

# **AraLat: A relational database for the development of bilingual Arabic dictionaries**

**Mark Van Mol\* & Hans Paulussen\*\***

\*Catholic University Leuven, Belgium

Mark.VanMol@ilt.kuleuven.ac.be

\*\* Lernout & Hauspie Speech Products, Belgium

Hans.Paulussen@lhs.be

## **Abstract**

In this paper we will discuss the relational database AraLat (Arabic vs. Latin script Languages) which has been developed at the Catholic University of Leuven for the generation of paper and electronic versions of Arabic bilingual dictionaries. The foreign language involved was Dutch, but the dictionary concept is open for any Latin script based foreign language (hereafter FL).

The database consists of four separate units. Two units, reserved for source and target language, contain all the grammatical and semantic information related to the language in question. The two other units contain the link Arabic-FL which makes it possible to generate dictionaries in both directions. The database offers the possibility to generate an Arabic dictionary in alphabetical order or in typical Arabic root order. Every unit contains four levels, which makes the compilation of the dictionary much easier. It contains a completely elaborated learners' dictionary consisting of 19,000 Arabic words translated in context and more than 10,000 illustrative sentences chosen from a representative corpus of 3,000,000 words. It is the first Arabic dictionary which is based on extensive corpus analysis, and for which modern corpus analysis tools have been developed. The user-friendly dictionary can be consulted on-line.

Among the new features for Arabic dictionaries is the use of a discriminating pointer. The structure of the dictionary is context-oriented so that the user cannot only find the correct translation, but also the correct collocations. The database includes an exporting feature to produce automatically a camera-ready copy for a paper dictionary. The presentation of this paper will be accompanied by several examples of structures of lemmata such as they have been conceived in the database.

## **1 Introduction**

In 1996, the Dutch Language Union started funding the Institute for Living Languages of the Catholic university Leuven for the realization of the compilation of a Learners' dictionary Modern Standard Arabic-Dutch and Dutch-Modern Standard Arabic (hereafter MSA). From the beginning we opted to base our dictionary on a large corpus of authentic Arabic text material.

Although there are a number of excellent Arabic dictionaries, such as Hans Wehr (1979) and Abdel-Nour (1983, 1995), we did not want to base the compilation of our dictionary on a translation of such a dictionary (which is an unfortunate common practice in lexicography). The two main reasons for not doing so are related to the problem of intuition based approaches and the lack of contextual information (Van Mol: 2000b).

In the first place, most existing dictionaries of MSA are not corpus but intuition based. Our corpus analysis revealed several biased translations in different existing dictionaries, precisely because a literal translation of a word was given instead of the word currently used in the target language. We found several terms in French and English which were literally translated in Arabic

dictionaries, which means that for some concepts often two Arabic concepts were to be found. Our corpus analysis, however, revealed that for many of these notions a separate Arabic word exists.

In the second place, the existing Arabic dictionaries only give in most cases a word-by-word translation. No attention at all is given to lexical phraseology. Seldom do they contain phrasal verbs, typical phrases, multi-word lexical units, collocations, etc. Nor is there any attempt made to include elements of pragmatics. This is partly due to the fact that only a thorough analysis of corpora can yield a systematic survey of collocations, idiomatic expressions and typical phrases of a language. Moreover, in order to explore such a corpus, specific tagging and corpus analysis techniques are required, which are uncommon for Arabic.

## **2 Strategies**

The challenge of compiling a new learners' dictionary implied two basic strategies, the first being the development of a relational database which is suited to treat Arabic and the second being the compilation of a representative corpus and the development of specific tools to explore the corpus in order to compile a dictionary enriched with many collocations, idiomatic expressions and illustrative phrases in an economic way.

### **2.1 The development of the database**

The development of a database for Arabic script and Latin script based languages was constrained in different ways. First of all, it had to be flexible enough to support both scripts. In the second place, the database had to be user-friendly which implies, for example, that the number of screens for data-input had to be limited. In the third place, the database had to be reversible. This means that the Arabic-FL data had to be transferred into a FL-Arabic database. And the last prerequisite was the possibility to produce a camera-ready copy from the database. All these elements will be discussed briefly.

#### **2.1.1 The adaptation of the database for the Arabic language**

The treatment of Arabic on computer is not self-evident. As Arabic is written from right to left, and European languages are generally written in the opposite direction, the combination of both scripts (and especially the embedding of one script in the other at different levels) can pose a number of computational problems that had to be solved beforehand. Moreover, a solution to this problem required a different approach in the paper and electronic versions of the dictionary.

A basic problem is the fact that a character code page for one language often intersects with the code page for another language. Although computers can support multiple languages, and are in that sense multilingual, this feature is usually restricted to one or a few languages only. Genuine multilinguality requires a multiscrypt solution, which only will be possible with multiscrypt character encodings, as proposed by Unicode. Unfortunately, the Unicode solution was not yet available at the start of this project (See also Paulussen: 2001).

After the evaluation of different database management systems (DBMS) the choice fell on Fourth Dimension (or 4D) of ACI (France). This DBMS can be operated both on a Macintosh and on a PC platform, be it that because of the differences in Arabic ASCII codes used on a PC and a Macintosh a conversion is necessary to make both systems compatible.

Besides, we discovered that even this DBMS was not completely ready to be used for the treatment of Arabic, because it did not take into account all the specifications of the Arabic diacritical signs. No difference was made, for instance, between the different Arabic vowels, which is, of course, a problem for frequency counts. In the early version of the DBMS, the search engine did not make any differentiation between certain Arabic consonants, which poses not only a problem for the searches but for frequency counts as well. The same problem occurred on the Arabic Macintosh system level. Two Arabic consonants (dāl and khā) seemed to have the same ASCII codes which provoked problems in sorting tables and in looking up words. In order to make the DBMS completely

adapted to the Arabic language we cooperated with the developers of the 4D of ACI in Paris. Both the problem of the system software and the problem of the precise recognition of Arabic characters by 4D have been solved.

### **2.1.2 Structure of the database**

In order to make all the data reversible in the most economical way possible, a structure of four units was designed. Every unit contains all the information appropriate for that unit. There were two units for the language specific information, one for MSA and the other for Dutch. The Dutch unit can be adapted for any other Latin language. The two other units are transfer units used as a link between the two languages analysed.

#### **2.1.2.1 The language specific FL unit**

The Dutch unit contains the following elements: The lemma such as it occurs in the dictionary, the gender of the word, the feminine form of the word, information about the infinitive and conjugation of verbs, fixed prepositions, plural forms of the word, degrees of comparison, grammatical categories which serves not only as a classification of the word but also for the ordering of the word in the dictionary and the region where this word is most typically used (e.g. different words used in Belgium vs. the Netherlands). The input of this kind of fixed data was only required once.

#### **2.1.2.2 The language specific Arabic unit**

The Arabic unit contains the following elements: The fully vocalized form of the word and, in a separate field, its orthographic variations. Because of corpus analysis purposes we also encoded every Arabic word in order to obtain a maximum degree of disambiguation between words. This encoding of Arabic words was carried out for the greater part automatically. The encoding of Arabic words allows making more intelligent searches in a text, thus avoiding garbage (Van Mol: 2000a). Further on, the root of the word was also given. In Arabic dictionaries, words are normally ordered by root. Only foreign words, coming from languages such as Persian, Turkish, but also English, Italian and French, are ordered alphabetically. This means that the traditional order of Arabic dictionaries is mixed. There is an alphabetical order of the word roots and the foreign words. However, all the Arabic words derived from a root are ordered within the root by grammatical category. That is why an option was provided for every word to be classified by root or by alphabetical order.

By using a boolean field to indicate whether the new found word did already appear in the reference dictionary of Hans Wehr, it was possible to make an inventory of all 'new' words, in the sense that they were not yet included in the existing dictionaries. This type of marking showed that approximately 5% of the words in our dictionary do not occur in the current dictionary of Hans Wehr. As our dictionary is based on a corpus of spoken language, a similar type of field was used for labeling dialectal words.

Further on the file Arabic word contains for a verb also its derivational forms, which serve also as a means of classification. A field was also provided for information about irregular conjugations of the verb in present and past tenses. In the case of a noun a choice can be made between the external Arabic plural and the internal Arabic plural. These fields were also encoded, because plural forms occur as such in a text. Information about irregular dual forms was also added. Other information includes: word gender and possible irregular feminine forms; prepositions occurring in collocations with certain nouns or verbs; irregular accusative and genitive forms. An extra word frequency field is provided, and software for specific frequency counts based on Arabic corpora is under development.

#### **2.1.2.3 Word order**

An innovative feature introduced in this project, is the possibility to order words within a root automatically by grammatical category. In order to do so the Arabic grammatical categories had to be elaborated. This approach was based on the structure of the Arabic f'l patterns. In Arabic, words can

be classified according to different traditional Arabic categories (or Western categories). In order to describe the grammatical category of words, they are in a way translated in a combination pattern of consonants and vowels. This pattern is in most cases used to define verb forms, verbal nouns or participles by making use of the three consonants fā', 'ayn and lām. By expanding this 'translation system' to all the Arabic nouns we were able to give every word its precise identification to be classified within the root. 200 different patterns were formed this way. Every Arabic entry in the dictionary was allocated by its convenient pattern. In giving these patterns a numerical identification, it became now possible to order all the words depending on the same root in the same order. Confronting our word order with the order of Hans Wehr revealed that the manual ordering by Hans Wehr contains several inconsistencies. Because of the precise definition of words, the ordering by our system is consistent. This system gives the database the possibility to produce an Arabic-FL dictionary in two ways, or by root order, or by alphabetical order.

The question which of those two classification orders is preferable for an Arabic dictionary is not so easy to answer. Both orders have advantages. It may seem that it is easier for a learner to look for words in a dictionary ordered alphabetically. On the other hand, when a dictionary is more concise the classification by root is more useful because it gives the possibility to the user to find also meanings of words that are not in the dictionary, because he can derive the meaning of these words by looking at the root or at related words. Besides, Arabic is a very mathematically structured language which structure has to be mastered for reading as well. Our database can be used both for the production of a dictionary in alphabetical order or a dictionary with root and grammatical classification. It is up to the editor to decide which structure fits most the dictionary users he intends.

#### **2.1.2.4 Consequences for dictionary production**

As previously mentioned, the two first units contain language specific information about words. All the information related to one of the two languages was stored in one of those two units. It is clear that in most dictionaries all this information ought to be given with the entry, e.g. with the source language. It is, however, not evident yet to give such information also for the target language. Indeed, in most quality dictionaries, you can find a lot of information about the word of the source language. Information about the words of the target language, however, is in most cases very limited. Our structure allows the editor to choose automatically which information he wants to add with every word in the target language. This choice is dependent on the specific purpose for which dictionary is to be produced. If we design a dictionary for receptive use only, fewer elements will be given with the words of the target language, but when a dictionary for productive use is to be published, more detailed information can be given. In the printed version of our dictionary Dutch-MSA, for instance, we added to every Arabic verb, the grammatical verb form, the vowel of the present tense, deviations in conjugation and fixed prepositions. For the nouns, we also added the plural forms (broken and external forms), but also fixed prepositions, exceptions in case, etc. The advantage of the database is that the editor himself can define which information is to be included into the dictionary and which information is not.

#### **2.1.2.5 The Arabic-FL translation unit**

The other two units formed the Arabic-Dutch structure and the Dutch - Arabic structure. As our mother tongue is Dutch, we preferred to start with the Arabic-Dutch part basing our translations on a corpus of approximately 3,000,000 words. Every word was translated in detail in context and filled in the Arabic-Dutch structure. Both the structures of the Dutch-Arabic unit and the Arabic-Dutch unit consist of four levels.

The first level contains a list of all the words that had already been inserted into the database, with their root, grammatical category and the most prominent meanings in Dutch. A click on a row shows the second level on the screen. This level is meant to regroup and to classify homomorph words of different grammatical categories. In this level the link is made with the first unit, which consists of the files containing language specific information. A click on the row of the word shows

the third level on the screen, which has basically a simple structure. On the one hand, there is the Arabic word linked to the first Arabic specific unit. On this level, translations in the FL can be added. When those translations are words in the FL, the link is made immediately with the second unit containing all the language specific information of the FL.

A new feature, which is not yet to be found in existing dictionaries, is the introduction of a pointer in the source language. For productive use, it is necessary to guide the user through the different possible meanings of words. Therefore, translated words were grouped according to their main common meaning. Such a common meaning was described in the source language in order to guide the user through the different possible meanings of words. Every group of words with a common meaning is linked with another field in which idiomatic expressions, multi-word items, collocations and sample sentences can be added.

The fourth and final level serves for the input of sample sentences and expressions, but contains also fields in which additional information can be stored. Some of this information is well structured, such as, for instance, the relation of certain words with complements or subjects. Other information is more globally structured, such as cultural information, pragmatics and additional descriptions that clarify the content of strange words. Also at this level, the editor can make a choice as for which elements he wants to include in a paper dictionary.

#### **2.1.2.6 The FL-Arabic translation unit**

The fourth unit is the Dutch-Arabic entity. As was mentioned before, first the Arabic-Dutch dictionary was compiled. After the translation in context of millions of words, it was decided to reverse the database into its Dutch-Arabic structure, which is a twin structure that is fairly similar to the structure described above. Before the conversion, however, it was important to indicate which meanings had to be converted. Verbal nouns, for instance, in Arabic and also a certain category of participles cannot be used as an entry in a Dutch dictionary. In this way, the conversion procedure was a crucial one, since it was at that moment that we had to decide which elements to be converted and included in the Dutch entry list and which not. Whereas the compilation of the Arabic-Dutch dictionary took several years, the automatic conversion into a Dutch-Arabic dictionary took about 72 hours.

Although the conversion procedure was an enormous gain of time in compiling the reversed part of the dictionary, it does not imply that the dictionary was ready for printing. The created conversion tables had still to be carefully checked with existing frequency lists of the Dutch language. The comparison of our macro structure with Dutch frequency lists revealed that there were many frequent words which were not in our database, in spite of the fact that the dictionary was based on a very large corpus.

This reason of this discrepancy resides in the fact that in Dutch and Arabic the world is perceived in a different way. This means, for example, that in the Dutch speaking areas words or concepts occur which do not exist at all in Arab countries. For those words, a suitable translation had to be provided. Very often it was not at all a simple task. It is obvious that the correct Arabic equivalent of a part of these words might be found by expanding the Arabic corpus. However, it is also known that after a certain size, the corpus has to be expanded extensively in order to find new meanings. On the other hand, there are also words for which no equivalent will be found, because the reality of both societies is perceived differently or simply differs. In those cases, we opted for a description of such words, which was printed in a different font.

As soon as the dictionary was finished, it still had to be printed. From the database, an export file was produced containing the necessary markup required by the desktop publishing program Pagemaker. The whole database could be entered into Pagemaker after which detailed editing could start. The main element which hampered the printing of the database was the fact that vowels and consonants were not sufficiently separated from each other. In those instances where there was an overlap in printing of vowels and consonants, a manual adaptation was necessary.

### 3 Perspectives

At this moment, a completely tagged corpus of spoken Modern Standard Arabic of 1,000,000 words is available. We intend to extend this corpus with written language as well, of which so far a corpus of 800,000 words has been tagged. In the near future, we intend also to develop tools to explore corpora for lexicographic purposes. The tools we have developed so far are not yet integrated in a database. The integration of those tools in a database will facilitate the compilation and updating of Arabic dictionaries in a considerable way.

### 4 Conclusion

We were able to develop a relational database, completely adapted for the treatment of Arabic, which makes it possible to print dictionaries on size. Because the database contains all relevant language specific information for Arabic, and because it had been adapted completely to be used with Arabic language, it can be used by other lexicographers for the production of other dictionaries that contain a combination of Arabic script and Latin script languages.

### 5 References

- Abdel-Nour, Jabbour (1983). *Dictionnaire Arabe-Français*, Beiroet, 1126 p.
- Abdel-Nour, Jabbour (1995). *Dictionnaire détaillé Français-Arabe*, Beiroet, 1110 p.
- Al-'Adnānī, Muḥammad (1989). *Mu ḡam al-'aglāṭ al-luḡawīya al-mu 'āsira*, Beiroet, 870 p.  
(added title: *A Dictionary of Common Mistakes in Modern Written Arabic*).
- Al Kasimi, A.M. (1977). *Linguistics and bilingual dictionaries*, E.J. Brill.
- Al-Munjid (1972). *Français-Arabe*, Beyrouth, 980 p.
- Fromm, W.D. (1982). *Häufigkeitwörterbuch der modernen arabischen Zeitungssprache: ein Mindestwortschatz*, Leipzig, 351 p.
- Kouloughli, Djamel-Eddine (1991). *Lexique fondamental de l'arabe standard moderne/Basic lexicon of modern standard arabic*, Paris, 287 p.
- Paulussen Hans (2001), "Character Encoding Standards: a Matter of Content and Form", in Temmerman, Rita & Madeline Lutjeharms (eds.), *Proceedings of the International Colloquium: Trends in Special Language & Language Technology*, Brussels, 29-30 March 2001, 105-117.
- Stetkevych, Jaroslav (1970). *The modern Arabic literary language, lexical and stylistic developments*, Chicago-London, UCP, 135 p.
- Thiry, Jacques (1985). *Arabe Moderne, Arabe classique. Rapport d'Activités de l'Institut de Phonétique*, 20, Université Libre de Bruxelles, pp. 95-126.
- Van Mol, Mark (2000a). "Exploring annotated Arabic corpora, preliminary results", In *Corpora and Natural Language Processing, proceedings of the International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications*, Monastir, pp. 94-98
- Van Mol, Mark (2000b). "The development of a new learner's dictionary for modern standard Arabic, the corpus linguistic approach", In *Proceedings of the Ninth EURALEX international Congress*, Stuttgart, 8-12 august 2000, pp. 831-836
- Van Mol, Mark & Berghman, Koen (in press). *Leerwoordenboek Modern Arabisch-Nederlands*, De Nederlandse Taalunie, Bulaaq, 520 p.
- Van Mol, Mark & Berghman, Koen (in press). *Leerwoordenboek Nederlands-Modern Arabisch*, De Nederlandse Taalunie, Bulaaq, 530 p.
- Wehr, Hans (1979). *A Dictionary of Modern Written Arabic*, Ed. J. Milton Cowan, Wiesbaden, xvii, 1301 pp.