Diacritization: A Challenge to Arabic Treebank Annotation and Parsing

Mohamed Maamouri, Ann Bies, Seth Kulick
Linguistic Data Consortium, University of Pennsylvania, USA
{maamouri,bies,skulick}@ldc.upenn.edu

Arabic diacritization (referred to sometimes as vocalization or vowelling), defined as the full or partial representation of short vowels, *shadda* (consonantal length or germination), tanween (*nunation* or definiteness), and *hamza* (the glottal stop and its support letters), is still largely understudied in the current NLP literature. In this paper, the lack of diacritics in standard Arabic texts is presented as a major challenge to most Arabic natural language processing tasks, including parsing. Recent studies (Messaoudi, et al. 2004; Vergyri & Kirchhoff 2004; Zitouni, et al. 2006 and Maamouri, et al. forthcoming) about the place and impact of diacritization in text-based NLP research are presented along with an analysis of the weight of the missing diacritics on Treebank morphological and syntactic analyses and the impact on parser development.

Keywords: Arabic NLP, Arabic diacritics, Diacritization, Modern Standard Arabic (MSA), Treebanks, Linguistic Annotation, Parsing

1 INTRODUCTION

Arabic NLP research, focusing mainly on Modern Standard Arabic (MSA) in this paper, faces two major challenges, not necessarily shared with many other natural languages: the first is its complex linguistic structure and the second, the specific features of its orthographic system.

Arabic is a highly inflected language which has as a rich and complex morphological system. Any given Arabic lemma has usually more than one word form to represent it, which includes a root, its internal structure, prefixes, suffixes and clitics. The internal structure itself includes short vowels and vocalic length, which together carry the bulk of the morphological and morphosyntactic structures, and a consonantal skeleton, which, as in other Semitic languages, bears the weight of the lexical (semantic) structure.

The Arabic orthographic system uses superscript and subscript diacritical marks (or diacritics) for the representation of the three short vowels (a, i, u), and four letters (ا ‘alif, ى ‘imaala, و waaw, and ي yaa’) to mark vocalic length. The ‘imaala, an undotted form of the letter yaa’, is used idiosyncratically for certain words ending in the long vowel [-aa] while the yaa’ and the waaw, in addition to being consonants in their own right, function as glides or semi-vowels and are used to represent long [-uw] and long [-iy]. Many of the MSA grammatical functions, such as verb passive forms and irregular noun plural forms, also use short vowels. Finally, short vowels are also used to indicate mood, aspect and voice endings for verbs and case endings for nouns. Moreover, long vowels are mostly used in derivation and word formation: as in kataba ‘to write’ vs. kAtaba ‘to correspond with’. The shadda (consonantal length or gemination) is another important diacritic which is used for the derivation of new words. The hamza is used

just to mark the existence of the glottal stop. Its major issue is its complex graphemic support system and the fact that the hamza is frequently omitted and sometimes misused. Finally, it is to be noted that no further mention will be made in this paper of the sukun, which uses a small superscript zero-shaped grapheme and is in fact nothing more than the absence of a vowel. The sukun does not add anything to the written text and is only used for syllabic identification in speech.

From the comprehensive description above, we see that Arabic is a relatively complex and difficult language to analyze, not so much because of its difficult linguistic structure but mostly because of how that structure is impacted and made more complex by the orthographic issues of its written form. The view presented in this paper is that the Arabic NLP scientific community should become more aware of some of these Arabic script issues and more discriminative of their role and responsibility in Arabic NLP work. The present paper will focus on the role and impact of diacritization on Arabic Treebank annotation and Arabic parsing.

2 REALITY OF ARABIC SPEECH AND TEXT

The issue of diacritization in Arabic arises as the result of a mismatch between the orthographic conventions that have developed for written MSA and the Arabic language itself, including spoken MSA, with respect to the amount of linguistic information represented. MSA is generally written without diacritics, but the language itself, and also spoken MSA, of course includes all of the features that the diacritics would represent (short vowels, consonantal gemination, etc.). When working with MSA for NLP purposes, several choices must be made. Will a particular effort focus on written or spoken MSA (the issue of dialects being largely beyond the scope of this paper)? Spoken MSA must be transcribed to facilitate downstream processes, so in either case, we will be dealing with text of some kind. Will the text be vocalized/diacritized or not (and if so, how will this be accomplished)? Will the text be in the Arabic script or in a transliterated/Romanized from? Each of these questions can be answered independently, and the answers will greatly affect the work to be done.

2.1 The Nature of Arabic Script-based Text

It is obvious that when people speak MSA (or any dialectal Arabic) they must use the linguistic features of Arabic including its short vowels and other diacritical marks. Starting from recorded speech (as in MSA Broadcast News corpora), one has available and can therefore just transcribe all the existing phonetic data. The LDC CallHome Egyptian Colloquial Arabic (ECA) Corpus provides, in a Romanized form, an ASCII representation, which includes short vowels and other information relating to that which is transcribed with diacritics in the Arabic script. The ECA CallHome corpus has in fact all the information required for any NLP task, especially automatic speech recognition (ASR). As noted in Vergyri & Kirchhoff (2004) and Maamouri, et al. (2004), a Romanized transcription of Arabic could probably be an excellent way of providing full linguistic information to all Arabic research tasks if it weren't for its difficulty and the prohibitive training and annotation costs that it would entail. This excessive difficulty and the need for huge volume of data led to the inevitable use of what is readily available (namely, written Arabic script data) and made newswire Arabic the favourite and most used source data. Moreover, most acoustic material for Arabic ASR is in text also and the move to Arabic script based text followed Vergyri & Kirchhoff (2004). A detailed account of the advantages and pitfalls of a two-prong Arabic orthography-based transcription is given in Maamouri et al. (2004b), with the

consonantal skeleton bare forms leading to a non-diacritized transcription (in the LDC Transcription using AMADAT's GREEN layer) and a fully diacritized one (in the same transcription tool's YELLOW layer). This last move by the Arabic NLP scientific community led to a closer look at diacritization.

2.2 Importance of Diacritics

Though the use of diacritics is extremely important in setting up grammatical functions leading to acceptable text understanding and correct reading or analysis, diacritical markings are rarely present in real-world/life situations. It is true that they are rarely visible in out-of-school written documents and they do not, as a rule, appear in most printed materials in the Arab region. This predominant/generalized practice of graphemic under-representation of linguistic information concerns not only the short vowels but also the shadda (consonantal length) and the hamza (glottal stop). The use of diacritics is usually restricted to the early years of formal education and the sacred Koranic text and seems to be limited to whatever length of time is considered sufficient for the learner to be initiated to reading without the missing information (Maamouri 1998). This generally amounts to about six years and most times more. Nowadays, vocalized Arabic text seems to be only used in pure deference to the needs of young and inexperienced learners. It is to be noted that diacritized MSA text does exist outside of the Koran in numerous sources, such as the rich and important heritage Arabic literature books. However, this source of diacritized data is not used by the NLP community, usually because neither the language nor the domains are of acceptable currency. Anywhere else, in Arabic newswire and in most other Arabic-script based transcriptions, the norm and 'real-world' data is Arabic non-diacritized text, whether MSA or dialectal.

2.3 Diacritics and Ambiguity

The loss of the internal diacritics (such as short vowels or shadda) leads to the following types of ambiguity, as exemplified in a given MSA lemma: علم Elm. The situation of this specific bare graphemic form is as follows:

(a) An ambiguity within 'core' POS tags, which distinguishes between the different lexical senses of the same 'core' POS tag. Example: The bare form علم Elm can be diacritized as عِلم Eilm (a noun meaning 'science, learning') or عَلَم Ealam, another noun meaning 'flag'.

(b) A second type of 'core' POS tag ambiguity distinguishes between different lexical senses leading to different core POS tags. The same bare form علم Elm, which was diacritized as two different nouns above, can additionally be diacritized as three different verb forms, all lexically and semantically connected. Example: عَلِمَ Ealima for 3rd Person Masculine, Singular, Perfective Verb (MSA Verb Form I) meaning 'he learned/knew'; عُلِمَ Eulima for 3rd Person Singular, Passive Verb (MSA Verb Form I) meaning 'it/he was learned' and عَلَّمَ Eal~ama for the Intensifying, Causative, Denominative Verb (MSA Verb Form II) meaning 'he taught.' Ambiguities (a) and (b) are lexical/morphological ambiguities.

(c) Finally, a huge amount of ambiguity occurs at the structural/grammatical level, where the use of short vowels is correlated with case (nominal) and mood/aspect (verbal) information. This information is rendered by the use of one of six possible diacritics. Using the above example, we have the following: عِلمُ/عِلمٌ Eilmu/EilmN

(NOM Noun + Definite and Indefinite), عِلمَ/عِلماً Eilma/EilmAF (ACCU Noun + Definite and Indefinite) and عِلمِ/عِلمٍ Eilmi/EilmK (GEN Noun + Definite and Indefinite).

The loss, even partial, of diacritics frequently leads to a significant increase of linguistic ambiguity (both structural and lexical), which can only be resolved by contextual information and an adequate knowledge of the language. In order to be able to read/transcribe, annotators have to provide their own grammatical interpretations and bring to task considerable additional knowledge of syntax, vocabulary, and sometimes contextual interpretation in order to obtain correct and meaningful vocalizations which will allow them to reach acceptable word recognition and sense disambiguation.

The considerable lexical ambiguity consequent to loss of diacritics is observed in Debili, et al. (2002), who calculate that an Arabic non-diacritized dictionary word form had 2.9 possible diacritized forms on average and that an Arabic text containing 23K word forms showed an average ratio of 1:11.6 (quoted in Vergyri & Kirchhoff 2004).

2.4 'Real-World' Arabic Text Data

When we look at the availability of Arabic text data, the situation boils down to the following:

(a) Unvocalized/non-diacritized Arabic text for MSA (and even for newly written dialectal Arabic) seems to be the most available material for the speech research community and the main data source for all other NLP research needs (mostly in newswire form).

(b) Since non-diacritized text prevails, the Arabic NLP community seems to have accepted using it as the de facto 'real world' information material without feeling an obligation to question its choice/use, even espousing the idea sometimes that the robustness of software algorithms can deal with the problem and reduce the negative effect of the missing information on their research.

(c) While it is possible, as noted in Vergyri & Kirchhoff (2004) to collect available MSA (or even dialectal) public data and to transcribe it manually with full or partial restoration of the missing diacritics, obtaining thus an acceptably diacritized form, this practise has not been continued. The prohibitive cost and the usually unequal and questionable quality of human/manual diacritization have led the scientific Arabic NLP community and its sponsors to focus more on volume of unvoweled data so far.

(d) Most NLP Arabic research – even research dealing with diacritization – makes use of text-based information only and makes little use of diacritics even when they exist. No significant use is made of diacritics in the acoustic data – even when work starts from a speech source.

3 DIACRITIZATION RESEARCH REVIEWED

The NLP scientific community is slowly becoming aware of the vital importance of diacritization in Arabic text-based research. The problem posed to the Arabic NLP community is how to diacritize text data in order to reach full linguistic information and better research results. In the remaining sections, a brief presentation of the place and impact of diacritization in text-based NLP research will be reviewed, and an analysis of

the weight of the missing diacritics on Treebank morphological and syntactic analyses and the impact on parser development will be given.

A look at current research shows that it is possible to restore/recover and provide much diacritization information automatically or semi-automatically, if it is not included in the transcription text – via manual annotation, from the annotator's 'virtual knowledge' or from available acoustic information (as heard in MSA Broadcast News). If the source is text-based information, diacritization could be determined from morphological and contextual knowledge. Whether one is working from the perspective of Acoustic Modelling – mainly ASR systems – or from that of Language Modelling, the available knowledge sources that could be used for most appropriate diacritization of a script-based Arabic text forms are the following:

- analysis of the morphological structure: segmentation of words/lemmas into stems, roots and patterns
- consideration of the syntactic context in which the word/lemma form occurs
- knowledge added from the context of speech/acoustic data accompanying the transcription.

3.1 Automatic Diacritization of Training Data /Arabic Text Research

A fully automatic approach to diacritization (reported in Vergyri & Kirchhoff 2004) was presented in Gal (2002). Gal used an HMM-based bigram model, which was used for decoding diacritized sentences from non-diacritized sentences. Gal applied this technique to the Koranic text, achieving 14% word error.

Kirchhoff et al. (2002) and Lamel (2003): Since the LDC CallHome ECA corpus was distributed with both Romanized and script-based transcriptions, the above work compared error rates of recognizers trained on both transcriptions. The results show that the loss of information due to training on script forms is significantly worse with a relative increase of 10% in Word Error Rate (WER). Kirchhoff et al. (2002) show that the MSA WER ranged from 9% to 28%, depending on whether or not case endings were counted.

Vergyri & Kirchhoff (2004) indicate that lower diacritization error rates were produced when more linguistic information is included (morphological and syntactic context) in combination with acoustics. Nonetheless, the best word error rate reported for diacritizing Egyptian Arabic was 41.6%. The authors mention that they intend to apply knowledge-poor diacritization procedures to dialects of Arabic for which morphological analyzers do not exist in their future work. Safadi, et al. (2006) report on an unsupervised method of diacritizing that builds on a combination of automatic tagging and manually written rules. They report an error rate of 10-20%, as determined by expert evaluation of their output.

Nelken and Shieber (2005) use a weighted finite-state transducers for diacritic restoration, including also the separation of clitics. They report a diacriticization error rate of 12.79% and word error rate of 23.61% when including the case endings. Disregarding the case endings for restoration and evaluation improves the scores to 6.35% and 7.33%.

Zitouni, et al. (2006) report on a maximum entropy approach to restoring a comprehensive list of diacritics, treating the problem as one of sequence classification.

They report a diacritic error rate of 5.1% (and a word error rate of 17.3%) in a fully diacritized setting. They also find a significant improvement in error rates when case endings are disregarded, with the scores improving to 2.2% and 7.2%. This study did not make any mention of the hamza, especially on whether its inconsistent and sometimes lacking representation had any appreciable impact on the final results. One notable aspect of this work is that they used a part of speech tagger to generate tags used as features in the model, and they report that this improves the score. Given the connection between some diacritic restoration and part of speech tags, as discussed in Section 2.3, this is not surprising. However, they do not discuss what POS tag set is used, making it difficult to understand precisely how their approach captures this relationship.

Messaoudi, et al. (2004) found that it was useful to transcribe short vowels in order to build acoustic models for Arabic broadcast news. They report a 10% improvement in Arabic WER with the inclusion of diacritics.

3.2 Semi-Automatic Diacritization of Arabic Text

In our Arabic Treebank effort at LDC, MSA text data has so far been morphologically and syntactically been annotated using the following combination of semi-automatic and manual/human annotation:

- Morphological knowledge provided by Buckwalter analyzer (stemmer) – 1.5% of unknown word forms (typos, named entities, 'not in lexicon,' etc.) (Buckwalter 2002)
- Probabilistic contextual knowledge provided by tagger (still in development)
- Word forms with full case endings provided by human annotation
- Syntactic (tree structure) output provided by the Bikel implementation of the Collins parser (publicly available at http://www.cis.upenn.edu/~dbikel/), which is being enhanced by a University of Pennsylvania/BBN team of researchers (Kulick, et al. forthcoming)
- Corrections, final trees and final word forms provided by human annotators

We did not initially have all the tools to address the problem of diacritization when the LDC Arabic Treebank project began at the end of 2001. We decided to annotate our first Arabic Treebank segment, ATB: Part 1 (also known as the AFP Corpus), by having annotators supply word-internal lexical identity vocalization only, because that is how people normally read Arabic – taking/assuming the normal risks taken by all Arabic readers, with the assumption that any interpretation of the case or mood chosen would be the acceptable interpretation of an educated native speaker/annotator.

In our second Arabic Treebank segment, ATB: Part 2 (also known as the UMAAH Corpus), we decided that it would improve annotation and the overall usefulness of the corpus to complete the vocalization of the morphologically analyzed texts by adding the necessary case and mood endings at the Part-of-speech + Gloss (MPG) level of annotation. Up to six case endings (-a, -u, -i for definite and N, AF, K for indefinite nominals) were automatically added in Tim Buckwalter's Morphological Analyzer (BAMA) output alternatives. Our annotators had to select not only the correct/appropriate word-internal lexical identity vocalization for each targeted token from the lexicon-based set of output analyses provided by BAMA, but they also had to make sure that the selection incorporated the correct case and mood endings.

For our third treebank segment, ATB: Part 3 (also known as the ANNAHAR Corpus), we decided to fully vocalize the text, adding the final missing piece, mood and voice endings for verbs in all of the alternatives presented by the BAMA output to our POS annotators.

The LDC experience shows that the Arabic Treebank research team had to adjust its initial MSA Treebank annotation in order to include a 'diacritization' which related mostly to the addition of inflectional endings on top of the full short vowel representation (mostly provided by BAMA) in its morphological layer of analysis (Maamouri & Bies 2004). The ensuing fully diacritized POS output was used by syntactic annotators and provided them with a single interpretation (which is the result of word-disambiguation) that they had to accept/confirm or contest for the syntactic layer. The availability of a fully diacritized text, always present in the syntactic annotation tool, made the syntactic annotation task easier and decreased the annotation responsibility load, leading hopefully to more annotation consistency, less time on the job and less annotation stress in general. It is our belief that the results of the morphological and POS annotation and word disambiguation used in all consequent segments of the Arabic Treebank led to a scientifically sound methodology for diacritizing bare MSA text. Although we did not originally set out with this goal, as a result of annotation necessity and through the morphological annotation process, we have now produced nearly a million words of diacritized MSA newswire text.

4 PARSER DEVELOPMENT: HOW DOES DIACRITIZATION IMPACT PARSING?

The first experiments on parsing the Arabic Treebank were reported in Bikel (2004), for a small section of the early work on the ATB from the AFP section, consisting of 149K tokens for training and 11K for testing. The results were a recall/precision of 75.4/76.0 for sentences of length <=40, and 72.5/73.4 for all sentences.

This work used a combination of the bare data, together with the morphological Parts of Speech tags resulting from the POS annotation step. The large size of the POS tagset was problematic for the parser, since it fragmented the data in a way that the current configuration of the parser was not well suited for. In addition, it was always possible for "new" tags that the parser had not seen in the training data to be created by a new combination of affixes. Therefore, the tagset was mapped down to a reduced number of tags by Ann Bies.

The reason given by Bikel for using the bare data is directly relevant to the concerns of this paper: "We only used the unvoweled data, because that would ultimately be necessary for any real-world [defined here as an NLP situation in which the text is all too often bare] parser. However, because the purpose was for bootstrapping treebank annotation, we used the gold-standard, Bies-mapped part-of-speech tags." (p. 80, Bikel 2004)

There are two points that should be made about this:

(1) It is not necessarily true that the bare data would be the input to the parser for a real-world parser. There could be a preprocessing step reconstructing the diacritics, or at least some of them. Also, the bare data and corresponding tags

are already the result of a certain amount of morphological analysis and tokenization, since various clitics have been separated from their tokens as they originally appeared in the "real-world" data, such as conjunction prefixes and pronoun suffixes.

(2) Even though the "Bies-mapped" tags are being used by the parser, the gold POS tags are still taken as the input before that mapping. The gold tags, however, correspond to the diacritized data. It is in fact reasonable to say that the bare data together with the gold POS tags carries the same information as the diacritized data. This therefore seems to us to be an inconsistency in the original parsing experiments. (Due to the lack of a part-of-speech tagger at the time, however, there was no choice but to use the gold part-of-speech tags.)

In general, the role of diacritics in a NLP pipeline that includes parsing is very much an open question. It seems fair to say that the end process of an NLP pipeline that aims to extract some sort of syntactic structure and semantic meaning from the initial text should have sufficient "understanding" that it can restore the diacritics to the text by the end of the pipeline. This however does not speak to the question of at what point in the pipeline diacritics should be restored, and the possibility that different diacritics perhaps should be restored at different times. The steps taken for diacritic restoration for manual annotation may not be the appropriate ones for an NLP pipeline.

There are two aspects to the problem of how the parser might utilize diacritic information. One question concerns what diacritic information might be useful for the parser. While the earlier work, as mentioned above, used the bare text, there has been very little work examining whether a parser can make use of vocalized text. The Arabic Treebank now gives us a corpus to carry out such experiments, as we discuss in the following subsection.

However, in addition to exploring which diacritic information is useful for the parser, we must also be concerned with what might be available to the parser outside the context of these experiments and outside the context of Treebank research. As discussed in Section 3, there is a growing body of work on diacritic restoration. However, Zitouni, et al. (2006) and Nelken & Shieber (2005) both report that it is much harder to restore the diacritics representing case information. This is not surprising, since as Nelken & Shieber write, "including case information naturally yields proportionally worse accuracy.
Since case markings encode higher-order grammatical information, they would require a more powerful grammatical model than offered by finite state methods".

While noting these concerns and issues, for the experiments reported here we continue to train and test the data as it is tokenized following the stage of POS annotation, and we continue to use the gold tags, while varying the level of vocalization in the words. We recognize that in any actual use of the parser on bare text, there would need to be a level of preprocessing for tokenization and perhaps some level of diacritic restoration. For the current work, however, our goal is to utilize the diacritization that is present in the ATB for understanding how different amounts of vocalization affect the parser.

4.1 Parser Experiments

We have shifted the parser experiments to use ATB3 (Annahar) instead of the data Bikel was working with. The ATB3 corpus is larger than ATB1, and was done after

more experience was gained with the annotation process. We did a balanced 80/10/10 split on the corpus for training/development/test, so that eight out of every 10 sentences was used for training, and so on.

The input to the parser was varied in four ways:

Bare: is using the bare data (ATB3 Annahar)

Full: is using the fully diacritized form of the words.

No Case: is using a modified form of the diacritized data, in which the case endings for nouns and adjectives have been deleted. This was based on the hypothesis that the extra case endings were causing data fragmentation. For example, suppose the bare word is أميركياً ">myrokyAF", with the diacritized form أمِيرْكِيّاً ">amiyrokiy~+AF" (American+ [acc.indef]) and tag ADJ+CASE_INDEF_ACC. For the **No Case** run, this would be treated as أمِيرْكِيّ ">amiyrokiy~".

No Case, Mood: is a more extreme form of the **No Case** run, in which the mood suffixes have been deleted as well.

	#Words	#Instances	%Unknown words	% Unknown instances
WSJ English	23333	296872	79.43%	10.61%
Bare Annahar MSA	32204	296748	80.37%	15.42%
Full	43141	296748	84.95%	21.24%
No Case	31524	296748	80.28%	14.94%
No Case, Mood	30454	296748	79.31%	14.45%

TABLE 1: Unknown word frequencies.

As discussed in Bikel (2004), the parser considers all words occurring below some threshold (here, 6) to be "unknown" words. This is somewhat of a misnomer, as they can be considered rare words, for which their lexical information is not stored the same way as with more common words. The idea is that this will help the handling of words during parsing that were not seen during training, which will also be treated as "unknown" words. Table 1 shows the # of words and instances (i.e., word types and word tokens) encountered during training, along with the % of unknown words and % unknown instances. To compare these numbers for the ATB to the Penn Treebank, we also trained the parser on a training size of the Penn Treebank roughly the same as our training size for the ATB.

One thing to note is the interestingly high percentage of unknown words, even for the Wall Street Jounal (WSJ). This is no doubt because of the large number of nouns that occur only one or two times, for the "Bare" Arabic run, the % of unknown words is roughly the same, while the % of unknown instances is significantly higher than for the WSJ, and indeed the number of words is significantly higher as well, although the number of instances is roughly the same. This indicates a somewhat flatter distribution of words in the ATB. Switching to the "Full" run, the situation worsens, as expected. Since the words now include the full tokenization, including the case endings, a single

unvocalized entry can now be shattered into many different fully vocalized forms. The # of words for the same # of instances increases from 32204 to 43141, with an increase to 21.24% of the unknown instances.

However, using the "No Case" version of the diacriticized data, the data sparseness is actually *less* than with the bare form, and the same is even more true when the mood suffixes are eliminated as well. Our suspicion is that this is because the diacritized form, in addition to adding the diacritics, also normalizes the orthographic variations.

For example, the lemma [<ihAnap_1] (insult/contempt) occurs 4 times with a determiner prefix and feminine singular ending, but with three different spellings: Al<hAnp (1 time), AlAhAnp (2 times), and Al<hAn_p (1 time). The diacritized versions of these are all the same, however: Al+<ihAn+ap (leaving aside the case endings). So there are likely many cases in which a word that appears in different orthographic forms, each of which is "unknown", can be unified as one entry in the diacriticized form and thus not treated as an "unknown word".

We also performed parsing runs for each of the four models, to see what sort of effect the different training and test input had on the overall score. In general, there is a significant difference in the score for sentences <=40 and all sentences, more so than in the initial Bikel results, perhaps because the sentences in ATB3 have more conjunction at the S level than in the ATB1, although we have not confirmed that. We include testing results only for sentences <=40, because it is then much faster to run the parser. Our concern is looking at differences in parsing results, with the assumption that the scores for including all sentences will continue to lag behind.

For the experiments, as mentioned above, we are continuing to use the gold POS tags in the "Bies-reduced" form, while recognizing the inherent limits of this approach, as discussed above. The results of the four runs are shown in Table 2.

	Recall/Precision
WSJ English	87.42/87.72
Bare Annahar MSA	77.43/79.42
Full	78.08/79.88
NoCase	77.87/79.72
No Case, Mood	77.83/79.69

TABLE 2: Results of parser experiments

As has been generally recognized, the scores for ATB are much lower than for the WSJ. The results for our baseline with the "Bare" data are somewhat different from those reported in Bikel (75.4/76.0), for two reasons: (1) Different corpora are being used – ATB3 instead of ATB1, and (2) We are taking advantage of various improvements to the parser that are discussed in a separate paper.

4.2 Discussion of Parser Experiments

As can be seen in Table 2 above, there is little difference in the parsing results. Given all of the open questions around the status of parsing Arabic, it is not entirely clear how to evaluate these results. Still, there is little change in the overall scores, and there is no correlation between the improvement in unknown word frequency and parsing improvement.

It is clear that the parser needs a substantial overhaul as to its handling of morphological information. For example, nouns with and without a determiner Al- prefix are treated as two separate words, with no relation to each other. This is clearly wrong. But even that aside, there are several ways to expand on this work.

One aspect that needs to be investigated more is the relation between the diacritics and the Part of Speech tags. As discussed in Section 2, diacritics can distinguish between different 'core' POS tags (and so different tags in the reduced tag set used by the parser.) It is likely that the gold Part of Speech tags, even in their reduced form, are masking some of the benefits of using the diacritized data. For example, "ktb" occurs 56 times in the corpus, 17 as NOUN, 6 as PV_PASS (passive perfective verb), and 33 as PV (perfective verb). As a NOUN, it has the vowels "kutub", as a PV_PASS it has "kutib", and as PV it has "katab". Therefore, in this case the vowelization does not add much beyond the Part of Speech tags. It is possible therefore that the full benefits of vocalization can only be seen in the context of a wider NLP pipeline than just the parser, including in particular the Part-of-Speech tagger. Related to this, Habash (2005) reports a drop in ambiguity when considering tokens only within the same Part of Speech tag. Also notable in the connection is the work of Habash & Rambow (2005), who describe an integrated approach to tokenization, Part of Speech tagging and morphological disambiguation. Of particular interest is the close connection between POS tagging and morphological disambiguation. While this connection is clearly related to the concerns expressed here, they do not include a step of diacritazation.

In our view, the categorization of the ambiguities resolved by diacritic restoration discussed in Section 2 deserves detailed empirical analysis based on the data the parser is using. As just discussed, there is a close connection between POS tags and some diacritic restoration (ambiguity (b) in Section 2.3). Another type of ambiguity is that within 'core' POS tags, distinguishing for example between two different nouns (ambiguity (a) in Section 2.3). Since so many nouns occur infrequently enough that they are categorized as "unknown" by the parser, this will probably make less of a difference. However, this may account for some of the flatness of the distribution of words in the ATB, as discussed above. The utility of the case and mood/aspect markers (ambiguity (c) in Section 2.3) for the parser is even more of an open question, and it seems reasonable that such restoration should take place as a byproduct of the parsing process, rather than as a preprocessing step.

It seems worthwhile to explore this point more, to determine not just what the ambiguity of unvocalized forms is, but also what kinds of ambiguity matter the most to our research.

5 CONCLUSION

The role of diacritization in the annotation process for the Arabic Treebank is now firmly established, and this data has been available and quite useful to the scientific community. In general, however, the correct way to utilize diacritization in various Natural Language Processing tasks is more of an open question. In Section 4 we have described some initial experiments exploring the role of diacritics in parsing. In our view, one of the primary tasks for this line of investigation is a more systematic investigation of the ambiguities that different diacritics resolve, and their interaction with the Part of Speech tags.

REFERENCES

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

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

Y. Gal. (2002). 'An HMM Approach to Vowel Restoration in Arabic and Hebrew'. In *ACL-02 Workshop on Computational Approaches to Semitic Languages*.

N. Habash. (2005). 'Introduction to Arabic Natural Language Processing'. *ACL Tutorial*.

N. Habash and O. Rambow. (2005). 'Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop'. *ACL 2005*.

K. Kirchhoff, et al. (2002). Novel Approaches to Arabic Speech Recognition – Final report of the JHU Summer workshop. John Hopkins University.

S. Kulick, R. Gabbard and M. Marcus. (2006, forthcoming). 'Parsing the Arabic Treebank: Analysis and Improvements'. University of Pennsylvania.

M. Maamouri. (1998). 'Arabic Diglossia and its Impact on the quality of education in the Arab region'. Discussion paper prepared for World Bank at "The Mediterranean Development Forum." World Bank, Marrakech, 3-6 September 1998.
http://www.worldbank.org/wbi/mdf/mdf2/papers/humandev/education/maamouri.pdf

M. Maamouri and A. Bies. (2004). 'Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools'. In *Proceedings of the 20th International Conference on Computational Linguistics* (COLING 2004).

M. Maamouri, A. Bies, T. Buckwalter and W. Mekki. (2004). 'The Penn Arabic Treebank: Building a Large-Scale Annotated Arabic Corpus'. In *Proceedings of NEMLAR 2004*.

M. Maamouri, A. Bies and S. Kulick. (2006, forthcoming). 'Diacritization in Arabic Treebank Annotation and Parsing'. University of Pennsylvania.

M. Maamouri, D. Graff, H. Jin and C. Cieri. (2004b). 'Dialectal Arabic Orthography-based Transcription & CTS Levantine Arabic Collection'. *EARS PI Meeting and RT-04 Workshop*, IBM Executive Conference Center, Palisades, NY, USA. November 7-11, 2004. www.sainc.com/richtrans2004/

A. Messaoudi, L. Lamel and J-L. Gauvain. (2004). 'The LIMSI RT-04 BN Arabic System'. Proceedings of the *EARS RT-04 Workshop*, 29.

R. Nelken and S. M. Shieber. (2005). 'Arabic diacritization using weighed finite-state transducers'. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*. Association for Computational Linguistics. Ann Arbor, Michigan.
URL {http://www.aclweb.org/anthology/W/W05/W05-0711}: 79-86.

H. Safadi, O. Dakkak and N. Ghneim. (2006). 'Computational Methods to Vocalize Arabic Texts'. *Second Workshop on Internationalizing SSML*, Crete, 30-31 May 2006.

D. Vergyri and K. Kirchhoff. (2004). 'Automatic diacritization of Arabic for Acoustic Modeling in Speech Recognition'. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages*. COLING 2004, Geneva: 66-73.

I. Zitouni, J. S. Sorensen and R. Sarikaya. (2006). 'Maximum Entropy Based Restoration of Arabic Diacritics'. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics. July 2006 Workshop on Computational Approaches to Semitic Languages*. Association for Computational Linguistics. Sydney, Australia. URL {http://www.aclweb.org/anthology/P/P06/P06-1073}: 577-584.