Exploring annotated Arabic corpora, preliminary results

Mark Van Mol

Institute of Living Languages, Catholic University of Leuven
Dekenstraat 6, B 3000 Leuven, Belgium
Mark.VanMol@ilt.kuleuven.ac.be

Abstract

In this paper, some results are given concerning the analysis of an Arabic corpus based on the use of a number of new techniques. Although necessary, the tagging of Arabic texts is hampered in three ways: the polysemy of the language, the fact that the language is normally not vocalised (thus enlarging the ambiguity of many words), and the fact that functional words are directly linked to content words, expressed as prefixes or suffixes.

We developed a classification system of Arabic words that enables us to define the words in the sentence separately from the affixes. This enables us in the first place to count the words in Arabic texts as independent units. Secondly it enables us to search for the word we are looking for, avoiding noise. The tagging, for instance, makes it possible to distinguish the verb form "hamma" from the conjunction "fa" followed by the verb "hamma", and also to distinguish the conjunction "fa" from the personal suffix "hum".

We compiled a tagged corpus of Arabic news broadcasts from the radio stations of Algeria, Egypt and Saudi Arabia. For each country a tagged corpus of approximately 80,000 words was compiled. Corpus analysis of the use of the particles of the future "sa" and "sawfa" show that no distinction is made between the near and the remote future.

Keywords: corpuslinguistics, arabic corpora, arabic tagger, lexicography

Introduction

With the availability of large amounts of corpora, also of the Arabic language, the interest in studying the language by analysing corpora is steadily growing. Indeed, we may obtain a completely new look, and new insights in the structure and the use of the Arabic language by analysing corpora. We might find new collocations, new meanings and even, it might be possible to identify some regional variation in Modern Standard Arabic (MSA). As far as we know a number of electronic Arabic text corpora have been compiled (cf. Ditters 1995: 123), but these corpora are raw, which means that the exploration of these corpora remains problematic. Some analysis which have been conducted on these corpora involve sometimes very limited data (cf. Ditters 1992: 42). Others have developed proficient word form analysers, such as the analyser by the Xerox European Research Centre, but the question remains whether these provide an adequate

solution for the exploration of Arabic tagged corpora. In this paper we will give a glimpse of some results we obtained in analysing a sample of the Arabic language by using a few new techniques.

Complexity of the analysis of Arabic corpora

In order to explore corpora in an efficient and in an economically reliable way, some preliminary operations ought to be made. As is generally known, analysing Arabic corpora is more complex than other corpora because of three main reasons. In the first place the Arabic language is very polysemic. While constructing an Arabic-Dutch and Dutch-Arabic learners' dictionary (Van Mol & Berghman, 2001) over the last 20 years we have noticed that the Arabic language is much more polysemic than, for example, Dutch. In fact in the Dutch language one way to create new words is by adding two words together in order to obtain a new word as a compound. These new words are very widespread, but are also identifiable by a computer in a simple way, i.e. by defining a word as a string of characters between two blanks. In the Arabic language new meanings for words are often given by expanding the older meaning of an existing word to a new one. This means that the external morphological form of the word does not change, in spite of the fact that the word carries a new meaning. In this way the word "miḍaḫḫa" does not only mean pump but also bicycle pump.

The fact that in Arabic new words are created on the basis of already existing stems or by elaborating the meaning of an existing word to another word makes it very polysemic. When we converted our database Arabic-Dutch of 15,000 words into a Dutch-Arabic database we discovered that we obtained a list of 20,000 words in Dutch, in spite of the fact that we made several restrictions before converting the database. E.g. all the maṣādir (verbal nouns) in Arabic that were translated by a verbal noun in Dutch were not converted into the Dutch database.

The polysemic character of the Arabic words makes the tagging of the words as a unity of the language more complex than in e.g. Dutch or English. The question then is, how far one has to go in tagging identical morphological forms into different categories according to the different meanings of the word.

A second element that makes analysis of Arabic more complex than other languages is the fact that the language is usually not vocalised, which means that the degree of ambiguity of words as separate units is much greater than e.g. in Dutch or English. Words, in their raw form, can belong to different grammatical categories as e.g. the string of characters "ktb" shows. This string of characters stands for the verb "kataba" (*to write*) as well as for the plural "kutub" (*books*). This complicates the searching for the words in a corpus of texts. Looking for the word "kataba", not only do we find also the plural form "kutub" but also a lot of other words that have nothing to do with the verb that we are looking for. We will for example also find the words "maktab" (*office*), "maktabīy" [*office* (in compounds)], and the word "maktaba", (*library*). This means that while we are searching for a word in an Arabic text corpus we find a lot of words of which we are in no need of. But the consequence is also that when examining for example my concordance program we lose a lot of time by reading sentences in which the wrong word is found.

To illustrate this point, let me give here a survey of the searches made in a raw corpus for the word "kataba". The searches were made on a very varied corpus of texts of approximately 1,500,000 words. Searching for the verb "kataba" gave the following results:

word	Percentage right finds	Percentage other finds	percentage rate lost time in searching
kataba	28		72%
kutub		18	
maktab		42	
maktaba		12	

This table shows that the searches on a raw corpus are very time-consuming. A more striking example is the search after the elative "ʿašadd". The table below shows that we have a success rate of only 20%, which means a loss of time of 80%.

word	Percentage right finds	Percentage other finds	percentage rate lost time in searching
ʿašadd	20		80%
nāšada		38	
surnames		34	
munāšadāt		6	
other		2	

The searches of a limited number of categories of words give a high rate of success. But even when we take a maṣdar of the second form we will obtain variable forms. When searching for the word "taʿlīq", we obtained a success rate of 100%. But these 100% comprised also forms such as the plural form "taʿlīqāt" and the accusative form "taʿlīqan", which, of course for lexicographical analysis are still useful in that context.

In most cases however success rates by searches are much lower. In our example above, the searches for the words "maktaba" give the highest rate for success because the word is very long, which makes it less ambiguous. That way it can not be mixed with words of another form or of another meaning. The word "maktab" however gives a success rate of 78%. This still means that by examining a large corpus we lose 22% of our time in finding the wrong words. Because lexicographical analysis is anyhow very time consuming, every loss of time has to be avoided. For words such as "kataba" the success rate is only 28% and for the plural form "kutub" the success rate is only 18%. Let alone that the verb "kataba" is still a comfortable form as there does not exist a 5th form of the verb, nor does there exist a maṣdar of the first form that completely matches with the verb form. When searching for the verb "rafaʿa" (*reject*), (72%) e.g. we also find the maṣdar "rafʿ" *(rejection*) (28%).

Some words such as the hollow verbs show an even poorer result. Somebody who wants to examine the function of the verb "kāna" (*to be*), will, when searching for the verb in a raw corpus

obtain the correct form in 84% of the cases. But these results are still very partial because they do not include at all other morphological forms of the verb "kāna" that cannot be omitted in such an investigation. Think for instance of forms of the present tense "akun", "takun" or "yakun" etc. but also of the forms of the jussive "yakun" etc. Let alone that until now it is not possible to do combined searches in which we may look after a certain verb e.g. "kāna" in all its forms or in some of its forms in combination with other verbs, for example in the past or the present tense.

One might try to solve the problem of the searches by vocalising the text. This solution however is partial, because it does not give an adequate division into grammatical categories in all cases. For example, the word "masʾūl" can be an adjective (responsible) as well as a substantive (the responsible person). Vocalising the word does not tell me whether we have to do with a "mafʿūl" form that is used independently as a noun or attributively as an adjective.

In the third place, the problem is complicated by the fact that in Arabic a number of prefixes and suffixes are directly linked to the word. This makes the searching by computer even more complex. For example the string of characters fhm can stand for the verb "fahima" (*to mean*), but it can as well stand for the particle and suffix "fahum" (*and they*) or for the particle and verb "fahamma" (*and he began*). Only conducting a search on the basis of the three characters "fhm" means that we will obtain a lot of words that we are not searching for. As already mentioned above there are of course, in the actual state of affairs, to a certain degree searches that can be successful. This however does not solve the real need that exists to examine and explore corpora in the Arabic language in a precise and thorough way. How, for instance, are we going to explore the use of the conjunction "fa" or the conjunction "wa". Impossible to find in a raw corpus because every word that begins with the consonant "fa" or the "wa" will show up.

Solution for the disambiguation of words by POS tagging

Therefore we were looking for a solution to identify words in Arabic texts, not only as words but also in their grammatical form. In advance we assumed that the tagging ought to be very transparent and that it ought to be set up in such a way that the tagged text could easily be transformed into a normal written text.

We developed a system of classification of Arabic words that enables us to define the words in the sentence separately from the prefixes and suffixes. This enables us in the first place to count the words in Arabic texts as independent units. Secondly it enables us to search for the word we are looking for, avoiding noise. The tagging, for instance, makes it possible to distinguish the verb form "fahima" from the conjunction "fa" followed by the verb "hamma" and also to distinguish the conjunction "fa" from the personal suffix "hum". Our system has indeed its limitations. In the case of the particle "fa" for example, the system enables us to find exclusively the particle "fa" in context. The particle "fa", however, can in itself also be divided into different subcategories according to the function it has in a sentence. This detailed subdivision is not provided in our system.

Another problem that we have encountered by exploring Arabic corpora is that the existing concordance programmes do not allow to explore text corpora in a fashionable way. In some of the concordance programs (KWIC Key word in Context-indexes), you can only define the

number of characters you want to see at the right side of the word and the number of characters you want to show up at the left side of the word. This technique may be useful for a certain amount of collocations of words. For instance, the examination of the collocations between nouns and adjectives. Indeed as an adjective in Arabic immediately follows the noun, it is possible to examine a number of collocations. We did this, for example, for the collocation of the word "šadīd" (*strong*). In spite of the fact that here we obtained a serious amount of garbage [e.g. the word "tašdīd" (*intensification*) is also shown] we nevertheless obtained an interesting range of possible collocations which were very useful for the dictionary we have compiled.

However, the search findings are of no use when we want to examine parts of speech in a sentence that does not follow immediately after a crucial functional word. This is especially true, for example, for the study of the conditional particles such as "ʾiḏā" (*when*), "ʾin" (*if*) and "law" (*if*). Here, in fact, the main clause may start tens of words after the particle studied, which makes an efficient analysis impossible. In our programming, therefore, we made the option to look for words in complete sentences. The program itself divides at first every text in sentences, and the concordance program always shows the word in a complete sentence. Larger connections are also useful. For the study of connectives certainly the paragraph might be the most interesting *unit* in the concordance program. As we do not yet work in a database, we have to limit ourselves at this stage of the research to the sentences as basic unit for investigation.

Software for frequency counts

The tagging of every word in context gave us the possibility to count the amount of words in a text. As far as we know, until now the counting of words in Arabic texts is limited to the counting of a string of characters between two blanks. With our system it is also possible to identify the particles that are attached to the words such as the connectives "fa" and "wa", but also the prepositions "ka", the particle of the future "sa" and so on, and to count them. We also can count the definite article. Further we can also count the possessive suffixes, but we refrained from doing so because, in our view, these form a unit with the word. The identification of particles in texts is the first step towards the compilation of computerised frequency word lists, which is an option for further research.

The compilation of a corpus of Arabic radio broadcasts

In what follows we will give a brief survey of a few results that the exploration of a tagged corpus can yield. We compiled a tagged corpus of the Arabic news broadcasts from the radio stations of of three different Arabic countries, viz.: Algeria, Egypt and Saudi Arabia. Algeria was chosen because of the presumed big influence of the French language in the Algerian society. Egypt, on the other hand, because of its presumed predominant position in the Arabic world, especially as far as language is concerned. And finally Saudi Arabia because of the presumed closed character of its society. For each country we compiled a tagged corpus of approximately 80,000 words.

The distribution of the particles *sa* and *sawfa*

A point of investigation was the use of the particle "sa" and "sawfa" in Modern Standard Arabic on the radio (Van Mol, 1998). We compared the actual use of these particles in MSA with the grammatical norm as pointed out in grammatical works of Yaʿqūb and ʾal-ʿUmarī. They point out that the particle "sa" refers to the near future, whereas the particle "sawfa" refers to the remote future. In the modern grammatical works of e.g. Blohm & Reuschel (1981: 462) this distinction in meaning is no longer mentioned. The same goes for the writings of Cowan (1973: 88), Cantarino and Stoetzer (1991: 182). Notwithstanding that it is still considered the norm by Arab grammarians in the Middle East. (Ziyān 1994)

The quantitative analysis of this particle in our corpus, however, shows the following distribution for the different countries:

News emissions						
particles of the future	Algeria		Egypt		Saudi Arabia	
	N	%	N	%	N	%
Total occurrences	254	100	263	100	270	100
sa	218	86	212	81	220	82
sawfa	4	2	25	9	26	10
lan	32	12	26	10	24	8

The application of the chi square test shows that in this table there is statistically a significant difference in the distribution of the particles of the future in the different countries. It is significant that the particle "sawfa" is much less used in Algeria than in the two other countries under investigation. The chi square test shows that the difference is significant. Statistically spoken, there is only no significant difference in frequency for the use of the particle "lan".

This table shows that the particle "sawfa" is statistically much less used in (the corpus of) Algeria than in the two other countries. Analysis of the corpus also shows that the negative combination "sawfa lā" did not occur. Negation of the future was exclusively expressed by the particle "lan".

The qualitative analysis shows only significant results for Egypt and Saudi Arabia. In the Algerian corpus, three of the four occurrences of "sawfa" do refer to a remote future and only one to the near future followed by the word "lāʾiqan". In the Egyptian corpus we could not determine a difference in time span between the use of the particle "sa" and "sawfa". Both particles do occur with exactly the same time indications. For Egypt, for instance, we found both "sa" and "sawfa" with the collocation "fi waqtin lāʾiqin" (*later on*). We found both particles in collocation with the word "ʾalyawma" (*today*). But also in Egypt both particles in collocation with the expression *in the next three months* and *in the next two months*. Both particles are used in the remote undefined future as well. We found some variation in time aspects in the corpus of Saudi Arabia, where the particle "sa" is used, for instance, in collocation with the expression *in the next nine months* and the particle "sawfa", in collocation with the expression *in the coming six months*.

Investigation of the corpus seems to indicate that in the news broadcasts in MSA, no distinction is made anymore in the meaning of the particles "sa" and "sawfa".

The distribution of the expression at the same time

As we saw, corpus analysis can throw some light on the tension that exists between actual language use and the norm as it is proposed. The next example shows that also regionally some variety may occur in the actual use of the language. The expression *at the same time* can be expressed as "fi-l waqt nafsihi". or as "fi nafs il-waqt". According to some Arabs, the expression "fi-l waqt nafsihi" is considered to be a flaw. Investigation of the corpus gives the following results:

	fi nafs il-waqt		*fi-l waqt nafsihi*		Total	
	N	%	N	%	N	%
Algeria	16	100	0	0	16	100
Egypt	21	33.8	41	66.2	62	100
Saudi Arabia	3	50	3	50	6	100
Total	40	47.6	44	52.3	84	100

This table shows that the construction "fi nafs il-waqt" in general does occur almost with the same frequency as the construction "fi-l waqt nafsihi". However, the differentiation by country learns that the construction "fi-l waqt nafsihi" has completely disappeared in the sample of Algeria, that we have the same distribution in the Saudi Arabian corpus, but that the construction "fi-l waqt nafsihi" is most intensively used on the radio in Egypt where in 2/3 of the occurrences this form is used.

Conclusion

Although the amount of data is rather small, these two examples give some indication of topics that might be investigated by examining tagged Arabic corpora. Tagging Arabic corpora of different periods, kinds and countries may learn us a lot about the evolution of e.g. MSA. Indeed, in the above case it might be also interesting to investigate whether the proclaimed difference in meaning of the particles "sa" and "sawfa" existed, and if so, to determine when the blurring in meaning of these particles started.

The second example is one illustration of regional differentiation in the use of MSA. Until now we only possess a rudimentary intrumentarium to investigate Arabic corpora. In the future we hope to explore Arabic corpora in a database that will be developed specifically for that purpose. In the future we hope to expand our search procedures from searches for a specific word, or specific word combinations to the investigation of relationships between grammatical categories and words.

Literature

ʾAl-ʾAdnānī Muḥammad (1989). Muʿjam al-ʾAlfāẓ al-luġawīya al-muʿāṣira, Beiroet, 870 p. (added title: A Dictionary of Common Mistakes in Modern Written Arabic).

Al-ʻUmarī (1993). Maṣābīḥ al-maʿānī fī ḥurūf al-maʿānī, Cairo, 701 p.

Blohm, Dieter, Reuschel, Wolfgang & Samarraie, Abed (1981). *Lehrbuch des Modernen Arabisch*, Leipzig: Verlag Enzyklopädie, Vol 2, 1112 p.

Cantarino, Vicente (1974-75). *Syntax of modern Arabic prose*, 3 Vol, Bloomington, London.

Cowan, D. (1970). *An Introduction to Modern Literary Arabic*, London, xi, 205 p.

Ditters, Everhard (1992). *A Formal Approach to Arabic Syntax: The Noun Phrase and the Verb Phrase*, Amsterdam, 475 p.

Ditters, Everhard, & Moussa, Ali Helmy (1995) "The compilation of a Corpus of Modern Standard Arabic", *Processing Arabic*, Report 8, pp. 113-132.

Stoetzer, Willem (1991). *Arabische Grammatica in schema's en regels*, Muiderberg, 337 p.

Van Mol Mark (1998). *Variatie in Modern Standaard Arabisch in radionieuwsbulletins, een synchronisch descriptief onderzoek naar het gebruik van complementaire partikels*, (English title: Variation in Modern Standard Arabic in radio news broadcasts, a synchronic descriptive investigation into the use of complementary particles), Ph.D. dissertation, Leuven, 292 p.

Van Mol, Mark & Berghman, Koen (2001) *Leerwoordenboek Modern Arabisch – Nederlands*, De Nederlandse Taalunie, Bulaaq, 506. (English title: Learners'dictionary Modern Standard Arabic – Dutch), The Dutch Language Union, Bulaaq, 506.

Van Mol, Mark & Berghman, Koen (1999) *Leerwoordenboek Nederlands - Modern Arabisch*, De Nederlandse Taalunie, Bulaaq, 529. (English title: Learners' dictionary Dutch - Modern Standard Arabic), The Dutch Language Union, Bulaaq, 529.

Yaʿqūb, ʾImīl Badīʿ (1988). Mawsūʿat al-ḥurūf, Beiroet, 662 p.

Ziyān ʾAḥmad ʾal-ḥājj ʾIbrāhīm, (1994). ṣafḥat fī-l-luġat, ʾal-qāfilat, January - February 1994, p. 48.