From paper dictionary to an elaborate electronic lexicographical database

Mark Van Mol

Keywords: *Arabic database driven lexicography*, *Arabic tagset development*, *online dictionaries*.

Abstract

At the 2000 Euralex conference we presented a paper on the development of a new learner's dictionary for Modern Standard Arabic, based on a corpus linguistic approach. In 2001 this dictionary was published in two volumes: a Dutch-Arabic volume and an Arabic-Dutch one. After the publication of the two dictionaries, we started new projects to work on both the existing corpus on which the dictionary was based (at that time 3,000,000 words) and the internal extension of the lexicographical database. We did not limit ourselves to additional lexical information and expressions, but included very detailed grammatical information. In recent years, the evolution of language technology has led to increased possibilities for lexicographical exploration of databases, especially in Arabic. In this paper we present the elements that we added to the contents of the lexicographical database: new words and expressions, 646 detailed POS tags, the technological changes it underwent (for example, the transformation from 4th Dimension (4D) to Mysql). This resulted this year in the development of the first full online consultable Arabic-Dutch/Dutch-Arabic dictionary. This Arabic dictionary is the first of its kind, not limiting itself to mere lexical information, but allowing a much greater variety of searches for all kinds of grammatical information. In this paper we offer an overview of some of the possible searches. One of the next challenges is the production of an online dictionary with a clear layout in order not to be forced to skip much of its detail and accuracy.

1. Introduction

After the publication of the Arabic-Dutch and Dutch-Arabic learner's dictionaries (Van Mol: 2001), we aimed to produce a database that would integrate lexical with corpus information. In order to make combined searches possible in both the corpus and the lexicographical database, we had to solve the problem of the agglutinary character of Arabic which hampers accurate searches. For that reason, a specific tagging system needed to be developed in order to match all the lexical elements from the corpus with those available in the database. For that purpose, a two-step tagging system was developed: a primary and a secondary tagging system. The primary tagging system was developed by using Arabic diacritical signs according to certain conventions that guarantee the total disambiguation of Arabic language. Accordingly only seven diacritical signs were used, and the Arabic corpus transcribers taught to apply them according to conventions settled in advance (Van Mol: 2003).

2. The development of the tagsets[1]

To illustrate the contrast between the ease of application of the primary tags and the complexity of the definitive tags linked to the preliminary tagset, we give two examples in transcription. In Arabic, the past tense of a verb in the first person singular is expressed by adding the suffix *tu*. For the verb *kataba* (to write), for example, this becomes *katabtu*. Because Arabic is usually written without diacritical signs, *ktbt* is written. This constellation of consonants, however, is not unambiguous. It can indicate four different singular persons, namely the first person, the second masculine and feminine or the third feminine. The convention for the primary tag is that only the last vowel is added, because this vowel precisely marks the distinction between the different persons. Completely vocalized the verb *katabtu* means I wrote, *katabta* you (masculine) wrote, *katabti* you (feminine) wrote and

katabat she (feminine) wrote. The transcribers only use the last diacritical sign for verbs in the past. Therefore, they write these words as follows: ktbt*u*, ktbt*a*, and ktbt*i* and ktbt°. The °, or so-called *sukun*, is a diacritical sign that is used when no vowel is placed. By means of these easily applied primary tags, we can yield complex tags when they are matched to the words in the database.

The second example is the plural form in a construct state of the word *mas'uul* (responsible or responsible persons) which in Arabic is written as *ms'uuluu*. The convention in diacritical signs to make a distinction between a noun and an adjective is as follows: An adjective is never voweled, in other words, no diacritical signs are applied to adjectives. To differentiate between nouns that are homonymous with adjectives, we vowelize the first consonant of the noun. This means that in the database the word *mas'uuluu*, which is ambiguous because both the noun as well as the adjective are written the same way, is filed as follows: the noun with a vowel that is *mas'uuluu*, and the adjective as *ms'uuluu*.

3. The elaboration of the secondary tags in the database

For the secondary tags, a special algorithm was developed that generated all derived forms of the primary tagged word forms. This approach differs from common practice whereby word forms are derived from a fully vocalized form. In this case, the forms are derived from a partially vocalized word form.

For 27,393 Arabic words, 594,941 tagged word forms were generated. However we did not, for example, write an algorithm merely to provide the conjugations of the verbs. In order to match the primary tagged words in the corpus with the words in the lexical database the word forms were generated and tagged following the same conventions as the tag conventions for the corpus. This meant that the above-mentioned persons of the verb *kataba* were generated in the database in an identical way, namely ktbt*u*, ktbt*a*, ktbt*i* and ktbt°.

As far as the secondary grammatical tags are concerned, a preliminary remark should be made. Grammatical categories, of course, are closely linked to the language itself. Arabic – in contrast to many other languages – has a long and rich grammatical tradition that goes back to more than a millennium. For that reason, we decided from the outset to store in the database both the Arabic grammatical categories and the Latin or European ones. Thus three kinds of POS data are stored in the database. There are 360 pure Arabic POS tags. Many of these do not correspond to the Latin POS tags. Arabic grammar makes a distinction between three word categories, *verbs, nouns* and *particles*.

From this basic division others are derived that do not correspond to any of the Latin grammatical categories. A distinction is made, for example, between 15 verb forms and 12 kinds of verbal nouns. This typical Arabic information can be searched for in the database. Arabic grammar also uses so-called *fi'l* patterns to describe word forms. All these patterns are attached to every Arabic word. This gives us the possibility of searches in the database based on Arabic categories. For example, the *maf'al* and *maf'il* patterns indicate nouns of places and the *mif'al* pattern nouns of implements. When we search, for example, for all the *maf'al* and *maf'il* form words and all the *mif'al* forms, we obtain the following results: *maf'al* 161 occurrences, *maf'il* thirty-six occurrences and *mif'al* ninenty-three occurrences.

This additional information - lacking in Arabic databases - which we only give as an example sheds new light on the Arabic language and is especially important for educational purposes. Since Arabic is written without vowels this information could teach students that the most probable vocalization for a place name is *maf'al* and for a utensil is *mif'al*. However, additional corpus investigations are necessary because frequent words, such as mosque - *masjid* - and house – *manzil* - follow the *maf'il* pattern.

The other tags we applied to every word and word form in the database were Latin. These are more limited in number, at fifty-six. These are the classical Latin POS tags extended with more detailed information, such as, *verb*, *noun*, or *adjective*. Finally, there are the combined Arabic-Latin tags which in contrast to the above-mentioned tags are not limited to words, but which encompass all word forms. These are the so-called secondary or definitive tags. In order to compose these tags in a well-organized way we opted for a maximum of nine detailed elements per tag. In the beginning we used numbers to tag all the different words. We presented them as a sequence of separate numbers. The first number always referred to the main POS tag of which there are eleven elements: 1. Noun; 2. Adjective; 3. Adverb; 4. Pronoun; 5. Number; 6. Verb; 7. Particle; 8. Interjection; 9. Conjunction; 10. Preposition; and 11. Article.

When we take, for example, the combination: 1.1.1.3.3.3.0.1.0., which is the combination for the plural of a noun such as, for example, *mas'uuluu*: the first number (one) indicates that the word is a noun. The numbers that follow give detailed information about the first element. This information, of course, depends on the first tag. When the first tag is number 1 (a noun) it is clear that there will not follow any information about conjugation (such as: first person) which will be the case when we deal with number 6 (a verb). The tags that follow are a combination of nine elements: After the first, which is the word class, follow: 2. Kind of noun; 3. Gender; 4. Grammatical subcategory; 5. Number; 6. Case; 7. Semantic information; 8. Language level and 9. Specific Arabic morphological elements. The contents of the detail information depend on the first element of the tag. The following table shows how the different detailed tags are compiled for nouns.

Table 1. Detailed combined tags for nouns.

Tag Nr	**First**	**Second**	**Third**	**Fourth**	**Fifth**	**Sixth**	**Seventh**	**Eighth**	**Ninth**
	Word Class	Kind of noun	Gender	Grammatical subcategory	Number	Case	Semantic information	Language Level	Specific Arabic morphology
1	**NOUN**	**COMMON**	**MASCULINE**	**GENERAL**	**SINGULAR**	*NOMINATIVE*	**CHEMICAL**	**STANDARD**	*FINAL HAMZA U*
2		**PROPER**	**FEMININE**	*MASDAR*	*DUAL*	*ACCUSATIVE-GENITIVE*	**COUNTRY**	**DIALECT**	*FINAL HAMZA I*
3		**FOREIGN**		**PARTICIPLE-GENERAL**	*SOUND-MASCULINE*	*NOMINATIVE CONSTRUCT STATE*	**CITY**		*MASCULINE TAA MARBUTA*
4				**PARTICIPLE-WEAK**	*PLURAL-BROKEN*	*ACCUSATIVE-GENITIVE CONSTRUCT STATE*	**MONTH**		*FEMININE NO TAA MARBUTA*
5				*ELATIVE*	*COLLECTIVE*	*ACCUSATIVE*			*ALIF DROPPED*
6				*DIMINUTIVE*	*SOUND-FEMININE*	*WEAK-INDEFINITE*			*SUFFIX TUMU*
7				*NISBAT*	*DOUBLE*	*WEAK-DEFINITE*			*FINAL ALIF MAQSURA - ALIF*
8				*ALIF-MAQSURA*	*PLURAL-WEAK*				*FOREIGN TAA MARBUTA*
9				*AL KHAMSA*	**PLURAL**				*DIALECTAL CONJUGATION STANDARD*
10					*PLURAL-MARBOUTA*				*DIALECTAL CONJUGATION DIALECT*
11					*MASCULINE-AAT*				

Tag number 1.1.1.3.3.3.0.1.0. for the word *mas'uuluu* (responsible people for) contains the following information: NOUN – COMMON – MASCULINE – PARTICIPLE GENERAL – SOUND MASCULINE – NOMINATIVE CONSTRUCT STATE – 0 (= no information) – STANDARD LANGUAGE – 0. As one can see in the table, the tags are mixed POS tags. Latin POS elements are used in combination with Arabic grammatical information. The further one moves to the right in the table the more specific Arabic information is found. We show the Arabic tag elements in cursive script whereas the other tag elements are in bold. Note that some elements are dual. The element singular is an item that occurs in both European and Arabic grammar.

Thus 646 tags are available. Because they all have their abbreviations, the word forms to which these tags apply can be searched for as a whole. For the above-mentioned tag the abbreviation is as follows: NOU-COM-MAS-PLG-SOM-NCS-0-STA. Since the tags are stored in separate fields, searches can also be done on separate elements of the tags. Someone might, for instance, want to search for all the participles, or the weak Arabic participles.

Whereas for the first person singular of the verb *kataba* we only had to add the vowel *u* to the *ktbt* cluster of consonants, we now obtain for the same person as secondary tag: 6.0.0.1.7.1.6.7.1. which stands for VERB-0-NEUTRAL-ACTIVE-SINGULAR-PERFECT-FIRST PERSON-STANDARD-0. The combination of primary and secondary tags makes the tagging of Arabic corpora much easier and within the reach of everybody. Researchers who want to tag available raw Arabic texts do not have to add all kinds of detailed and complex information after every word. The only thing they have to do is to apply the diacritical signs according to the convention. Every educated Arabic speaker is able to apply vocalization accurately, even when based on a convention, but adding complex tag sets consciously by reflecting on every separate word not only demands much more time, but also leads to many mistakes and misinterpretations.[2]

4. The extension of the corpus: a brief description[3]

Because of the relative simplicity of the primary tagset, we were able to compose a large primary tagged corpus of more than 12,000,000 words. Given that the selective addition of diacritical signs by a native speaker is done almost spontaneously, this work could be done in an economical way because the transcriber did not have to waste time on adding complex tagging information. Nevertheless, it still took eleven person-years to arrive at the current size of the corpus.

5. The technological changes

The initial database from which the twin dictionaries were printed was 4th Dimension (4D). We chose this database because it was the most suitable for Arabic language at that time. We also cooperated with the developers of 4D in Paris to adjust the database throughout for Arabic. In the meantime new possibilities have arisen. One of those was the mysql database format. This database format is quite interesting because it can be used in combination with php, which makes web applications possible. The main question was whether we could transform the data in 4D (more than 30 years of lexicographical work) to a mysql format without any loss of data. The data in 4D were in Macintosh ASCII format. When transferring the data they had to be converted to Unicode utf-8. To export the database we used three formats: pipe, tab and xml. Finally we managed to export all the data and transform them into a Unicode utf-8 format without a loss of data. The only problem we had to solve was the

spaces. In Macintosh ASCII, there is a different encoding for Latin and Arabic spaces. Using Latin spaces between Arabic words means that the order of these Arabic words will be reversed. This was an important issue for the expressions in the database because they always contain spaces. When Latin spaces were replaced by Arabic spaces, the problem was solved.

The transformation from a 4D platform to a Mysql platform opens many possibilities. In the first place, the researchers can make queries themselves and do not have to depend on specialized programmers any more as was the case for the 4D platform. In addition, building the database further does not need a specialist programmer. This is how we were able to develop the definitive tags without the aid of specialized programmers.

6. Conclusion and further prospects

So far, we have managed to develop an experimental online version of the printed dictionaries, making use of a combination of mysql, php, html and css elements to build the web page. For the presentation of the information in the dictionary we currently make use of tables. These, however, cannot show all the details available in the dictionary. This is why we are currently working on an interface that will show all the information available in the dictionary. Another step to take is the conversion of the corpus, which is also in Macintosh ASCII format. The first tests have revealed that the transformation of these data from Mac ASCII to utf-8 is more complex. All consonants and vowels are correctly transformed, but the punctuation marks, a quite essential element in texts, are all reproduced the same way. A conversion program by my colleague Hans Paulussen (KULAK) to match these demands has already been developed, but needs further testing.

Notes

[1] For a recent survey on the automatic tagging of words see: Sawalha, M., & Atwell, E., *Linguistically Informed and Corpus Informed Morphological Analysis of Arabic*, In: Proceedings of the 5th Corpus Linguistics Conference. CL2009, University of Liverpool, UK. Lancaster University Centre for Computer Corpus Research on Language, University of Liverpool.: http://eprints.whiterose.ac.uk/42634/.

[2] See, for instance, the work of Khoja who developed the first tagset for Arabic. He writes the following: Although corpora are widely available for English (some for free), there is very little available for the Arabic language. Also, although some of these corpora are marked-up with XML or SGML tags, none of them are POS tagged. I have manually tagged Arabic newspaper text that I can provide freely for research purposes. I have two corpora: 50,000 words of tagged newspaper text. The words are as being either definite or indefinite noun, verb, particle, punctuation or number and: 1,700 words of tagged newspaper text. The words are tagged with more detailed tags using gender, number and so on. Details of the tagset used can be found in this paper.
http://zeus.cs.pacificu.edu/shereen/research.htm,

[3] See for a survey on other corpora:
http://www.comp.leeds.ac.uk/eric/latifa/arabic_corpora.htm

References

A. Dictionaries

Van Mol, M., and K. Berghman 2001. *Leerwoordenboek Nederlands – Arabisch* (Learners' Dictionary Modern Arabic – Dutch). Amsterdam: Bulaaq.

Van Mol, M., and K. Berghman 2001. *Leerwoordenboek Arabisch – Nederlands* (Learners' Dictionary Dutch - Modern Arabic). Amsterdam: Bulaaq.

B. Other literature

Van Mol, M. 2003. *Variation in Modern Standard Arabic in Radio News Broadcasts, A Synchronic Descriptive Investigation in the use of complementary Particles*. Leuven: Peeters.