The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus

Mohamed Maamouri, Ann Bies, Tim Buckwalter, Wigdan Mekki
Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
maamouri@ldc.upenn.edu, bies@ldc.upenn.edu, timbuck2@ldc.upenn.edu, wmekki@ldc.upenn.edu

Abstract

From our three year experience of developing a large-scale corpus of annotated Arabic text, our paper will address the following: (a) review pertinent Arabic language issues as they relate to methodology choices, (b) explain our choice to use the Penn English Treebank style of guidelines, (requiring the Arabic-speaking annotators to deal with a new grammatical system) rather than doing the annotation in a more traditional Arabic grammar style (requiring NLP researchers to deal with a new system); (c) show several ways in which human annotation is important and automatic analysis difficult, including the handling of orthographic ambiguity by both the morphological analyzer and human annotators; (d) give an illustrative example of the Arabic Treebank methodology, focusing on a particular construction in both morphological analysis and tagging and syntactic analysis and following it in detail through the entire annotation process, and finally, (e) conclude with what has been achieved so far and what remains to be done.

1.0 Introduction

1.1 Why a '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. Many researchers in the field attest that fully automated fundamental Arabic NLP tools such as Base Phrase Chunkers are still not available for Arabic (Diab et al., 2004). On the other hand, there has been an increasing demand for high quality Arabic language resources and need for greater volumes of sophisticated annotated text in Arabic.

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 desperately needed linguistic resources in multiple areas of HLT research and development, including natural language processing, information extraction and summarization. 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.

1.2 Why an 'Arabic Treebank' at Penn?

The Linguistic Data Consortium (LDC) and the University of Pennsylvania have played a key role in the development, production and sharing of linguistic resources and treebanks in English, Chinese, and Korean. As of fall 2001, Arabic was added to that list with the creation of an Arabic Treebank team. This is not surprising, as Penn and LDC were a very appropriate birthplace and environment for this effort. They bring to bear a rich academic institutional framework and a unique experience in the creation of large-scale linguistic resources. The unique skills developed over the past decade, which proved very efficient once again in the Arabic Treebank experience, are embodied in such important principles as: (a) empirical methods providing portability to new languages, (b) a pragmatic mixture of manual, semi-automatic, and fully automatic annotation methods, and (c) robust tools such as morphological analyzers and parsers for bracketing annotation, which increasingly automate tasks and speed up the annotation process.

The Penn Arabic Treebank (ATB) began in the fall of 2001 (Maamouri and Cieri, 2002) and has now completed three full releases 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; (2) Arabic Treebank: Part 2 v 2.0, LDC Catalog No. LDC2004T02, roughly 144K words from Al-Hayat distributed by Ummah Arabic News Text (New features of annotation in the UMAAH corpus, so named as UMmah's Arabic Al-Hayat, include complete vocalization including case endings, lemma IDs, and more specific part-of-speech tags for verbs and particles.), and (3) Arabic Treebank: Part 3 v 1.0, LDC Catalog No. LDC2004T11, roughly 350K words of newswire text from An-Nahar morphologically annotated (150K of which have been treebanked in ATB: Part 3(a) v 1.1 LDC2004E71). The ATB corpora are annotated for morphological information, part-of-speech, English gloss (all in the "part-of-speech" phase of annotation), and for syntactic structure (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 is generalizable 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.

1.3 Arabic Language Issues

1.3.1 What is Modern Standard Arabic (MSA)?

A comprehensive description is given in Maamouri and Bies (2004) of 'Modern Standard Arabic' (MSA) as the 'language' mostly targeted by Arabic NLP research and, therefore, by the Penn Arabic treebank annotation which has so far only focused on Arabic newswire text. The term MSA is commonly used among linguists and computational linguists, although there is often little agreement on its definition. MSA, nobody's native or first language, though there exists a 'living' writing and reading MSA community, is mainly the language of written discourse and is used in formal communication both written and oral with a well-defined range of stylistic registers. A more convenient term than MSA would have been 'Modern Written Arabic' if it were not for the ambiguity of mixing together written MSA with written dialectal occurrences, though this mix is more and more evident mainly in MSA broadcast news (in all Arab countries) and sometimes in MSA newswire text (mostly in Egypt, Lebanon and a few other Middle-Eastern countries). Another term which is also appearing on the Arabic NLP scene is 'Modern Conversational Arabic.' Though perhaps useful and acceptable as a generic term, MCA is problematic because it lacks the required specificity to pin it down to one specific dialect identification, as there is no standard 'coverall' dialectal Arabic in the Arabic Language Continuum spectrum (as defined by Dell Hymes, 1973).

1.3.2 Impact of Arabic Language Specificities on Corpus Annotation

The description of Arabic language idiosyncrasies and their impact on the annotation process and methodology (Maamouri and Bies, 2004) can be summarized in the following points:

(1) Leaving out the short vowel markers, consonantal length (shadda), inflection and word-final case and mood markings is typical in most written Arabic. Vocalized MSA text is scare and limited to a small number of literary, religious or school-related titles.

(2) Most Arabic NLP applications seem to do away with all diacritics working from a graphic representation, which is stripped of many significant linguistic features mostly relating to grammatical marking.

(3) The reader reads the text and interprets its meaning by mentally providing the missing grammatical information (vocalization process) that leads to its/an acceptable interpretation. This amounts to an additional manual/human annotation with decisions that may have a non-trivial impact on the overall annotation routine in terms of both accuracy and speed.

(4) The graphemic representation of vocalization diacritics is not absolutely necessary for Treebank annotation. However, its presence completes the text and enhances the quality of the linguistic analysis of the targeted corpus. Morphological annotators provide a first reading and an interpretation of the bare text based on an internalized knowledge of the required vowels and case/mood endings ('mental vocalization'). They produce a vocalized output, shown as a text with full vocalic diacritics in the text box of the TreeEditor tool, for syntactic analysis. Syntactic annotators can either accept or challenge the interpretation shown.

(5) Providing a vocalized text for annotation will decrease the amount of additional ambiguity produced by the lack of grammatical and lexical markings. However, it should be made clear that Arabic, like any other language, will continue to have the usual amount of linguistic ambiguity.

(6) Since readers have to provide the missing MSA grammar towards understanding and annotating the targeted newswire corpora and since the level of internalized MSA grammar differs drastically sometimes from annotator to annotator, there is an added degree of grammatical inconsistency which will negatively impact inter-annotator agreement rates.

2.0 Methodological Choices

2.1 Issues of Data

We use newswire for many data production and annotation projects at LDC because it is easily available in electronic format and in significant volume, we have been able to develop IPR agreements that allow use of the data, and it is an on-going source of current, topical and new linguistic data. For Arabic in particular, it represents the bulk of written Arabic currently being produced (including the new lexical items that journalists must coin when faced with new realities), and using newswire avoids potential IPR and other issues that might arise with the use of religious, educational or literary texts. The data in our releases to this point is from Agence France Presse, Ummah (Al-Hayat), and An-Nahar.

Over time, we have developed a number of tools and pre-processing procedures that handle the technical issues involved with Arabic script, such as bidirectionality and ligatures.

2.2 Choice of Morphological Annotation Style

The output from the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002) is 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, as in the following example:

Example 1:

```
INPUT STRING: الغاز
SOLUTION 1: >alogAz
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN
     GLOSS: mysteries/enigmas
SOLUTION 2: >alogAzu
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN+u/CASE_DEF_NOM
     GLOSS: mysteries/enigmas + [def.nom.]
SOLUTION 3: >alogAza
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN+a/CASE_DEF_ACC
     GLOSS: mysteries/enigmas + [def.acc.]
SOLUTION 4: >alogAzi
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN+a/CASE_DEF_GEN
     GLOSS: mysteries/enigmas + [def.gen.]
SOLUTION 5: >alogAzN
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN+N/CASE_INDEF_NOM
     GLOSS: mysteries/enigmas + [indef.nom.]
SOLUTION 6: >alogAzK
  LEMMA_ID: lugoz_1
       POS: >alogAz/NOUN+K/CASE_INDEF_GEN
     GLOSS: mysteries/enigmas + [indef.gen.]
SOLUTION 7: AlgAz
  LEMMA_ID: gAz_1
       POS: Al/DET+gAz/NOUN
     GLOSS: the + gas
SOLUTION 8: AlgAzu
  LEMMA_ID: gAz_1
       POS: Al/DET+gAz/NOUN+u/CASE_DEF_NOM
     GLOSS: the + gas + [def.nom.]
SOLUTION 9: AlgAza
  LEMMA_ID: gAz_1
       POS: Al/DET+gAz/NOUN+a/CASE_DEF_ACC
     GLOSS: the + gas + [def.acc.]
SOLUTION 10: AlgAzi
   LEMMA_ID: gAz_1
        POS: Al/DET+gAz/NOUN+i/CASE_DEF_GEN
      GLOSS: the + gas + [def.gen.]
```

From 2002 to 2004 three corpora were analyzed and over half a million Arabic word tokens were annotated and tagged (see Table 1). The tagged AFP, UMAAH, and ANNAHAR corpora were published as "Arabic Treebank: Part 1 v 2.0" (Maamouri et al., 2003), "Arabic Treebank: Part 2 v 2.0" (Maamouri et al., 2004a), and "Arabic Treebank: Part 3 v 1.0" (Maamouri et al., 2004b), respectively, and are available from the LDC website <http://www.ldc.upenn.edu>

Corpus	Arabic Word Tokens
AFP	123,810
Ummah	125,698
Annahar	293,035
Total	542,543

Table 1: Arabic newswire corpora

The results of each pass were recycled through the system in order to fill gaps in the lexicon and make modifications to the POS tag set in order to meet the requirements of treebanking that was performed subsequently at the LDC with the same annotated and POS-tagged newswire data. The accuracy of the morphological analyzer output and the lexicon coverage statistics improved with each cycle (see Table 2).

Corpus	Accurate analyses
AFP	90.63%
Ummah	99.24%
Annahar	99.25%

Table 2: Arabic lexicon coverage statistics

Statistics were compiled of all the cases where the Morphological Analyzer failed to provide an accurate analysis. By far the most frequent problem (38% of cases) was the absence of non-Arabic proper names, place names, and company names (e.g., Andreotti, Zurich, Airlines). False-positives were also a recurring problem, as some foreign names are mistakenly identified as valid Arabic words, such as *huwa* (Ho) and *minhu* (Minh). Missing Arabic proper names (15% of cases) are also identified as common nouns (e.g., Adil, Ansari, Bani, Abbad). Incorrect vocalization (21%) typically involved failure to identify the passive voice or provide the proper verbal prefix or suffix. Cases of incorrect POS assignment (12%) usually involved tagging as ADJ items that also function as NOUN, such as *'amaliyya* ("practical", "operation") and *diblumasi* ("diplomat", "diplomatic"). Remaining problems involved improper English glosses (8%), missing Arabic common noun entries (3%), and typos in the original (3%).

The Morphological Analyzer algorithm itself underwent some changes in order to adapt to the various orthographic challenges posed by each of the three corpora being tagged. Initially, the Analyzer looked up orthographic variants of the input word only if the first lookup attempt resulted in a "not found." However, this approach does not work in cases where the misspelled word is a valid word, such as when the preposition *'ala* is spelled with *ya'* instead of *alif maqsura* (which is quite frequent now on Egyptian websites, especially that of *al-Ahram*). The initial version of the Analyzer would accept these words at face value and return the corresponding analyses. The UMAAH corpus in particular contained a large number of words that ended in *alif maqsura* but which had been spelled with *ya'* instead: some of these words could be analyzed on second lookup, and others generated a correct analysis for exactly what was written, but which was not useful for morphological tagging because these words had to be flagged as typos. To remedy this problem, we modified the Analyzer algorithm to look up two variant forms every time the input string ended in either *ya'* or *alif maqsura*. Consequently, the words *'ali* and *'ala* now generate identical analyses (although ordered differently).

2.3 Choice of Syntactic Annotation Style

When the Penn Arabic Treebank project began and we had to choose a style of syntactic annotation, we considered both using a traditional Arabic grammar style and using the Penn Treebank style. Annotating according to traditional Arabic grammar 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-style annotated data. In addition, as there have been no large-scale annotation projects in the traditional Arabic grammar style, we would need to develop and refine all guidelines from scratch. As speed was important to the project, we chose to take advantage of methodologies already in place for treebanks of other languages at Penn.

The long history in the NLP and computational linguistics communities of using the Penn Treebank annotation style (Marcus et al., 1993; Marcus et al., 1994; Xue, Chiou & Palmer, 2002; Kingsbury, Xue & Palmer, 2004; Han et al., 2001) led us to adopt a closely related style of annotation for the Penn Arabic Treebank. We were able to take advantage of guidelines already developed for several languages for questions of general structure and annotation policy (though of course, it was necessary to revise them to be appropriate for Arabic). There are also a number of processing tools that are already optimized to the Penn Treebank style of annotation and structure (for example, Dan Bikel's parsing engine (Bikel, 2002). There has been so much work in the area of automatic parsing and tagging using the Penn Treebank style (Chiang and Bikel, 2002; Brill 1993; Collins, 1997) that we were able to take advantage of the existing understanding of how to manipulate treebank structures and get results quickly.

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 the 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.

2.4 Treebank Annotation Specifications and Traditional Grammar Concepts and Rules

The question we had to face in the early stages of ATB was how to develop a Treebank methodology – an analysis of all the targeted syntactic structures – for MSA represented by unvocalized written text data (Maamouri and Bies, 2004; Fassi Fehri, 1993). Since all Arabic readers – Arabs and foreigners – go through the process of mentally providing/inserting the required grammatical rules which allow them to reach an interpretation of the text and consequent understanding, and since all of our recruited annotators are highly educated native Arabic speakers, we accepted going through our first corpus annotation with that premise. Our conclusion was that the two-level (morphological/POS and syntactic/TB) annotation was possible, but we noticed that because of the extra time taken hesitating about case markings at the TB level, TB annotation was more difficult and significantly more time-consuming. This led to including all possible/potential case endings in the POS alternatives provided by the morphological analyzer (Buckwalter, 2002). Our choice was to make the two annotation passes equal in difficulty by transferring the vocalization difficulty to the POS level. We also thought that it is better to localize that difficulty at the initial level of annotation and to try to find the best solution to it. So far, we are happy with that choice. We are aware of the need to have a full and correct vocalization for our ATB, and we are also aware that there is no extensive vocalized Modern Standard Arabic corpus available, except for the Koranic text, some classical literary landmarks, and most schooling materials below Grade 8. The challenge was and still is to find annotators with a very high level of grammatical knowledge in MSA, and that is a tall order here and even in the Arab region.

2.5 Training Annotators, ATB Annotation Characteristics

The two main factors, which affect the training of annotators in our ATB experience are both related to the specific 'stumbling blocks' of the Arabic language.

(1) The first factor, which affects annotation accuracy and consistency, pertains to the annotators' educational background (their linguistic 'mindset') and more specifically to their knowledge – often confused, and not clear – of the traditional MSA grammar. Some of the important obstacles to POS training come from the confusing overlap which exists between the morphological categories as defined for Western language description and the MSA traditional grammatical framework. The traditional Arabic framework recognizes three major morphological categories only, namely NOUN, VERB, and PARTICLE. This creates an important overlap which leads to mistakes/errors and consequent mismatches between the POS and syntactic categories. We have noticed the following problems in our POS training: (a) the difficulty that annotators have in identifying ADJECTIVES as against NOUNS in a consistent way; (b) problems with defining the boundaries of the NOUN category presenting additional difficulties coming from the fact that the NOUN includes adjectives, adverbials, and prepositionals, which could be formally nouns. In this case, the NOUN category then overlaps with the adverbs and prepositions of Western languages, and this is a problem for our annotators who are linguistically savvy and have an advanced knowledge of English and, most times, a third Western language; (c) particles are very often indeterminate, and their category also overlaps with prepositions, conjunctions, negatives etc.

(2) The second factor, which affects annotation accuracy and speed, is the behemoth of grammatical tests. Because of the frequency of obvious weaknesses among very literate and educated native speakers in their knowledge of the rules of inflection (i.e., case ending marking), it became necessary to test the grammatical knowledge of each new potential annotator, and to continue occasional annotation testing at intervals to maintain consistency. While we were able to take care of the first factor so far, the second one seems to be a very persistent problem because of the difficulty level encountered by Arab and foreign annotators alike vis-a-vis the use of case-ending rules.

3.0 Annotation Procedures

3.1 Part-of-Speech and Morphological Annotation

3.1.1 Pre-processing and tools

The current procedure for POS annotation includes the following steps: we begin by segmenting the raw input text and we apply the Buckwalter Arabic Morphological Analyzer to generate a list of candidates for each Arabic segment (i.e., token or word). Human annotators then go through the alternatives for each word and select the appropriate POS if it is present in the list provided (pass1). Next, human annotators check the work of other annotators (pass2). We may, as needed, perform a third pass to correct some particular errors and improve overall quality.

The morphological annotation process is performed by means of SelectPOS, an annotation application developed at LDC that displays the various morphology analysis solutions provided by the Morphological Analyzer. The human annotator must carefully review the available choices, and then accept one of the solutions, but only if it meets the following criteria: (a) the POS tag is correct; (b) the identified sequence of morphemes (word segmentation) is correct; (c) the vocalization (short vowels and diacritics) is correct; and (d) the English gloss is accurate.

For further details on the POS annotation system, see Maamouri and Cieri (2002).

3.1.2 Human Intervention is Necessary in POS

Human annotators perform the necessary task of disambiguating many orthographically identical forms. For example, active verbs may have the same input string as passive verbal forms, and prepositions cliticized with nouns (*bi*-noun) may have the same input string as pure nouns or verbs (noun or verb starting with *b*), as in Example 2 below. The morphological analyzer will give all of the possibilities allowed by the orthography (nine potential solutions, only two of which are shown below).

Example 2:

```
INPUT STRING: باسم
SOLUTION 1: bAsim
  LEMMA ID: bAsim_1
       POS: bAsim/NOUN_PROP
     GLOSS: Basem/Basim
SOLUTION 9: biAisomi
  LEMMA ID: {isom_1
       POS: bi/PREP+{isom/NOUN+i/CASE_DEF_GEN
     GLOSS: by/with + name + [def.gen.]
```

In this example, the correct choice is the proper name "bAsim" (SOLUTION 1) but the choice of "bi-{isom" with the prepositional clitic "bi" is also available (SOLUTION 9), since that is one of the possible analyses of the text string. The POS annotator must choose the correct analysis. An incorrect choice will lead to drastically different syntactic tree structures (a noun phrase vs. a prepositional phrase), and one of them is clearly (to the human reader or annotator) incorrect. At the moment, these are distinctions that are hard for automatic tools to make, so human annotation is necessary.

3.1.3 POS Annotator Decision Process

Annotators use the following criteria for making POS decisions:

- Correctness/acceptability is a decision concerning each of the following ordered set, provided by the Morphological Analyzer's output: (a) POS tag, (b) morphological segmentation, (c) vocalization including case and mood endings, and (c) English gloss
- If all four criteria are met in one of the displayed solutions, the annotator chooses that solution and is automatically moved on to the next item in the displayed text.
- If the first three criteria are acceptable but the English gloss is defective, the annotator may still choose the appropriate solution, but should also choose the "Gloss Problem" option and type an explanation in the Comment field (e.g., "Gloss Problem: should be…").
- If any of the first three criteria is unacceptable (i.e., wrong POS tag, or wrong segmentation, or wrong vocalization), the annotator must choose the "No match" option, and then enter the appropriate explanation in the Comment field (e.g., "Adj should be Noun").
- When no solution options are provided by the morphological parser, the annotator has two options: (1) if the word is a proper name, the annotator chooses the "X-Solution" option, which displays one or more morphology analyses based on prefix/suffix analyses only, and (2) if the word is *not* a proper name, the annotator chooses the "No match" option and enters the appropriate explanation in the comment field. In these cases the annotator should attempt to provide all the information that the parser did not, namely: the POS, morphological segmentation, and vocalization of the word.

Our plan is to develop and train an automatic morphological/POS tagger for Arabic in the near future, so that the initial selection will be made automatically, and the annotators can switch to a correction task. The decision process will remain the same, but it is hoped that the automatic tagger will get a significant number of the tags right.

3.2 Treebank Annotation

3.2.1 TB Pre-Annotation Processing

Our annotation procedure is to use the automatic tools we have available to provide an initial pass through the data. This allows the annotators to focus on correcting the automatic output.

Once POS annotation is done using the SelectPOS tool, clitics are automatically separated based on the POS selection in order to create the segmentation necessary for treebanking. Then, the data is automatically parsed using

Dan Bikel's parsing engine for Arabic. Treebank annotators correct the automatic parse and add semantic role information, empty categories and their coreference, and complete the parse. The annotation is done using the TreeEditor tool developed at LDC. After that is done, we check for inconsistencies between the treebank and POS annotation. Many of the inconsistencies are corrected manually by annotators or automatically by script if reliably safe and possible to do so.

3.2.2 On Cliticization

The prevalence of cliticization in Arabic sentences of determiners, prepositions, conjunctions, and pronouns led to a necessary difference in tokenization between the POS files and the treebanking files. Clitics that play a role in the syntactic structure are split off into separate tokens (e.g., object pronouns cliticized to verbs, subject pronouns cliticized to complementizers, cliticized prepositions, etc.). Clitics that do not affect the structure are not separated (e.g., determiners). Since the word boundaries necessary to separate the clitics are taken from the POS tags, and since it is not possible to show the syntactic structure unless the clitics are separated, correcting the POS tagging (the second POS pass) is extremely important in order to be able to properly separate clitics prior to treebanking.

3.2.3 Human Intervention is Necessary for Treebanking

Since we are already using an automatic parser to provide the first parsing pass on our data (Bikel, 2002; Chiang and Bikel, 2002; Maamouri and Bies, 2004), our treebank annotators already have the advantage that they correct the automatic parse given rather than having to start each tree from scratch. Example 3 shows the output of the automatic parser on a simple sentence.

Example 3:

```
(S (VP EAd+at    عَادَت
      (NP EaqArib+u    عَقَارِبُ
         (NP Al+zaman+i    الزَمَنِ
            (NP fajo>+ap+F    فَجْأَةً)))
      (PP <ilaY    إلَى
         (NP Al+warA'+i    الوَرَاءِ))))
```

عَادَت عَقَارِبُ الزَمَنِ فَجْأَةً إلَى الوَرَاءِ

the hands of time turned suddenly backwards
returned hands time suddenly to the-back
[An-Nahar 20020415.0042.12]

The initial automatic parse is helpful and time-saving, but a number of corrections must be made. The parser does not provide any information on functional category (dashtags) or on empty categories, so all such information must be added by our annotators. The parser also does not get all constituency or dependency relationships correct, and these must be provided by our annotators as well. Example 4 shows the same tree, after the treebank annotator has hand-corrected the parse. Note that the annotator corrected an error in constituency, by moving the adverbial noun phrase out of the subject noun phrase (its incorrect placement from the automatic parser) and into the verb phrase, where it is shown as modifying *turned*. It is also marked as temporal (TMP), and so clearly not a required argument of the verb. In addition, the semantic function tags marking the subject (SBJ) has been added by the treebank annotator. These distinctions are a crucial improvement on the initial automatic parse, since they provide the necessary information about the argument structure of the sentence.

Example 4:

```
(S (VP EAd+at    عَادَت
      (NP-SBJ EaqArib+u    عَقَارِبُ
            (NP Al+zaman+i    الزَمَنِ))
      (NP-TMP fajo>+ap+F    فَجْأَةً)
      (PP <ilaY    إلَى
         (NP Al+warA'+i    الوَرَاءِ))))
```

عَادَت عَقَارِبُ الزَمَنِ فَجْأَةً إلَى الوَرَاءِ

the hands of time turned suddenly backwards
returned hands time suddenly to the-back
[An-Nahar 20020415.0042.12]

We plan to continue development of the parser, eventually adding initial automatic inclusion of the functional tags and empty category information, and we will re-train the parser as each new corpus of human-corrected parses is completed, improving the initial parse.

However, a number of typically ambiguous syntactic constructions are difficult for automatic parsers to get right, and human annotation will be needed for these constructions in any case. These include PP attachment (does a given prepositional phrase modify the sentential verb or a lower noun phrase), NP-internal modification (which noun does a relative clause go with, e.g.) and the distinction between arguments and modifiers, especially in noun phrases. Example 5 is an example of PP attachment ambiguity.

Example 5:

```
 (S (VP yu+Tamo}in+u    يُطَمْئِنُ
      (NP-SBJ *)
      (NP-OBJ (NP Al+lAji}+iyna    اللاجِئِينَ )
              (PP-LOC fiy    فِي
                    (NP brAzafiyl    بَرَازَفِيل )))))
```

يُطَمْئِنُ اللاجِئِينَ فِي بَرَازَفِيل

(he) reassures the refugees in Brazzaville

This example is structurally ambiguous, and factually ambiguous also, since both prepositional phrase attachment interpretations are possible (*refugees in Brazzaville* vs. *reassures in Brazzaville*). However, the context resolves the ambiguity, and the NP attachment is preferred because "in Brazzaville" was a more relevant description to the annotator of the refugees themselves in the context (although the speaker himself may have been located in Brazzaville at the time of this statement).

Example 6 is an example of a similar attachment ambiguity for adjectives within noun phrases.

Example 6:

```
 (PP li-    لِ
    (NP (NP -waziyr+i    وَزِيرِ
            (NP Al+difAE+i    الدِفاعِ))
        (ADJP Al+EiraAqiy~+i    العِرَاقِيِّ)))
```

لِوَزِيرِ الدِفاعِ العِرَاقِيِّ

to the Iraqi minister of defense
to minister the-defense the-Iraqi

Again, the example above is structurally ambiguous, although the context resolves the ambiguity here: it is an Iraqi minister rather than Iraqi defense, and the treebank annotation represents this interpretation. Such cases of structural ambiguity require human annotation, since automatic parsers perform poorly in resolving ambiguity.

3.2.4 TB Annotator Decision Process

Annotators use the following criteria for making syntactic treebank decisions:

- The first task is for the annotator to determine the correct interpretation (what does the sentence mean?) and resolve any ambiguities in the interpretation.
- The next step is to get the constituent boundaries correct – constituent structure should reflect the interpretation and should be chosen to accurately represent any potential ambiguities.
- Additional structure is added as necessary to represent modification and argument structure in noun phrases.
- Function tags are added to verb phrase constituents – the argument structure of the sentence is shown through function tags on every argument or modifier of the verb.
- Empty categories (pro-drop subjects, passive and Wh-traces, etc.) and their co-reference are added – all argument positions of the verb should be filled, and “moved” constituents (such as Wh-words, topicalized noun phrases, extracted sub-constituents) should be co-referenced to the correct empty category in the proper/original/interpreted position.

In addition, simply getting enough human-annotated data in Arabic to make it possible to train automatic tools is important. Such annotated data did not exist before this project, and does exist to some extent now.

4.0 A Practical Illustration of the Arabic Treebank Methodology

The stages of the annotation process are as follows:

1. The plain Arabic text is acquired from the newswire source.
2. The text is run through the automatic morphological analyzer, and the initial lexicon possibilities are provided.
3. The POS/morphological annotator’s choice and selection leads to the fully vocalized form, including case endings, etc.
4. The clitics with independent syntactic function are automatically separated.
5. The text and POS information are run through the automatic parser, and the initial parse is provided.
6. The treebank annotator’s decisions and annotation lead to the final tree.

4.1 The Original Newswire Text

The original text as received from the An-Nahar newswire source is not vocalized:

Example 7:

وقالت إن إجراءات الحماية هذه ستتخذ بناء على طلب حكومة سيول

and (he) said that the measures of this protection will be taken according to the request of the Seoul administration

[An-Nahar 20020215.0117.16]

4.2 Output of Morphological Analyzer for the Verb

The morphological analyzer provides the potential solutions for each word in the sentence. For example, there are twelve potential solutions provided by the morphological analyzer for the verb in this sentence (the two most relevant of which are shown below).

Example 8:

```
INPUT STRING: ستتخذ
SOLUTION 1: satat~axi*u
  LEMMA ID: {it~axa*_1
       POS: sa/FUT+ta/IV3FS+t~axi*/IV+u/IVSUFF_MOOD:I
     GLOSS: will + it/they/she + take/adopt + [ind.]
SOLUTION 4: satut~axa*u
  LEMMA ID: {it~axa*_1
    POS: sa/FUT+tu/IV3FS+t~axa*/IV_PASS+u/IVSUFF_MOOD:I
   GLOSS: will+it/they/she+be taken/be adopted+[ind.]
```

4.3 POS Annotation

The annotator’s task is to choose among the potential solutions provided by the morphological analyzer. For this verb, the annotator must choose SOLUTION 4 (`POS: sa/FUT+tu/IV3FS+t~axa*/IV_PASS+u/IVSUFF_MOOD:I`) in order to get the correct interpretation for this sentence.

4.4 Automatic Clitic Separation

At this stage, the clitic conjunction *wa* is split from the verb *qAl+at.* This allows the annotator to correctly represent the verb as heading the VP verb phrase independent of the conjunction.

4.5 Output of Bikel Parsing Engine

Example 9:

```
 (S wa-    وَ
  (VP –qAl+at   قالَت
      (SBAR <in~a   إِنَّ
            (S (NP <ijorA’+At+i    إِجْراءاتِ
                   (NP Al+HimAy+ap+i h’*ihi  الحِمَايَةِ هذِهِ))
               (VP sa+tu+t~axa*+u    سَتُتَّخَذُ
                   (NP (NP binA’+F    بِناءً)
                       (PP EalaY    عَلَى
                           (NP Talab+i    طَلَبِ
                               (NP Hukuwm+ap+i    حُكُومَةِ
                                   (NP siyuwl    سِيُول
                                           ))))))))))
```

وَ قالَت إِنَّ إِجْراءاتِ الحِمَايَةِ هذِهِ سَتُتَّخَذُ بِناءً عَلَى طَلَبِ حُكُومَةِ سِيُول

and (he) said that the measures of this protection will be taken according to the request of the Seoul administration

[An-Nahar 20020215.0117.16]

4.6 Treebank Annotation

The final hand-corrected annotation, provided by the treebank annotator is below. Note that the semantic function tags marking the subject (SBJ) and object (OBJ), as well as the pro-drop subject, passive trace and topicalization information have been added by the treebank annotator. In addition, the adverbial noun phrase is marked as ADV, and so clearly not a required argument of the verb.

Example 10:

```
(S wa-    وَ
  (VP –qAl+at    قالَت
     (NP-SBJ *)
     (SBAR <in~a    إنَّ
          (S (NP-TPC-1 <ijorA’+At+i    إجْراءاتِ
                (NP Al+HimAy+ap+i h’*ihi  هذِهِ الحِمَايَةِ))
             (VP sa+tu+t~axa*+u    سَتُتَّخَذُ
                (NP-SBJ-1 *T*)
                (NP-OBJ-1 *)
                (NP-ADV (NP binA’+F    بِناءً)
                     (PP EalaY    عَلَى
                        (NP Talab+i    طَلَبِ
                          (NP Hukuwm+ap+i    حُكُومَةِ
                             (NP siyuwl    سِيُول
                                          ))))))))))
```

وَ قالَت إنَّ إجْراءاتِ الحِمَايَةِ هذِهِ سَتُتَّخَذُ بِناءً عَلَى طَلَبِ حُكُومَةِ سِيُول

and (he) said that the measures of this protection will be taken according to the request of the Seoul administration

[An-Nahar 20020215.0117.16]

5.0 Conclusions and the Future

Our already annotated ATB corpora give us a solid foundation for experimenting with new techniques for bracketing additional text as semi-automatically as possible. We will be able to test our proposed new tools and techniques in successive passes. The previous ATB annotated corpora will provide training data and be a testbed at the same time for our new tool developments.

As in our earlier work, our goal is to allow rapid, efficient annotation with a highly eclectic approach to the linguistic uniqueness of two new corpora from diverse regional sources in the Arab region. Our intention is to improve the power and efficiency of our automated and semi-automated tools in order to substantially increase the rate at which manual correction can be performed.

Our upcoming Levantine Dialectal Arabic pilot Treebank will allow us to test the porting of the MSA experience to a linguistically different though highly related language. We will be looking very closely at how to initially ‘leapfrog’ our annotation level of effort by ‘porting’ adjusted annotation guidelines and completing a challenging conversational dialectal Arabic corpus in a shorter period of time than our previous MSA segments.

6.0 References

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

Bikel, D. (2002). Design of a multi-lingual, parallel-processing statistical parsing engine. In Proceedings of the Human Language Technology Workshop.

Brill, E. (1993). Automatic Grammar Induction and Parsing Free Text: A Transformation-Based Approach. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics.

Buckwalter, T. (2002). Buckwalter Arabic Morphological Analyzer Version 1.0. Linguistic Data Consortium, catalog number LDC2002L49, ISBN 1-58563-257-0.

Chiang, D. & Bikel, D. (2002). Recovering Latent Information in Treebanks. In Proceedings of COLING 2002.

Collins, M. (1997). Three Generative, Lexicalised Models for Statistical Parsing. In Proceedings of ACL-1997.

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

Fassi Fehri, A. (1993). Issues in the Structure of Arabic Clauses and Words. Dordrecht: Kluwer.

Han, C., Han, N. & Ko, E. (2001). Bracketing Guidelines for Penn Korean Treebank. Technical Report, IRCS-01-10.

Hymes, D. (1973). Speech and Language: On the Origins and Foundations of Inequality among Speakers. Deadalus, 59—86.

Kingsbury, P., Xue, N. & Palmer, M. (2004). Propbanking in Parallel. In Proceedings of the Workshop on the Amazing Utility of Parallel and Comparable Corpora, in conjunction with LREC'04. Lisbon, Portugal.

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

Maamouri, M., Bies, A., Jin, H. & Buckwalter, T. (2003). Arabic Treebank: Part 1 v 2.0. Linguistic Data Consortium, catalog number LDC2003T06, ISBN: 1-58563-261-9.

Maamouri, M., Bies, A., Buckwalter, T. & Jin, H. (2004a). Arabic Treebank: Part 2 v 2.0. Linguistic Data Consortium, catalog number LDC2004T02, ISBN: 1-58563-282-1.

Maamouri, M., Bies, A., Buckwalter, T. & Jin, H. (2004b). Arabic Treebank: Part 3 v 1.0. Linguistic Data Consortium, catalog number LDC2004T11, ISBN: 1-58563-298-8.

Maamouri, M., Bies, A., Buckwalter, T. & Jin, H. (2004c). Arabic Treebank: Part 3(a) v 1.1. Linguistic Data Consortium, catalog number LDC2004E71.

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

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

Marcus, M., Santorini, B. & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics.

Xue, N., Chiou, F. & Palmer, M. (2002). Building a Large-Scale Annotated Chinese Corpus. Proceedings of COLING 2002.