Diacritic Annotation in the Arabic Treebank and Its Impact on Parser Evaluation

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

Motivation

❖ Data released for Arabic Treebank by Linguistic Data Consortium in several formats, including
- "vocalized" (with-vowel)
- "unvocalized" (without-vowel)

❖ Parsing work so far on unvocalized data
- More closely represents "real-world"?

❖ What do unvocalized and vocalized really mean in ATB?

What are diacritics?

❖ Superscript and subscript marks in the Arabic orthographic system

عَلِمَ   عَلَّمَ عِلْمٌ / Ealima; Eal~ama; EilmuN

❖ Represent the three short vowels (a, i, u)

كَتَبَ كُتِبَ كُتُبٌ / kataba; kutiba; kutubuN

❖ Mark vocalic length: four letters (ا ‘alif, ى ‘imaala, و waaw, and ي yaa’)
- The ‘imaala, an undotted form of the letter yaa’, is used idiosyncratically for certain words ending in the long vowel [-aa]
  عَلِي Ealiy   عَلَى EalaY   عَلَيَّ Ealay~a
  (indiscriminate use by Egyptians of عَلَى for both ي y and ى Y)
- The yaa’ and the waaw, in addition to being consonants in their own right, function as glides or semi-vowels and are used to represent long [-uw] and long [-iy] and dipthongs [-ay] and [-aw]
  zAra زَارَ zuwruN زُورٌ وَزِيرٌ waziyruN زَيْتٌ zaytuN زَوْرَقٌ zawraquN

And what do they indicate?

❖ Short vowels:
- MSA grammatical functions, such as verb passive forms and irregular noun plural forms
  كُتُبٌ يُكْتَبُ   يَكْتُبُ   كُتِبَ   كَتَبَ
  kutubuN   yuktabu   yaktubu   kutiba   kataba
- Mood, aspect and voice endings for verbs
  لم يَكْتُبْ   لَنْ يَكتُبَ (lam yaktub -- lan yaktuba)
- Case endings for nouns
  لِقِرَاءَةِ النَصِّ لِقِرَاءَتِهِ النَصَّ

❖ Long vowels are mostly used in derivation and word formation: as in كَتَبَ kataba ‘to write’ vs. كَاتَبَ kAtaba ‘to correspond with’

Non-vocalic diacritics

❖ Shadda (consonantal length or gemination) is used for the derivation of new words.
عَلِمَ Ealima (he knew) / عَلَّمَ Eal~ama (he taught)

❖ Hamza marks the existence of the glottal stop.
- Complex graphemic support system
- Frequently omitted and sometimes misused

آ - ا أ أ إ ئ ؤ ء (hamzat waSl)

❖ Sukun (a small superscript zero-shaped grapheme) is nothing more than the absence of a vowel. The sukun is used for syllabic identification and to mark the imperative, the jussive verb forms and 5 nouns
لَنْ يَكْتُبَ / لَمْ يَكْتُبْ

Ambiguity without short vowels

❖ A single string in the text (bAsm, for example) can mean many different things, depending on what the short vowel morphology is. Some strings can be ambiguous among 120 or more possible solutions!

INPUT STRING: باسم

SOLUTION 1: **bAsim**
LEMMA ID: bAsim_1
POS: bAsim/NOUN_PROP
GLOSS: Basem/Basim

SOLUTION 9: **biAisomi**
LEMMA ID: {isom_1
POS: bi/PREP+{isom/NOUN+i/CASE_DEF_GEN
GLOSS: by/with + name + [def.gen.]

Diacritics and ambiguity: Lexical senses

The loss of the internal diacritics (such as short vowels, hamza, or shadda) leads to the following types of ambiguity, as exemplified in a given MSA lemma: علم Elm

1. An ambiguity within ‘core’ part-of-speech (POS) tags, distinguishing different lexical senses.
   - عِلْم Eilm (a noun meaning ‘science, learning’)
   - عَلَم Ealam (another noun meaning ’flag’)

Diacritics and ambiguity: Different POS

2. A second type of ‘core’ POS tag ambiguity, distinguishing between different core POS tags (lexically and semantically connected). Example:
   - عَلِمَ Ealima for 3rd Person Masculine, Singular, Perfective Verb (MSA Verb Form I) meaning ‘he learned/knew’
   - عُلِمَ Eulima for 3rd Person Singular, Passive Verb (MSA Verb Form I) meaning ‘it/he was learned’
   - عَلَّمَ Eal~ama for the Intensifying, Causative, Denominative Verb (MSA Verb Form II) meaning ‘he taught.’

Diacritics and ambiguity: Inflectional endings

3. Structural/grammatical level, where the use of short vowels is correlated with case (nominal) and mood/aspect (verbal) information.
   - عِلْمُ/عِلْمٌ Eilmu/EilmuN (NOM Noun + Definite and Indefinite)
   - عِلْمَ/عِلْماً Eilma/EilmaAF (ACC Noun + Definite and Indefinite)
   - عِلْمِ/عِلْمٍ Eilmi/EilmiK (GEN Noun + Definite and Indefinite)

Role of diacriticization and the life of a token

❖ **Source Token:** Source text consists of words treated as whitespace-delimited tokens, usually lacking diacritic information

❖ **POS Token:** Annotator’s choice of Buckwalter Arabic Morphological Analyzer solution, includes morphological segmentation of the word and vocalization/diacritization of each segment

❖ **Treebank Token:** POS tokens including separation of clitics as necessary

Parser evaluation

❖ Question of evaluation framework (rather than parser results)

❖ Use unvocalized or vocalized forms?

❖ Unvocalized sometimes assumed to be “real-world” but is not because
- Not an accurate representation of “real-world” data
- Unvocalized = vocalized with diacritics stripped out (not necessarily unchanged input data)
- Vocalized = diacritics not in input data, plus some orthographic normalization

❖ Roughly 3.7% of tokens include some form of orthographic normalization

Example: Added consonants in unvocalized data

❖ The white-space delimited input string (source token) is llqDA’ للقضاء. The l- لـ is a prefix for the preposition ”li” لِـ, and the chosen solution from the morphological analyzer
- **POS token =** li/PREP + Al/DET + qaDA’/NOUN

❖ The preposition is split off, creating two (vocalized) treebank tokens:
- **Treebank tokens =** li لِـ and AlqaDaA’ القَضَاء

❖ Second word includes additional consonant A, in addition to the insertion of the short vowels/diacritics i in the first word and a in the second word

Example: Orthographic normalization

❖ **Source token** = Aly الي

❖ **POS token** = vocalized: <ilay/PREP+˜a/PRON_1S

❖ **Treebank tokens =** vocalized: two segments, <ilay/PREP and ˜a/PRON_1S

❖ Preposition:
- Vocalized treebank token = <ilay, which not only adds the short vowels I and a, but also corrects for the ”missing hamza” problem by normalizing A to <
- Unvocalized form of this token results from stripping out the short vowels, and so is <ly.
- But -- original text file, which did not have the correct hamza placement, with A instead of <

Linking unvocalized and vocalized trees in ATB

❖ Modified release format of ATB with explicit the links between unvocalized and vocalized trees → maximum flexibility for experimentation
- Pointers to original source file to relate different annotation levels
- Trees with complex terminals including
  - Source token, vocalized and unvocalized forms, lemma and gloss
  - Brings together information previously available only by accessing multiple release formats

Conclusion

❖ To evaluate a parser on “real-world” data, the unvocalized form is insufficient
- Already tokenized
- Orthographically normalized

❖ Future work on parser evaluation must
- Take both issues above into account
- Disclose what degree of diacritization is chosen for parsing and parser evaluation