Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash and Ramy Eskander
{maamouri,bies,skulick,mciul}@ldc.upenn.edu,{habash,reskander}@ccls.columbia.edu

Parallel Development of an Egyptian Arabic Treebank and a Morphological Analyzer for Egyptian Arabic

- Egyptian Arabic = new dialect for treebank annotation
- Morphological analyzer for Egyptian Arabic (CALIMA) needed to mediate between written text and segmented, vocalized form used for syntactic trees
- Necessity of feedback loop between treebank team and analyzer team, as improvements in each area were fed to the other
- Led to close cooperation between annotation team and tool development team throughout this process, to their mutual benefit

Informal Dialectal Arabic Data

- Arabic dialects are not written or standardized → challenges for both morphological annotation and morphological analyzer
- Scarcity of normalized written Arabic dialectal resources
- Ad hoc orthography often used: significant degree of noise and high level of inconsistency in spelling, whether in Arabic script or in a Romanized representation
- Previous experience showed that Arabic dialects have to be treated as new and separate languages

Egyptian Arabic Linguistic Features

- **Phonology**: Egyptian is characterized specifically by /q/ and /ǰ/ being replaced by glottal stop /ʔ/ and /g/
  - قطن /ʔuṭn/ *cotton*, and جمل /gamal/ *camel*
- **Morphology**: Egyptian has future pro-clitics h+ and ħ+ (as opposed to the MSA equivalent s+)
- **Lexicon**: Significant lexical differences between Egyptian Arabic and MSA, with no etymological or cognate relationship
  - Egyptian Arabic بص /buSS/ *look* is أنظر /'unZur/ in MSA
- **Syntax**: Overall syntactic structures are available in both MSA and Egyptian Arabic

→ Development of specialized Egyptian Arabic Morphological Annotation Guidelines

Development of Egyptian Morphological Analyzer (CALIMA)

- CALIMA was bootstrapped using the LDC Egyptian Colloquial Arabic Lexicon (ECAL) and the CALLHOME Egyptian Arabic (CHE) corpus, developed in the 1990s
  - ECAL entries (66K entries) converted into diacritized Arabic script words & lemmas (from phonological form & undiacritized orthography)
  - Finite-state transducer (FST) implemented to map phonological form to multiple possible diacritized Arabic script forms
- Manual linguistic mapping rules followed by manual checking and correction
  - Converted ECAL examples used to construct databases of morphological analyzer
  - Manually specified orthographic variants of prefixes & suffixes used to add entries automatically

FEEDBACK LOOP

Goal = making the analyzer and treebank annotation in sync as much as possible; morphological solutions in the annotation should exactly match a solution in CALIMA

1. Interface: Annotation Process & Analyzer

- Reorganization of CALIMA tables to allow bidirectionality between words and POS tags → modified tables to translate CALIMA into a FST
- Generation of wildcard solutions for annotation of solutions not (yet) in CALIMA
  - Stem not in CALIMA → wildcard solutions, in which the stem for an open-class word (noun, etc.) would be unvocalized, but the prefixes and suffixes exactly matched the possibilities elsewhere in CALIMA
  - Closed-class items and morphemes (pronouns, etc.) should not have missing solutions
  - Restrict annotators' entry for missing solutions

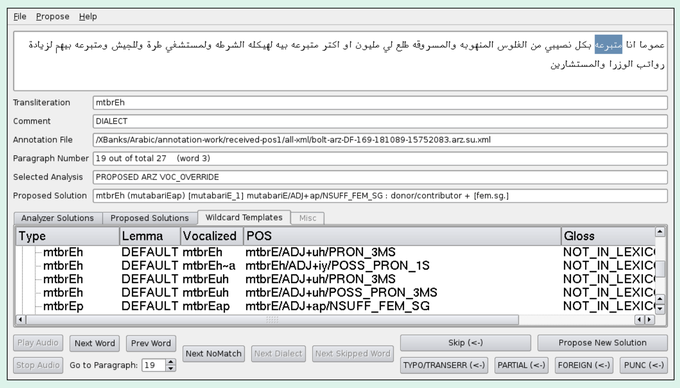

Figure 1. Wildcard annotation in the Egyptian Arabic morphological annotation tool

- First initial pilot annotation, leading to…

2. CALIMA Revision

- Some annotated solutions did not match CALIMA after first initial annotation. Non-matching solutions included both wildcard solutions and fully manual solutions
- Arbitration (sometimes requiring further joint discussion by the treebank and analyzer teams) and normalization, before entering new solutions into the CALIMA tables
- Feedback/collaboration between LDC annotation team & CALIMA development team
- Integration of new CALIMA solutions into further annotation, leading to…

3. Treebank Revision and Further Annotation

- New CALIMA version → integrated into POS/morphological annotation stage of treebank annotation process
- Fewer "holes" in each new CALIMA version → improved annotation process, more often the desired solution was available for the annotator, reducing wildcard or manual solutions
- Cycle repeats, as remaining new solutions are sent to the analyzer team, which creates a new version of the analyzer, which is sent back to the treebanking team, and so on…

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Nos. HR0011-11-C-0145 and HR0011-12-C-0014. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Improved Synchronization of CALIMA and Treebank Annotation

- Increasing coverage of the tokens in the treebank over the three CALIMA versions
- The CALIMA system used here is a restricted version of CALIMA, where only Egyptian Arabic is present. However, there are richer CALIMA versions where SAMA and CALIMA are combined together (CALIMA-SAMA-ADAM) to cover both Egyptian Arabic and Modern Standard Arabic. The more extended version of CALIMA is used in the tools developed at Columbia University for Egyptian Arabic POS tagging and morphological disambiguation.

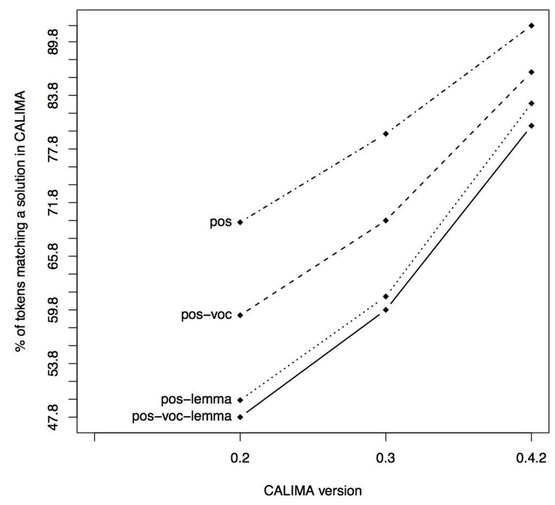

Figure 2. Improvement in synchronization between successive CALIMA versions and Egyptian Arabic morphological annotation

1) **pos** – The POS tag in the treebank is the same as the POS tag for at least one solution in CALIMA (for this source token string)
2) **pos-lemma** – Both the POS tag and lemma in the treebank match the POS tag and lemma for at least one solution in CALIMA
3) **pos-voc** – Both the POS tag and vocalization in the treebank match the POS tag and vocalization for at least one solution in CALIMA
4) **pos-voc-lemma** – The POS tag, vocalization, and lemma in the treebank match the POS tag, vocalization, and lemma for at least one solution in CALIMA

Corpus section	%NO_FUNC
ARZ Part 1	1.8%
ARZ Part 2	1.6%
ARZ Part 3	1.7%
ARZ Part 4	1.5%
ARZ Part 5	1.5%
ARZ Part 6	1.7%
ARZ Part 7	1.3%
ARZ Part 8	1.0%

Table 1: Improvement in CALIMA coverage over successive Egyptian Arabic corpus segments

Conclusions

- Developing the morphological analyzer and the treebank annotation in parallel was successful, showing improvement from one segment to the next for both the analyzer and the annotation
- Contacts between the CALIMA team and the LDC Treebank team were crucial to solving nagging issues and meeting common goals
- Collaboration on this type of challenge, where tools and resources are limited, proved to be remarkably synergistic, and opens the way to further fruitful work on Arabic dialects

This data has been treebanked and released as e-corpora and will be published in the LDC Catalog

Mohamed Maamouri, Ann Bies, Seth Kulick, Sondos Krouna, Dalila Tabassi, and Michael Ciul. (2012). *Egyptian Arabic Treebank DF Parts 1-8*. Linguistic Data Consortium, Catalog Nos.: LDC2012E93, LDC2012E98, LDC2012E89, LDC2012E99, LDC2012E107, LDC2012E125, LDC2013E12, LDC2013E21