Issues in Arabic Orthography and Morphology Analysis

Tim BUCKWALTER
Linguistic Data Consortium
University of Pennsylvania
Philadelphia, PA 19104 USA
timbuck2@ldc.upenn.edu

Abstract

This paper discusses several issues in Arabic orthography that were encountered in the process of performing morphology analysis and POS tagging of 542,543 Arabic words in three newswire corpora at the LDC during 2002-2004, by means of the Buckwalter Arabic Morphological Analyzer. The most important issues involved variation in the orthography of Modern Standard Arabic that called for specific changes to the Analyzer algorithm, and also a more rigorous definition of typographic errors. Some orthographic anomalies had a direct impact on word tokenization, which in turn affected the morphology analysis and assignment of POS tags.

1 Introduction

In 2002 the LDC began using output from the Buckwalter Arabic Morphological Analyzer (Buckwalter, 2002), in order to perform morphological annotation and POS tagging of Arabic newswire text. From 2002 to 2004 three corpora were analyzed and over half a million Arabic word tokens were annotated and tagged (see Table 1).[1]

Corpus	Arabic Word Tokens
AFP	123,810
Ummah	125,698
Annahar	293,035
Total	542,543

Table 1: Arabic newswire corpora

The author was responsible for developing and maintaining the Analyzer, which primarily involved filling in the gaps in the lexicon and modifying the POS tag set in order to meet the requirements of treebanking efforts that were performed subsequently at the LDC with the same annotated and POS-tagged newswire data.

2 Lessons from the AFP corpus

During the tagging of the AFP data, the first corpus in the series, the Buckwalter Analyzer was equipped to handle basic orthographic variation that often goes unnoticed because it is a common feature of written Arabic (Buckwalter, 1992). This orthographic variation involves the writing (or omission) of *hamza* above or below *alif* in stem-initial position, and to a lesser extent, the writing (or omission) of *madda* on *alif*, also in stem-initial position. In both cases use of the bare *alif* without *hamza* or *madda* is quite common and goes by unnoticed by most readers. What took the LDC morphology annotation team by surprise was to find that in the AFP data the common omission of *hamza* in this environment had been extended to stem-medial and stem-final positions as well, as seen in the following words from that corpus: بدات تاييد متاخر تاسيس تستانف رات لتاكيد بشان.

This type of orthographic variation was not attested to the same extent in the two subsequent corpora, Ummah and Annahar, which leads us to conclude that some orthographic practices might be restricted to specific news agencies. It is important to note that most of the native Arabic speakers who annotated the AFP data using the output from the Analyzer did not regard these omissions of *hamza* on *alif* in stem-medial and stem-final positions as orthographic errors, and fully expected the Analyzer to provide a solution.

3 Lessons from the Ummah corpus

During the tagging of the Ummah data, a different set of orthographic issues arose. Although the Buckwalter Analyzer was equipped to handle so-called "Egyptian" spelling (where word-final *ya'* is spelled without the two dots, making it

[1] The tagged AFP, Ummah, and Annahar corpora were published as "Arabic Treebank: Part 1 v 2.0" (Maamouri 2003), "Arabic Treebank: Part 2 v 2.0" (Maamouri 2004), and "Arabic Treebank: Part 3 v 1.0" (Maamouri 2004), respectively, and are available from the LDC website <http://www.ldc.upenn.edu >

identical to *alif maqsura*), the Ummah corpus presented the LDC annotation team with just the opposite phenomenon: dozens of word-final *alif maqsura*'s spelled with two dots.[2] Whereas some of the affected words were automatically rejected as typographical errors (e.g., القري الأعلي موسي متي أخري), others where gladly analyzed at face value (e.g., لدي إلي علي). Unfortunately, this led to numerous false positive analyses: for example علي was analyzed as '*ali* and '*alayya*, but not as '*ala*. Initially, these words were tagged as typographical errors, but their pervasiveness led the LDC team to reconsider this position, upon which the author was asked to modify the Analyzer algorithm in order to accommodate this typographic anomaly. As a result, all words ending in *ya*' were now re-interpreted as ending in either *ya*' or *alif maqsura*, and both forms were analyzed, as seen in the following (abridged) output:[3]

```
<token_Arabic>علي
  <variant>Ely
    <solution>
      <lemmaID>EalaY_1</lemmaID>
      <pos>Ealay/PREP+ya/PRON_1S</pos>
      <gloss>on/above + me</gloss>
    </solution>
    <solution>
      <lemmaID>Ealiy~_1</lemmaID>
      <voc>Ealiy~N</voc>
      <pos>Ealiy~/ADJ+N/CASE_INDEF_NOM</pos>
      <gloss>supreme/high + [indef.nom.]</gloss>
    </solution>
    <solution>
      <lemmaID>Ealiy~_2</lemmaID>
      <voc>Ealiy~N</voc>
      <pos>Ealiy~/NOUN_PROP+N/CASE_INDEF_NOM</pos>
      <gloss>Ali + [indef.nom.]</gloss>
    </solution>
  </variant>
  <variant>ElY
    <solution>
      <lemmaID>EalaY_1</lemmaID>
      <voc>EalaY</voc>
      <pos>EalaY/PREP</pos>
      <gloss>on/above</gloss>
    </solution>
  </variant>
</token_Arabic>
```

4 Lessons from all three corpora

The Annahar corpus presented no orthographic surprises, or at least nothing that the LDC annotation team had not seen before. The Annahar data did contain some additional orthographic features that we now identify as being common to all three corpora, as well as corpora outside the set we have annotated at the LDC.

The first orthographic feature relates to the somewhat free interchange of stem-initial *hamza* above *alif* and *hamza* below *alif*. With some lexical items the orthographic variation simply reflects variation in pronunciation: for example, both '*isbaniya* (with *hamza* under *alif*) and '*asbaniya* (with *hamza* above *alif*) are well attested. But in cases involving other orthographic pairs, more interpretations are possible. Take, for instance, what we called the "*qala 'anna*" problem. This problem was identified after numerous encounters with constructions in which *qala* was followed by '*anna* rather than '*inna*, and for no apparent linguistic reason. Initially this was treated as a typographical error, but again, its pervasiveness forced us to take a different approach.

One solution we considered was to modify the Analyzer algorithm so that instances of stem-initial *hamza* on *alif* would also be treated as possible instances of *hamza* under *alif*, very much in the spirit of the approach we used for dealing with the *alif maqsura* / *ya*' free variation cited earlier. However, there is compelling evidence that the orthography of *hamza* in stem-initial position is a fairly reliable indication of the perceived value of subsequent short vowel: *a* or *u* for *hamza* above *alif*, and *i* for *hamza* below *alif*. In other words, there is no free variation. The decision was taken to regard "*qala 'anna*" constructions as grammatically acceptable in MSA.[4]

5 Concatenation in Arabic orthography

The second, and more serious, orthographic anomaly we encountered in all three corpora is what we call the problem of Arabic "run-on" words, or free concatenation of words when the word immediately preceding ends with a non-connector letter, such as *alif*, *dal*, *dhal*, *ra*, *za*, *waw*, *ta marbuta*, etc.

The most frequent "run-on" words in Arabic are combinations of the high-frequency function words *la* and *ma* (which end in *alif*) with following perfect or imperfect verbs, such as *la-yazal*, *ma-yuram*, and *ma-zala* (مازال مايرام لايزال). The *la* of "absolute negation" concatenates freely with nouns, as in *la-budda*, *la-shakka* (لاشك لابد). It can be argued that these are lexicalized collocations, but their spelling with intervening space (لا يزال –

[2] It is not entirely clear whether these "dotted" *alif maqsura*'s were produced by human typists or by an encoding conversion process gone awry. It is possible that the original keyboarding was done on a platform where word-final *ya*' and *alif maqsura* are displayed via visually identical "un-dotted" glyphs, so it makes no difference which of the two keys the typist presses on the keyboard: both produce the same visual display, but are stored electronically as two different characters.

[3] A key to the transliteration scheme used by the Analyzer can be found at <http://www.ldc.upenn.edu/myl/morph/buckwalter.html>

[4] Badawi, Carter and Gully regard "*qala 'anna*" constructions as grammatical but restricted to contexts "where the exact words of the speaker are not used or reported" (Badawi, Carter and Gully 2004, p. 713). This assertion could be investigated in the LDC corpora.

لا بد – ما زال) is just as frequent as their spelling in concatenated form.

Proper name phrases, especially those involving the word *‘abd* (عبدالرحمن عبدالله) are also written either separately or in concatenated form. Part of the data annotation process at the LDC involves assigning case endings to tokenized words, but there is currently no mechanism in the Analyzer to assign two case endings (or several pairs of POS tags) to what is being processed as a single word token. As a result of this, the phrase ‘*abd allah* is assigned a single POS tag and case ending when it is written in concatenated form, but two POS tags and two case endings when written with intervening space.

The problem of assigning more than one case ending and POS tag to concatenations is more obvious in fully lexicalized concatenations such as *khamsumi’atin*, *sittumi’atin*, *sab’umi’atin*, etc (سبعمائة – ستمائة – خمسمائة). When these numbers are written with intervening space (خمس مائة – ست مائة – سبع مائة), two case endings and two POS tags are assigned by the Analyzer. But when they are written in concatenated form only one case ending and POS tag is assigned, and the “infixed” case ending of the first token is left undefined: *khamsmi’at*^{in}, *sittmi’at*^{in}, *sab’mi’at*^{in}, etc.[5]

So far we have discussed relatively controlled concatenation, involving mostly high-frequency function words and lexicalized phrases. But concatenation extends beyond that to random combinations of words—the only requirement being that the word immediately preceding end with a non-connector letter. These concatenations are fairly frequent, as attested by their Google scores (see Table 2).

It is important to note that these concatenations are not immediately obvious to readers due to the characteristics of proportionally spaced Arabic fonts. Most of the native readers of Arabic at the LDC did not consider concatenations such as these to be typographical errors. Their logic was best expressed in the statement: “I can read the text just fine. Why can’t the Morphological Analyzer?”

Concatenation	Google Frequency
مديرعام	846
وزيرالخارجية	719
ملياردولار	162
الدكتورمحمد	158
عضومجلس	138
وقدتم	130
واشارالى	99
كماتم	77
عددكبير	54

Table 2: Arabic Concatenations and their Google Frequencies (sample taken March 25,2004)

6 Conclusion

There are several levels of orthographic variation in Arabic, and each level calls for a specific response to resolve the orthographic anomaly. It is important that the output analysis record which method was used to resolve the anomaly. The methods used for resolving orthographic anomaly range from exact matching of the surface orthography to various strategies of orthography manipulation. Each manipulation strategy carries with it certain assumptions about the text, and these assumptions should be part of the output analysis. For example, an analysis of علي obtained by exact matching in a text known to contain suspicious word-final *ya*’s (that may be *alif maqsura*’s) does not have the same value as an analysis of the same word, using the same exact matching, but in a text where word-final *ya*’s and *alif maqsura*’s display normal character distribution frequencies.

The problem of run-on words in Arabic calls for a reassessment of current tokenization strategies, including the definition of “word token” itself.[6] It should be assumed that each input string represents one or more potential word tokens, each of which needs to be submitted individually for morphology analysis. For example, the input string فقدتم can be segmented as a single word token, yielding two morphological analyses (*faqad-tum* and *fa-qud-tum*) or it can be segmented as two word tokens (*fqd tm*), yielding several possible analysis pairs (*faqada* / *fuqida* / *faqd* / *fa-qad* + *tamma*).

[5] We regard these as “fully lexicalized” concatenations because the first of the two constituent tokens ends in a connector letter. In other word, their concatenation is deliberate and not an accident of orthography.

[6] By “tokenization” we mean the identification of orthographically valid character string units that can be submitted to the Analyzer for analysis. The Analyzer itself performs a different kind of “tokenization” by identifying prefixes and suffixes that are bound morphemes but which may be treated as “word tokens” in syntactic analysis.

Syntactic analysis would be needed for determining which morphology analysis is most likely the correct one for each tokenization (*fqdtm* and *fqd tm*).

7 Acknowledgements

Our thanks go to the Arabic annotation team at the LDC, especially the team of native speaker informants that provided the author with daily feedback on the performance of the Morphological Analyzer, especially in areas which led to a reassessment and better understanding of orthographic variation, as well as tokenization and functional definitions of typographical errors.

References

Elsaid Badawi, M.G. Carter, and Adrian Wallace. 2004. *Modern Written Arabic: A Comprehensive Grammar.* Routledge, London.

Tim Buckwalter. 1992. "Orthographic Variation in Arabic and its Relevance to Automatic Spell-Checking," in *Proceedings of the Third International Conference on Multilingual Computing (Arabic and Roman Script),* University of Durham, U.K., December 10-12, 1992.

Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer Version 1.0.* Linguistic Data Consortium, catalog number LDC2002L49 and ISBN 1-58563-257-0. < http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002L49 >

Mohamed Maamouri, et al. 2003. *Arabic Treebank: Part 1 v 2.0.* Linguistic Data Consortium, catalog number LDC2003T06 and ISBN: 1-58563-261-9. < http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T06 >

Mohamed Maamouri, et al. 2004. *Arabic Treebank: Part 2 v 2.0.* Linguistic Data Consortium, catalog number LDC2004T02 and ISBN: 1-58563-282-1. < http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T02 >

Mohamed Maamouri, et al. 2004. *Arabic Treebank: Part 3 v 1.0.* Linguistic Data Consortium, catalog number LDC2004T11 and ISBN: 1-58563-298-8. < http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004T11 >