Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools

Mohamed MAAMOURI
LDC, University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104, USA
maamouri@ldc.upenn.edu

Ann BIES
LDC, University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104, USA
bies@ldc.upenn.edu

Abstract

In this paper we address the following questions from our experience of the last two and a half years in developing a large-scale corpus of Arabic text annotated for morphological information, part-of-speech, English gloss, and syntactic structure: (a) How did we 'leapfrog' through the stumbling blocks of both methodology and training in setting up the Penn Arabic Treebank (ATB) annotation? (b) How did we reconcile the Penn Treebank annotation principles and practices with the Modern Standard Arabic (MSA) traditional and more recent grammatical concepts? (c) What are the current issues and nagging problems? (d) What has been achieved and what are our future expectations?

1 Introduction

Treebanks are language resources that provide annotations of natural languages at various levels of structure: at the word level, the phrase level, and the sentence level. Treebanks have become crucially important for the development of data-driven approaches to natural language processing (NLP), human language technologies, automatic content extraction (topic extraction and/or grammar extraction), cross-lingual information retrieval, information detection, and other forms of linguistic research in general.

The Penn Arabic Treebank began in the fall of 2001 and has now completed two full releases of 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; and (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 (UMmah Arabic Al-Hayat) corpus include complete vocalization (including case endings), lemma IDs, and more specific part-of-speech tags for verbs and particles. Arabic Treebank: Part 3 is currently underway, and consists of text from An-Nahar. (Maamouri and Cieri, 2002)

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)

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.

2 Issues of methodology and training with Modern Standard Arabic

2.1 Defining the specificities of 'Modern Standard Arabic'

Modern Standard Arabic (MSA), the natural language under investigation, is not natively spoken by Arabs, who acquire it only through formal schooling. MSA is the only form of written communication in the whole of the Arab world. Thus, there exists a living writing and reading community of MSA. However, the level of MSA acquisition by its members is far from being homogeneous, and their linguistic knowledge, even at the highest levels of education, very unequal. This problem is going to have its impact on our corpus annotation training, routine, and results. As in other Semitic languages, inflection in MSA is mostly carried by case endings, which are represented by vocalic diacritics appended in word-final position. One must specify here that the MSA material form used in the corpus data we use consists of a graphic representation in which short vowel markers and other pertinent signs like the 'shaddah' (consonantal germination) are left out, as is typical in most written Arabic, especially news writing. However, this deficient graphic representation does not indicate a deficient language system. The reader reads the text and interprets its meaning by 'virtually providing' the missing grammatical information that leads to its acceptable interpretation.

2.2 How important is the missing information?

Our description and analysis of MSA linguistic structures is first done in terms of individual words and then expanded to syntactic functions. Each corpus token is labeled in terms of its category and also in terms of its functions. It is marked morphologically and syntactically, and other relevant relationship features also intervene such as concord, agreement and adjacency. This redundancy decreases the importance of the absence of most vocalic features.

2.3 The issue of vocalization

The corpus for our annotation in the ATB requires that annotators complement the data by mentally supplying morphological information before choosing the automatic analysis, which amounts to a pre-requisite 'manual/human' intervention and which takes effect even before the annotation process begins. Since no automatic vocalization of unvocalized MSA newswire data is provided prior to annotation, vocalization becomes the responsibility of annotators at both layers of annotation. The part-of-speech (POS) annotators provide a first interpretation of the text/data and a vocalized output is created for the syntactic treebank (TB) annotators, who then engage in the responsibility of either validating the interpretation under their scrutiny or challenging it and providing another interpretation. This can have drastic consequences as in the case of the so-called 'Arabic deverbals' where the same bare graphemic structure can be two nouns in an 'idhafa (annexation or construct state) situation' with a genitive case ending on the second noun or a 'virtual' verb or verbal function with a noun complement in the accusative to indicate a direct object. In Example 1, genitive case is assigned under the noun interpretation, while accusative case is assigned by the same graphemic form of the word in its more verbal function (Badawi, et al., 2004, cf. Section 2.10, pp. 237-246).

Example 1[1]

Neutral form: <xbArh Al+nb> إخباره النبا

Idhafa: <ixbAruhu Al+naba>i إخبارُهُ النَبَأِ

his receipt (of) the news [news genitive]

Verbal: <ixbAruhu Al+naba>a إخبارُهُ النَبَأَ

his telling the news [news accusative]

These are sometimes difficult decisions to make, and annotators' agreement in this case is always at its lowest. Vocalization decisions have a non-trivial impact on the overall annotation routine in terms of both accuracy and speed.

Vocalization is a difficult problem, and we did not have the tools to address it when the project began. We originally decided to treat our first corpus, AFP, by having annotators supply word-internal lexical identity vocalization only, because that is how people normally read Arabic – taking the normal risks taken by all readers, with the assumption that any interpretation of the case or mood chosen would be acceptable as the interpretation of an educated native speaker annotator. In our second corpus, UMAAH, we decided that it would improve annotation and the overall usefulness of the corpus to vocalize the texts, by putting the necessary rules of syntax and vocalization at the POS level of annotation – our annotators added case endings to nouns and voice to verbs, in addition to the word-internal lexical identity vocalization. For our third corpus, ANNAHAR (currently in production), we have decided to fully vocalize the text, adding the final missing piece, mood endings for verbs. In conclusion, vocalization is a nagging but necessary "nuisance" because while its presence just enhances the linguistic analysis of the targeted corpus, its absence could be turned into an issue of quality of annotation and of grammatical credibility among Arab and non-Arab users.

3 Reconciling Treebank annotation with traditional grammar concepts in Arabic

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. Since all Arabic readers – Arabs and foreigners – go through the process of virtually providing/inserting the required grammatical rules which allow them to reach an interpretation of the text and consequent understanding, and since all 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 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 more time-consuming. This led to including all possible/potential case endings in the POS alternatives provided by the morphological analyzer. 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

[1] For the transliteration system of all our Arabic corpora, we use Tim Buckwalter's code, at http://www.ldc.upenn.edu/myl/morph/buckwalter.html

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 will never be an existing extensive vocalized corpus – except for the Koranic text – that we could totally trust. 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.

So, having made the change from unvocalized text in the 'AFP Corpus' to fully vocalized text now for the 'ANNAHAR Corpus,' we still need to ask ourselves the question of what is better: (a) an annotated corpus in which the ATB end users are left with the task of providing case endings to read/understand or (b) an annotated ATB corpus displaying case endings with a higher percentage of errors due to a significantly more complex annotation task?

3.1 Training annotators, ATB annotation characteristics and speed

The two main factors which affect annotation speed 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 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 prepositions, which could be formally nouns in particular functions (e.g., from fawq فوق NOUN to fawqa فَوقَ PREP "above" and fawqu فَوقُ ADV etc.). 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 '<iErAb' (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 in order to maintain consistency.

While we have been 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 in reaching a consistent and agreed upon use of case-ending annotation.

4 Tools and procedures

4.1 Lexicon and morphological analyzer

The Penn Arabic Treebank uses a level of annotation more accurately described as morphological analysis than as part-of-speech tagging. The automatic Arabic morphological analysis and part-of-speech tagging was performed with the Buckwalter Arabic Morphological Analyzer, an open-source software package distributed by the Linguistic Data Consortium (LDC catalog number LDC2002L49).

The analyzer consists primarily of three Arabic-English lexicon files: prefixes (299 entries), suffixes (618 entries), and stems (82158 entries representing 38600 lemmas). The lexicons are supplemented by three morphological compatibility tables used for controlling prefix-stem combinations (1648 entries), stem-suffix combinations (1285 entries), and prefix-suffix combinations (598 entries).

The Arabic Treebank: Part 2 corpus contains 125,698 Arabic-only word tokens (prior to the separation of clitics), of which 124,740 (99.24%) were provided with an acceptable morphological analysis and POS tag by the morphological parser, and 958 (0.76%) were items that the morphological parser failed to analyze correctly.

Items with solution	124740	99.24%
Items with no solution	958	0.76%
Total	125698	100.00%

Table 1. Buckwalter lexicon coverage, UMAAH

The ANNAHAR coverage statistics after POS 1 (dated January 2004) are as follows:

The ANNAHAR Corpus contains 340,281 tokens, of which 47,246 are punctuation, numbers, and Latin strings, and 293,035 are Arabic word tokens.

Punctuation, Numbers, Latin strings	47,246
Arabic Word Tokens	293,035
TOTAL	**340,281**

Table 2. Token distribution, ANNAHAR

Of the 293,035 Arabic word tokens, 289,722 (98.87%) were provided with an accurate morphological analysis and POS tag by the Buckwalter Arabic Morphological Analyzer. 3,313 (1.13%) Arabic word tokens were judged to be incorrectly analyzed, and were flagged with a comment describing the nature of the inaccuracy. (Note that 204 of the 3,313 tokens for which no correct analysis was found were typos in the original text).

Accurately analyzed Arabic Word Tokens	289,722	98.87%
Commented Arabic Word Tokens/ items with no solution	3,313	1.13%
TOTAL	**293,035**	**100.00%**

Table 3. Lexicon coverage, ANNAHAR

COMMENTS ON ITEMS WITH NO SOLUTION		
(no comment)	1741	52.55%
MISC comment	566	17.08%
ADJ	250	7.55%
NOUN	233	7.03%
TYPO	204	6.16%
PASSIVE_FORM	110	3.32%
DIALECTAL_FORM	68	2.05%
VERB	37	1.12%
FOREIGN WORD	34	1.03%
IMPERATIVE	24	0.73%
ADV	9	0.27%
GRAMMAR_PROBLEM	9	0.27%
NOUN_SHOULD_BE_ADJ	7	0.21%
A_NAME	6	0.18%
NUMERICAL	6	0.18%
ABBREV	5	0.15%
INTERR_PARTICLE	4	0.12%
TOTAL	3313	100.00%

Table 4. Distribution of items with no solution, ANNAHAR

4.2 Parsing engine

In order to improve the speed and accuracy of the hand annotation, we automatically pre-parse the data after POS annotation and before TB annotation using Dan Bikel's parsing engine (Bikel, 2002). Automatically pre-parsing the data allows the TB annotators to concentrate on the task of correcting a given parse and providing information about syntactic function (subject, direct object, adverbial, etc.).

The parsing engine is capable of implementing a variety of generative, PCFG-style models (probabilistic context free grammar), including that of Mike Collins. As such, in English, it gets results that are as good if not slightly better than the Collins parser. Currently, this means that, for Section 00 of the WSJ of the English Penn Treebank (the development test set), the parsing engine gets a recall of 89.90 and a precision of 90.15 on sentences of length <= 40 words. The Arabic version of this parsing engine currently brackets AFP data with recall of 75.6 and precision of 77.4 on sentences of 40 words or less, and we are in the process of analyzing and improving the parser results.

4.3 Annotation procedure

Our annotation procedure is to use the automatic tools we have available to provide an initial pass through the data. Annotators then correct the automatic output.

First, Tim Buckwalter's lexicon and morphological analyzer is used to generate a candidate list of "POS tags" for each word (in the case of Arabic, these are compound tags assigned to each morphological segment for the word). The POS annotation task is to select the correct POS tag from the list of alternatives provided. Once POS is done, 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. 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.

4.4 POS annotation quality control

Five files with a total of 853 words (and a varying number of POS choices per word) were each tagged independently by five annotators for a quality control comparison of POS annotators. Out

of the total of 853 words, 128 show some disagreement. All five annotators agreed on 85% of the words; the pairwise agreement is at least 92.2%.

For 82 out of the 128 words with some disagreement, four annotators agreed and only one disagreed. Of those, 55 are items with "no match" having been chosen from among the POS choices, due to one annotator's definition of good-enough match differing from all of the others'. The annotators have since reached agreement on which cases are truly "no match," and thus the rate of this disagreement should fall markedly in future POS files, raising the rate of overall agreement.

5 Specifications for the Penn Arabic Treebank annotation guidelines

5.1 Morphological analysis/Part-of-Speech

The guidelines for the POS annotators are relatively straightforward, since the task essentially involves choosing the correct analysis from the list of alternatives provided by the morphological analyzer and adding the correct case ending. The difficulties encountered by annotators in assigning POS and case endings are somewhat discussed above and will be reviewed by Tim Buckwalter in a separate presentation at COLING 2004.

5.2 Syntactic analysis

For the most part, our syntactic/predicate-argument annotation of newswire Arabic follows the bracketing guidelines for the Penn English Treebank where possible. (Bies, et al. 1995) Our updated Arabic Treebank Guidelines is available on-line from the Linguistic Data Consortium at: http://www.ldc.upenn.edu/Catalog/docs/LDC2004T02/

Some points where the Penn Arabic Treebank differs from the Penn English Treebank:

- Arabic subjects are analyzed as VP internal, following the verb.
- Matrix clause (S) coordination is possible and frequent.
- The function of NP objects of transitive verbs is directly shown as NP-OBJ.

We are also informed by on-going efforts to share data and reconcile annotations with the Prague Arabic Dependency Treebank (two Prague-Penn Arabic Treebanking Workshops took place in 2002 and 2003). Some points where the Penn Arabic Treebank differs from the Prague Arabic Dependency Treebank:

- Specific adverbial functions (LOC, TMP, etc.) are shown on the adverbial (PP, ADVP, clausal) modification of predicates.
- The argument/adjunct distinction within NP is shown for noun phrases and clauses.
- Empty categories (pro-drop subjects and traces of syntactic movement) are inserted.
- Apposition is distinguished from other modification of nouns only for proper names.

In spite of the considerable differences in word order between Modern Standard Arabic and English, we found that for the most part, it was relatively straightforward to adapt the guidelines for the Penn English Treebank to our Arabic Treebank. In the interest of speed in starting annotation and of using existing tools to the greatest extent possible, we chose to adapt as much as possible from the English Treebank guidelines.

There exists a long-standing, extensive, and highly valued paradigm of traditional grammar in Classical Arabic. We chose to adapt the constituency approach from the Penn English Treebank rather than keeping to a strict and difficult adherence to a traditional Arabic grammar approach for several reasons:

- Compatibility with existing treebanks, processing software and tools,
- We thought it would be easier and more efficient to teach annotators, who come trained in Arabic grammar, to use our constituency approach than to teach computational linguists an old and complex Arabic-specific syntactic terminology.

Nonetheless, it was important to adhere to an approach that did not strongly conflict with the traditional approach, in order to ease the cognitive load on our annotators, and also in order to be taken seriously by modern Arabic grammarians. Since there has been little work done on large data corpora in Arabic under any of the current syntactic theories in spite of the theoretical syntactic work being done (Mohamed, 2000), we have been working out solutions to Arabic syntax by combining the Penn Treebank constituency approach with pertinent insights from traditional grammar as well as modern theoretical syntax.

For example, we analyze the underlying basic sentence structure as verb-initial, following the traditional grammar approach. However, since the verb is actually not the first element in many sentences in the data, we adopt a topicalization structure for arguments that are fronted before the verb (as in Example 2, where the subject is fronted) and allow adverbials and conjunctions to appear freely before the verb (as in Example 3, where a prepositional phrase is pre-verbal).

Example 2

```
(S (NP-TPC-1 Huquwq+u حُقُوقُ
         (NP Al+<inosAn+i الإنْسَان ))
   (VP ta+qaE+u  تَقَعُ
      (NP-SBJ-1 *T*)
      (PP Dimona ضِمْنَ
         (NP <ihotimAm+i+nA إهْتِمامِنا )
       )))
```

حُقُوقُ الإنْسَان تَقَعُ ضِمْنَ إهْتِمامِنا
human rights exist within our concern

Example 3

```
(S (PP min مِن
      (NP jih+ap+K جِهَةٍ
          >uxoraY أُخرَى ))
   (VP ka$af+at  كَشَفَت
      (NP-SBJ maSAdir+u  مَصادِرُ
              miSoriy~+ap+N  مِصْرِيَّةٌ
              muT~aliE+ap+N مُطَّلِعَةٌ ))
      (NP-OBJ Haqiyqata حَقِيقَةَ
             (NP Al->amri الأمر )))
```

مِن جِهَةٍ أُخرَى كَشَفَت مَصادِرُ مِصْرِيَّةٌ مُطَّلِعَةٌ حَقِيقَةَ الأمرِ
from another side, well-informed Egyptian sources revealed the truth of the matter

For many structures, the traditional approach and the treebank approach come together very easily. The traditional "equational sentence," for example, is a sentence that consists of a subject and a predicate without an overt verb (*kAna or "to be"* does not appear overtly in the present tense). This is quite satisfactorily represented in the same way that small clauses are shown in the Penn English Treebank, as in Example 4, since traditional grammar does not have a verb here, and we do not want to commit to the location of any potential verb phrase in these sentences.

Example 4

```
(S (NP-SBJ Al-mas>alatu المَسألَةُ )
   (ADJP-PRD basiyTatuN بَسِيطَةٌ))
```

المَسألَةُ بَسِيطَةٌ
the question is simple

5.3 Current issues and nagging problems

In a number of structures, however, the traditional grammar view does not line up immediately with the structural view that is necessary for annotation. Often these are structures that are known to be problematic in a more general sense for either traditional grammar or theoretical syntax, or both. We take both views into account and reconcile them in the best way that we can.

5.3.1 Clitics

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 TB files. Such cliticized constituents are written together with their host constituents in the text (e.g., Al+<inosAn+i الإنْسَان "the person" and بِقِراءَةِ bi+qirA'ati "with reading"). 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.), so that their syntactic roles can be annotated in the tree. 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, correct POS tagging is extremely important in order to be able to properly separate clitics prior to the syntactic annotation.

In the example below, both the conjunction *wa* "and" and the direct object *hA* "it/them/her" are cliticized to the verb and also serve syntactic functions independent of the verb (sentential coordination and direct object).

Example 5

وستشاهدونها
wasatu$AhiduwnahA
wa/CONJ+sa/FUT+tu/IV2MP+$Ahid/VERB_IMPERFECT+uwna/IVSUFF_SUBJ:MP_MOOD:I+hA/IVSUFF_DO:3FS
and + will + you [masc.pl.] + watch/observe/witness + it/them/her

The rest of the verbal inflections are also regarded as clitics in traditional grammar terms. However, for our purposes they do not require independent segmentation as they do not serve independent syntactic functions. The subject inflection, for example, appears readily with full noun phrase subject in the sentence as well (although in this example, the subject is pro-

dropped). The direct object pronoun clitic, in contrast, is in complementary distribution with full noun phrase direct objects. Topicalized direct objects can appear with resumptive pronouns in the post-verbal direct object position. However, resumptive pronouns in this structure should not be seen as problematic full noun phrases, as they are parasitic on the trace of movement – and in fact they are taken to be evidence of the topicalization movement, since resumptive pronouns are common in relative clauses and with other topicalizations.

Thus, we regard the cliticized object pronoun as carrying the full syntactic function of direct object. As such, we segment it as a separate token and represent it as a noun phrase constituent that is a sister to the verb (as shown in Example 6 below).

Example 6

```
(S wa-  -و
   (VP sa+tu+$Ahid+uwna-  سَتُشاهِدُون
       (NP-SBJ *)
       (NP-OBJ –hA  ها ))
```

وستشاهدونها
and you will observe her

5.3.2 Gerunds (Masdar) and participials

The question of the dual noun/verb nature of gerunds and participles in Arabic is certainly no less complex than for English or other languages. We have chosen to follow the Penn English Treebank practice to represent the more purely nominal *masdar* as noun phrases (NP) and the *masdar* that function more verbally as clauses (as S-NOM when in nominal positions). In Example 7, the *masdar* behaves like a noun in assigning genitive case.

Example 7

```
(PP bi- -بِـ
   (NP qirA’ati   قِراءَةِ
       (NP kitAbi   كِتابِ
           (NP Al-naHwi النَحو ))))
```

بِقِراءَةِ كِتابِ النَحو
with the reading of the book of syntax
[book genitive]

In Example 8, in contrast, the *masdar* functions more verbally, in assigning accusative case.

Example 8

```
(PP bi-  -بِـ
   (S-NOM (VP qirA’ati  قِراءَةِ)
              (NP-SBJ fATimata فاطِمَةَ -)
              (NP-OBJ Al-kitAba الكِتابَ
                    ))))
```

بِقِراءَةِ فاطِمَةَ الكِتابَ
with Fatma’s reading the book
[book accusative]

This annotation scheme to allow for both the nominal and verbal functions of *masdar* is easily accepted and applied by annotators for the most part. However, there are situations where the functions and behaviors of the *masdar* are in disagreement. For example, a *masdar* can take a determiner ‘Al-‘ (the behavior of a noun) and at the same time assign accusative case (the behavior of a verb).

Example 9

```
(PP bi  -بِـ
  (S-NOM
      (VP Al+mukal~afi  المُكَلَّف
          (NP-SBJ *)
          (NP-OBJ <injAza  إنجازَ
               (NP Al+qarAri  القَرار
                   Al+mawEuwdi
                       المَوعُودِ )))))
```

بالمُكَلَّفِ إنجازَ القَرار المَوعُودِ
with the (person in) charge of completion (of) the promised report [completion accusative]

In this type of construction, the annotators must choose which behaviors to give precedence (accusative case assignment trumps determiners, for example). However, it also brings up the issues and problems of assigning case ending and the annotators’ knowledge of Arabic grammar and the rules of ‘<iErAb.’ These examples are complex grammatically, and finding the right answer (even in strictly traditional grammar terms) is often difficult.

This kind of ambiguity and decision-making necessarily slows annotation speed and reduces accuracy. We are continuing our discussions and investigations into the best solutions for such issues.

6 Future work

Annotation for the Arabic Treebank is on-going, currently on a corpus of An-Nahar newswire (350K words). We continue efforts to improve annotation accuracy, consistency and speed, both for POS and TB annotation.

Conclusion

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 other languages. 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 and to have credibility with regards to the attitudes and respect for correctness known to be present in the Arabic world as well as with respect to the NLP and wider linguistic communities. The creation and use of efficient tools such as an automated morphological analyzer and an automated parsing engine ease and speed the annotation process. These tools helped significantly in the successful creation of a process to analyze Arabic text grammatically and allowed the ATB team to publish the first significant database of morphologically and syntactically annotated Arabic news text in the world within one year. Not only is this an important achievement for Arabic for which we are proud, but it also represents significant methodological progress in treebank annotation as our first data release was realized in significantly less time. Half a million MSA words will be treebanked by end of 2004, and our choice of MSA corpora will be diversified to be representative of the current MSA writing practices in the Arab region and the world. In spite of the above, we are fully aware of the humbling nature of the task and we fully understand and recognize that failures and errors may certainly be found in our work. The devil is in the details, and we remain committed to ironing out all mistakes. We count on the feedback of our users and readers to complete our work.

8 Acknowledgements

We gratefully acknowledge the tools and support provided to this project by Tim Buckwalter, Dan Bikel and Hubert Jin. Our sincere thanks go to all of the annotators who have contributed their invaluable time and effort to Arabic part-of-speech and treebank annotation, and more especially to our dedicated treebank annotators, Wigdan El Mekki and Tasneem Ghandour.

References

Elsaid Badawi, M. G. Carter and Adrian Gully, 2004. *Modern Written Arabic: A Comprehensive Grammar*. Routledge: New York.

Daniel M. Bikel, 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. *Proceedings of the Human Language Technology Workshop*.

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

Mohamed Maamouri and Christopher Cieri, 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.

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

M. Marcus, B. Santorini and M.A. Marcinkiewicz, 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics.*

Mohamed A. Mohamed, 2000. *Word Order, Agreement and Pronominalization in Standard and Palestinian Arabic*. CILT 181. John Benjamins: Philadelphia.

Zdenek Žabokrtský and Otakar Smrž, 2003. Arabic Syntactic Trees: from Constituency to Dependency. *EACL 2003 Conferenceompanion*. Association for Computational Linguistics, Hungary.