Natural Language processing and Arabic: the Leuven tandem approach

Mark Van Mol (1) and Hans Paulussen (2)

(1) Katholieke Universiteit Leuven
Institute for Living Languages – Faculty of Arts
Dekenstraat 6 – B-3000 Leuven – Belgium
mark.vanmol@ilt.kuleuven.ac.be
(2) Katholieke Universiteit Leuven – Campus Kortrijk
Etienne Sabbelaan 53 – B-8500 Kortrijk – Belgium
Hans.Paulussen@kulak.ac.be

Abstract

In order to develop a computer system for the analysis of Arabic as a natural language, the problem of ambiguity of Arabic words and strings of characters has to be solved. In order to do so we designed a bi-directional approach which departs from both a lexical database and a lexicon. In the lexical database, a generator created all possible minimal basic forms (morphological patterns), which are all the forms of the words which can be derived directly form the stem of every word. In the corpus, which consists of both written and spoken samples of MSA, a specific primary part-of-speech tagging was applied in order to identify the maximal basic forms of the words which involve all the affixes which can be added to Arabic words, but which as such do not form part of the word. In this way, a more complex and detailed form of tagging of Arabic texts, which is still under development, will be made possible in the future.

Keywords

Arabic lexicography - Arabic corpus linguistics - tagging of corpora

1 Introduction

As is generally known, the Arabic language is complicated for natural language processing because of two main language characteristics. The first is the agglutinative nature of the language and the second is the aspect of the vowellessness of the language which causes problems of ambiguity at different levels. In order to be able to analyse Arabic by computer we first have to disambiguate the Arabic words. This demands a pre-treatment of the language.

2 The agglutinative nature of the language

The first problem is the identification of words in sentences. As in most European languages, Arabic words can, to a certain degree, be identified in computer terms as a string of characters between blanks. Two blanks in a text serve as a marker for the separation of strings of characters, but those strings of characters do not always coincide with words. Some Arabic grammatical categories which are considered words in other languages appear to be affixes. Those affixes are directly linked to the words in Arabic. This means that a string of characters between two blanks can contain more than one word so that multiword combinations are found which are not separated by blanks.

As far as the analysis by computer is concerned, Arabic does have an advantage, because of the fact that words never split into two separate units, a phenomenon which occurs both in English and in Dutch, for instance, in the case of passive verb forms, where the verb *to shut*, becomes *was shut*. In English, one could consider both entities as two separate words, but in this way passive verb forms cannot be identified. In Dutch, there is the problem of the separable affixes. Although forming an inherent part of the word, the affixes are often split off from the head word, and placed quite a number of words away. Unlike the two previous languages, Arabic has the advantage that all words remain together between the blanks. On the other hand, the string of characters between two blanks can contain more than one word. The question remains then how to separate the agglutinated words. In most cases the added words between two blanks are affixes (see also Dichy, 2002). The reason for this agglutination —as far as the prefixes in Arabic is concerned— is, among others, the grammatical rule, that all words containing only one consonant and one vowel must be linked to the following word.

2.1 Prefixes

A small inventory of the prefixes in Arabic yields the following grammatical categories: The definite article *'a*l (the), the connectives *fa* and *waw* (and), the prepositions *bi* (with), *li* (for, to), *ka* (as), the particle of the future used by verbs *sa* and the conjunctive particle *li* (in order to) and the interrogative particle *alif-hamza*. When we include the spoken varieties, we also have to take into account the verbal particles *ha* (future) and *bi* (general) for the Middle East and the particle *ka* and *ga* for North Africa (especially Morocco and Algeria). Although we wrote all the previous particles with a vowel, we must take into consideration that in written language use the vowel is omitted. This means that, in practice, those grammatical categories are reduced to no more than one consonant which is written onto the word. Hence the above mentioned series becomes (with exception of the definite article), the consonants *f*, *w*, *b*, *l*, *k*, *s*, *g*, *h* and *hamza*. This is why we prefer to call these prefixes *consonant particles*. They are, of course particles, but they consist of one sole consonant which complicates the identification in a text. Indeed, many words in Arabic do start by one of these consonants. It is true that many Arabic words are composed of three consonants, but this is not always the case. It might be possible to identify those prefixes by comparing prefixed words to a huge database of lexical forms in order to define which words contain prefixes and which do not. The outcome of this process, however, is not clear at all.

There are of course morphological constraints in the use of these prefixes as far as the grammatical category of the following word is concerned. It is obvious that the particle of the future *sa* cannot be used before a noun, nor before an adjective or any other grammatical

category except verbs in the present tense. This might diminish the scope of words which can be prefixed with this particle. For computer analysts this particle is easier identifiable, because it is followed by a limited number of consonants. As particle of the future, it can only be followed by the consonants *alif*, *ta*, *ya* or *nun*. But here too, we encounter words which are no verbs at all that start with this combination of consonants. Many nouns without vocalisation do resemble verbs. Besides, in order to check those prefixes with a lexicon, all kinds of derivations need to be generated in order to obtain a maximum of possible word forms in the database. In our view, it is hardly feasible to identify these prefixes by computer, unless they are marked one way or another.

In this respect, we propose to identify the affixes in a specific manner. This is, of course, a manual treatment which demands quite a lot of energy, but once a corpus is primarily tagged that way, it might reveal to be rewarding. A complicating factor is that the same consonant prefix (particle) can have more than one function. According to Al-Batal (1990, 239), for instance, the particle *fa* can have different meanings, viz.: "causal (because), conclusive (therefore), consequential (accordingly), discourse switch (so then, for instance), succession (in time: subsequent to this) and Topic introduction: *'amma fa* (as for)." This holds also true within the traditional Arabic approach (Ya'qub, 1986, al-Dahdah 1990 and al-'Umari 1993). It is clear that it is not very easy to implement such a fine variegation.

The first aim of our research is to identify particles as such (e.g. the connective *fa*). Once the particles are identified in a text, the analysis can start. The function of some other consonant particles, however, differs so much, that a different *encoding* seems to be necessary. When the different functions of one consonant particle are too divergent, we propose to differentiate between the different functions by creating a specific encoding for every function. This is especially the case for the consonant *l* which can be prefixed (1) to a verb in the present tense, meaning *in order to*, such as, for instance in, *liyaktuba*: *in order to write*, (2) to a verb in the past tense as the particle *la* which is used after the conditional particle *law* and (3) to a noun when used as a preposition, such as, for example in *li-ahmad*: *for Ahmad*. Because these functions are too divergent, we propose to mark those prefixes one way or another, so that they can be identified by a parser in a computer program.

However, the above mentioned prefixes do also occur in combination. This means that in practice two or three prefixes can be linked to a word. We registered a total of 31 combinations of prefixes, when we take also the spoken varieties of Arabic into account. The three most frequently used combinations of prefixes are: (1) a combination between a connective and a preposition (for instance: *wa-bi*, in written language *wb*, meaning: *and with*), (2) a combination between a preposition and the article (for instance: *bi-al*; in written language *bal*, meaning: *with the*) and (3) a combination of three particles, which is most commonly the combination between a connective, a preposition and the article (for instance: *wa-bi-al*, in written language *wbal*, meaning *and with the*).

Those combinations might also be identified by comparing a lexicon to specific words in a text, but this will result in many mistakes, because a series of words have the same combination of consonants at the beginning of the word. A special case is the combination of the particle *li* with the article *al* which in Arabic script becomes the combination of *ll* (i.e. twice the consonant *lam*). Since no Arabic word starts with this combination, it is a prefix which is easily detectable by computer, except for those words of which the first consonant is *lam*, which again might be a source of mistakes in identification.

The other combinations of particles are less easy retraceable by computer, because there are many words which start with the same combination of consonants. Anyhow in order to determine prefixes in Arabic words a huge lexical database ought to be compiled.

2.2 Suffixes

As for the suffixes, there are 17 used as possessive suffixes. Besides, there is the suffix of the *alif* which is used as an undefined accusative and there is the suffix of the energetic, the *nna* which in Modern Standard Arabic is used rather sparingly. The possessive suffixes consist of one or two consonants. It is obvious that one consonant suffix is more difficult to identify than two. Moreover, there will always remain combinations which are ambiguous. The suffix *ha*, for instance, of the third person singular feminine can easily be mixed up with the undefined accusative of a word ending with the consonant *h*.

The most tricky part concerning the suffixes is the distinction between adjectives and adverbs. Adjectives in the accusative case end with an *alif*, but also adverbs end up with an *alif*. It is very important to make a distinction between those two grammatical categories. It is possible to make this distinction in a lexical database, but in corpora, only context can serve as effective measure to determine the right grammatical category.

3 The vowelless nature of the Arabic language

The second problem in tagging Arabic corpora is the vowellessness of the words in sentences. This causes problems not only on the previous mentioned multiword combinations, but also on word level. The vowellessness not only affects the meaning of words but also the grammatical labelling of words. This is especially the case for verbs. The different persons of the verb form, both in the present and past tenses, are in most cases only identifiable by means of vowels which are omitted. The verb form *ktbt*, for example, can refer to four possible persons: i.e. the first person singular, the second person singular masculine, the second person singular feminine and the third person singular feminine.

It is almost impossible for a computer program to determine the person of these verbs. Only the context can help in defining the correct persons of a verb form. In this respect some help might be expected from a minimal form of text categorisation. Indeed, in newspaper text, the first person singular is less likely to occur, whereas in literature this person might occur more abundantly. Nevertheless it seems quite difficult to tag texts automatically when they are not vocalised or when the larger context cannot be taken into account.

On word level there is also interference with multiword combinations. In these cases ambiguity often occurs. As an example, we take the string of characters consisting of two consonants, viz. the *lam* and the *kaf*. Immediately, a reader will identify these two as *laka* (*for you*). However, there is also the verb *lakka* which means: *to hit with the fist*. Another example is the combination *kl*. An Arab reader will identify this combination immediately as the noun *kull (all)*. However there also exists a seldom used verb *kalla* (*to become tired*), which has the same outlook as the noun *kull*. Another example of such an ambiguity is the frequently used demonstrative pronoun *hadha* (meaning *this*) which form completely coincides with the verb *hadha* (with the meaning: *to talk irrationally*).

Incidentally, we may remark that in all the above mentioned cases the first of the two possibilities is more frequent than the other. In 99% of the cases the combination *lk* will refer to the prefix. Indeed, in most cases *la* is followed by the possessive suffix *ka*. The same goes for the combination *kl*, which in 99% of the cases will refer to the noun *kull* and also for the demonstrative pronoun. There are, however, other words where the choice between two or more possible forms lies around the 50%. In those cases reference to previous and following words might be a strong indication for one choice or the other, but this too presupposes that those following and preceding words have been identified previously.

At this point statistics can become to a certain level very important and of great help in tagging corpora together with text categorisation which might also prove to be very valuable in determining the different categories of forms and words. This, of course, presupposes the existence of corpora which have been encoded and on which statistical analyses can be conducted. In order to do so, Arabic corpora will have to be annotated first and analysed to determine the statistical occurrence of words and grammatical categories.

4 The combination of vowellessness and agglutination

In Arabic, vowellessness and agglutination go together in such a way that the combination of both complicates the identification of words. In their new book Badawi, Carter and Gully (2004) give also a glimpse of different possible interpretations for one set of characters between blanks. They add: “The above are merely hints at the disambiguation strategies practiced unconsciously by the native reader: they require a complete knowledge of all the possible morphological and syntactical combinations, and an awareness of the lexical and contextual factors.” In order to help the scientist tag Arabic corpora, the Arabic department of the university of Leuven proposes a tandem approach.

5 The Leuven tandem approach

In the Leuven tandem approach we make a distinction between the lexical approach and the corpus approach.

5.1 The lexical approach

For both approaches we make a distinction between minimal basic forms of words and maximal basic forms. The nucleus of the minimal basic form of a word is the lemma such as it is found in Arabic dictionaries. This vocalised form is in our view not suitable for corpus analysis, for two reasons. In the first place, because a vocalised form is seldom used in Arabic texts. Only Coranic Arabic is vocalised, as well as schoolbooks for primary schools. The latter, however, are not available in electronic form. The least one can observe about vocalisation is that it is in normal texts, such as, newspapers, magazines and literature sparingly used and often not in a consistent way. In the second place a vocalised form is not suitable because vocalised words are, contrary to what one might think, not completely unambiguous. To a certain extent vocalisation does away with the ambiguity of words but not in all cases, especially as far as noun and adjective or adjective and adverb are concerned.

5.1.1 Minimal basic forms

In order to disambiguate words, a tag has to be added to the word in question. This can be done in two ways, by adding the tag to the word while keeping the Arabic word itself unaffected, or by marking the Arabic word itself and in addition adding the tag to a word. In Leuven we chose the last option. This option has the advantage that a tag can be read immediately from the Arabic word itself. In other words, the encoding is marked in the Arabic word, and visible before the more expletive tag is added to the word. This pre-tagging of Arabic words is done in Leuven by a systematic selective use of the diacritical signs based on a convention which we have developed for Arabic.

In order to disambiguate the main grammatical categories in Arabic (i.e. verb, noun and particle) we made the following convention. Verbs are never vocalised, whereas for nouns the first consonant is always vocalised. This 'rule' makes it possible for a reader to disambiguate immediately between a noun and a verb. In this way, the three consonants *shrf* denote a verb (*to be highborn*), whereas the combination of consonants in which the first one bears a vowel denotes a noun, for instance *sharf* (*dignity*).

Note that the minimal basic form should not be confused with the root of the word, although both forms may coincide, as illustrated in *shrf* in the previous paragraph. Whereas the root is the theoretical stem of an Arabic word (which is often used as the principle entry in Arabic dictionaries), the minimal basic form is the encoded form of the morphological pattern of an Arabic word as it occurs in spoken or written language. For example, the word *maktb* (office) is a minimal basic form which does not coincide with its root *ktb*.

Other grammatical categories within the minimal basic forms are also indicated by using the diacritical signs in a selective but systematic way. Prepositions, for instance, are indicated as such by vocalising the last consonant in a systematic way (Van Mol, 2002). In this respect the consonants *khlf* denote a verb (*to be the successor*), the combination with a vowel on the first consonant *khalf* denotes a noun, whereas the combination with a vowel on the last consonant *khlfa* denotes a preposition. This pre-tagging makes it possible for the reader to identify immediately the grammatical category of a word on the one hand, but on the other hand it gives way for an electronic device, such as a computer, to detect the prepositions or nouns which were marked by these conventions.

Most of this primary part-of-speech tagging of Arabic words had to be done manually. In this way, we have composed a lexical database of until now 26,613 Arabic words which have been manually disambiguated. These words were not copied from an existing dictionary. Within a dictionary project of the Dutch Language Union, a corpus of 4,000,000 Arabic words of texts all dating after 1980 was read and translated in detail and entered in a database, after which the encoding was added to every single word. The dictionary was also made available in printed form (Van Mol, Berghman, 2001). This means that the corpus of words in the lexical database are not a theoretical sample of all kinds of Arabic words, but are words registered in the current use of Modern Standard Arabic. The corpus on which the lexical database was founded contained both oral and written sources.

On the other hand, the simple disambiguation of lemmata does not suffice as a reference scheme for the tagging of Arabic corpora. Indeed, words also have derived forms which, as such, occur in texts. In this respect, we developed a generating system for all words in the database, which generated all possible derived forms of the stem of a word. For nouns, these

are among others the sound and broken plurals (the broken plurals had to be added manually), but also the dual forms in both nominative and accusative case. These additional elements, which we consider as belonging to the minimal basic forms of the words, were also encoded in applying diacritic signs in a selective but systematic way.

For the verbs, we programmed a generating system which generated the different conjugated forms of the verbs according to their patterns and stems. This quite complex program makes also use of the diacritical signs in a selective way. For the different persons of the verbs in the past tense, the diacritical sign on the last consonant was added, in order to disambiguate between the different persons. For example, the first person of the verb *ktb* is written *ktbtu*, the second person masculine *ktbta*, the second person feminine *ktbti* and the third person feminine *ktbt*° with the *sukun*. In total 581,516 minimal basic forms were generated from the existing 26,613 Arabic words. This, however, does not mean that all those forms do occur in current language use. Only the comparison of those forms with a text corpus can give a further indication on occurrence and even frequency of the generated forms.

5.1.2 Maximal basic forms

The maximal basic form coincides also with a string of characters between two blanks, but is more than a minimal basic form. A minimal basic form departs from the word itself and consists of the nucleus, which is the Arabic word as found in a dictionary, or the derivations of the stem. In this way, all the additions to the minimal basic form are word related and are in a sense predictable. That is why they can be generated from the stem.

The maximal basic form, however, consists of a minimal basic form with the addition of other elements which do not directly relate to the word itself, but are elements which are in a way independent from the word. Those elements comprise all the affixes and combinations of affixes which are linked to the word because of the agglutinative nature of the language. They remain, however, independent elements which can be analysed as such.

It is of course possible to predict, to a certain level, the occurrence of these affixes to words. It is, for example, clear that there are verb particles which are only used in the present tense, such as, for example, the future particle *sa*. It is also obvious that some other affixes, such as the preposition *bi*, are never placed before a verb. It was possible to generate for all the words in the database other forms which consist of every generated minimal basic form plus the theoretical possible prefixes which might be added to it. As we registered a total number of 31 possible combinations of prefixes, we considered that another generation program for all the possible maximal basic forms would make the database too large and too burdensome in use. The more we generate, the more data will be available, which have to be checked.

For the indication of the affixes which belong to the maximal basic forms we also made use of the diacritic signs in a selective but systematic way. All the prepositions, for example, are indicated by means of the kashida. We first composed a test corpus of 320,000 words on which we conducted a linguistic investigation on the particles. The linguistic study we gave conducted on the test corpus shows revealing results (Van Mol 2003). For the first time in history of Arabic linguistics we were able to conduct a statistical analysis on Arabic particles and to examine the difference in use between different Arabic countries.

5.2 The corpus approach

In order to define the affixes which belong to the maximal basic form, we preferred to base ourselves on a primary tagged corpus. Besides the lexical database, which has been built up in a bidirectional relational database 4D for Apple Macintosh, we chose to build up a corpus of Arabic texts which was encoded according to the principles of the selective systematic use of diacritical signs. In this way, a corpus was compiled of 8,000,000 words all encoded according to the developed conventions. The corpus is still under development. The corpus consists of both oral and written Arabic and contains both literature (fiction and non-fiction) and media Arabic.

The next phase is precisely to confront the large corpus containing both minimal and maximal basic forms to the generated lexical forms in the database. It is clear that this confrontation will give us a more realistic view on the occurrences of minimal and maximal basic forms. All minimal basic forms which the program encounters in a text will be marked as well as the maximal basic forms. In this way, we will obtain a selection of the more than half a million minimal basic forms which do occur in reality; whereas others might never occur.

6 Conclusion

The data which will be gathered by comparing real corpus data with the lexical database will give us sufficient material to serve as a reference point for the part-of-speech tagging of raw corpora. This might be a step forward in the automatic tagging of corpora. The comparison of both databases will, as we hope, give us more insight into the structure of Arabic, especially at word level, as well as on the divergence in use of lexical items.

References

Al-Batal, M. (1990), Connectives as Cohesive Elements in a Modern Expository Arabic Text. Papers from the Second Annual Symposium on Arabic Linguistics. In Mushira Eid & John McCarthy (Eds.), *Perspectives on Arabic Linguistics II*, Amsterdam, pp. 234-268.

Ad-Dahdah, A. (1990), *Mu'jam qawa'id al-'arabîya al-'alamîya*, (added title: *A dictionary of Universal Arabic Grammar*), Beirut, Librairie du Liban.

Al-'Umarî (1993), *Masabih al-ma'anî fî hruf al-ma'anî*, Cairo, Dar al Manar.

Badawi, E., Carter, M. G., Gully A. (2004), *Modern Written Arabic, A comprehensive Grammar*, Routledge, 812 p.

Dichy, J. (2002), Arabic lexica in a cross-lingual perspective, In *Arabic Language Resources and Evaluation - Status and Prospects, Workshop proceedings LREC 2002*, 7 p.

Van Mol, M., Berghman, K. (2001), *Leerwoordenboek Arabisch - Nederlands (Learners' Dictionary Modern Arabic – Dutch)*, Amsterdam, De Nederlandse Taalunie, Bulaaq, 520 p.

Van Mol, M., Berghman, K. (2001), *Leerwoordenboek Nederlands – Arabisch (Learners' Dictionary Dutch - Modern Arabic)*, Amsterdam, De Nederlandse Taalunie, Bulaaq, 530 p.

Van Mol, Mark (2002), The semi-automatic Tagging of Arabic Corpora, In *Arabic Language Resources and Evaluation - Status and Prospects, Workshop proceedings LREC 2002*, 4 p.

Van Mol, Mark (2003), *Variation in Modern Standard Arabic in radio news broadcasts, a synchronic descriptive investigation into the use of complementary particles*, Peeters, 324 p.

Ya'qub, I. B. (1986), *Mawsu'at an-nahw wa-l-harf wa-l-'i'rab*, Beirut, Dar al Jil.