Expanding Arabic Treebank to Speech: Results from Broadcast News

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

Abstract

Treebanking a large corpus of relatively structured speech transcribed from various Arabic Broadcast News (BN) sources has allowed us to begin to address the many challenges of annotating and parsing a speech corpus in Arabic. The now completed Arabic Treebank BN corpus consists of 432,976 source tokens (517,080 tree tokens) in 120 files of manually transcribed news broadcasts. Because news broadcasts are predominantly scripted, most of the transcribed speech is in Modern Standard Arabic (MSA). As such, the lexical and syntactic structures are very similar to the MSA in written newswire data. However, because this is spoken news, cross-linguistic speech effects such as restarts, fillers, hesitations, and repetitions are common. There is also a certain amount of dialect data present in the BN corpus, from on-the-street interviews and similar informal contexts. In this paper, we describe the finished corpus and focus on some of the necessary additions to our annotation guidelines, along with some of the technical challenges of a treebanked speech corpus and an initial parsing evaluation for this data. This corpus will be available to the community in 2012 as an LDC publication.

Keywords: Arabic Treebank, Broadcast News, treebank construction

1. Introduction

Treebanking a large corpus of relatively structured speech transcribed from various Arabic Broadcast News (BN) sources has allowed us to begin to address the many challenges of annotating and parsing a speech corpus in Arabic. The now completed Arabic Treebank BN corpus consists of 432,976 source tokens (517,080 tree tokens)[1] in 120 files of manually transcribed news broadcasts. This corpus will be available to the community in 2012 as an LDC publication.

We raised a variety of preliminary issues of metadata, transcription, audio signal, and SU annotation in Maamouri et al. (2010b). In this paper, we describe the finished corpus and focus on some of the necessary additions to our annotation guidelines, along with some of the technical challenges of a treebanked speech corpus and an initial parsing evaluation for this data.

2. Challenges of Spoken Language in Arabic Broadcast News Data

Using transcribed BN data as a source for Arabic treebank annotation allowed us to face challenges inherent in all speech data and also to begin to investigate the impact of Arabic dialect issues.

Because news broadcasts are predominantly scripted, most of the transcribed speech is in Modern Standard Arabic (MSA). As such, the lexical and syntactic structures are very similar to the MSA in written newswire data. However, because this is spoken news, cross-linguistic speech effects such as restarts, fillers, hesitations, and repetitions are common.

There is also a certain amount of dialect data present in the BN corpus, from on-the-street interviews and similar informal contexts. This data represents a variety of Arabic dialects, and presents a range of issues.

In addition, the (manual) process of transcription itself (and potential transcription errors inherent in the process) affects downstream annotation in Arabic-specific ways.

2.1 Cross-linguistic speech effects

Speech effects occur in a similar way across languages (Shriberg, 1994): restarts, repetitions, hesitations, unfinished constituents, etc. all occur in both English and Arabic, for example, and they have a similar distribution with respect to the syntax. Annotation guidelines for the treebanking of speech in English were developed for the Penn Treebank (Taylor, 1996). These guidelines were used as the basis for the development of Arabic Treebank (ATB) annotation guidelines for speech, focusing on the transcribed BN corpus at hand (Maamouri et al., 2009). Some of the annotation solutions developed are discussed in Section 3.1.

[1] "Source tokens" are the whitespace/punctuation-delimited tokens (offset annotation) on the source text that receive a morphological analysis through the SAMA analyzer. The "tree tokens" result from splitting up these source tokens into subsequences as appropriate for the annotation of syntactic structure. See Kulick, Bies and Maamouri (2010) for a detailed description of the difference between source and tree tokens in the Arabic Treebank.

2.2 Arabic dialect issues in BN data

Out of the 517,080 tree tokens in this corpus, 4,941 (less than 1%) received the DIALECT part-of-speech (POS) tag. (These 4,941 tree tokens arose from 4,760 source tokens with at least one tree token with a DIALECT tag.) We labeled all such tokens in this corpus with the simple POS tag "DIALECT" regardless of function. This was feasible because the percentage of dialect tokens is quite small in BN. However, even the limited dialect speech that occurs in this corpus has allowed us to begin an investigation into the relatively complex challenges that arise with Arabic dialect data. Some annotation solutions are discussed in Section 3.2.

2.3 Transcribed spoken Arabic

The Arabic BN data was collected and manually transcribed, as described in Paulsson, et al. (2009). Some of the issues specific to the transcription of Arabic as they relate to treebank creation are described as well in Kulick, Bies and Maamouri (2010). Any errors in transcription are given the part-of-speech tag "TRANSERR" in this corpus (similar to the TYPO part-of-speech tag that is used for text corpora).

For example, the token *mtwvrp* متوثـــرة occurs in the corpus and is a transcription error (it should be *mtwtrp* متوتـــرة (tense)). The token itself is left as transcribed for consistency with other annotation work on the same transcribed corpus. In the treebank annotation, the token is given the part-of-speech tag TRANSERR, but it is syntactically annotated as if it were written correctly.

```
(NP-SBJ
   (NOUN+NSUFF_FEM_PL+CASE_INDEF_NOM
             EalAq+At+N )
   (TRANSERR  mtwvrp)
   (ADJ+NSUFF_FEM_SG+CASE_INDEF_NOM
             EadA}iy~+ap+N ))
                                  علاقاتٌ متوثـــرة
   ((Tense)) relations
```

The transcribers also had the option of marking a token as an incomplete, or partial, word (with a trailing hyphen "-" on the transcribed token itself) when the speaker produced an incomplete word. Tokens of this type are all given the part-of-speech tag "PARTIAL" in the treebank, along with the lemma "partial," and they do not have glosses or vocalized forms. They are typically contained in a tree node that is marked –UNF for unfinished (see Section 3.1), and they are included as Status #4 with respect to SAMA (see Section 4.1). The partial word below was spoken probably as an incomplete form of *TbyEy* طبيعـــي (normal).

```
(ADJP-UNF (PARTIAL -طـبي Tby-,nogloss)
```

2.3.1 Can transcribed data be better than written data? A case in point: The annotation of initial *hamza*

The issues of transcription for Arabic may interact with downstream annotations, including the morphological and syntactic annotation of the Treebank. The annotation of initial hamza, for example, is an interesting annotation issue as the issues involved change from written text to transcribed speech.

In BN, all initial *hamzas* or glottal stops are heard and transcribed with either *<i-* or *>a-* (leading to different words, as in for example, إنَّ *<in~a* (is indeed) or أنَّ *>an~a* (that)). Because the distinction is made in the spoken vowels and transcribed as such, virtually no instances of the orthographically neutralized *An* ان , where the lexical distinction is not made in the written form, occur in this half a million word BN corpus.

However, in written newswire (NW) data, the neutralized *An* ان form is quite common (1.5% of the tokens in ATB3 have an *An*), forcing treebank annotators to make the distinction between the *<i-* or *>a-* forms based on context. The two forms require different part-of-speech and syntactic annotations. This is a difficult distinction in some respects, and additional morphological and syntactic annotation guidelines are necessary to distinguish the إنَّ *<in~a* vs. أنَّ *>an~a* usage of the neutralized form.

Because this distinction is part of the original transcribed speech in BN, the burden does not fall on the treebank annotators, and the annotation of these forms is quite consistent in the BN corpus. There are 984 cases of إنَّ *<in~a* (including *<in~ahu* إنَّه (it is indeed) and *li<in~ahu* لِإنَّه (because it is indeed)) and 3577 cases of أنّ *>an~a* (including أنَّه *>an~ah* (that which) and لأنَّه *li>ana~h* (for that which)), which are annotated accordingly in the BN corpus. In the case of initial hamza, then, the transcribed speech data actually presents fewer issues for downstream annotation than written NW data.

However, there are instances of transcription errors as well, and in these cases the tree should be annotated correctly, and the token that is in error is given the part-of-speech tag TRANSERR, as in the following example.

```
(S (ADJP-PRD (ADJ_COMP akovar)
             (PP (PREP min)
                 (ADVP (ADV kdh))))
   (SBAR-SBJ (TRANSERR <n)
             (S (NP-SBJ
                   (NP (NOUN+CASE_DEF_ACC
                          Hizob+a)
                       (NP (DET+NOUN+CASE_DEF_GEN
                          Al+tajam~uE+i)))
                   (NP (PRON_3MS huwa)))
                (NP-PRD (DET+NOUN+CASE_DEF_NOM
                           Al+Hizob+u)
                        (DET+ADJ+CASE_DEF_NOM
                           Al+waHiyd+u)))))
```

أكثر مِن كِدَه حِزبَ التَّجَمُّع هُوَّ الحِزبُ الوَحِيدُ ...

More than that, the Reunification Party is the only party…

3. Treebank Annotation Solutions

3.1 Cross-linguistic speech effects in the trees

Descriptions of cross-linguistic speech effects and their annotation guidelines for Arabic BN are available in Maamouri et al. (2009). These guidelines were based in large part on the annotation guidelines developed for the treebanking of cross-linguistic speech effects in the English Penn Treebank (Taylor, 1996), so as to annotate similar speech effects across languages in a similar way. Below are example trees for selected speech effects from the Arabic BN corpus.[2]

Unfinished constituents: The dashtag **-UNF** marks 'unfinished' spoken constituents, including partial words, phrases, clauses and sentences.

```
(S (NP-TPC-1 أنـا ·>anA·I)
   (VP أقُـولُ ·>a+quwl+u·I+say+[ind.]
       (NP-SBJ-1 *T*)
       (NP-TMP الآنَ ·Al+|n+a·the+time/moment)
       (SBAR أنَّ ·>an~a·that
             (INTJ أه ·>ah·uh!)
             (S-UNF (NP-SBJ المُفَـوَّضِيَّـةَ ·
                       Al+mufaw~aDiy~+ap+a
                          ·the+delegation
                          العُـلْيـا ·
                          Al+EuloyA
                          ·the+highest)
                    (INTJ أه ·>ah·uh)))))
```

أنا أقول الآن أنّ المفوضيّة العليا أه

I say now that the high delegation is uh…

Filled pauses are marked as interjections, such as the (INTJ أه ·>ah·uh) in the above example.

Restarts and repetitions: The tree node label **EDITED** is used to show the repetition and restarting of constituents that are repaired by subsequent speech.

```
(S (INTJ أه •<h·uh)
   (EDITED (EDITED (EDITED وع •wE·NO_GLOSS)
                   وَ-·wa-·and)
           (VP-UNF -استخدم •-Astxdm·
                                NO_GLOSS)
           (INTJ وَ-·wa-·and))
   (VP -يَسْتَخْدِمه •-yasotaxodimh·NO_GLOSS
       (NP-SBJ *)
       (NP-OBJ (NP المُسْتَقِـلِّيْنَ •
                   Al+musotaqil~+iyna·
                   the+independent)
               (NP-ADV خـارِجَ •xArij+a·
                           outside
                       (NP الإخْـوانِ •
                           Al+<ixowAn+i·
                          the+brothers)))))
```

أه وعـ- واستخدم ويستخدمه المستقلون خارج الإخوان

Uh, and the independent candidates other than the Brothers, use- use- used it

3.2 Annotating speech and dialect constructions

BN data includes constructions that are specific to spoken language, to broadcast style as opposed to written style, certain novel MSA usages, and some dialectal constructions.

For example, يعني *yaEoniy* (he/it means) is a frequent discourse filler in spoken Arabic, with a discourse function much like "you know" in English, and it is therefore similarly also annotated as a parenthetical whenever it occurs, whether the surrounding sentence is dialectal or MSA:

```
(PRN (S (VP yaEoniy يَـعْـنِي
            (NP-SBJ *))))
```

True dialect constructions also occur, and those received novel syntactic analyses accordingly. For example, the Levantine word بد *bid~* (wish) functions as a verb, and its tree therefore includes a subject and either an S complement or an object, even though *bid~* does not morphologically inflect like a verb.

```
(S (VP bid~
       (NP-SBJ-1 w)
       (S (VP ySiyr
              (NP-SBJ-1 *)
              (NP-PRD muhAmiy)))))
```

بدّو يصير محامي

He wants to become a lawyer

Additional dialect constructions encountered in BN and annotation guidelines developed for them are available in Maamouri et al. (2009).

[2] We use the Buckwalter transliteration system: http://www.qamus.org/transliteration.htm

BN data also includes some categories that do not require new constructions to be annotated, but where the annotation of speech data is actually simpler than the annotation of newswire (NW) text data, such as the *hamza*/glottal stop annotation discussed in Section 2.4.

4. Technical Challenges and Solutions

4.1 Status of BN corpus integration with SAMA

Arabic NLP pipelines make crucial use of the tight connection between the morphological analysis from SAMA (Maamouri et al., 2010c) and the ATB POS annotation. In the BN corpus, we have continued the enhancement from the revised NW data of including a status flag for each source token to make explicit this connection (Kulick et al., 2010).

SAMA STATUS		# BN source tokens	%	ATB3
#1	INCLUDED	415924	96.1%	84.6%
#2	LIMITED	735	0.1%	0.3%
#3	PENDING	3474	0.8%	1.3%
#4	EXCLUDED	12843	3.0%	13.9%
	TOTAL	432976		

Table 1: SAMA status in BN corpus

Table 1 summarizes the SAMA status flag results for both the current BN corpus and the ATB3 NW corpus (Maamouri et al. 2010a). We give exact numbers and percentages for the current corpus, and the percentages for ATB3 for comparison (ATB3 has 339,710 source tokens).

We briefly summarize the description of each status type while discussing the reasons for the differences in the token breakdown.

Status #1 INCLUDED IN SAMA. The source token annotation exactly matches a SAMA solution for that source token. The larger percentage of such cases in BN is due to the corresponding decrease in tokens of type Status #4.

Status #2 LIMITED SOLUTION is not a SAMA solution, but is of very limited format, i.e. without vocalization information. The percentage of tokens with this status is similar for BN and ATB3. However, in part this is due to a change in which tokens are included in Status #4, as discussed below.

Status #3 PENDING SAMA SOLUTION is not a SAMA solution, but is a manually-vocalized solution. These solutions will be subject to further review and eventual inclusion in SAMA.

Status #4 EXCLUDED FROM CHECK WITH SAMA is used for source tokens that are not expected to have a solution in SAMA. Because of the nature of the two genres, the distribution of tokens that make up Status #4 in the two corpora is entirely different. For the NW ATB3 corpus, this status consists entirely of punctuation and numbers written as digits, which are classes of tokens that essentially do not occur in the BN corpus because they are not part of the transcription specifications. For the BN corpus, the 12,843 Status #4 source tokens include 4,760 tokens with a DIALECT tag, 3,001 tokens with a TRANSERR (transcription error) tag, and 4,765 tags with a PARTIAL tag. The DIALECT tag is practically non-existent in the NW corpus, and by definition the TRANSERR tag is as well, while the analogous TYPO (typographical error) is very rare. The PARTIAL tag indicates that the token contains a metadata marker (the hyphen) signifying that the token represents a word that was unfinished in speech, and thus is also not present in the NW data. In addition, numbers in BN are transcribed as written out words rather than as digits, and so are not included as Status #4 in BN.

4.2 Parsing evaluation

During treebank construction, we parsed the BN data using the Bikel parser [3] trained on ATB3 data. With the completion of the BN corpus, we can now evaluate training and parsing on this corpus. For comparison, we used the training/test split from ATB3[4] and used a corresponding amount of data from BN for training, and a test section. We ran the parser in two modes – either free to choose a part-of-speech tag for each word, or forced to use the gold tags (see Kulick et al., 2006).

While there are many parsers available, some with better results than the Bikel parser, the Bikel parser is currently used in the annotation project – and for current purposes what we are interested in here is the comparison of the parsing of the BN and ATB3 corpora, which we would expect to have a similar comparison even using other parsers. The parsing experiments here are based only on the tree tokens, not the source tokens. This means that the tokens used for parsing assume gold tokenization and part-of-speech tags (although as discussed in this section, the parser is run in two modes, forced to use the gold tags or not). The gold tags used here are a mapped-down version of the full complex ATB tags, as described in, e.g., (Kulick et al., 2006). For some work on the problem of integrating a parser with tokenization and POS tagging, see (Kulick, 2011) and (Green and Manning, 2010).

Table 2 shows the results for sentences of length <=40 words. The BN corpus test section has many more

[3] http://www.cis.upenn.edu/~dbikel/software.html

[4] http://nlp.stanford.edu/software/parser-arabic-data-splits.shtml

sentences of length <=40, and thus more of the test section is included (91.2% compared to 74.5% for ATB3). As discussed in Section 3, the corpus contains the EDITED node to indicate repetition and restarts, which are not present in ATB3 and hard for the parser. We therefore retrained and parsed after eliminating all EDITED subtrees, with the results shown in the third row[5] of table of Table 2.

As can be seen comparing rows 1 and 3, the results for the BN corpus are somewhat below that of newswire. This is obviously an area requiring further investigation in future work, but we suspect that the decrease arises from the different nature of the corpus, as discussed above (even with EDITED subtrees removed, -UNF, DIALECT, and TRANSERR are frequent). We are pleased that the scores are so close to those for newswire, despite the difficult nature of the corpus.

	#words	Parser chooses tags	Parser uses gold tags
ATB3	17854	78.2	79.6
ATB-BN	28058	76.1	77.8
ATB-BN (EDITED removed)	28378	77.2	78.9

Table 2: Initial BN parsing results

In future work, nodes with the –UNF could be deleted as well, and we would expect the results to increase, closer to ATB3.

5. Conclusion

We have presented a number of annotation and technical lessons learned from this first large corpus of treebanked Arabic speech. Perhaps somewhat surprisingly, in addition to the challenges posed by speech data, in some respects (hamza annotation and SAMA inclusion, for example) the BN data is actually more consistent than NW data.

These lessons will inform our methodologies as we continue to expand the Arabic Treebank into less formal speech and web text domains, where a greater impact from dialects and vernacular usage is expected.

6. Acknowledgements

This work was supported in part by the Defense Advanced Research Projects Agency, GALE Program Grant No. HR0011-06-1-0003. The views, opinions and/or findings contained in this article/presentation are those of the author/presenter and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

We would also like to thank the Arabic Treebank annotators at LDC and MediaNet (Tunis, Tunisia) for their many contributions.

[5] Eliminating the EDITED subtrees resulted in more sentences of length <=40, and so #words increased.

7. References

Spence Green and Christopher Manning. (2010). Better Arabic Parsing: Baselines, Evaluations, and Analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (*COLING 2010).

Seth Kulick. (2011). Exploiting Separation of Closed-Class Categories for Tokenization and Part-of-Speech Tagging. In *ACM Transactions on Asian Language Information Processing (TALIP).* Volume 10, Issue 1, March 2011.

Seth Kulick, Ryan Gabbard and Mitchell Marcus. (2006). Parsing the Arabic Treebank: Analysis and Improvements. In *Proceedings of the 5th International Conference on Treebanks and Linguistic Theories* (TLT 2006).

Seth Kulick, Ann Bies and Mohamed Maamouri. (2010). Consistent and Flexible Integration of Morphological Annotation in the Arabic Treebank. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010).

Mohamed Maamouri, Ann Bies, Fatma Gaddeche, Sondos Krouna, and Dalila Tabessi Toub. (2009). *Guidelines for Treebank Annotation of Speech Effects and Disfluency for the Penn Arabic Treebank, v1.0*. http://projects.ldc.upenn.edu/ArabicTreebank/. Linguistic Data Consortium, University of Pennsylvania.

Mohamed Maamouri, Ann Bies, Seth Kulick, Fatma Gaddeche, Wigdan Mekki, Sondos Krouna, Basma Bouziri. (2010a). *Arabic Treebank part 3 - v3.2*. Linguistic Data Consortium, Catalog No.: LDC2010T08.

Mohamed Maamouri, Ann Bies, Seth Kulick, Wajdi Zaghouani, David Graff and Michael Ciul. (2010b). From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010).

Mohamed Maamouri, David Graff, Basma Bouziri, Sondos Krouna, Ann Bies, Seth Kulick. (2010c). *Standard Arabic Morphological Analyzer (SAMA) Version 3.1*. Linguistic Data Consortium, Catalog No.: LDC2010L01.

Niklas Paulsson, Khalid Choukri, Djamel Mostefa, Denise DiPersio, Meghan Glenn and Stephanie Strassel. (2009). A Large Arabic Broadcast News Speech Data Collection. In *Proceedings of the MEDAR Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April 22-23, 2009.

E. E. Shriberg. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis. University of California at Berkeley.

Ann Taylor. (1996). *Bracketing Switchboard: An addendum to the TREEBANK II Bracketing Guidelines*. Penn Treebank Project, University of Pennsylvania.