The development of a new learner's dictionary for Modern Standard Arabic: the linguistic corpus approach

Mark VAN MOL, Leuven, Belgium

Abstract
This paper reports on the development of a new Arabic-Dutch/Dutch-Arabic learner's dictionary, which has been compiled on a geographically distributed computer corpus of written and spoken Arabic. In the field of Arabic lexicography, it is the first dictionary of its kind. Although the use of computer corpora has become a well-accepted approach for many languages ever since the first publication of the COBUILD dictionary (1987), no such dictionary has been compiled before for Arabic. The three million words corpus provides the lexicographer with useful contexts of contemporary usage, giving information on, for example, collocations and fixed prepositions. Since part of the corpus is not vocalised, a special encoding system has been developed to facilitate corpus exploration. The compilation of the dictionary and the exploration of the corpus has brought new insights in lexicographic research of Modern Standard Arabic, the results of which will be used for the development of an electronic version of this dictionary.

1 Introduction

Twenty years ago I decided to start with the compilation of a learner's dictionary for the Arabic language. From the outset it was decided to base the dictionary on an extensive corpus of Arabic texts. At that time there was only one Arabic-Dutch dictionary, and one Dutch-Arabic dictionary. The Dutch-Arabic dictionary was merely a kind of word list of approximately 10,000 words of which the Arabic meaning was given without vocalisation ([Derwish 1988]). As is generally known, Arabic words are not vocalised in plain text. However, in a learner's environment, Arabic words are indeed vocalised, precisely to aid the student to pronounce the word correctly. Arabic dictionaries ought, always to be vocalised; if not, these dictionaries remain of very limited use, as only very learned Arabs would be able to take advantage of these, and even they will have their doubts. In the above mentioned dictionary words just follow each other without any indication of their specific meaning, let alone specifications about the use of the words. We conclude the Dutch-Arabic dictionary was most inadequate. The Arabic-Dutch dictionary [Amien 1980] was also lacking. It contained a large amount of faulty translations into Dutch. That means that there was a great need for a dictionary that not only gave the complete vocalisation of the Arabic words, and would take into account accurate translations, but that also would provide appropriate collocations for both languages.

2 Origin of the project

As many people proposed, it would be possible to translate an existing excellent dictionary such as e.g. the famous Arabic-German dictionary of Hans Wehr, [Wehr 1979]. The advantage of compiling a new dictionary on the basis of another dictionary might be the serious amount of time that might be saved in this way. This, however, is an illusion as we shall point out. It is clear that an already existing dictionary may form a stable basis for a new dictionary, on the condition that it is of excellent quality and that the target language is the same language as is used in the existing dictionary. We want to draw the attention to the fact that in our case, we had to start from zero. The low quality of existing Arabic-Dutch, Dutch - Arabic dictionaries excluded them as a basis for a new dictionary.

On the other hand, a very serious disadvantage of the 'add on' approach is that it is unfeasible to define the exact meaning of words without a clarifying context. The method of translating an existing dictionary would surely lead to inaccurate translations from the original words.

Take for example the Arabic verb *'amala*. In the dictionary of Hans Wehr we found 36 English meanings for this verb. By analysing a corpus, of these 36 meanings we retained only 8 English meanings that were clearly appropriate in context. But in turn these 8 meanings of the dictionary gave us a list of 257 words in Dutch. Without a context it is an impossible task to define which of these 257 words give an accurate equivalent of the Arabic language. By analysing the word in context we obtained only 31 out of the 257 Dutch words that were applicable.

This means that only 12 % of the meanings found in the English-Dutch dictionary were useful and that, on the other hand, 88 % of the meanings were not useful and hence in most cases did not represent the correct meaning of the word in Arabic. This was one of the main reasons why we decided to base the dictionary almost exclusively on corpus analysis.

3 Advantages of a corpus analysis

Only by analysing corpora were we able to accurately define the Dutch meanings of the Arabic words. It is clear that nowadays dictionaries can hardly keep up with the development of new words. Especially in European languages the number of new words is very large. I have the impression that in the Arabic language the creation of new words is a more gradual one. Nevertheless, comparing our corpus with the famous dictionary of Hans Wehr, we found that about 5% of frequent new words and meanings were not found in that dictionary. On the other hand, we found that the great majority of the words in the dictionary of Hans Wehr are not frequently used (anymore?) in Modern Standard Arabic.

The corpus approach also has the advantage that it gives the possibility to find new meanings that are not represented in the available dictionaries and to define more precisely the modern vocabulary of a language. It is generally known that the Modern Standard Arabic language has undergone a serious evolution and transformation over the last seventy years, especially as far as the creation of new words is concerned. Arabic academies have proposed a lot of words for new terms. Many of these proposals did not survive in the actual use of the language. The compilation of a contemporary representative corpus gives, indeed, an interesting indication about the actual use and acceptance of a word. Al-Šihābī (cited in [Stetkevych 1970, 28]), for instance, mentions eleven neologisms coined for the word *brake*. The analysis of our corpus reveals that in actual language use only two of these proposed words are still currently used and that an additional third word has come to light.

Another important aim was to give a deep insight in the context in which the words are used in the Arabic language. The traditional dictionaries lack additional information like, for instance, all kinds of collocations, fixed prepositions, and of course telling example sentences. In the existing Arabic dictionaries only the meaning of a given word is available, without any specification about its use in context. From a productive point of view, context or collocations are of great importance. One might, for instance, on the one hand easily find the Arabic word for *snow* in the dictionary, and on the other hand, the Arabic word for the verb *play*. But this leaves the user helpless in constructing a practical sentence such as *the children play in the snow*. When the user of the dictionary is looking for the Arabic word for the preposition *in* he is likely to take the most frequent word *fī* to use in translation. Corpus analysis, however, shows that the correct preposition in that sentence is not the preposition *fī* but the preposition *'alā* which is generally translated in English as *on*. Because all this information is lacking in the existing Arabic dictionaries the compilation of a corpus of Arabic texts seemed essential.

As I mentioned above the basic aim was to compile a learner's dictionary that covers the basic vocabulary of the Arabic language. The macro-structure of the dictionary ought to be limited, but the micro-structure had to stay open for an optimal and thorough elaboration. In order to define the basic vocabulary of the Modern Standard Language, we followed a certain strategy in developing the corpus. The most crucial question was the selection criteria of texts in order to find the core vocabulary of the language. Core vocabulary in its broadest sense, because we wanted to create a dictionary which, in spite of the fact that the macro-structure was limited, would serve as a useful tool to translate or understand every Arabic text.

4 Composition of the corpus

Finally, the corpus was based on three main sources. I presumed that the spoken and written to be read language, such as it was found in the media ought to provide the most relevant kind of vocabulary. Indeed, when speaking, people do not have much time to grapple for words, hence it was presumed that rare words would not so easily appear in a spoken corpus. We therefore started with the transcription of radio and television broadcasts. We also tried to cover the whole geographical Arabic area. As a basis for the corpus three countries were initially chosen. Algeria, because of the presumed major influence of the French language in 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. At first only news programs were transcribed. Later on I also transcribed other programs such as documentaries, talk shows, all kinds of interviews, speeches, radioplays, press conferences, etc. After some while I expanded the corpus to include other Arab countries from the Middle East and North-Africa. Eventually, we ended up with a corpus of the spoken language of approximately 700,000 words.

After the compilation of the spoken corpus, I immediately started, the detailed translation of the corpus, word by word and sentence by sentence. In the beginning this work went very slowly. Only one sentence was translated per hour. I wanted to work as accurately as possible. Therefore, every word in the sentence was looked up in an Arabic dictionary, also when there was no doubt about the translation. Most of the time I used the Arabic-English dictionary of Hans Wehr, but also the Arabic-French dictionary of Abdel-Nour, [Abdel-Nour 1983]. Every

English or French word that was at first sight suitable in the context of the Arabic sentence, had to be looked up in a Van Dale English-Dutch or French-Dutch dictionary [Van Dale 1991] in order to define the exact range of each word. Every Dutch word that matched with the English or French word was checked in the Arabic context. Only when a word matched a hundred per cent was it accepted. We also paid a lot of attention to the corresponding prepositions and collocations. Through this method a lot of new collocations were found and inserted in the dictionary.

In order to include texts of the written language in the corpus, we expanded the corpus with the handbooks for acquiring the Arabic language used in primary schools of nine Arabic countries. This means that altogether we compiled a corpus of ca. 50 textbooks. I chose these because they form the basis of the vocabulary such as it is presented by the authorities in the different countries to their children. The texts in the handbooks also cover a very great variety of subjects and situations. All these textbooks were translated in detail. Also the handbook for the Arabic language of the Arabic League was translated. One of the advantages of compiling the corpus this way, was that all the Arabic words were completely vocalised. When working on normal Arabic texts that are not vocalised such as magazines, novels or newspapers there is always some doubt about the exact pronunciation of the words. By using vocalised texts we excluded all doubts as to the vocalisation of the words. We even found that the vocalisation in reality sometimes differs from the vocalisation in existing dictionaries. In the current dictionaries, for example, the word *mfdy* is vocalised as *mafdiy*, whereas in the news media this word was always pronounced as *mufadda.*

Moreover, some word forms are not identifiable when the text is not vocalised. This goes, for example, for the Arabic verb forms of the second and the fourth form, and to some extent even for the verbs of the first form. If a text is not vocalised, a non-native speaker, and even an untrained native speaker could not possibly define which form is intended. By translating vocalised texts this problem was completely avoided.

The third sample of texts on which the dictionary was based consisted of non vocalised texts from magazines and newspapers, a large part of which were taken from the internet. In all, an Arabic corpus of 3,000,000 words was compiled, of which one fourth was taken from oral sources.

5 The problems of exploration in a raw Arabic corpus

Precisely because of the fact that the Arabic language is not vocalised the exploration of a raw corpus in Arabic is even more time consuming than in an other language. The degree of ambiguity of words as separate units is much greater than e.g. in the Dutch or the English language. 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 noun *kutub* (*books*). This complicates the search for words in a corpus of texts. When I want to look for the word *kataba* not only do I also find the plural form *kutub* but also a lot of other words that have nothing to do with the verb that I am looking for. I will, for example, also find the words *maktab* (*office*), *maktabiy* (*office-*), and the word *maktaba* (*library*). This means that while I am searching for a word in an Arabic text corpus I find a lot of redundant words. Consequently when examining, for example, my concordance program I lose a lot of time by reading sentences in which the wrong word is found.

To illustrate this point, let me give a survey of the searches made in a raw corpus for the word kataba. Searches on a raw Arabic corpus are very time consuming. Only for some categories of words do I obtain a high rate of success. When searching, for instance, a masdar (verbal noun) of the second form, such as the word *ta‘lĪq*, we had a success rate of 100%. In most cases, however, success rates by searches are much lower. Especially for verb forms such as *kataba* the success rate is only 28%, and for the plural noun *kutub* (*books*) the success rate is only 18%. Notwithstanding that the verb *kataba* is still a comfortable form as there does not exist a 5th form of the verb, nor is there a masdar (verbal noun) of the first form that completely matches the verb form. This means that when exploring a corpus in that way for every word examined up to 82% of time may be lost by finding the wrong word.

6 The tagging of an Arabic corpus

Therefore, I developed an encoding system for the Arabic language that eliminates the ambiguity of the words to a great extent. This not only grants important timesaving when exploring Arabic corpora, but it also rewards investment in time by providing every word with the correct tag. When first using the programme the balance of investment in time and time saving is equal, but after a while a tagged corpus presents a lot of advantages. At the point of writing we can find the exact word we are looking for in a text. In the future, however, we hope to develop the searches in order to make combined searches. Indeed, the larger the corpus the more sentences will show up while searching. This is why we are also developing a system to perform combined searches in order to obtain the most relevant collocations in a corpus and to group them.

The searches of the corpus made it not only possible to refine the translations, but they also gave an interesting survey of the importance of the translations per word. The concordance files of every Arabic word in context quickly gave an interesting survey of the meanings of a given word that were more predominant than others. That way it was possible to order the different meanings of a given entry acording to its prevalence.

7 Contents of the dictionary

After years of intensive teamwork, a corpus of approximately 3,000,000 words was translated for the greater part word by word in context, but also by computer searches in a concordance program. This resulted into two learner’s dictionaries. One Arabic-Dutch of 17,000 Arabic entries, and one Dutch-Arabic of ca. 20,000 entries. [Van Mol 2000] Samples of different texts point out that this learner’s dictionary covers 99% of the vocabulary of any average text. This means that in spite of the limited macro-structure, (the large dictionary of Hans Wehr, contains approximately 45,000 words), we cover almost the whole range of the actual vocabulary. It also means that a learner ought to be able to understand every modern Arabic text in using this dictionary.

8 Conclusion

To conclude I want to mention important innovations that we introduced in this dictionary. In the first place there are the discriminating pointers. In the available Arabic dictionaries, a list of meanings with each of the entries may be found that are most of the time even typographically not very sharply delineated. If a meaning differs from an other meaning, the lists of words are in most cases separated by a comma. This means that the user of the dictionary has to search through the whole list of words for the appropriate meaning. Moreover the Arabic user encounters many difficulties in finding the right meaning of a word, because of the fact that, as is the case in most dictionaries, discriminating pointers are lacking.

The new feature in our dictionary is that there are a great variety of discriminating pointers which help the Arabic user to search for the right word. The second new feature is that there is also a typographical distinction between the most prominent meaning and the following synonyms. The last important feature is that the dictionary contains over 10,000 illustrative contexts. The problem of the exemplary sentences is that this takes a lot of space. Exemplary sentences do reveal a lot about the actual use of a word but, on the other hand, they take up the greatest part of the dictionary. Therefore, sentences were chosen for their relevance in relation to the translation. Special attention was paid to the contrastive use of the prepositions.

At this moment we are working on an electronic version of the dictionary. Thanks to the tagging of the Arabic corpus, it might be possible to look for the translation of a word, by clicking on the word in a text. However, this demands a detailed operation to the corpus as well as to the tagging of the words in the dictionary.

References

[Derwish 1988] DERWISH, H.H. (1988) Kramers woordenboeken, *Nederlands Arabisch*, Elsrevier, 367 p.

[Amien 1980] AMIEN, (1980) *Arabisch Nederlands woordenboek*

[Wehr 1979] WEHR, Hans (1979). *A dictionary of Modern Written Arabic*, Ed. J. Milton Cowan, Wiesbaden, xvii, 1301 p.

[Stetkevych 1970] STETKEVYCH, Jaroslav (1970), *The Modern Arabic Literary Language , lexical and stylistic developments*, Chicago - London, UCP, 135 p.

[Abdel-Nour 1983] ABDEL-NOUR, Jabbour (1983) *Dictionnaire Arabe-Français*, Beiroet, 1126 p.

[Van Dale 1991] VAN DALE (1991) *Groot woordenboek Engels - Nederland*s, 1691 p.

[Van Mol 2000] VAN MOL, Mark & BERGHMAN, Koen (2000) *Leerwoordenboek Modern Standaard Arabisch - Nederlands,* De Nederlandse Taalunie, Bulaaq, 500 p.

[Van Mol 2000b] VAN MOL, Mark & BERGHMAN, Koen (2000) *Leerwoordenboek Nederlands - Modern Standaard Arabisch*, De Nederlandse Taalunie, Bulaaq, 500 p.