From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News

Mohamed Maamouri, Ann Bies, Seth Kulick, Wajdi Zaghouani, David Graff, Michael Ciul

Linguistic Data Consortium

Arabic Treebank Annotation of Broadcast News Speech

- Coordination necessary for Broadcast News (BN) data at LDC across
  - Arabic Treebank (ATB)
  - English translation of Arabic BN transcripts
  - word-level alignment of Arabic and English data
  - Corresponding English Treebank
- Robust enough to account for the new genre of data?
  - Newly improved ATB annotation pipeline
  - Revised ATB annotation guidelines
- Yes! (with some adaptation)

Issues of Broadcast News Data

- **Metadata → Do Not Annotate in Treebank**
  - Metadata to convey several kinds of information in addition to the text of what each speaker is saying
  - Coughs, laughter, background noise or music, etc.
  - Speech in a language other than Arabic, or a colloquial dialect of Arabic rather than MSA
- **Speech Effects → Annotate in Treebank**
  - Discourse markers, hesitation sounds, word fragments, mispronunciations and other disfluencies
- **Indistinct Audio Signal → Annotate if possible in Treebank**
  - ((text)) = speech could be heard but not understood, or could only be guessed at from context rather than from the audio signal
  - Cascading impact on higher-level annotations

Tool Development for ATB BN Data

- Tools adapted to filter out the metadata that ATB would ignore, while preserving the ability to align the annotation results to the initial transcripts
- Give POS annotators access to the original audio files when necessary to disambiguate doubtful words in the transcript
  - Transcribed typo “zbr” زبر ‘to prune’ in place of “brz” برز ‘to appear’
  - Transcribed typo “lmE” لمع ‘to shine’ in place of “Elm” علم ‘to learn’

Guidelines Development for BN Data

- Syntax: adapted the Penn English Treebank Switchboard annotation guidelines for use with Arabic BN data
  - Treatment of speech effects, disfluencies and metadata is not language-specific
  - Arabic-specific dialect-related structures were addressed
  - Revised and enhanced Arabic Treebank Syntactic Guidelines for general syntax
- Morphology/POS: Dialect words given DIALECT tag
  - Dialect is low frequency in highly monitored BN speech
  - Dialect guidelines under development for other speech genre

ATB Annotation Pipeline

- ATB annotation and processing pipeline improved overall
- Adapted to support the production of treebanked broadcast news corpora
- Several components devoted to handling word forms outside the vocabulary and grammatical repertoire of SAMA
  - Feedback to upgrade SAMA’s lexicon and morphotactic tables
  - Careful vetting of POS labels and glosses assigned to novel terms
- **Pipeline stages, including consistency checking at every stage**
  1. **Speech Transcription and SU Annotation**
  2. **Morphological Analyzer and Morphological/ Part-of-Speech Annotation**
  3. **Clitic Separation, Parsing, and Syntactic Annotation**
  4. **Quality Control Searches and Corrections**

1. Speech Transcription and SU Annotation

- **Speech Transcription**
  - XTrans transcription tool
  - Transcription guidelines: audio segmented into time-stamped regions
- **SU Annotation**
  - One phrasal “sentence/semantic unit” (SU) per time-stamped region of audio
  - Identify SUs
  - Segment them coherently
  - Assigning final punctuation to indicate SU type (statement, question, or incomplete)
  - SUs needed for independent downstream annotations done in parallel (e.g., translation of the Arabic transcripts into English, and treebanking of the English translations)

2. Morphological Analyzer and Morphological/ Part-of-Speech Annotation

- **Morphological Analyzer**
  - Standard Arabic Morphological Analyzer SAMA, an expansion of the Buckwalter Arabic Morphological Analyzer used in previous ATB corpora
  - For each Arabic word token and morpheme segment, lists all known/ possible annotation solutions, diacritic marks, morpheme boundaries, and Part-of-Speech (POS) labels and glosses
  - Watch out for “out-of-band” characters, such as the Persian character “keheh” mistakenly for the MSA letter “kaf” (same shape in some contexts)
  - Skip over the metadata annotations
- **Morphological/Part-of-Speech Annotation**
  - SelectPOS tool for selecting morphological/part-of-speech analysis
  - Annotators choose from SAMA solutions
  - NO_MATCH tool for words with no correct SAMA solution – limited or pending annotation entered and flagged for possible later integration into SAMA
  - New version of SelectPOS tool currently in development, to allow for proposed solutions to be entered directly for NO_MATCH tokens

3. Clitic Separation, Parsing, and Syntactic Annotation

- **Clitic Separation**
  - Clitics separated automatically according to the POS annotation
  - Segmentation necessary for treebanking phase

```
kutub/NOUN/books+i/CASE_DEF_GEN/def.gen+hi/POSS_PRON_3MS/its-his
                    →
kutub/NOUN/books+i/CASE_DEF_GEN/def.gen
hi/POSS_PRON_3MS/its-his
```

  - Dialect tokens with clitics also split

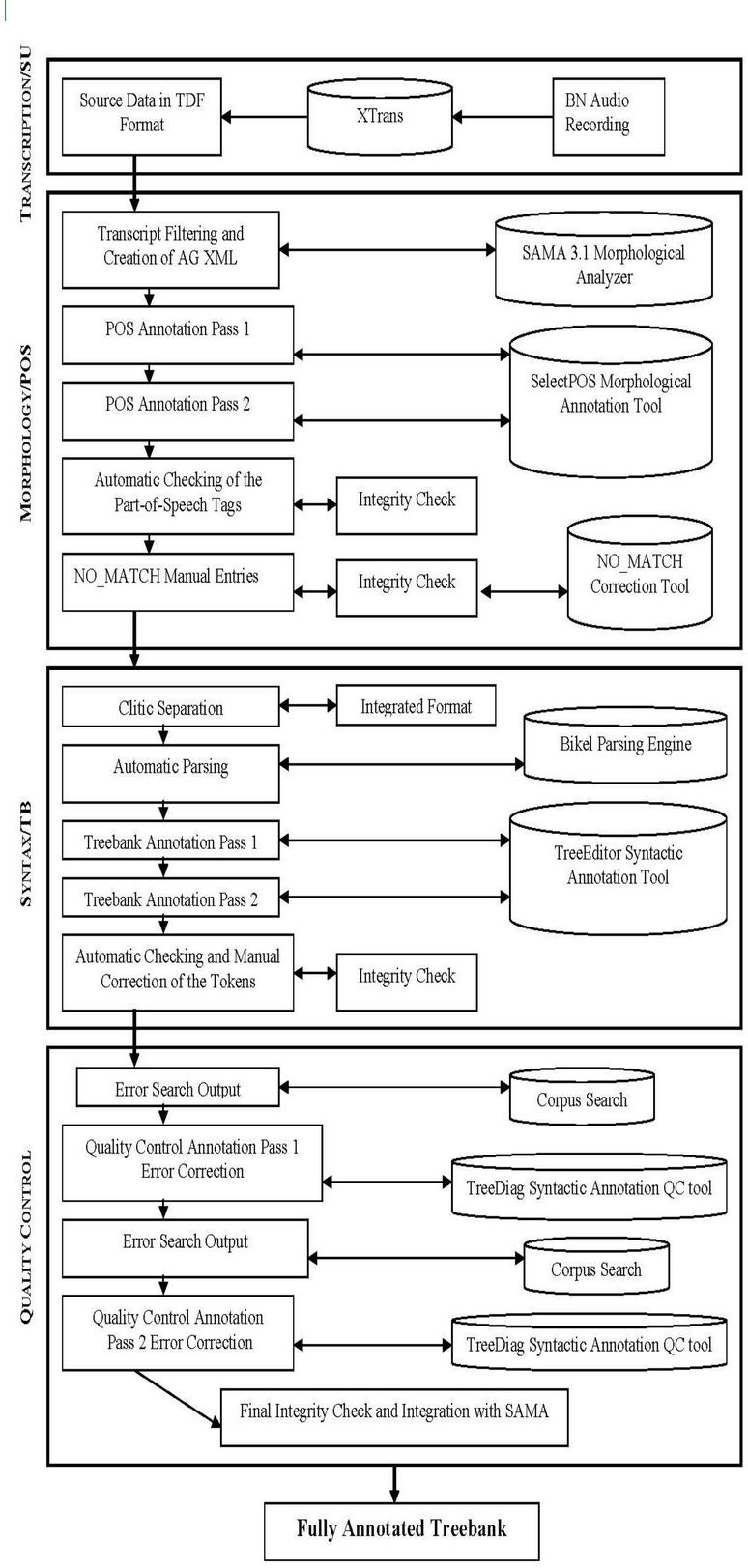
```
wa/CONJ/and+rAH/DIALECT/(he) went, started
                    →
wa/CONJ/and
rAH/DIALECT/(he) went, started
```

- **Parsing**
  - Bikel parser used to automatically create syntactic trees for treebanking
  - Input = “gold” POS annotation, as split for the treebank tokens, and “unvocalized” form of the token
- **Syntactic Annotation**
  - TreeEditor Annotation tool
  - Treebank annotation
    - Correct the parse output when necessary
    - Add function tags not included by the parser (most adverbial tags)
    - Add empty categories with appropriate co-indexing
  - Tool displays as separate listings
    - Vocalized tree tokens
    - Full morphological analysis
    - Arabic script source tokens
    - English gloss for each token
  - Limited modification of POS tags possible
  - Mark speech disfluencies (repetitions and restarts, etc.), according to the BN syntactic annotation guidelines

4. Quality Control Searches and Corrections

- **Quality Control Searches**
  - Corpus Search tool
  - 93 error-search queries to locate known problems involving improper patterns of tree structures and node labels
- **Corrections**
  - TreeDiag annotation tool: displays search results and allows annotators to click through directly to affected portion of each tree
  - Errors found are hand corrected by treebank annotators

Arabic Treebank Annotation Pipeline Stages



Conclusions

- **New challenges posed by Broadcast News data**
- **ATB’s improved pipeline and revised annotation guidelines robust enough to require few changes for BN data**
- **Similar adaptations planned in the future to account for additional new data genres (webtext and dialectal speech, …)**
- **Expect current pipeline will continue to prove flexible and robust enough to accommodate the morphological and syntactic annotation of the necessary data**