

Arabic language and vocabulary acquisition

Mark Van Mol
Catholic University Leuven

In order to develop a strategy for the teaching of Modern Standard Arabic (hereafter MSA) in the coming millennium, and to design an efficient curriculum for the future, we have to obtain a thorough insight into the nature and the current use of MSA; secondly we have to look for new methods of teaching. In order to fully grasp the current needs, I would like to refer to the study programme at our university a few decades ago. At that time, students were only trained to decipher classical Arabic texts, and a limited number of modern texts. The emphasis of the training was on grammatical skills. What was asked from the student of Arabic, was the ability to translate written material using a dictionary. This training was, indeed, very useful in the beginning and up to the middle of this century, because the most important sources of Arabic materials at that time were written texts. Not much attention was paid to the thorough assimilation of a sizeable vocabulary. The students just devoted their time to translating texts. On graduation, the lucky few might retire to their study to spend a few years translating a manuscript in order to defend a doctoral thesis, for instance.

There were no Arabic broadcasts that could be received on the radio in Europe, or in America. There was no internet, no television, no fax, no telex, no satellites and even the telephone connection was very limited, certainly between the Arabic world and the west. When people chose to go to the Arabic world, they had to choose a specific region. Indeed, also the means of transport were limited, so that almost nobody would stay in the mašreq for some time and afterwards in the mağreb. When people wanted to communicate in the Arabic world, they had to learn the specific dialect of the region. And some people, indeed, obtained a very high proficiency level in a specific dialect, as e.g. the well-known Lane.

However, since then, times have changed dramatically. In Europe today, we receive more than twenty Arabic television satellite stations and even more radio stations. Every day newspapers and magazines from all over the Arabic world are available. In fact in the Arabic world there is an explosion of information. Besides, MSA is used more and more in many more situations. We also find a much higher degree of schooling in the Arabic world, which leads to an increased intelligentia. Consequently more and more of these people, are, and will be able to approach MSA in their formal oral communication. In my Arabic corpus I observed that, depending on the country, MSA is used by a larger number of people in interviews. This means that nowadays the auditive and even the verbal component of the language training is becoming more and more important. The old Arabic study curriculum turned out people that were well-trained to decipher Arabic texts, but most of them were by no means capable of conversing in Arabic, nor were they able to understand Arabs.

The changing situation, not only of the Arabic language as such, but also the impact of technology, demands another approach to the teaching of Arabic. Not only do we have more

direct contact with the Arabic world by means of television, radio and press in general, we also have more opportunities to be physically present in the Arabic world. This means that the goal of language training can no longer be restricted to passive written knowledge. Teaching should focus on real participation in the Arabic world. The goal of teaching Arabic, in my opinion, ought to be to train students to become independent operators able to participate in the Arabic world. Participation in full means the ability to write, read, listen and talk at near-native level. If a graduate student needs a whole day to decipher and understand the first page of a newspaper, then participation in the Arabic world is handicapped. A graduate student who can digest all the important items of a newspaper between half an hour and one hour will be able to participate in the Arabic world. The same goes for radio and television programmes in MSA. In my opinion, a trained arabist ought to be capable to digest, auditively as well as in written form, the flood of information from the Arabic world, in order to define what is really important and then, analyse in a more thorough way certain aspects or developments of the Arabic society.

Concerning MSA and the study of this language nowadays, one of the crucial points we have to make here, is the distinction between two kinds of sources. There are urgent and non-urgent sources. The urgent ones ought to be consulted in no time while others can wait. Newspapers and magazines ought to be consulted in the same way as Arabs do. This means that the only and most effective way is to read and understand these sources without a dictionary. Moreover, if we want our students to understand radio and television programmes, then they will have to work without a dictionary. Other sources, like novels and literary works, and certainly classical texts, can be consulted in a more tranquil way.

Pragmatic vocabulary acquisition strategies are essential for the students to identify which words they need to know in order to read or understand the most urgent sources in the most productive way near native manner. In the European community, and also in the United States, language proficiency is divided into different levels expressed in numbers of hours. These are:

A1	Breakthrough Basic User (number of hours 80-100)
A2	Waystage Basic User (number of hours 180-200)
B1	Threshold Independent User (number of hours 375-400)
B2	Vantage Independent User (number of hours 500-600)
C1	Effective Proficiency Proficient User (number of hours 1000)
C2	Mastery Proficient User (number of hours 1500)

At these different levels nothing is said about the amount of vocabulary that ought to be assimilated at each level. From experience, I would make a distinction between three levels, without referring to the number of hours needed, but to the amount of vocabulary that ought to be assimilated. Our investigation has pointed out that a student does not have to learn by heart the complete vocabulary of the dictionary of Hans Wehr, about 45,000 words, in order to read and understand current MSA fluently. Nevertheless, there is a certain minimum. The absolute minimum level of words to be assimilated is situated between 8,500 and 10,000 words for newspapers and magazines and radio and television programmes. The proficiency level for novels and literature is situated between 12,000 and 14,000 words. At that level a student is able to understand all sources in MSA automatically without a dictionary.

Our findings are based on the study of a geographically distributed computer corpus of written and spoken Arabic of 3,000,000 words. One quarter of the corpus consists of oral material (such as transcribed radio and television programmes from all over the Arabic world), and three quarters of the corpus consists of written texts, in the first place newspapers, magazines (including written texts from the internet), but also school manuals of the Arabic language. No text in the corpus is older than 20 years. On the basis of this corpus, I compiled a learner's Arabic-Dutch, Dutch-Arabic dictionary of the 15,000 most frequent Arabic words. Samples point out that this vocabulary in most cases covers 99 % of the vocabulary used in current MSA material.

Nevertheless, I would like to make a distinction between three levels.

Level A	preparatory level	3,000 words
Level B	oral and written media-level	8,500 - 10,000 words
Level C	novels and literature level	12,000 - 14,000 words

In order to participate e.g. as a diplomat or an observer in the Arabic world Level B seems to me obligatory. The main aim of the preparatory level is to teach students grammatical skills, but also to help them use the basic vocabulary in all kinds of circumstances. This training is focused on the productive use of the language. The way in which this vocabulary has to be assimilated at the preparatory level, and also in general needs consideration. As a preliminary remark, I'd like to mention that the degree of transparency between the Arabic language and most European languages is almost zero. The assimilation of the vocabulary is therefore much more complicated than e.g. with familiar languages. For example, it is generally known that more than 50 % of the English vocabulary is derived from French. Lots of words and constructions are transparent. Take, for example, the English word *politics* that corresponds to the French *politique* compared to the completely strange "siyāsa", or the English word *economy* that corresponds to the transparent French word *économie* compared to the Arabic "iqtisād". The lack of transparency is one of the main features that makes the study of the Arabic language more tiresome. Moreover, the rare transparent words are often not recognized by the majority of students. For example many students are not able to identify the words "bīdāgūjīya" (*pedagogy*) and "hībūfisīya" (*hypophysis*) even in a clear context. For languages such as Arabic we cannot permit ourselves in training to teach superfluous words. The volume of vocabulary has to be chosen effectively.

The first 3,000 words are those found in textbooks. These words ought to be integrated in a language manual and trained in a large number of exercises where the students learn to use the words in correct grammatical sentences, written and orally. In our university these words are trained in a gradual, repetitive and progressive way. However, assimilation of these words and even a proficient ability to use them actively, does not yet make the student a real participant in the Arabic world.

This can only be obtained via the second level. At this level, the student has to be trained through reading comprehension and also listening comprehension with exposure to a large number of appropriate texts in which the next 8,500 or 10,000 words occur. The question is, in how far it is possible to teach this number of words to students, and how much time it would take. Moreover, what are the possibilities or techniques to reduce the study time; in other words, to make that

study time more efficient. In my opinion, one of the basic stimuli and tools for learning a language is repetition. A child, when learning its mother tongue, is constantly confronted with the same words in the same and varied contexts. One of the major features of the teaching of vocabulary ought to be the constant repetition of vocabulary.

On the preparatory level, this kind of drill is feasible, albeit also a complex matter. On the higher level, this constant repetition is much more difficult to realise. When students study Arabic texts, they are confronted with words that are not vocalised. The larger part of their study time consists in looking up words in the dictionary. We calculated that, after two years of study, an average student needs about one and a half minutes to look up a word in a dictionary, which means that five hours are needed to translate and comprehend a text of ca. 200 words. The question is, whether we can limit this process of dictionary consultation, so that the student can spend time on the real study and assimilation of the vocabulary. Every word that the student assimilates, that is, every word of which the correct meaning is known by heart, will mean valuable time saved. Moreover, when listening to radio or TV, only the words assimilated, can be recognised and correctly interpreted, every unknown word is not retained. Moreover it is not simple to define the appropriateness and the level of an oral or written text in a study curriculum.

In order to teach the required amount of words effectively to our students at the Catholic University of Leuven (Belgium), we have developed an encoding system for Arabic words that successfully eliminates the ambiguity of many Arabic words. 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. We have 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 "fahima" from the conjunction "fa" followed by the verb "hamma", and also to distinguish the conjunction "fa" from the personal suffix "hum".

This encoding system opens the way to make more efficient frequency lists with computer programs and it will also give us the possibility to define the level of difficulty of different kinds of texts. Linked to the electronic dictionary we have developed, this encoding system will also make it possible for students to search for the meaning of words in a text without using a printed dictionary. A simple click on the mouse will give them the right meaning, which save a significant amount of time in the learning process. Students will no longer lose precious time by looking up the meaning of a word in a dictionary, and will have more time to study and assimilate the meanings of the words in context. With our system, the student will find the correct meaning of the word in two seconds. As I have already mentioned, we calculated that a student needs about five hours to translate 200 unknown words in context. With our system it will only take 10 minutes. This means a time saving of 96 %. In other words, the student has more than four hours' time to study the new words, but can read more texts within that time. They will have the opportunity to read new texts, and can constantly check their knowledge by clicking on the words in the text. Thus a cycle of repetition is inbuilt.

However, in the whole training, a distinction must be made between oral and written sources. The written sources demand another approach than the oral ones. The translator can always translate a text at their own pace. This is not the case for oral texts. The listener has to understand the words and sentences on the level of the speaker. This implies three problems: speed, dialectal interference and ungrammatical sentences. We calculate that in news broadcasts, for example, Arabic speakers produce between 130 and 180 words per minute. This means that a student has to be able to identify a vocabulary of between 1,300 and 1,800 words in ten minutes' time. Moreover, as previously mentioned, it is impossible to identify this amount of words in ten minutes, when the words have not yet been assimilated. Only words that are known by the student will be recognised. Another problem is that in oral texts words form clusters, which are sometimes only identifiable as clusters, but no longer as words. While transcribing Arabic texts, the non native beginner finds it impossible to identify words at the first listening, and even at the second and the third listening, because they are sandwiched between two other words. Finally sentences in oral Arabic, even in MSA are often constructed differently from written sentences, simply because of the fact that speakers are reasoning while they are talking; hence sentences are not finished completely and are cut by other sentences, and sometimes essential words are omitted etc.

Most of the oral sources are also more complicated than the written ones, because of the dialectal interference. Seldom do we find an Arabic speaker who speaks pure MSA. In the mašreq, and more specifically in Egypt, learned speakers still frequently use dialectal words and also grammatical constructions such as the prefix "bi" before a verb. But also in the Mağreb, we find dialectal interference such as the function word "bāš" in order to instead of "li" or "likay".

In order to train students to digest oral media, we train them at the preparatory level to understand all kinds of sentences with the basic vocabulary of 3,000 words. The students are exposed to the sentences which are run at normal speed. At the advanced level, students have to listen to radio and television programmes and to transcribe the contents.

Corpus analysis may teach us which dialectal words and constructions are frequently used. These words and constructions also ought to be included in the curriculum, in order to train students to identify them when listening. In future, not only do we hope to insert tagged written texts in the database, directly linked to the electronic dictionary, but we also intend to insert real life audio material to the tagged texts. In this way, the student can not only read the text, but will be able to hear the sentences as they are spoken in real life situations.