

The semi-automatic tagging of Arabic corpora

Mark Van Mol

Katholieke Universiteit Leuven

ILT

Dekenstraat 6

B 3000 Leuven

mark.vanmol@ilt.kuleuven.ac.be

Abstract

At the Institute of Living Languages of the Katholieke Universiteit Leuven we developed a system to encode Arabic corpora which enables us to identify strings of characters and to analyse them and disambiguate words. At the institute we developed two kinds of databases, one word-oriented and one sentence-oriented. The word-oriented database contains until now 26,000 Arabic lemmata with all the grammatical information. The second database contains a text corpus of approximately 4,000,000 tagged words of which 1,200,000 from spoken Arabic language resources. Both databases will be used in the future in order to develop a semi-automatic tagger of raw Arabic corpora. In order to make Arabic electronic corpora useful for a large variety of purposes a pre treatment seems to be necessary. The treatment comprises three main phases. In the first place the uniformisation of Arabic corpora. The second phase involves the identification of strings of characters and the third phase involves the disambiguation of words on the basis of information coming from both sources. Once the corpus is tagged this way, it contains enough detailed information to make scientific searches and analyses.

1. Introduction

Recent years more and more attention has been paid to gather Arabic corpora. Besides the ELRA initiative, numerous groups take the initiative in compiling all kinds of corpora. Many of these initiatives however, suffice with the compilation of raw corpus materials. Unlike other languages, however, the Arabic written language is ambiguous in many respects. The ambiguity of Arabic lies in the first place in the fact that the language is not vocalised. Often it is stated that languages with a rich morphology open much more facilities for tagging. The first problem in Arabic however is, that written texts are not vocalised except in school books from primary schools and Coranic texts. All other material remains unvocalised which, of course, raises the level of ambiguity.

2. Levels of ambiguity in Arabic

The ambiguity of Arabic lies on different levels. The first level is the core word itself. Many core words can contain in itself different grammatical categories. Below we give a few examples of possible combinations of grammatical categories of unvocalised words. Even if we limit ourselves to the main part of speech categories we find many ambiguous words.

2.1. First level: Core word

2.1.1. Noun - adjective.

Many Arabic word patterns can stand both for a noun and an adjective. Without being exhaustive we mention first the pattern **فَعِيل** which is most often an adjective, but which can also be a noun. For example, the word **صَغِير** (*small*) which, of course, more exceptionally, can be used as a noun meaning *the small one*. The predictability of the grammatical category of those patterns is not always self-evident. One might suppose that the word pattern **فَعِيل** most often stands for an adjective, but this is not always the case. Take for instance the word **رئيس** of which it is quite clear, at first sight, that it is a noun, but which in Modern Standard Arabic (especially in North Africa) is often used as an adjective meaning *principal*.

The same goes for almost all the words ending in a so-called *nisba*. Indeed, most of those words are, as far as their Arabic pattern is concerned, unpredictably a noun or an adjective. Take, e.g., the word **سياسي** which means *politician* (noun) as well as *political* (adjective). Of course, a completely trained tagged corpus might shed some light on the chance rate of those grammatical categories, but the pattern itself does not say anything on the grammatical category of the word, except that it excludes to some extent the labeling of other categories, such as a verb or a particle. But even this remains in many cases problematic, because the *nisba* characteristic, is in many cases not sufficient to exclude other grammatical categories, such as the verb or the particle, especially when unvocalised words are involved. Indeed, forms such as verbs ending with *ya'*, for instance, the verb **بقي** (*to stay*) or the particle **أي** (*any*) could on the basis of the ending characteristic wrongly be interpreted as being an adjective or a noun.

Other ambiguous word patterns which cover both nouns and adjectives are the patterns **فَعْلَان** (for example **سكران** noun: *drunk* - adjective: *intoxicated*), **فَعَّال** (for example: **بنا** noun: *mason* - adjective: *constructive*), **فَعُول** (for example **كسول** noun: *idler* - adjective: *lazy*)

2.1.2. Participles

Another word pattern which covers both nouns and adjectives is the pattern of both active and passive participles **فَاعِل**, **مَفْعُول** and derivatives. These cases are sometimes even more complicated because they can also be classified from time to time as a preposition (for example: **داخل** *within*) but even sometimes as an participle with the function of a verb, such as in **هو داخل** *he is going inside*.

2.1.3. Verb - adjective.

Many verbs have the same shape as adjectives. Often an unvocalised verb with three radicals has the same pattern as an adjective. The three radicals فـرـح, for example, can both stand for the verb فَرِحَ and the adjective فَرِحَ.

2.1.4. Verb - noun.

The most important mingling of word patterns between verbs and nouns occurs with the verbal nouns (masdar). The verbal nouns of the fifth and the sixth form often raise confusion. For example تدخل (Vth form) can both be a verb (*to meddle*) and a noun (*interference*) and also تعاون (VIth form) can both be a verb (*to help*) and a noun (*cooperation*). However, the verbal nouns of the Vth and VIst form are easily detectable in a written text. The verbal nouns of the Ist form, on the other hand, are much more difficult to define as verbal noun, because these forms can often also be used as a noun. But also nouns are mixed up with verbs, such as, for example, the shape وفد which can be a noun (*delegation*) or a verb (*to arrive*).

2.1.5. Verb - noun - adjective

The pattern أفعل is even more complicated. This pattern offers at least three possibilities, viz. a noun, an adjective or a verb. The word أبيض, for instance, means both *white* as a *white* (a member of the white race). However, it can also have the function of a verb in the sentence ما أبيض وجهه *what is his face white!* in which, according to the Arab grammarians, the أفعل form is considered to be a verb.

2.1.6. The taa marbuta element

One morphological element which might seem to help to disambiguate words is the taa marbuta (Khoja 2002) which is considered in grammar to be the indication of a feminine noun par excellence. There are however exceptions, for instance, the rare فعالة forms, such as سعادة and فخامة (*excellence*) which are masculine and the pattern علامة in فعالة which represents an adjective meaning *very learned*.

The above elements show that it is not sufficient to take a lexicon and tag it. Many ambiguities are not resolved that way. Only the completely unambiguous forms will be tagged, but it is clear that most of the others will not. This does not mean that the tagging of words in a lexicon is not helpful. One might suppose that when going into more detail, word patterns which can contain two or more grammatical categories, and which are for that reason ambiguous, lose in quite a number of cases this ambiguity when they are translated in their practical word form. The above mentioned word غموس for instance is clearly an adjective.

This however remains very tricky, because a word in Arabic, which in its concrete form is clearly an *adjective* but of which the theoretical form is ambiguous, can always by one

Arab author or another be used as substantive. Arab authors often renew the style of the language and the language itself precisely by enlarging the meaning of already existing forms. The case of رئيس illustrates this clearly. As one can discover in the dictionary of Hans Wehr, the word رئيس (*president*) is definitely only a noun. No other meanings are given in this dictionary. However, corpus analysis of radio texts of Algeria revealed that this word in this pattern is often used as an adjective, meaning *principle*. Even when basing ourselves on existing lexicons, we can not guarantee the distinct definition of parts of speech for Arabic words.

2.2. Second level: Derived word forms or conjugated forms

Not only on the level of the core forms there are many ambiguities, the same goes for derived word forms or conjugated forms. Due to the lack of vocalisation the conjugation of verbs yields many ambiguous word patterns which are quite difficult to interpret without any valid context. كتبت can have four possible meanings: *I wrote, you wrote (m. and f.) or she wrote*. New ambiguities arise with the conjugated forms of verbs, not only within a conjugational level but also between different conjugational levels. The verb forms in the past tense of the first person singular, the second person masculine singular, the second person feminine singular and third person feminine singular all have the same shape. But new ambiguities arise between, for example, imperative forms and indicative forms. The shape اكتب can mean both *I write* or the imperative form *write*, but it can also be the third person in the past tense of a verb of the IVth form اكتب (*to dictate*).

Also, these derived forms interfere often with similar forms from other words which makes the correct indication of the tag even more complex. Here too different grammatical categories mix up. In some cases the أفعل form does not only lead to the confusion mentioned above, but can even have a form which goes beyond the grammatical categories of a verb such as an elative.

The character combinations of nouns also can have the same shape of conjugated verbs. For example, the first person of the jussive form of the verb بنى (*to build*) which becomes ابن and hence is a pattern of consonants which mixes up with the noun ابن (*son*). Derivated forms of adjectives too can have the same shape as nouns. Many feminine forms of adjectives ending with the *nisba* do correspond in their shape with feminine nouns. For example the feminine adjective شخصية (*personal*) which corresponds to the noun شخصية (*personality*).

2.3. Third level: Agglutinative forms of words

Not only isolated morphological forms can be dubious, but the agglutinative character of the language provokes unexpected ambiguities between strings of characters between two blanks. The combination of the conjunction و with the particle قد corresponds to the

verb *وقد* (*to take fire*). These new ambiguities can occur with all combinations of particles or conjunctions which are being written directly to the word. The combination of the conjunction *ف* with the verb *حش* (*to cut*) corresponds to the verb *فحش* (*to be detestable*). Both the particle *ب* and *ل* provoke the same kind of ambiguities. For example, the preposition *ب* in combination with *يد* (*hand*) which corresponds to the subordinate conjunction *بيد* (*however*). And the preposition *ل*, for example, in combination with *حظ* (*part*) which corresponds to the verb *لحظ* (*to regard*).

3. Automatic vocalisation, a solution?

One might argue that the vocalisation of an Arabic corpus might solve the problems of tagging. This is only true to a certain extent. First of all, the above shown ambiguities indicate that it is not at all self-evident to make a tagger which disambiguates Arabic raw texts by vocalising them. Even then, algorithms will have to be written in order to apply the correct grammatical categories to the different lemmas in a text. But even so, in a completely vocalised text, ambiguities remain, as far as grammatical part of speech tagging is concerned, be it that overall ambiguity in a vocalised text is quite lower than in an unvocalised text. It is clear that on all three discussed levels a degree of ambiguity remains.

3.1. Ambiguities on the first level

On the first level, which is the level of the core word, ambiguity remains in the forms *فَعِيل*, *فَعَال*, *فَعُول* and in all the words which have the form of a participle, both active and passive, such as those of the form *فَاعِل*, *مَفْعُول* and all their derived forms. All those forms can be both an adjective or a noun even when they are completely vocalised. On the same level ambiguities remain also between, for example, some verbs and nouns, such as the noun *كُلٌّ* in the meaning of *totality*, and the verb *كُلٌّ* with the meaning *to be tired*.

3.2. Ambiguities on the second level

On the second level this is valid for many word forms ending in a nisba followed by a taa marbuta. The complete vocalised word *شخصية*, for instance, does not give any more information on the exact grammatical part of speech to apply. The same goes for every word with this pattern.

3.3. Ambiguities on the third level

On the third level also, new problems arise. Word forms which were not ambiguous on the first level in their core form become ambiguous and mix up with other words. The very frequent collocation *وفي* (conjunction - preposition) (*and in*) has the same shape as the adjective *وفي* (*faithful*). There are many other agglutinated word combinations which mix up with existing core word forms. Another example is *وسم* (verb = *to brand*) which mixes up with *وسم* (conjunction + verb = *and he poisoned*). Exclusively basing the tagging of

Arabic texts on a lexicon is therefore not sufficient. Indeed, the analysis of in detail tagged corpora gives additional information which might be of great use for the tagging of raw corpora.

4. Information to be derived from tagged corpora

The additional information, to be derived from tagged corpora, is both a statistical one and a grammatical one. Both kinds of information can make a high contribution for the tagging of corpora. Both remain to a certain extent probabilistic.

4.1. Statistical information

The statistical element is evident. Many core word forms in Arabic are no longer used in MSA. The dictionary of Hans Wehr contains many words which seem to be out of use nowadays. In compiling our dictionary MSA-Dutch * Dutch-MSA (Van Mol & Berghman 2001), which is based on a corpus of 3,000,000 words both from oral and written resources, we discovered that many root patterns did not occur in our corpus. Our corpus contains only words in texts dated from 1980 onward. For instance, no word relating to the stem *فره* occurs in the corpus.

The fact that in MSA some word forms are less used is very important for the automatic tagging of corpora. Let us take the form *لك* as an example. At first sight this shape is a preposition and a personal pronoun meaning *for you*. The shape *لك* occurs very frequently in MSA. There is, however, an identical form which is a verb meaning *to hit with the fist*. This form however did not occur at all in our corpus. This means that a count of words in a complete disambiguated corpus can give much relevant information as far as automatic disambiguation of words is concerned or at least it can give a hint about the probability in tagging certain shapes according to the word count statistics.

4.2. The Leuven approach

4.2.1. The encoding system

In order to make preparations for the automatic tagging of Arabic corpora, we developed at the Instituut of Levende Talen of the Katholieke Universiteit Leuven a system to encode Arabic corpora. This system not only enables us to identify strings of characters and to analyse them, it also disambiguates words and makes it possible to label all kinds of strings of characters by the appropriate grammatical information. The disambiguation of words is made by using the Arabic diacritical signs in a special structured systematic way.

As an example we take the root *قبل*. This shape can be a verb (*قَبِلَ to accept*), an adverb (*قَبْلُ before*), a noun (*قَبْلُ front part*), a preposition (*قَبْلَ near*) or another preposition (*قَبْلُ before*) and even a verb of the second form, if the *sjadda* is omitted (*قَبِلَ to kiss*). In order to disambiguate between these different shapes we apply the diacritical signs according to a systematic description. Basically these rules can briefly be summarized as follows: the basic form of a verb is never vocalized (e.g. *قَبِلَ*). The first consonant of a noun is

always vocalised (e.g. قُبِلَ). The last consonant of a preposition is always vocalised (e.g. قَبِلَ). If there is more than one preposition with the same shape, the second consonant is vocalised as well (e.g. قَبِلَ). Adverbs normally take the *alif*, if not, the last consonant is vocalised, such as in قَبِلُ. Derived verb forms, such as those of the second form are always written with the *sjadda*. More details on this system of disambiguation can be found in (Van Mol - in print).

4.2.2. The lexical database

At the institute we develop two kinds of databases, one word-oriented and one sentence-oriented. The word-oriented database contains up to now 26,000 Arabic lemmata with all the relevant grammatical information. The words in this database were all disambiguated by way of our encoding system. After every word has been disambiguated by using the diacritical signs in a selective way, the grammatical categories are allocated for those words. Until now this has been done for approximately 20,000 words in this database. Linked to this database is a dictionary Arabic - Dutch v.v. which has recently been published also in book form (Van Mol & Berghman: 2001)

4.2.3. The corpus

The second database contains a text corpus of approximately 4,000,000 tagged Arabic words of which 1,200,000 from spoken Arabic language resources. Both databases will be used in the future in order to develop a semi-automatic tagging of raw Arabic corpora. In order to do so several steps have to be taken, the first of which is *uniformisation*.

4.2.4. The tagging preparations

In order to open up Arabic electronic corpora for a large variety of purposes a pre-treatment seems to be necessary. This treatment comprises three main phases.

4.2.4.1. The uniformisation of Arabic corpora

Uniformisation means that all possible shapes of one word in a raw text are reduced to one identical shape. In the first place the *uniformisation* of Arabic corpora. Due to the fact that, contrary to other languages, there is quite a large freedom in typing Arabic language, the ambiguity or the variety in writing Arabic is quite large. As the computer only recognises ASCII codes, a minimal amount of standardisation seems to be prerequisite.

Even when we have a detailed lexical database of which there are minimum two kinds of information, viz. all the words in their vocalized form, but also in their neutral unvocalised form, it is not always self-evident to find the right matches between words occurring in the database, and words occurring in a raw corpus. A few examples can make this clear. One of the problematic Arabic characters is the *alif*. The *alif* can be written, without a *hamza* or with a *hamza*. This means that when in a database the word أَوْلَادُ (*boys*) is stored as the unvocalised word form for the vocalised form أَوْلَادُ, it is not always certain that this will match with a corresponding word form in the raw corpus, such as, for example, the form

او لاد, because the ASCII code of both *alif*-forms differs.

Another example is the use of the *alif maqsura*, for example in the word *فى*, in Egyptian newspapers, whereas in most other countries the *ya'* is used (e.g. *في*). Other elements which ought to be uniformized is the use of the *sjadda*. Words of which the word pattern contains both *alif* and *sjadda* are even more complicated. In those cases uniformization is not a matter of reducing two forms to one. Without counting the vowels, words with *alif* and *sjadda* can have up to eight different forms in a raw text. For example, the word *اروبيّ* (*European*), which can be written *أوربيّ, أوربيّ, أوربيّ, أوربيّ, أوربيّ, أوربيّ, أوربيّ* or *أوربيّ*.

4.2.4.2. The identification of strings of characters

The second phase involves the *identification* of strings of characters. In order to do so we develop a two-level approach.

The first approach departs from the word-oriented database from which all possible minimal basic forms of words are generated. For every word we generate all possible, what we might call, minimal basic forms. The minimal basic form contains all the possible prefixes and suffixes which can be added to a word, but which still are part of it. On the other hand we also produce for every word all, theoretic possible, maximal basic forms, which correspond to the possible *word* combinations between two blanks. All of these forms are given a minimal encoding so that every added linguistic element has an unambiguous shape.

The second approach departs from the sentence-oriented database from which all possible maximal basic forms are retrieved.

The third phase involves the *disambiguation* of words on the basis of information coming from both sources. Once the corpus tagged this way, it contains enough detailed information to make scientific searches and analyses.

Conclusion

In our view, the best way to make preparations for the automatic annotation of Arabic corpora will be by using a completely in detail annotated corpus which will give a more detailed insight in the distribution of the different Arabic word patterns and their corresponding grammatical category. We hope to give in the near future much more details on the degree of ambiguity on the three word levels, the core word level, the derived word forms or conjugated forms and agglutinative forms of words. Those data will be compared with the data retrIEved FROM the annotated test corpus.

References:

Cantarino, V. (1974). *Syntax of modern Arabic prose*, 3 Vol, Bloomington, London.

Ditters, E. (1992). *A formal approach to Arabic Syntax: the noun phrase and the verb phrase*, Amsterdam.

Van Mol, M. & Berghman, K. (2001a). *Leerwoordenboek Modern Arabisch - Nederlands*, The Dutch Language Union, Amsterdam, Bulaaq

Van Mol, M. & Berghman, K. (2001b). *Leerwoordenboek Nederlands - Modern Arabisch*, The Dutch Language Union, Amsterdam, Bulaaq.