Enhancing the Arabic Treebank: A Collaborative Effort toward New Annotation Guidelines

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

Abstract

The Arabic Treebank team at the Linguistic Data Consortium has significantly revised and enhanced its annotation guidelines and procedure over the past year. Improvements were made to both the morphological and syntactic annotation guidelines, and annotators were trained in the new guidelines, focusing on areas of low inter-annotator agreement. 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 f-measure scores already and has also improved parsing results.

1. Introduction[1]

The Arabic Treebank (ATB) team at the Linguistic Data Consortium (Maamouri and Bies, to appear) has significantly revised and enhanced its annotation guidelines and procedure over the past year. 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 f-measure scores already and has also improved parsing results.

2. Motivation

The 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.

Many of the inconsistencies derived from an improper partitioning of the work between different levels, both conceptually and in the actual annotation procedure. Conceptually, subordinating syntactic to semantic needs in certain constructions led to inconsistencies in annotation, as different annotators gave higher priority to one or the other. For example, a quantifier-noun sequence such as "every collection" in Arabic is traditionally expressed in terms of an إضافة/idafa construction, in which the noun is considered a complement of the quantifier, which itself is treated as a noun:

```
(NP every/all/each_one |-kul~u | - كُلُّ
    (NP collection/group | majomuwEapK | مَجْمُوعَةٍ))
```

However, in earlier ATB work, this structure was treated as flat

```
(NP every/all/each_one|-kul~u|- كُلُّ
    collection/group |majomuwEapK|مَجْمُوعَةٍ)
```

in order to make what is often thought of as the "semantic head" (here, "collection") more easily accessible to users. However, annotators applied both interpretations, and such structures were inconsistently annotated in the Treebank.

The resolution of this type of inconsistency among others led to substantial revision of the annotation guidelines.

3. Improvements to Annotation Guidelines and Procedures

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.

In the actual annotation procedure, we have made two major decisions regarding the morphological/part-of-speech (POS) tags, both in an effort to ensure that the POS tags are more helpful for the syntactic annotation. First, as a practical matter, the Treebank annotators are now able to correct the POS tags chosen at the previous level of annotation, removing a source of a significant amount of previously-identified "mismatch" between the POS tags and syntactic structure.

[1] 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.

Second, we have overhauled some key aspects of the POS guidelines, such as making new tags for comparatives and quantifiers, since tagging these simply as NOUN, was not informative enough for either the Treebank annotators or the parser.

Both POS and Treebank guidelines were revised in several respects, balancing the goals of (1) representing more finely-grained distinctions, and (2) aligning more closely with traditional grammar concepts already familiar to annotators.

3.1 Morphological/Part-of-Speech Level

The POS tags for nouns and adjectives in particular were revised to be more fine-grained. In addition to NOUN_PROP (proper name), the core POS tag of NOUN is now further distinguished as

- NOUN (common noun)
- NOUN_NUM (number)
- NOUN_QUANT (quantifier)

The core POS tag of ADJ is also further distinguished as

- ADJ (common adjective)
- ADJ_NUM (ordinal number)
- ADJ_COMP (comparative adjective)

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 of those lexical items now being categorized as nouns, or "prepositional nouns"), 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 four different POS tags available, following its four traditional categories:

(a) CONJ for *fa Al-EaTf*/فاء العطف (the *fa* of coordination), used for the coordination of words and sentences, marking a temporal sequence between them and glossed as *and*

(b) CONNEC_PART for *fA' Al-rabT/*فاء الربط (the *fa* of connection), used to introduce the comment after the focus particle >*am~A*/أمّا and glossed as *well (then)*

(c) RC_PART for *fa Al-jazA'*/ *فاء الجزاء* (the *fa* of reward, response conditional), used in conditional constructions in the main clause to introduce the result of the preceding conditional clause, and glossed as *then* or *so*

(d) SUB_CONJ for *fa Al-sababiy~ap*/فاء السببية (the *fa* of causality), used to introduce a subordinate result clause

A new POS category of pseudo-verbs has been added, to account for the verbal behavior of certain Arabic particles. These are "the sisters of إِنَّ <inna" (with the exception of أَنَّ ">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.

3.2 Syntactic/Treebank Level

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).

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

grammar book

vs.

```
(NP every/all/each_one |-kul~u | -كُلُّ
    (NP collection/group | majomuwEapK | مَجْمُوعَةٍ))
```

However, in idafa structures that are syntactically headed by quantifiers (as in the "every collection" example above), the semantic head of the noun phrase is not the quantifier at all, but its complement noun. The interaction of this idafa structure with the new, more fine-grained POS tags allows the difference in semantic and syntactic heads to be captured. The syntactic/Treebank annotation is based on the syntactic head (the quantifier, "every"). However, the semantic head (the complement noun, "collection") is still easily accessible to end-users based on the POS tag NOUN_QUANT on the quantifier.

It has been a pleasant outcome that the interaction of the changes in POS and syntactic annotation results in an overall conceptual and practical improvement. For example, while we were driven to give primary concern to the syntactic aspect of the annotation, as in the quantifier example above, we were at the same time still concerned about losing the indication of the semantic head ("collection"). However, the more fine-grained nature of the POS tags resulting from our analysis of the annotation procedure means that this information is still easily available to the end user, since "every" would be clearly marked as a quantifier. Thus, a simple algorithm can recover the necessary semantic information, and inter-annotator agreement is higher.

As with the revision of POS guidelines, the revision of the syntactic annotation guidelines also served to more closely align the Treebank annotation with traditional Arabic grammar categories for several constructions. These include the treatment of comparatives, numbers and numerical expressions and the treatment of several particular pronominal constructions such as separating pronouns/Damiyr Al-faSl/ضمير الفصل and anticipatory pronouns/Damiyr Al$a>n/ضمير الشأن .

Further revisions include a more careful and complete classification of verbs and their argument structure and a thorough treatment of gerunds and participles. Gerunds and participles in Arabic, as in many other languages, often have a dual verbal and nominal role, behaving as verbs with respect to their objects (assigning accusative case, for example) and complements, but as nominals with respect to the rest of the sentence (occupying canonically noun phrase positions, for example, such as the subject of the sentence, or the complement of a preposition). When it arises, this dual nature is represented in the Treebank by annotating the gerund or participle as heading a VP with a complete internal argument structure, while at the same time labeling its parent S as nominal, or "S-NOM."

```
(S (VP rafaDat رَفَضَت
       (NP-SBJ Al+suluTAtu السُلُطاتُ)
       (S-NOM-OBJ (VP manoHa مَنْحَ
                      (NP-SBJ *)
                      (NP-DTV Al>amiyri الأمير
                              AlhAribi الهارب)
                      (NP-OBJ (NP jawAza جَوازَ
                                 (NP safarK سَفَرٍ))
                          (ADJP dyblwmAsy~AF
                                    ديبلوماسيّاً)
                                        )))))
```

رفضت السلطات **منحَ الأمير الهارب جوازَ سفرٍ ديبلوماسياً**

The authorities refused to give the escaping prince a diplomatic passport

This representation was part of the Treebank guidelines prior to the revisions for gerunds or participles with accusative complements (as in the example above), but the revised guidelines are more comprehensive with respect to the contexts in which the gerund or participle has a verbal reading. For example, a maSdar/gerund, active participle or passive participle followed by a PP complement to the regular verb form (PP-CLR) is now shown with a verbal reading.

maSdar/gerund with PP-CLR complement:

```
(S (VP <iHotafala إحْتَفَلَ
       (NP-SBJ Alfariyqu الفَريقُ )
       (PP-CLR bi بِ
              (S-NOM (VP fawzi فَوز
                         (NP-SBJ hi هِ )
                         (PP-CLR bi بِ
                           (NP ka>osi كَأْسِ
                             (NP Al>aboTAli الأبْطالِ)
                                        )))))))
```

إحتفل الفريق **بفوزه بكأس الأبطال**

The team celebrated winning its champions cup

Active participle with PP-CLR complement:

```
(NP (NP Almawoqifu المَوْقِفُ )
    (SBAR (WHNP-1 0 )
          (S (VP muEab~iru المُعَبِّرُ
                 (NP-SBJ-1 *T*)
                 (PP-CLR Ean عَن
                    (NP ra}iyi رأي
                      (NP Al>aglabiy~api الأغلَبِيَّةِ)
                                  ))))))
```

الموقف **المعبر عن** رأي الأغلبية

The attitude which shows the opinion of the majority

The verbal reading indicated by the PP-CLR complement can override other nominal indications, such as the presence of a determiner on a participial, in specific contexts. Participial relative clauses where the participle has a verbal reading are now shown as SBARs with null relative pronouns, even if the participle has a determiner. In this example, the verbal reading for the participial is forced by the PP-CLR complement, even though the determiner *Al* is present on the participial *AlmuEab~iru*.

```
(NP (NP Almawoqifu المَوْقِفُ )
    (SBAR (WHNP-1 0 )
          (S (VP AlmuEab~iru المعبر
                 (NP-SBJ-1 *T* )
                 (PP-CLR Ean عَن
                    (NP ra}iyi رأي
                      (NP Al>aglabiy~api الأغلَبِيَّةِ)
                                  ))))))
```

الموقف **المعبر** عن رأي الأغلبية

The position which shows the opinion of the majority

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

3.3 Corrections of Previous Annotation Level

The initial POS annotation is still selected from the morphologically analyzed alternatives provided by the Buckwalter morphological analyzer (BAMA 2004). However, crucial to reducing the number of mismatches between POS tags and syntactic structures is the ability of Treebank annotators to correct POS tags from the earlier annotation level. The available corrections include correcting the core POS tag (changing an active verb tag to a passive verb tag, for example), correcting tokenization errors, and correcting case endings. The annotation tool has been revised so that Treebank annotators now have the ability to correct case endings and specific POS tags such as CONJ → ADV or PREP → NOUN.

As reported in Maamouri, Bies and Kulick (2008), a number of experiments in automatically correcting POS tags along these lines have also been carried out, allowing for improved parsing results even before full hand-correction of the POS tags can be completed.

4. Improvements in inter-annotator agreement and training

Intensive annotator training focused on agreement and consistency and led to an improvement of inter-annotator agreement scores from an initial f-measure of 86.98% to the current f-measure of 94.3%.

The ATB production workflow includes both automatic and manual error correction, along with on-going annotator training, and it is hoped that these measures will continue to improve the agreement further. In order to maintain a high rate of inter-annotator agreement, approximately 10% of each corpus is dual blind annotated during production, put through the full workflow, and the final output is compared using evalb scores.

The initial agreement score was considered to be too low for the purpose of training statistical parsers on the ATB data. The goal was to approach the reported score of 93.8% for the Chinese Treebank. This goal has now been met or surpassed, and data produced with this level of agreement is expected to support on-going work on improving parsing results.

5. Error analysis and parsing improvement

As noted above, parsing results were significantly lower for the Arabic Treebank than for other treebanks of roughly similar size and complexity, the (English) Penn Treebank and the Chinese Treebank. Parsing work reported in Kulick, Gabbard, Marcus (2006) and related experiments have noted two particular problems for parsing the ATB: (1) inconsistent or unexpected Part-of-Speech tags for particular syntactic configurations, and (2) inconsistent annotation of syntactic structures. The analysis of these problems informed the early direction of the Treebank revision work described here, and continues to be at the core of the ongoing work.

One utility of Part-of-Speech tags is their value in “bootstrapping” the parsing process. It is therefore not surprising that inconsistent POS annotation makes parsing more difficult. For example, Kulick, Gabbard, Marcus noted that 5% of the VPs in the ATB have a head with a non-verbal tag (e.g., a VP headed by a mASdar/gerund, active participle or passive participle), and that changing the POS tags for such heads to a new tag (“DV”, automatically added for verbal readings of gerunds and participles) resulted in a 0.6 increase in the f-measure score.

Inconsistencies were also found in the annotation of a quantifier followed by a noun phrase. As noted above in Section 2, the annotation structure for this used to be a flat structure and now is a complement structure. In fact however, an analysis of the previous version of the Treebank showed that even with the older guidelines, 15% of such quantifier-NP sequences were annotated as idafa constructions, thus causing more problems for the parser.

Revised data following the improved guidelines and including automatic POS corrections as in Maamouri, Bies and Kulick (2008) is now in production. Using this revised and enhanced annotation as training data, there is in fact a preliminary improvement in statistical parsing results[2].

Initial parsing results for 100K words of the pre-revision Arabic Treebank Part 3 v. 2.0 were 75.1 f-measure. On the same 100K words after the annotation revision and the automatic POS corrections (Arabic Treebank Part 3(a) v. 2.6), the parsing score improves to 76.2 f-measure. Significantly, with more data (200K words of revised annotation and automatic POS corrections, in Arabic Treebank Part 3(b) v. 2.7), the score improves further to 79.7 f-measure. It is expected that on-going error detection, quality control and other work will improve the data further – and that using such further improved data will bring parsing results up to the immediate goal of matching the parsing results for the 230K word Chinese Treebank at 82.7 f-measure (Bikel, 2004).

The expected conclusion is that significant errors in annotation cause problems for the parser. While this indeed must be the case, it is noteworthy that many of these errors can in many cases be corrected via fairly simple automatic tree transformations, and in some cases already have been. This observation forms the core of the current analysis/revision work, in which tree fragments of different syntactic structures are being extracted and evaluated for consistency for tree structure, POS tags, and case information.

6. Conclusions

The initial application of the improved ATB annotation guidelines is to revise the annotation of the ATB3 corpus, a 350,000 word corpus of newswire data from *Annahar*. This revision is currently in progress. Revising the

[2] The parser is the Bikel Statistical Parsing Engine (Bikel 2004), available at http://www.cis.upenn.edu/~dbikel/software.html#stat-parser. For details on how it was adapted for Arabic, see Kulick, Gabbard and Marcus (2006).

annotation of this corpus to reflect the newly updated guidelines will provide a significantly improved resource to the community. Additional annotation of new data in the improved guidelines style will follow.

The improved ATB guidelines, improving inter-annotator agreement scores, and an expected continuing improvement in parsing scores are the result of a fruitful collaboration between data producers and end users, along with the support and time to effect the change. It is hoped that such collaboration will continue to benefit both annotation production and NLP applications in the future.

7. Acknowledgements

We would like to thank Fatma Kaddeche, Basma Bouziri and all of the Arabic Treebank annotators for their many contributions.

8. References

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

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

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

Arabic Treebank: Part 3(b) v. 2.7. (to appear, 2008). Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki.

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

Buckwalter, T. (2004). *Buckwalter Arabic Morphological Analyzer Version 2.0.* LDC Catalog No.: LDC2004L02.

Kulick, S., Gabbard, R. and Marcus, M. (2006). Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of Treebanks and Linguistic Theories 2006*.

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

Maamouri, M., Bies, A. and Kulick, S. (2008). Enhanced Annotation and Parsing of the Arabic Treebank. INFOS2008.