The receptive vocabulary size of Moroccan learners of Modern Standard Arabic*

Ahmed Ech-Charfi  Youssef Bouaddi
Mohammed V University in Rabat  Morocco
a.echcharfi@um5r.ac.ma  bouaddiyussuf@gmail.com

Abstract

This study is an attempt to measure the receptive vocabulary size of Moroccan learners of Modern Standard Arabic (MSA). Although these learners, just like other Arab learners, are often assumed to be native speakers of MSA, not much is known about the development of their proficiency in the language and, consequently, it is not clear whether their vocabulary size should be compared to that of native or non-native speakers of other languages. To contribute to a better understanding of this issue, this study used a Yes/No test based on Buckwalter and Parkinson’s (2011) list of the 5000 most frequent words in the language. 121 participants were conveniently selected from the 5th, 7th, and 9th grades. The results show that they recognized a mean of around 3500 words and that the difference between levels was not always significant. This vocabulary size is relatively less than the vocabulary size of native speakers in non-diglossic societies, thus pointing to weak proficiency in language skills.

Keywords: diglossia, Modern Standard Arabic, vocabulary size, Arabic Yes/No test

*We would like to thank the two referees for their remarks on an earlier version of this paper.
1 Introduction

During the last two decades, studies on vocabulary have risen exponentially (cf. Milton, 2009; Nation, 2001; Schmidt, 2010). The realization that vocabulary size accounts for performance in reading, writing, listening and speaking encouraged both researchers and stakeholders to use information on the vocabulary component to observe the development of learners’ language proficiency. Most research in this area, however, tends to focus on English as L2 and, consequently, not much is known about the L1/L2 acquisition of other languages, which is needed if a comparative model of vocabulary acquisition is to be developed.

One language for which vocabulary studies are of value is Modern Standard Arabic (henceforth MSA). Although MSA is often claimed to be spoken natively by more than 340 million people,¹ and learned by millions of others as a foreign language for religious or other purposes, research on the acquisition of its vocabulary by native or non-native speakers is alarmingly scarce. Besides, Arabic is a diglossic language in the sense that the standard and the colloquial, though being historically related, are fairly distinct. Therefore, it is not clear whether Arab learners’ vocabulary size is similar to that of native speakers of English, for example, or not. Research on MSA will certainly have implications for other cases of diglossia.

This paper will be articulated as follows: Section 2 will discuss the status of MSA in the Arab World in order to highlight the major factors that facilitate or hinder its learning; Section 3 will deal with the structure of the Arabic word with the view of evaluating the cons and pros of the lemma and word family as measurement units; Section 4 will report the study and Section 5 will discuss its findings.

2 Learning Arabic and the consequences of diglossia

Arabic is a diglossic language. According to Ferguson (1959, p. 336), diglossia is:

a relatively stable language situation in which, in addition to the primary dialects of the language (which may include a standard or regional standards), there is a very divergent, highly codified (often grammatically more complex) superposed variety, the vehicle of a large and respected body of written literature, either of an earlier period or in another speech community, which is learned largely by formal education and is used for most written and formal spoken purposes but is not used by any sector of the community for ordinary conversation.

This definition fits Arabic well, and it is probably for that reason that the language is considered to be the epitome of diglossia. More specifically, MSA is highly codified, enjoys a prestigious body of written literature, and is learned essentially through schooling because, unlike the colloquial, it is not acquired natively by any sector of the population. Additionally, it is usually used for writing purposes as well as for some formal spoken communication. In

¹See (https://www.ethnologue.com/language/ara)
comparison, the colloquial dialects are used basically for ordinary, but also some formal, conversation. In terms of attitudes, MSA is viewed as the correct version of the language while the dialects are viewed as corrupt versions of it. For this reason, Ferguson labels the standard in diglossia in general as High (H) and the dialects as Low (L). Obviously, the status of MSA as an H variety has consequences for its acquisition and, arguably, for the education system in the Arab World as a whole.

Although not a first language and is typically not used in informal contexts, the positive attitudes toward the standard can provide some motivation for its learning. Written Arabic enjoys a long literary tradition, spanning over about fourteen centuries. It should be pointed out that, although Western scholars distinguish between MSA and Classical Arabic, Arabic speaking peoples make no such distinction; for them, the language of the Quran is the same as that they use today, except for what is viewed as minor differences that should not disrupt its continuity or its unity. This language is called *al-fuṣḥā* ‘the eloquent language’, as opposed to *al-ʿāmmiyah* or *al-dārija* ‘the common language’. The Arabic heritage includes literary, scientific, religious and other writings that encapsulate the spirit of the Arab-Islamic civilization and define its culture. It is this civilization that is viewed as the source of pride of the Arab nation and, consequently, continues to bind its members together. Learning the Arabic language is the major means to have access to that culture and enjoy one’s pride in it. The crucial role that Islam plays in Arabs’ life also incentivizes learning the language in order to be able to read not only the Quran, which is recited in prayers on a daily basis, but also books of the prophet’s *Ṣirah* (i.e. biography) and his sayings, exegesis, jurisprudence and other related fields. Without some knowledge of *al-fuṣḥā*, even radio or television programs intended to transmit religious knowledge and practices cannot be fully understood. In brief, as long as the classical Arab-Islamic culture continues to dominate Arabs’ life, there will always be a need to learn the written variety of Arabic, as defined by the grammatical and the lexicographic tradition of the 9th century, despite the changes it is continually undergoing.

Whether MSA and the colloquial form a single language or two distinct languages is not clear, either to native speakers or to specialized scholars (cf. Badawi, 1973; Hary, 1996; Mejdell, 2006; Mitchell, 1986, among others). In fact, the whole issue may be an ideological one that is unlikely to be solved through scientific knowledge (Brustad, 2017; Ech-Charfi, 2019). But the question that is of more relevance to the measurement of Arabic vocabulary size concerns the formal character of the acquisition of MSA. More specifically, will the vocabulary size of Arab learners of MSA be similar to that of native speakers of other languages or to that of non-native speakers? The question also concerns the historical relation between the standard and the colloquial varieties; in other words, will knowledge of the colloquial contribute any significant role in learning MSA vocabulary? To formulate the question in analogical terms, is the relation between MSA and the colloquial similar to, e.g. Standard English and its dialects, or to related languages like French and Spanish. We would expect that learners’ vocabulary size would be larger in the first case than in the second. These questions, however, are still pending future research.
Before the design of the study and its results are presented and discussed, the following section will deal with the nature of the Arabic word while the next one will review some previous studies on the topic.

3 Units of measurement of Arabic vocabulary

Arabic belongs to the Semitic language family. The most salient characteristic of Semitic languages is that their lexicon tends to be generally constituted of consonantal roots, which define the abstract semantic field that individual related words concretize, in addition to vocalic melodies that determine the words’ grammatical category and their morphological meaning. For example, the root k-t-b denotes the field of writing, but it never occurs as an independent word on its own. Rather, it manifests only in words that combine the root and a vocalic melody like *katab* (wrote), *kutib* (was written), *kātib* (writer), *kitāb* (book), *kitābah* (writing), *maktabah* (library, bookshop), *maktab* (office, desk), *kuttāb* (Quranic school), etc. (cf. Ryding, 2005).

As is clear from the examples above, the vowels and the affixes form morphological material that specifies word meaning. But they can also be used with other roots to derive words of similar morphological and syntactic characteristics. For the example, the template of the active participle of tri-consonantal roots in the verb basic pattern is CāCiC. Examples include *xādim* (servant), *Ṣālim* (scholar), *zāhil* (ignorant), *sāʔiq* (driver), etc. With derived and un-derived verbs of quadri-literal patterns, the template of the active participle is muCaCCiC, as in *mutarʒim* (interpreter), *muʕallim* (teacher) and *muhandis* (engineer). As is the case with most derived forms, the rules for deriving the active participle are fairly general.

However, some derivations can be semantically opaque and/or lexically idiosyncratic. This can be illustrated by some verbal forms. For instance, Form I verbs are basic while the others are augmented by the addition of an affix. The formal aspect of this derivation is generally regular, but not all augmented forms are attested. For example, Form IV verbs are derived by the affixation of *ʔa*- to Form I verbs; e.g. *raẓaʕ* (to come back) vs. *ʔarʒaʕ* (to return sth/sb back). But not all basic forms can be augmented in this way. For instance, there is no such form as *ʔašraf* corresponding to *ṣaraʃ* (to know), or *ʔadras* corresponding to *daras* (to study). Instead, these verbs take Form II, which has more or less the same meaning as Form IV, viz. *ṣarraf* (to make sth/sb known) and *darras* (to teach). On the other hand, there are Form IV verbs for which no corresponding Form I is attested. Examples of such verbs include *ʔahabb* (to love), *ʔarsal* (to send), *ʔaʃtā* (to give), to name but a few. In other cases, both the basic and the augmented forms may be attested, but their meaning cannot be easily related to each other. For instance, *ẓamaʃ* means “to assemble”, but *ʔazmaʃ* denotes “to reach consensus”. The two meanings may be connected, but not in the same way Form I and Form IV generally are. Because of this opacity, the rule of formal derivation may not be of much help to the learner and cases like *ʔazmaʃ* must be learned independently, though some lexical inferencing may be possible if enough contextual clues are provided.
Like derivation, inflection also exhibits regularity as well as idiosyncrasy. Verbal conjugation, for example, is fairly simple, with the same affixes appearing almost on every verb. The so-called weak verbs, however, exhibit some complexities and, for that reason, they are often practised in grammar lessons. Thus, a hollow\textsuperscript{2} verb like \textit{qāl} (to say) manifests in the perfective aspect as \textit{qul-} in the first and the second person, and as \textit{qāl} in the third person. In the imperfective, it surfaces only as \textit{qūl-}, but takes other affixes related to mood, thus becoming another favorite topic for schoolteachers!

Noun inflection can also be challenging. Number, in particular, exhibits a lot of complexities, with “broken” plurals showing different patterns the derivation of which cannot be reduced to simple rules. There does exist a simple rule for “sound” plurals, namely the affixation of \textit{ān/ān} for masculine nouns and \textit{āt} for feminine nouns, but these are limited to derived nouns, which appear to be less numerous than non-derived nouns. Most non-derived nouns, however, have plural forms that must be learned independently. For instance, the plural of \textit{razul} (man) is \textit{rāzāl}, a template similar to the plural of \textit{kalb} (dog) (viz. \textit{kīlāb}) although the two words have different singular forms. Similarly, the plural of \textit{walad} (boy) is \textit{ḍawlād}, just like \textit{ḍarbāb} (owners), though its singular form \textit{rabb} is different from \textit{walad}. As pointed out earlier in this section in relation to verbs, these plural forms may be recognized receptively, perhaps by reliance on some guessing strategies, but they are unlikely to be used productively if they have not been learned before.

After this brief but illustrative discussion of the Arabic lexicon, we can now tackle the issue of the most optimal unit of measurement. In corpus and applied linguistics studies, the notion of “word” is often operationalized in terms of either the lemma or the word family. A lemma “includes a headword and its most frequent inflections, and this process must not involve changing the part of speech from that of the headword” (Milton, 2009, p. 10). A verb lemma in Arabic, following this definition, would include all its different forms marked for person, number, gender, tense and mood inflections. Similarly, a noun lemma would include related forms marked for gender, number and case. Undoubtedly, this operationalization is not only convenient when counting the number of different words that occur in a text or that a learner knows, but also distinguishes between two apparently different constructs: lexical knowledge and grammatical knowledge. While the first involves the memorization of lexical idiosyncrasies that cannot be predicted by any general rule, grammar involves the abstraction of patterns that apply to lexical material. In formal language education, rules of grammar are often taught explicitly with the view of applying them recurrently.

But if some forms cannot be predicted by the general rules of grammar, it is not clear whether they should be included under the same headword. In English, for example, irregular plurals such as \textit{children} and \textit{mice} are often treated as independent words from the corresponding singular nouns. The same thing applies also to irregular verbs. The reason is that these forms are idiosyncratic and, consequently, must be learned independently (cf. Milton, 2009; Nation,

\textsuperscript{2}Hollow verbs are verbs that a long vowel as a second radical in the surface form, but which is underlyingly a semi-vowel, i.e. \textit{w} or \textit{j}.\textsuperscript{115}
If this argument is tenable, it follows that many Arabic broken plurals, for example, should be treated as different lemmas rather than classified along with their singular forms (cf. Familiar, 2021). But it seems that there is no single optimal option; it all depends on the objective of the study. It was pointed out above that broken plurals may be recognized as related to their singular forms, but may not be produced when needed for communication. This implies that for the measurement of receptive knowledge, the singular and the plural forms can be considered as constituting a single word, but not for the measurement of productive knowledge (for more on this, see Gardner, 2007). Obviously, this hypothesis needs testing.

The same point can be made in relation to word family. As defined by Bauer and Nation (1993), “a word family consists of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately” (p. 253). When applied to the Arabic lexicon, this definition would imply that all the forms sharing the same root would arguably be considered as constituting a single family and counted as one word. Obviously, a word family is more inclusive than a lemma and, consequently, its use in the estimation of a learner’s vocabulary size would yield a smaller amount than if the lemma is used instead. For instance, the k.t.b word family would include not only the different inflected forms of the verb *katab* (to write), but also all the forms with the same root, including the various augmented verb forms, the active and the passive participles, the nouns of place and time, etc.

For the purposes of measuring the vocabulary size of Arabic speakers, it is perhaps advisable to rank derivational affixes according to complexity. Bauer and Nation (1993) propose a number of criteria to measure this complexity; these include frequency, productivity, predictability, regularity of the written form of the base, regularity of the spoken form of the base, regularity of the spelling of the affix, regularity of the spoken form of the affix, and regularity of function. Obviously, these criteria have been formulated on the basis of the English lexicon and, therefore, their relevance to Arabic remains to be determined, which falls out of the scope of this paper.

In brief, the notions of lemma and word family, as defined in relation to English vocabulary for example, may not be easily applied to Arabic vocabulary. Some of the flaws in previous studies can be ascribed to decisions made on this basis, as will be argued in the following section.

### 4 Previous studies

Not many studies have investigated the acquisition of MSA vocabulary. Concerning the measurement of receptive MSA vocabulary size, we have encountered only three, two of

---

3One of the reviewers pointed out that speakers of MSA may be able to derive various forms of the same root, but that does not necessarily mean that they will be able to understand the meaning of all the forms. While this may be true, the issue is an empirical one; the word family in MSA is still pending to be defined in such a way that it can measure and predict vocabulary knowledge of the language users.
which target foreign language learners and only one targets “native” speakers. These are presented and discussed briefly here.

Of the two studies on learners of Arabic as a foreign language, Baharudin et al. (2014) aim primarily to develop a test of Arabic vocabulary to estimate the vocabulary size of Malaysian learners. The researchers used the frequency list in Landau (1959) to randomly select 50 items from each of the first four 1000 frequency bands. But, as they suspected that learners might find the sample items challenging, they relied on other criteria all of which were intended to weigh the difficulty of those items against their presence (or absence) in the Malaysian teaching materials of Arabic. A further difficulty index and an items discrimination index were used to eliminate challenging items. The final form of the test included a sample of 40 words only, which were supposed to represent the 4000 most frequent words in Landau (1959). The researchers found that the test was reliable both in terms of internal consistency and in discriminating between learners with different proficiency levels.

On his part, Ricks (2015) also aims at developing tests for the estimation of the receptive vocabulary size of learners of Arabic as a foreign language. He proposes different test formats that target not only breadth, but also depth of vocabulary knowledge. Regarding breadth, which is the focus of interest in this paper, the researcher opts for the YES/NO format of Meara (1992) as well as that of Nation’s (1983) Vocabulary Levels Test. The two tests draw on Buckwalter and Parkinson’s (2011) frequency dictionary extracted from a 30 million words corpus. As the dictionary’s authors explain, 90% of the corpus consists of written material representing the main domains in which MSA is used. These include literature and fiction, academic writings, newspaper editorials, daily newswire, and discussions on the internet. The other 10% consists of speech data representing the major dialect groupings. The tests were also found to be valid and reliable.

As for studies on native speakers, Masrai and Milton (2019) is the only one we could find. The researchers developed what they call an “Arabic-Lex” test based on a 50,000 word list extracted from the Arabic Internet Corpus by Sawalha and Atwell (cf. Sawalha and Atwell, 2013). Two words were randomly selected from each 1000 word band, making 100 test items. To these were added another 20 pseudo-words to control for guessing. The test takers had to tick the words they knew only.

Although Masrai and Milton’s (2019) work marks a significant contribution to the measurement of native speakers’ receptive vocabulary size, it has some limitations. The most serious of them relates to the word list which the researchers relied on to design the vocabulary test. To begin with, the list includes a lot of words of non-Arabic origin, a fact which can easily be proven by querying some of these words in the website mentioned above. For example, when the word “آث” occurring in the 3822nd rank in the list is queried in the concordancer, most of the results of the query will turn out to be non-Arabic, though written in Arabic script. As explained by Sharoff (2006), internet corpora are harvested from open access webpages

---

4The list is available at the following website: http://corpus.leeds.ac.uk/query-ar.html
at a first stage, and are later on sifted to eliminate, among other things, texts that are not in
the target language. It seems, however, that undesired texts did manage to slip into the final
corpus against the compilers’ wish. Second, the Arabic Internet Corpus is composed not only
of Standard Arabic texts, but also of colloquial texts pertaining to various regional dialects.
An illustrative example is the verb šūf (to see), which ranks 2134th in the list, though it is
undoubtedly of colloquial origin. This remark paves the way for the third and major weak-
ess of the Leeds list, which has to do with the notion of lemma assumed there. Specifically,
although Sawalha and Atwell (2013) claim that the list is constituted of lemmatized words, it
does not seem to use any consistent definition of lemma. For instance, many forms of the verb
šūf are listed as distinct words, including šuft (I saw), šūfū (look.pl), šūfī (look.fem.sing),
among others. The same remark holds for most of the items in the list. In fact, there are many
instances in which the same form is listed many times. In brief, a close examination indicates
that the list is basically a list of word types and, even as such, it is still needs trimming.

Another serious shortcoming of the Leeds list has its origin in the Arabic writing itself. Prob-
ably because of the nature of Arabic morphology, as described briefly in Section 3 above, the
Arabic writing system does not include short vowels, nor does it mark gemination. These
are sometimes added as diacritics above or below consonants. Besides, the letters ֡ and ֨,
representing the long vowels ֭ and ֶ, are also used to symbolize the semi-vowels ֝ and ֝, and
the letter ֨, generally standing for long ֖, is sometimes used to carry the diacritic representing
the glottal stop. As a consequence of the ambiguity of this system, one written word form can
stand for different phonetic and phonological realities. The form ﮇلما, for example, can be
read as ﮇلما (he knew), ﮇلما (was known), ﮇلما (he taught), and ﮇلما (knowledge),
and only context can help determine which of the four lexemes is intended. Since corpora
used written texts that generally do not include diacritics, calculating the frequency of differ-
tent lexemes (i.e. different phonological forms represented by the same written word) would
be challenging.

In brief, the greatest merit of Masrai and Milton’s (2019) study is that it is the first of its kind
to address the issue of the vocabulary size of native speakers of Arabic. Its findings, however,
should be treated with extreme caution mainly because of the deficient list on which it was
based, a weakness that future studies are likely to suffer from as long as there are no adequate
frequency lists adopting a satisfactory definition of the Arabic lemma.

5 Study

5.1 Test construction

As pointed out in the previous section, the biggest challenge for constructing an Arabic vo-
cabulary size test for native speakers is the lack of an adequate word list extracted from a
representative and balanced corpus. As explained in the same section, the notions of lemma
and word family, as applied to the measurement of English vocabulary, may not be appro-
appropriate for a Semitic language like Arabic. Since there is no consensus on the most optimal lexical unit(s) of Arabic, the development of Arabic corpus linguistics – which is “still in its infancy” (McEnery, Hardie, and Younis, 2019, p. 1) – is likely to take longer than we would hope. Therefore, there is no alternative but to use the available resources, despite their deficiency.

In this regard, the list provided in Buckwalter and Parkinson (2011) is undoubtedly the best, given that it uses consistently a specific definition of the Arabic lemma. For example, all the different inflected forms of a verb are included under a single headword, and the frequency is computed accordingly, unlike the case of the Leeds list. Besides, great efforts were made to disambiguate homographs and assign them to different lemmas when necessary. The corpus from which it was extracted is also large (30 million tokens), fairly balanced and representative, as 90% of it “came from written sources, divided into five text or genre types of equal size” (Buckwalter and Parkinson, 2011, p. 3). All these characteristics make of it the most adequate list available.

But Buckwalter and Parkinson’s (2011) frequency list also has drawbacks when intended to develop vocabulary tests for native speakers of Arabic. The major flaw is probably the limited number of words. Generally, lists of the 5000 most frequent words in a language are intended for foreign learners and, on this basis, Buckwalter and Parkinson’s list would not be appropriate for native speakers of Arabic, who are expected to have a larger vocabulary size. Another drawback concerns the inclusion of colloquial words. The corpus from which the list was extracted consisted of 10% of transcribed conversation from different parts of the Arab World. This spoken corpus is likely to include a large number of words and expressions that have no counterparts in MSA, and some of them may be specific to the variety from which they were taken. As a case in point, we find in the first letter of the alphabetical list the words izzay (how’s come), illi (who, which), aku (there is), āni (I), intī (you.fem.sg), intū (you.pl), inzīn (Ok), among others, all of which come from Egyptian, Gulf or Iraqi Arabic and, consequently, may not be familiar to speakers of other varieties like Moroccans, nor will they be of any particular interest to them. For these speakers, there seems to be no point in examining their knowledge of such colloquial items since they may never need that knowledge.5

The notion of lemma used in the list as well, though being consistent, is not without flaws. In particular, there are cases in which the notion does not seem to be used systematically. For instance, while gender is generally treated as an inflection and, accordingly, masculine and feminine forms of a noun are considered as forms of the same lemma, there are cases in which this practice does not apply. As a case in point, the nouns ibn (son), bint and ʔibnah (daughter) are treated as separate lemmas, although they are clearly masculine and feminine forms of the same word. Besides, both sound and broken plurals are treated as inflections

5One of the referees rightly points out that these colloquial forms may be needed when reading texts, like fiction, which make use of them for various reasons, e.g. verisimilitude. In view of this fact, it seems that knowledge of other colloquial varieties must be included at some level of proficiency, viz. the advanced level
and, consequently, listed under the same lemma as the singular. But if this decision can be appropriate for sound plurals because they are regular, it certainly is not for the highly unpredictable broken plurals. In other words, a learner who knows the singular form of a noun will not necessarily be able to recognize its broken plural form. Therefore, any test based on such a lemma as a unit of measurement will result in overestimation of vocabulary size. All these and similar issues need to be taken into consideration when interpreting the results.

Buckwalter and Parkinson’s list served as a basis for the construction of a receptive vocabulary test. Meara’s (1992) YES/NO format was adopted as a convenient way of tapping into receptive vocabulary knowledge for its practicality, despite its limitations. Accordingly, 40 items were selected randomly from each band of 1000 words. A set of 20 non-words were also added for each band to correct for guess work. The non-words were carefully designed to sound phonologically similar to the other items in the frequency band, usually by replacing a letter of a real word by another. The total number of real and non-words in the test amounted to 300 items all of which were vocalized to avoid the potential ambiguities of Arabic writing. This procedure was repeated twice to develop two versions of the same test in order to allow for parallel forms reliability testing. Both versions included questions about name, age and study level, though only the effect of study level will be reported in this paper.

As to scoring, we have chosen the simplest method among all those reported in the literature (cf. Schmidt, 2010). This method consists in subtracting the number of false alarms (i.e. ticked non-words) from the number of hits (i.e. ticked real words). That is, the final score of a learner in a given level is the number of real words for which the learner ticked ‘yes’ (hits), minus the number of non-words for which the learner ticked ‘yes’ (false alarms). Since each real word represents 25 words in each band, the score was multiplied by 25 to obtain the final score of each level. Finally, the scores obtained in the five levels were added up to estimate the final score of the Arabic receptive vocabulary size of the learner.

5.2 Sample and procedure

A total of 121 participants from public schools in Morocco took part in this study. They belonged to three different educational levels, namely: the 5th, 7th and 9th grades. The first level is the pre-final stage of primary school while the other two are, respectively, the first and the last years of middle school. The study targeted two groups from each level, with an average of 20 participants per group, which amounts to 6 groups in total.

The two versions of the YES/NO test were administered to each group in each level separately. The first group from each level received version A of the test two consecutive times in order to check the test-retest reliability, while the second group was administered version A during the first meeting and version B on the second with the objective of checking parallel forms reliability. A period of one week separated the two paper-and-pencil administrations of the tests. On both occasions, the participants were asked to write their real names so that both
versions by the same testee can be identified. They were instructed that there was no time limit to the test, although it was supposed to take 15 to 20 minutes to be completed. They were also warned that the test included non-words and false responses to these words would be penalized. All these instructions were provided in Arabic on top of the test paper so as to avoid any possible confusion.

After the tests have been scored, the data were coded in SPSS by means of which the statistical analyses were run.

5.3 Results

The first step in the presentation of the results will be the reliability of the tests. It should be recalled that two versions of the YES/NO test were constructed and that, in each study level, one group was given version A twice while the second received versions A and B. The question to be addressed now is whether the participants scored more or less similarly on the two occasions.

The results indicate that both the test-retest reliability and the parallel forms reliability were satisfied. When scores of the two administrations of the version A were compared by means of a related-samples Wilcoxon Signed Rank Test, the difference was found to be non-significant, thus indicating that the participants received more or less the same scores on both occasions. The Pearson correlation between the two scores, however, though significant at the .01 level, barely reached a coefficient of .55. This coefficient was rather unsatisfactory, given that the participants were administered the same test. But when two outliers were deleted, the coefficient rose to .74. Regarding the second group, a related-samples Wilcoxon Signed Rank Test also indicates that there was no significant difference between scores from version A and those from version B. A Pearson correlation between the two sets of scores gave a moderate coefficient of almost .80. With a moderate reliability of the tests guaranteed, we can proceed to data analysis.

The first result to be discussed concerns the effect of frequency. It should be recalled that the tests were based on the assumption that the more frequent words will be learnt before the less frequent ones. Figure 1 overleaf indicates that this assumption is apparently born out.

The bands in this graph slope gradually as we move from the first to the fifth band, thus indicating that more high frequency words are recognized than low frequency words. When the band scores are submitted to a related-samples Friedman’s two way analysis of variance by ranks, the difference is found to be statistically significant at a p-value less than .01. This finding is not unexpected, given that research on vocabulary acquisition tends to converge on the crucial role of frequency on vocabulary learning.

What is unexpected about our data, however, relates to the effect of study level. The results

---

6In what follows, the use of a non-parametric test is an indication that at least one data set is not normally distributed
are given in the following tables, with the first line in each cell referring to data from the first administration of the test and the second to data from the second administration:

<table>
<thead>
<tr>
<th>Table 1: Group 1</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td><strong>Test Admin</strong></td>
<td><strong>Mean</strong></td>
<td><strong>S.D.</strong></td>
<td><strong>Maximum</strong></td>
</tr>
<tr>
<td>5th grade</td>
<td>1st</td>
<td>3259,21</td>
<td>707,55</td>
<td>1750</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>2866,26</td>
<td>1135,95</td>
<td>1300</td>
</tr>
<tr>
<td>7th grade</td>
<td>1st</td>
<td>3368,75</td>
<td>599,93</td>
<td>2175</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3363,75</td>
<td>651,99</td>
<td>2050</td>
</tr>
<tr>
<td>9th grade</td>
<td>1st</td>
<td>3615,47</td>
<td>558,93</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3872,61</td>
<td>480,29</td>
<td>2900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Group 2</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number</strong></td>
<td><strong>Test Admin</strong></td>
<td><strong>Mean</strong></td>
<td><strong>S.D.</strong></td>
<td><strong>Maximum</strong></td>
</tr>
<tr>
<td>5th grade</td>
<td>1st</td>
<td>3426,31</td>
<td>747,33</td>
<td>2025</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3110,05</td>
<td>851,62</td>
<td>1400</td>
</tr>
<tr>
<td>7th grade</td>
<td>1st</td>
<td>3276,31</td>
<td>580,67</td>
<td>1875</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3460,52</td>
<td>669,48</td>
<td>2100</td>
</tr>
<tr>
<td>9th grade</td>
<td>1st</td>
<td>3480</td>
<td>787,17</td>
<td>1625</td>
</tr>
<tr>
<td></td>
<td>2nd</td>
<td>3566,66</td>
<td>655,62</td>
<td>1925</td>
</tr>
</tbody>
</table>

In both these tables, the maximum score could sometimes be almost three times the minimum score. This variation is reflected in the high values of the standard deviations, which reached more than 1000 words in the second administration of Test A for the 5th grade participants, for example. This variation, however, is not an exceptional phenomenon in vocabulary acquisition, as it has been reported in data from various languages in L2 as well as in L1 acquisition (cf. Milton, 2009).
What seem surprising are rather the relatively low mean scores. For all the three levels, the mean scores rarely rose above the level of 3500 words. In the case of the 5th grade, the mean score in the second administration of the test was even lower than 3000. This low score is more reminiscent of L2 than L1 vocabulary size. It should be pointed out that these participants had studied Arabic in school for at least five years. In fact, even the 9th grade students, who had been studying the language for at least eight years not only as a school subject but also as a means of instruction, barely surpassed a mean score of 3500 words. It is true that some of our participants had a score close to 5000, but as groups, they performed less than expected if they are considered native speakers of the language.

Before submitting the data to statistical analysis, mean scores were calculated for the results of the two administrations of the tests. These are reported in the following tables:

**Table 3: Mean scores of Group 1**

<table>
<thead>
<tr>
<th>Number</th>
<th>Mean</th>
<th>S.D.</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th grade</td>
<td>19</td>
<td>3062.73</td>
<td>851.37</td>
<td>1525</td>
</tr>
<tr>
<td>7th grade</td>
<td>20</td>
<td>3366.25</td>
<td>604.49</td>
<td>2137</td>
</tr>
<tr>
<td>9th grade</td>
<td>21</td>
<td>3744.04</td>
<td>484.37</td>
<td>2450</td>
</tr>
<tr>
<td>Mean total</td>
<td>60</td>
<td>3409.79</td>
<td>694.47</td>
<td>1525</td>
</tr>
</tbody>
</table>

**Table 4: Mean scores of Group 2**

<table>
<thead>
<tr>
<th>Number</th>
<th>Mean</th>
<th>S.D.</th>
<th>Maximum</th>
<th>Minimum</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th grade</td>
<td>19</td>
<td>3270.78</td>
<td>754.73</td>
<td>1812.50</td>
</tr>
<tr>
<td>7th grade</td>
<td>19</td>
<td>3368.42</td>
<td>613.38</td>
<td>1987.50</td>
</tr>
<tr>
<td>9th grade</td>
<td>21</td>
<td>3525.69</td>
<td>689.01</td>
<td>1775</td>
</tr>
<tr>
<td>Mean total</td>
<td>59</td>
<td>3392.95</td>
<td>684.62</td>
<td>1775</td>
</tr>
</tbody>
</table>

The scores in these tables refer to the mean scores of the two administrations of Test A to the first group in Table 3, and to the mean scores of Test A and Test B for the second group in Table 4. These scores do not differ much from those of the individual tests reported in Table 1 and 2 above.

When the three study levels are compared, the results turn out to be inconsistent. For the first group, an increase is recorded of around 300 words from the 5th to the 7th grades, and more than 400 words from the 7th to the 9th grades. In the case of the second group, an increase is observed of around 100 words between the first and the second levels, and more than 200 words between the second and the third levels. When these differences are submitted to statistical analysis, however, the difference turns out to be significant for the first group but not for the second. An independent-samples Kruskal-Wallis test indicates the difference between the group means in Table 3 are significant at the level of .01, suggesting that students do add more vocabulary as they move from one study level to another. In comparison, the difference between the means in Table 4 is not significant, indicating that the recorded difference is due merely to chance.
Given these inconsistent results, it is not clear how they can be appropriately interpreted. A closer look at the data in Table 1 and 2 indicates that the behavior of Grade 5 and Grade 9 students of the first group may be at the origin of this inconsistency. In particular, the 5th graders recorded a sharp drop of scores between the first and the second administrations of the test while the 9th graders performed remarkably well on the two occasions. While these 9th graders probably had a strong background in the Arabic language for reasons not included in the design of this study, the 5th graders must have been unmotivated to take the second test. This possibility is reflected in the number of students (viz. 5) who failed to complete the fourth and the fifth bands in comparison with the other participants (viz. none). The high S.D value recorded by the 5th graders in the second test is also suggestive of this fact, indicating that some participants performed significantly less or more than the average. The 5th graders in the second group also recorded a drop in the mean scores, but the drop was not as sharp as in the first group. As to the other group levels, the participants performed relatively consistently on both tests, generally recording a slight increase on the second. Therefore, we can conclude that the tendency shown by our data is that, although there is a slight increase of vocabulary size in correlation with study level, the increase is generally not significant as far as the 5000 high frequency words are concerned.

This finding is rather unexpected, given that previous studies on vocabulary have all found that proficiency level does have an effect on vocabulary size. In this study, however, proficiency was defined in terms of study level. This operational definition may not be valid and is probably at the origin of these results. In Table 4 above, we notice for instance that the minimum was scored by a 9th grader. Similarly, the maximum score for the 5th graders was higher than that of the 7th graders. These and similar facts are clear indications that learners in a higher study level may be less proficient in Arabic than others in a lower level. Other possible interpretations are discussed in the next section.

6 Discussion

During the last two decades, a large body of literature has grown around the relation of vocabulary size and language proficiency. Studies on the acquisition of English as a second or a foreign language in particular have found a strong correlation between the number of words learners know and their performance in reading, listening, speaking and writing (cf. Schmidt, 2010). Another important finding in this research area relates to the role of word frequency both in vocabulary learning and in performance in the four skills. In particular, learners have repeatedly been found to master more high frequency words than low frequency words. This should come as no surprise, assuming that learners are exposed more frequently to the first type of vocabulary than to the second. What is less accessible to intuition, however, is the role of word frequency in text coverage, whether the text is written or spoken. For instance, the first 1,000 most frequent words in English generally constitute 74% of authentic texts; addition of the following 1,000 words in the frequency list will raise coverage to 81%, etc. But reasonable comprehension is argued to require knowledge of about 95% of the words in a
text or more, which would necessitate a vocabulary size of at least 12,000 words (cf. Milton, 2009; Nation, 2001; Schmidt, 2010). Some researchers go even further and advance that 98% of the running words should be familiar to the reader for higher-order comprehension (e.g. inferencing) to happen, a figure which undoubtedly requires a much richer vocabulary (cf. Laufer, 1989, 1992). Similar arguments have been advanced in relation to the other skills, though listening and speaking may be less demanding in terms of vocabulary knowledge. All these findings indicate that no progress can be made in language learning without familiarity with the most frequent words in the target language and, consequently, textbook development and language teaching must set standards as to how many words should be learned at each study level (cf. Milton, 2009).

Given the scarcity of research on vocabulary size and language proficiency in Arabic, we will limit the discussion to reading. Although reading itself has not received due attention in relation to vocabulary, at least there is some basis on which some speculation can be made. This relates to lexical coverage. In this regard, Masrai and Milton (2013) and Ech-Charfi (2023) are probably the only studies that could be found on the topic for reasons mentioned previously. Masrai and Milton (2013) selected 20,000 lemmas from Sawalha and Atwell’s (2013) list and calculated the text coverage of a set of frequency ranges. The problem, however, is that this list suffers from a number of weaknesses already discussed in Section 4. To get around this problem, Ech-Charfi (2023) opted for Landau’s (1959) list, although it was extracted from a small corpus of less than 300,000 tokens. The corpus, however, was balanced to a certain extent and included only written texts of modern Arabic prose. Besides, the notion of word lemma relied on was well-defined and used consistently throughout the list after it has been reorganized by the researcher. Text coverage according to both studies is given in the following table:

<table>
<thead>
<tr>
<th>Number of words</th>
<th>Coverage (%)</th>
<th>Coverage (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(Masrai and Milton, 2013)</td>
<td>(Ech-Charfi, 2023)</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>19,45</td>
</tr>
<tr>
<td>100</td>
<td>34</td>
<td>38,62</td>
</tr>
<tr>
<td>1,000</td>
<td>66</td>
<td>70,96</td>
</tr>
<tr>
<td>2,000</td>
<td>76</td>
<td>82,62</td>
</tr>
<tr>
<td>3,000</td>
<td>82</td>
<td>88,7</td>
</tr>
<tr>
<td>4,000</td>
<td>86</td>
<td>92,33</td>
</tr>
<tr>
<td>5,000</td>
<td>89</td>
<td>94,71</td>
</tr>
<tr>
<td>9,000</td>
<td>95</td>
<td>98,98</td>
</tr>
<tr>
<td>14,000</td>
<td>98</td>
<td></td>
</tr>
</tbody>
</table>

As the table shows, there is some discrepancy between the figures advanced by the two studies. Besides, the lists were not equal in size, with Landau’s list consisting of just about 12,000 words, a fact that explains the empty cell at the bottom right of the table.

Notwithstanding this discrepancy, the vocabulary size of our participants seems to be defi-
cient. With a mean of less than 4000 words, they can barely recognize between 86% and 88% of an authentic text. If Arabic is comparable to English, a reasonable text comprehension requires the recognition of at least 95% of its words (cf. Laufer, 1992). In other words, our participants need a vocabulary size of about 9000 lemmas to reach this objective. But since there is a difference of more than 5000 words between this threshold and the recorded mean, we can conclude that our participants will perform poorly in any reading test. This prediction is repeatedly borne out by international (e.g. OECD, 2019) as well as national reports (e.g. CSEFRS, 2021), which all rank Moroccan pupils among the least able in the reading skill. It could be objected that the tests used for this study targeted only the first five bands and, consequently, nothing can be said about the words they knew beyond this level. But given the Zipf\textsuperscript{8} distribution and the crucial role of word frequency, as has been demonstrated repeatedly in many studies including the present one, it is unlikely that learners will know more low frequency words than high frequency words. In other words, the participants’ vocabulary size is unlikely to be significantly larger than that recorded even if the tests targeted five more other bands, since the percentage of words recognized gets lower and lower in less frequent word bands. Besides, given the notion of Arabic lemma used by Buckwalter and Parkinson (2011), the results of the tests based on this list are likely to overestimate the vocabulary size of learners. It was noted in Section 3 above that the singular and the plural forms of a noun are considered as part of the same lemma. But Arabic broken plurals are highly idiosyncratic and it is not clear that learners who can recognize the singular form can also recognize the plural form or vice versa. More research is needed to shed light on this and similar issues.

On the basis of these remarks, it becomes doubtful whether Moroccan learners are native speakers of Arabic. According to Nation (2001), adult native speakers of English know about 20,000 word families, a figure that corresponds to more than 30,000 lemmas.\textsuperscript{9} He claims that they add about 1,000 word family each year, starting from their third year (see also Goulden, Nation, and Read, 1990; Zechmeister et al., 1995). If these figures are taken as benchmarks, we should expect our participants, whose age ranges between 11 and 15 years old, to know between 12,000 and 19,000 words. But they are far from that. Of course, the number of words in English and Arabic is not the same, nor is their lexical coverage similar, but the difference is not too stark to ban any reasonable comparison. (See Ech-Charfi, 2023 for a comparison between the two languages). Therefore, we can conclude that our groups of learners do not manifest a lexical knowledge similar to that of native speakers.

This conclusion has interesting implications for a theory of diglossia. Although not unfamiliar to researchers of Arabic, not many studies have attempted to measure Arab learners’ proficiency in MSA, Parkinson (1993) being a rare exception. But if these learners’ knowledge of H is more similar to that of a second or even a foreign language, how can a theory of diglossia continue to assume that H and L are “not so close as to be readily mutually intelligi-

\textsuperscript{8}Zipf’s Law predicts that a large percentage of natural language texts will be constituted by a small set of high frequency words, as exemplified by Table 5 above.

\textsuperscript{9}We multiplied the number of word families by 1.6, following a suggestion made in Milton (2009), though this formula yields only a very rough estimation.
ble, but also not so distant as to be unhesitatingly regarded as separate languages” (Hudson, 2002, p. 9)? As pointed out earlier in Section 2, advocates of classical diglossia seem to grant special status to the fact that diglossic speech communities believe that H and L form a single language. But this belief is not well-understood and seems to be paradoxical: individuals may express this belief overtly on some occasions, but assert on others that H is difficult to learn and that only a minority of erudite scholars can really master it. If this belief is apparently responsible in part for the stability of diglossic situations, it is challenged by the stark reality of learners’ school performance, especially in modern times where educational standards are set and continuously checked both nationally and internationally. In such a situation, while the socio-politics of diglossia direct attitudes toward the unity of H and L, its pedagogical consequences shift attention to the differences between the two varieties, as the discourse conflict between pro-MSA and pro-colloquial in the Arab World illustrates (cf. Ech-Charfi, 2016, 2019; Moustaoui Srhir, 2016).

From a scientific viewpoint, however, many issues remain to be resolved before it can be claimed that MSA is more of L2 than a native language as far as vocabulary is concerned. The most urgent among them is the measurement of a larger vocabulary size of Arabic speakers than we did in this study, which will be possible only if an adequate frequency list is available. In relation to this issue, the notion of Arabic lemma needs to be defined adequately for language learning purposes and, until that is done, the previous objectives cannot be attained. The role of the Arabic colloquial varieties in the acquisition of MSA is also not clear; if knowledge of these varieties turn out to facilitate the learning of MSA vocabulary, Arabic speakers may not be considered as Arabic L2 learners either. These and similar questions are still pending answers.

7 Conclusion

This paper has attempted to measure the Arabic vocabulary size of Moroccan primary and middle schoolers. The major finding of the study is that these learners have a relatively poor vocabulary that is unlikely to enable them to achieve a reasonable understanding of authentic texts. But since research in this area is still in its infancy, more studies are needed not only to measure learners’ language proficiency and its relation with vocabulary size, but also to provide more accurate estimations of vocabulary size itself. This objective cannot be attained unless more progress is achieved in Arabic lexicography (by providing a pedagogically oriented definition of lemma and word family) and corpus linguistics (by developing adequate word frequency lists). This endeavor is expected to deepen our understanding of the pedagogical consequences of diglossia as well as the interaction of the pedagogical and the social aspects of it. If it turns out that diglossia is a real obstacle for educational improvement, attention will turn toward its ideological power as the main factor behind its resilience.
References


