Data Consortium, CatalogID: LDC2009E72.

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.

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.

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.

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Seth Kulick. (2009b). Standard Arabic Morphological Analyzer (SAMA) Version 3.1. Linguistic Data Consortium, Catalog No.: LDC2009E73.

Ann Taylor. (1996). *Bracketing Switchboard: An addendum to the TREEBANK II Bracketing Guidelines*. Penn Treebank Project, University of Pennsylvania.