Developing and Using a Pilot Dialectal Arabic Treebank

Mohamed Maamouri[a], Ann Bies[a], Tim Buckwalter[a], Mona Diab[b], Nizar Habash[b], Owen Rambow[b], Dalila Tabessi[a]

[a]Linguistic Data Consortium
University of Pennsylvania
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
{maamouri,bies,timbuck2,dtabessi}@ldc.upenn.edu

[b]Center for Computational Learning Systems
Columbia University
475 Riverside Drive, Suite 850, MC7717
New York, NY 10115 USA
{mdiab,habash,rambow}@cs.columbia.edu

Abstract

In this paper, we describe the methodological procedures and issues that emerged from the development of a pilot Levantine Arabic Treebank (LATB) at the Linguistic Data Consortium (LDC) and its use at the Johns Hopkins University (JHU) Center for Language and Speech Processing workshop on Parsing Arabic Dialects (PAD). This pilot, consisting of morphological and syntactic annotation of approximately 26,000 words of Levantine Arabic conversational telephone speech, was developed under severe time constraints; hence the LDC team drew on their experience in treebanking Modern Standard Arabic (MSA) text. The resulting Levantine dialect treebanked corpus was used by the PAD team to develop and evaluate parsers for Levantine dialect texts. The parsers were trained on MSA resources and adapted using dialect-MSA lexical resources (some developed especially for this task) and existing linguistic knowledge about syntactic differences between MSA and dialect. The use of the LATB for development and evaluation of syntactic parsers allowed the PAD team to provide feedback to the LDC treebank developers. In this paper, we describe the creation of resources for this corpus, as well as transformations on the corpus to eliminate speech effects and lessen the gap between our pre-existing MSA resources and the new dialectal corpus.

1. Introduction

The Arabic language is a collection of spoken dialects and a standard written language. The dialects show phonological, morphological, lexical, and syntactic differences, although the standard written language is the same throughout the Arab world: Modern Standard Arabic (MSA). MSA is also used in some scripted spoken communication (news broadcasts, parliamentary debates). MSA is not a native language (children do not learn it from their parents but in school). Most native speakers of Arabic are unable to produce sustained spontaneous MSA. The most salient variations among the dialects are geographic and social.

The multidialectal situation has important negative consequences for Arabic natural language processing (NLP): since the spoken dialects are not officially written, it is very costly to obtain adequate corpora, even unannotated corpora, to use for training NLP tools such as parsers. While it is true that in unofficial written communication, in particular in electronic media such as web logs and bulletin boards, often ad hoc transcriptions of dialects are used (since there is no standardized orthography), the inconsistencies in the orthography reduce the value of these corpora. Furthermore, there are almost no parallel corpora involving one dialect and MSA.

The 2005 Johns Hopkins University (JHU) summer workshop on Parsing Arabic Dialects (PAD) took up the challenge to develop techniques for parsing Arabic dialects that do not rely on the presence of large, indeed any, dialect treebanks. The approach taken was to leverage the large available MSA resources by exploiting MSA/dialect similarities and addressing known differences. The creation of the small Levantine Arabic Treebank (LATB) we discuss here was intended to provide development and test sets for the JHU workshop – not training data. Additionally the JHU workshop approaches (Rambow et al., 2005) required the development of dialect-MSA lexicons. These lexicons were developed in tandem with the dialectal treebank to ensure orthographic consistency.

In this paper, we describe the creation of the Levantine Arabic Treebank and the associated lexical resources as well as transformations on the corpus to eliminate speech effects and lessen the gap between our pre-existing MSA resources and the new dialectal corpus.

2. Dialectal Arabic

Because of its socio-political characteristics, highly complex morphology and significant dialectal differences, Arabic continues to challenge the NLP community. Speakers of Arabic use a variety of mutually intelligible dialects, which vary phonologically, morphologically, syntactically, lexically, geographically, and socially, but are rarely written and therefore without stable writing conventions. The dialect used in this corpus is Levantine Arabic (LA), specifically Jordanian. This corpus of Levantine Conversational Telephone Speech was collected in 2004 at the Linguistic Data Consortium (LDC) for the EARS project (http://www.ldc.upenn.edu/Projects/EARS/Arabic).

The use of Arabic speech enables new kinds of questions to be examined objectively and for the first time. For example, how important is the role of vowels in dialectal Arabic word recognition? What happens to the syntactic structure of Arabic once all case endings and mood markings are dropped? Even ideas of educated professionals in this domain are based on guesswork. In addition, independent of the language at hand, spoken conversational language contains significant disfluency, which requires adjustments in both morphological and syntactic annotation, as well as further development to automatic tools such as parsers and taggers.

3. Morphological Analysis of Levantine Arabic

The morphological analysis of LA using an experimental output from a slightly modified version of the Buckwalter MSA morphological analyzer[1] failed. One important outcome of this experience was the design and use of a new morphological analysis and annotation routine. Because of time constraints, the development of LA tools was impractical. Since modified MSA tools did not produce output that could be used to assist annotation, all of the annotation for the LATB was manual (with the addition of a few automatic post-annotation consistency checks). The basis of this approach was to use a wordlist of the LA data sorted by frequency, to manually annotate the most frequent surface forms first, and then to perform pattern matching operations to identify potential new prefix-stem-suffix combinations among the remaining un-annotated words in the list. Through this work and future work we hope to develop tools for the morphological analysis of dialectal Arabic so that future annotation can be partially automated.

Morphological/Part-of-Speech/Gloss (MPG) tagging included (1) morphological analysis, (2) part-of-speech tagging, and (3) glossing, where we provided each morph with an English gloss, as in the following example:

INPUT STRING: **بيجوز**
LOOK-UP WORD: byjwz
SOLUTION 1: (biyjuwz) [jAz-u_1]
biy/IV3MS+juwz/IV
GLOSS: he/it + be allowed/be possible

MPG tagging of dialectal Arabic differed in numerous respects from the MPG tagging of MSA newswire, as the source language for LA is speech rather than text. In MSA newswire data, for example, diacritics representing case endings and other short vowel inflection are not in the text itself, and so must be decided upon by annotators during MPG annotation (Maamouri & Bies, 2004). In dialectal Arabic speech data, on the other hand, the short vowel inflections are present in the speech itself and to some extent in the transcript, so MPG tagging must be performed in agreement with both the speech data and its transcript.

The following areas posed significant challenges: (a) Creation of a dialectal Arabic MPG tagset. A preliminary MPG tagset was created from our previous MSA analyses and then hand-annotated. (b) The vocalization is not consistent in the transcription of the LA data. In the MPG tagging of the unvocalized transcript of dialectal Arabic, the short vowels and diacritics provided by the extra layer of careful MSA-based orthographic transcription of dialectal speech were used to represent the canonical vocalization. However, this vocalization may differ on a regional or even individual basis. (c) Major modifications were observed in the dialectal Arabic verb system. The MSA Morphological Analyzer was unable to handle the morphology of the dialectal verb system, especially because of differences in the set of verbal affixes and also in passive verbal forms. (d) While it was true that there were significant Dialectal Lexicon differences, they only needed to be compiled and translated. Finally, (e) false starts and pauses in speech resulted in an important number of incomplete words, and these words were transcribed with the disfluencies marked by means of the appropriate meta-language tags. Incomplete words were analyzed by the morphological analyzer as either orthographic errors ("word not found") or as false positives, which can only be discovered through human scrutiny of the analyzed output.

We have omitted more detailed morphological information in this presentation in order to concentrate on the dialect-specific syntax.

4. Syntactic Analysis of Levantine Arabic

In treebanking LA data, we investigated and developed syntactic annotation guidelines to accommodate novel structures. We started with the treebank guidelines for MSA (Bies & Maamouri, 2003) and adapted them as necessary to account for issues of spoken and dialectal language. We also address a number of syntactic issues that emerged with LA when compared with MSA, investigating whether the LA structures parallel the MSA structures (as with the question of whether dialectal Arabic maintains the same underlying VSO word order as MSA) or if LA required novel structural analysis (as with the dialectal use of present/active participles).

4.1. Conversational Speech Effects and Disfluencies

The disfluencies, restarts, and speech constructions found in LA conversational telephone speech closely parallel those found in English speech, with the addition of issues specific to Arabic conversational speech such as frequent inconsistencies based on the lack of standardized written forms and a lack of norms for writing practices. The syntactic analysis of disfluencies followed closely the style adopted for the English Switchboard Treebank (Meteer/Taylor, 1995).

The tree in (1) shows an example of a treebanked sentence from the LATB. The tree is simplified over what is actually in the treebank in that we have omitted detailed morphological information for reasons of presentation. We see a false start (marked by EDITED; in this case, the false start is repeated verbatim).

1. Eurs b- Eurs bnt xAltiy qryb
عُرس ب- عُرس بنت خالتي قريب

```
(S (EDITED (NP Eurs عُرس wedding
                (NP-UNF b))
                +)
    (NP-SBJ Eurs عُرس wedding
            (NP bint بِنت daughter
    (NP xAl+at- - خَالَت aunt
                -iy))) يِ my
    (ADJP-PRD qariyb))) قَرِيب near
```

"The wedding of my aunt's daughter is near"

[1] Buckwalter, Tim (2004): Buckwalter Arabic Morphological Analyzer Version 2.0, LDC Corpus Catalog No. LDC2004L02.

4.2. Levantine Arabic Active Participles

Developing a dialectal Arabic Treebank for the first time raised new syntactic issues, including the treatment of active participles. In MSA and the MSA Treebank, nearly all active participles could be treated straightforwardly as adjectives. One of the major differences we found in developing the dialectal LATB was the frequent occurrence of active participles with verbal behavior. This led us to a dual treatment of active participles as either adjectival or verbal depending on context, similar to our dual treatment of MSA gerunds and participles as nominal or verbal depending on context (Maamouri & Bies, 2004).

Our default treatment of active participles in LA is as adjectives, due to their predominantly adjectival behavior (negation patterning with adjectives rather than verbs, the lack of person agreement or tense, word order with respect to subject noun phrases). Active participles behave like adjectives with *mi$* in pre-word position as in (2a) and (b) and not like matrix verbs which are negated with a circumfix of the prefix *mA* and the suffix *-$* as in (2c):

2a. أنا مش عارفة <ana mi$ EArfa
"I + (am) not + knowing"

b. الصَوت مِش واضَح Al+Sawt mi$ wADiH
"The + voice + (is) not + clear"

c. ما يشتغلش mA+ya$tagil+$
" not + (He does)work + (not)"

3. أنا مش عارفة

```
(S (NP-SBJ <anA أنا) I
   (ADJP-PRD (PRT mi$ مش) not
             EArf+ap عارفة)) knowing
```

"I don't know"

LA active participles were treated as verbal only if they exhibit specific verbal behavior – occurring either with an accusative direct object (as in (4) below) or in a raising verb construction (as in (5) below). However, because even the active participles with explicit verbal features also exhibit the canonical adjectival features, we treat them as a secondary verbal predicate (S-PRD) rather than as the matrix verb of the sentence.

4. إنتو سامعيّني

```
(S (NP-SBJ-1 <intuw إنتو) you (plural)
   (S-PRD (VP sAmEiyn سامعين hearing
              (NP-SBJ-1 *)
              (NP-OBJ niy ني)))) me
```

"You are hearing me."

5. رايح أشترك فيه

```
(S (NP-SBJ-1 *)
   (S-PRD (VP rAyiH رايح going
              (NP-SBJ-1 *)
              (S (VP >a$tarik أشترك
                        ( I am) take part
                    (NP-SBJ *)
                    (PP fiy في in
                      (NP –h ه))))))
                                    it
```

"I am going to take part in it."

4.3. VSO vs. SVO Word Order in LA

According to Mohammad (2000), and in part as a result of the loss of case endings in all Arabic dialects, Palestinian Arabic, a Levantine dialect, allows 3 possible word orders: VSO, VOS and SVO. In our analysis of LA, we opted for VSO as the underlying word order (as in MSA) in spite of claims to the contrary (Eid, 1990). We made this choice primarily because existing research does not provide conclusive evidence that demonstrates the clear dominance of either VSO or SVO as the underlying word order in LA or any other Arabic dialects. In their final report on the JHU Summer Workshop on "Parsing Arabic Dialects," Rambow et al. (2005) show that MSA and LA allow both word orders. They clearly indicate that the choice of an SVO order in LA sentences is 'not a strict requirement, but a strong preference.' Other corpus-based linguistic studies (such as Brustad, 2000) argue the frequency of both typologies. In fact, the situation often varies from text to text and from context to context. Our choice of having a VSO word order for both MSA and LA was also motivated by the methodological consequences that may incur from an SVO underlying order when we are frequently also confronted with VSO sentences in the targeted corpus. An example from the corpus of a VSO sentence with a full NP subject is (6) below:

6. و الله بتعني كثير العيلة إلي

```
(S (PRN (PRT wa- و)  and
        (NP All~`h الله)) Allah
   (VP bi+ti+Eniy بتعني  (it) means
       (NP-OBJ kaviyr كثير) very much
       (NP-SBJ Al+Eiyl+ap العيلة) the family
       (PP <il- إل  to
           (NP –iy ي))))) me
```

"By God, the family means a lot to me"

Surface SVO word order is shown as the topicalization of the subject in the Levantine Arabic Treebank (in (7) below), as in the MSA Treebanks:

7. أنا عم بدرس طبّ بشري

```
(S (NP-TPC-1 >anA أنا) I
   (VP (PRT Eam عم)  currently
       ba+-LRB-null-RRB-+drus بدرس study
       (NP-SBJ-1 *T*)
       (NP-OBJ Tib~ ba$ariy~ طبّ بشري)))
                                human medicine
```

"I currently study human medicine"

5. Deployment and Use of the LATB

The goal of PAD team at the JHU summer workshop was to work on syntactic parsing of Arabic dialect, specifically using existing resources in MSA, including corpora and tools (Diab et al., 2004; Habash & Rambow, 2005; Bikel, 2004), and to modify them using various strategies. We specifically did not want to annotate a dialect corpus and train new tools on the new dialect corpus. The motivation for this approach was to investigate to what extent resources from closely related language variants can be used (in our case, MSA for the

dialects), since the large number of dialects (and their uncodified nature) makes it unlikely we will ever be able to develop sufficiently large treebanks for all dialects. Thus, we used the LATB only for development and testing purposes, not for training. For our research, the LATB was thus crucial. However, this means that our training data (MSA) differs from our testing data (LA) in three dimensions:

- The linguistic differences between MSA and the dialect (lexical, morphological, and syntactic differences).
- Different domain: the MSA data is primarily in the politics and sports domains, while the domain of the LA data covers issues such as family, and the purpose of the data collection.
- Different genre: the MSA data is newswire, while the LA data is spoken telephone conversations.

For our purposes we are interested in the linguistic differences, and how to overcome them. We discuss lexical differences below, and refer to (Chiang et al., 2006) for a discussion on our treatment of morphological and syntactic differences. It is impossible to overcome domain differences, and these affected our performance. However, we attempted to mitigate the genre differences by transforming the LATB to look more like the MSA Treebank, as discussed below.

6. Treebank Transformations

Since the LATB data is a speech genre, it was annotated for speech effects such as disfluencies. Therefore, in order to bridge the gap in genre, our goal was to render the LATB closer to the MSA text on which the parsers are trained. Accordingly, we applied a series of transformations to the LATB. We essentially removed the speech effects, and checked for syntactic well-formedness and consistency. The speech effects which we removed were

- Parentheticals marked as PRN in the LATB annotations. An example of a parenthetical would be the word *yEny* which is a speech filler equivalent to *like* or *you know* in English as well as oath type words such as *wAllh* meaning 'by God'. Subtrees rooted in nodes with non-terminal PRN were removed.
- Interjections, marked as INTJ. An example of an interjection is "|" which is an alif with a glottal stop indicating some form of stuttering. Subtrees rooted in nodes with non-terminal INTJ were removed.
- Constituents that have unfinished nodes or leaf nodes, marked with the dashtag -UNF. An example of such unfinished constituents is the following:

  (PP (PREP l- لـ for)
  (NP-UNF (PARTIALWORD -l>s لإس))

  NB we do not provide a translation for the partial word since we do not know what the speaker intended to say. In this case the whole PP constituent was removed to satisfy our wellformedness conditions.
- Speech repairs, whose reparandum (the part which was replaced by subsequent speech) is marked with the non-terminal EDITED. An example of an edited node is the following:

  (EDITED
  (PP (PREP min من from)
  (NP (NEG_PART gayr غير without)
  (NOUN+NSUFF_FEM_SG >usr+ap اسرة family)))
  (DISFL +))

  "Without family"

  The beginning of an EDITED node is marked with EDITED and the end of this constituency is marked with a pre-terminal DISFL label. Moreover, we removed the resulting singleton trees from the LATB. All subtrees rooted in EDITED nodes are removed from the trees.

We used four tree tools for these transformations, *cat-tree*, *clean-tree* (Chiang, pc), Tregex and Tsurgeon (Levy & Galen, 2006). *Cat-tree* checks for consistency and well-formedness of the syntactic trees, and it separates out multiple trees in an utterance. This latter feature came in handy since the annotation style preferred faithfulness to the utterance turn in rendering the trees, thereby allowing multiple trees per line. *Clean-tree* removes resulting null sentences and null constituents. Tregex and Tsurgeon are used to specify constituents and either remove them or transform them, as discussed above.

The LATB contained 6639 trees. After running all of these clean-ups and transformations, the LATB was reduced to 3979 trees with no speech effects and no singleton trees. In the process, we came across anomalies resulting from the rapid manual quality check by the LDC. These anomalies were reported back and fixed in a subsequent release. Moreover, using Tsurgeon, we were able to render trees composed of several embedded sentences into several trees while maintaining a consistent treebank.

In the process of using the treebank for development purposes, we scrutinized many of the syntactic structures in some detail in order to understand the behavior of our prototype parsers. This provided an excellent opportunity to provide feedback on the syntactic annotation. As an example, sentential subjects in MSA have an obligatory complementizer (or subordinating conjunction), while this is not the case in LA. The complementizer-free sentential subjects in LA were initially annotated as S-SBJ, but sometimes also as an SBAR-SBJ with an empty complementizer. The annotation was standardized in a subsequent release.

7. Levantine-MSA Dictionary

The approaches used in the JHU Parsing Arabic Dialect workshop required the presence of a LA-MSA dictionary. The dictionary was used in translating LA sentences to MSA or in MSA Treebank conversion to LA depending on the parsing approach (Rambow et al., 2005; Chiang et al., 2006).

This task was more complicated than typical creation of machine readable dictionaries because of the total lack of LA-MSA resources, whether parallel text or paper dictionaries. Natural parallel MSA-Dialect material doesn't exist because of the Arab perception of these two as being one language that is used in different contexts. Paper dialect dictionaries are usually for non-Arabic speakers, e.g. Levantine-French or Egyptian-English.

7.1. Dictionary Format

To minimize the overhead of morphological analysis and generation needed in the translation process and to have a single dictionary format used by all approaches, the LA-MSA dictionary created was in the morphologically inflected form of the Arabic Treebank and the LATB tokens (Bies & Maamouri, 2003; Buckwalter & Maamouri, 2004). Thus, proclitics (e.g., و+ w+ 'and') and enclitics (e.g., ها+ +hA 'her') separated in Treebank creation were not included in dictionary-inflected forms. Verbal forms varying in gender, number and person were included, however. Nominal and adjectival forms including the definite article (ال+ Al+ 'the') were also included since the definite article was not tokenized off in the LATB. Table 1 includes a sample of entries with added transliterations. The English glosses are for LA.

Levantine		POS	MSA		English
ايه	<yh	UH	نعم	nEm	yes
إنتو	<ntwA	PRP	أنتم	<ntm	you (pl.)
+كي	+ky	PRP	ك+	+k	her
كمان	kmAn	RB	ايضاً	AyDAF	also
كمان	kmAn	RB	كذلك	k*lk	also
اللي	Ally	WP	الذي	Al*y	who
اللي	Ally	WP	التي	Alty	who
شو	$w	WP	ماذا	mA*A	what
كيف	Kyf	WRB	كيف	kyf	how
شلون	$lwn	WRB	كيف	kyf	how
بحكي	bHky	VBP	أتكلم	>tklm	I speak
أحكي	>Hky	VBP	أتكلم	>tklm	that I speak
منحكي	mnHky	VBP	نتكلم	Ntklm	we speak
حكيت	Hkyt	VBD	تكلمت	Tklmt	I spoke
العيلة	AlEylp	NN	العائلة	AlEA}lp	the family
عيلة	Eylp	NN	عائلة	EA}lp	Family

Table 1: Sample MSA-Levantine dictionary entries

7.2. Dictionary Creation

We investigated four parallel paths for dictionary creation that produced four sub-dictionaries: Automatic-Bridge, Egyptian-Cognate, Human-Checked and Simple-Modification.

A. The Automatic-Bridge dictionary was created by using English as a bridge language between MSA and LA. English glosses for MSA were provided by the Buckwalter analyzer and its extension to LA described in Section 3 above. We only used the lexemes that were used in the LATB, as opposed to all the lexeme choices produced by the analyzer. The reason for this is that the LA analyzer included a lot of irrelevant MSA readings that were not chosen in the LATB.

B. The Egyptian-Cognate dictionary was a subset of Levantine-Egyptian cognate words in an Egyptian-MSA lexicon (2,500 lexeme pairs corresponding to 1800 Egyptian lexemes) developed at Columbia University as an extension to the monolingual lexicon of the LDC's Egyptian CallHome project.

C. The Human-Checked dictionary was created by a human lexicographer who cleaned a portion of the noisy union of the first two dictionaries. The lexicographer removed incorrectly-assigned entries and added missing MSA entries. Due to time constraints, only 600 LA lexemes were checked. These correspond to the most common 200 verbs and 400 nouns. The total number of lexeme pairs is over 4,700. The reason for the difference in lexeme-pair/lexeme ratio between the LA human-checked dictionary and the Egyptian human-checked dictionary was a result of the process of dictionary creation. In the case LA, the lexicographer was given a large noisy automatically generated dictionary to prune. Whereas, in the case of the Egyptian dictionary, the lexicographer created the translations directly since there was no Egyptian-English dictionary that could have been used for creating a noisy bridge dictionary.

The use of the lexeme level of representation speeded up the process of dictionary cleaning by (a) reducing the number of entries from all present surface forms to underlying forms and (b) minimizing word ambiguity decisions for the lexicographer in a principled way by removing morphological ambiguity and focusing on lexeme homonymy. The disadvantage of using lexemes is that morphological analysis and generation are required to map from inflected LA to inflected MSA.

A combined lexeme-based dictionary was created from the above three dictionaries. In case of overlapping entries, the preference order was Human-Checked > Egyptian-Cognate > English-Bridge. An inflected dictionary for all the LA words appearing in the development data was created using this lexeme-based dictionary and a simple mapper of LA inflectional features to MSA inflectional features. The inflectional features for LA words were provided by the Buckwalter analyzer. The generation of the MSA forms was done using the MSA generation system Aragen (Habash, 2004).

D. Finally, **The Simple-Modification** dictionary was created by minimal modification to the LA inflected forms to look more MSA-like. This dictionary covered all 2190 types in the development data. Around 24.5% of the types were not in any way MSA-like. Examples of modifications include

(a) Orthographic normalization such as ta-marbuta restoration after tokenization: (طاولت TAwlt 'table) is mapped to (طاولة TAwlp);

(b) Word form modification such as Hamza insertion: (أغنيا >gnyA 'rich pl.') is mapped to (أغنياء >gnyA');

(c) LA morphology modification: (بشرب b$rb 'I drink') is mapped to (أشرب >$rb);

(d) In extreme cases, the word was fully changed (translated) since there was no word in MSA similar to it. For example, (كمان kmAn 'also') is mapped to (ايضا AyDAF).

The majority of changes done (for the 24.5% of non-MSA-like types) were morphological (46.1%). Orthographic changes were 22.8%. Word form modifications were 18.6%. And Translation cases were 12.5%. This dictionary was created in 8 hours using one lexicographer.

Additionally, the dictionary was enriched with induced translation probabilities (Rambow et al., 2005).

7.3. Dictionary Use in Dialect Parsing

Three experimental settings were used to test the contribution of the LA-MSA dictionary: no dictionary, small dictionary and large dictionary. Both small and large dictionaries were subsets of the union of the four sub-dictionaries described in the previous section. We took the subsets to create unbiased conditions for comparing between development and test data, since many decisions for creating the sub-dictionary were influenced by observations from the development data. Moreover, we filtered the large dictionary to exclude pairs with MSA-like words on the LA side and pairs with MSA words that do not appear in the MSA Treebank used for training the parsers. The filtering was done to limit dictionary size without affecting its contribution.

The small dictionary comprised 321 LA-MSA word form pairs covering LA closed-class words and a few frequent open-class words. The large dictionary contained the small dictionary and an additional 1,560 Levantine-MSA word pairs.

The results of the parsing experiments show that the LA-MSA dictionary was the biggest contributor to the improved parsing accuracy. Using the small dictionary improved the F1 labeled constituent score for both dialect parsing conditions of using no part-of-speech (POS) tags in the input, and gold POS tags on the input. We reported more than a 10% reduction on F1 labeled constituent error for the test set when using the small dictionary as opposed to the baseline of using no dictionary. Higher contribution was seen on the development set. A further improvement was gained when using the large lexicon for parsing LA in the 'no POS tags in the input' condition, but this improvement disappears when we use the large dictionary with gold POS tags. We suspect that the added translation ambiguity from the large dictionary is responsible for the drop.

8 Conclusion

Our experience shows that rapid development of dialectal treebanks is feasible, and that guidelines and resources for annotation of the Standard language can be adapted with less effort than for an entirely new language. When the construction of a large treebank (for standard machine learning) is impossible, a small, rapidly developed treebank is crucial in developing NLP tools for dialects. First, the treebank serves as a source of insight on the phenomena that need to be addressed; second, the treebank serves as a development corpus to aid the NLP tool developer in choosing among possible alternatives. Finally, our experience also points to the necessity of creating small dialect-Standard dictionaries.

9 References

Bies, A. and Maamouri, M. (2003). *Penn Arabic Treebank Guidelines*. URL: http://www.ircs.upenn.edu/arabic/Jan03release/guidelines-TB-1-28-03.pdf

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

Brustad, K. (2000). *The Syntax of Spoken Arabic: A comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti dialects*. Washington, DC.: Georgetown University Press.

Buckwalter, T. and Maamouri, M. (2004). *Guidelines for the Transcription of Arabic Dialects (EARS)*. URL: http://www.ldc.upenn.edu/Projects/EARS/Arabic/Guidelines_Levantine_MSA.htm.

Chiang, D., Diab, M., Habash, N., Rambow, O. and Shareef, S. (2006). Arabic Dialect Parsing. In *Proceedings of the European chapter of the Association of Computational Linguistics EACL*.

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

Eid, M. (1990). *Perspectives on Arabic Linguistics I. Papers from the First Annual Symposium on Arabic Linguistics*. Philadelphia: John Benjamins.

Habash, N. (2004). Large scale lexeme based Arabic morphological generation. In *Proceedings of Traitement Automatique du Langage Naturel (TALN-04).* Fez, Morocco.

Habash, N. and Rambow, O. (2005). Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *Proceedings of the Conference of American Association for Computational Linguistics (ACL'05).* Ann Arbor, Michigan.

Levy, R. and Galen, A. (2006). Tsurgeon & Tregex. In *Proceedings of LREC-2006*. Genoa, Italy.

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

Meteer, M. et al. (1995). *Dysfluency Annotation Stylebook for the Switchboard Corpus*. Revised by Ann Taylor, 1995. Ms., University of Pennsylvania.

Mohammad, M. (2000). *Word Order, Agreement and Pronominalization in Standard and Palestinian Arabic*. Current Issues in Linguistic Theory. Philadelphia: John Benjamins.

Rambow, O., Chiang, D., Diab, M., Habash, N., Hwa, R., Lacey, V., Levy, R., Nicols, C., Shareef, S., Simaan, K. (2005). *Parsing Arabic Dialects*. Technical Report, The Johns-Hopkins University, 2005 Summer Research Workshop.