Enhancing the Arabic Treebank: 
A Collaborative Effort toward New Annotation Guidelines

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

Enhanced and revised Arabic Treebank Guidelines

- Arabic Treebank (ATB) at Linguistic Data Consortium
- Revised and enhanced annotation guidelines and procedure over the past year
- Now being applied in annotation production
- More complete and detailed annotation guidelines overall
- Period of intensive annotator training
- Inter-annotator agreement f-measure scores improved to 94.3%
- Parsing results improved to 79.7 f-measure

Motivation

- Examination of inconsistencies in annotation
  - Relationship between Part-of-Speech (POS) tags and the syntactic (TB) annotation
  - Priority of semantic vs. syntactic structures in some syntactic constructions
- Lower than expected initial parsing scores

Additional goals

- Representing more finely-grained distinctions
- Aligning more closely with traditional grammar concepts already familiar to annotators

Morphological/Part-of-Speech level: More fine-grained distinctions

- NOUN (common noun)
- NOUN_NUM (number)
- NOUN_QUANT (quantifier)
- NOUN_PROP (proper noun)
- ADJ (common adjective)
- ADJ_NUM (ordinal number)
- ADJ_COMP (comparative adjective)

POS: Follow Arabic traditional grammar

- List of prepositions strictly limited to traditional grammar tags (most lexical items previously PREP now categorized as NOUN, or "prepositional noun")
- Particles now given several POS alternatives: ف (fa), ك (ka), م (ma), أ (a), ب (ba), بـ (ba), كـ (ka), فـ (fa), كـ (ka)

Syntactic/Treebank level: Quantifiers

- Simultaneous dual verbal and nominal role possible
  - Behaving as verbs with respect to their objects (assigning accusative case, for example) and complements
  - Behaving as nominals with respect to the rest of the sentence (occupying canonically noun phrase positions, for example, such as the subject of the sentence, or the complement of a preposition)

Representation in Treebank

- (NP [NP Head] [Adjective])
- (NP [NP Noun] [Preposition])
- (NP [NP Preposition] [Noun])
- (NP [NP Preposition] [Noun])
- (NP [NP Preposition] [Noun])
- (NP [NP Preposition] [Noun])

Enhanced guidelines: Gerunds and participles

- More comprehensive about contexts for verbal readings
  - Greater grammatical precision increased inter-annotator agreement
- For example, a ma:SA:aGerund, active participle or passive participle followed by a PP complement to the regular verb form (PP-CLR) is now shown with a verbal reading

Verbal participle example

- Active participle with PP-CLR complement:
  - (NP [NP [NP [NP Head] [Adjective]] [Preposition] [Noun]])
- For the re-creation of Arabic Treebank Part 3.
- For the re-creation of Arabic Treebank Part 3.
- For the re-creation of Arabic Treebank Part 3.
- For the re-creation of Arabic Treebank Part 3.
- For the re-creation of Arabic Treebank Part 3.
- For the re-creation of Arabic Treebank Part 3.

Fruitful collaboration

- Improved ATB guidelines + improved inter-annotator agreement scores + expected continuing improvement in parsing scores
- Result of fruitful collaboration between data producers and end users, along with the support of sponsors and time to effect the change
- It is hoped that such collaboration will continue to benefit both annotation production and NLP applications in the future
- For a more complete description of the new annotation policies, see the Arabic Treebank Morphological and Syntactic Annotation Guidelines (2006)
  - http://projects.ldc.upenn.edu/ArabicTreebank/