International Journal of Arabic Linguistics

Introduction:

Measuring Arabic vocabulary acquisition

Brahim Ait Hammou
Ministry of Education
Morocco
hammou76@gmail.com

Ahmed Ech-Charfi
Mohamed V University in Rabat
Morocco
a.echcharfi@um5r.ac.ma

Starting from the 1970s and as a result of the shift of focus from competence to performance, the role of a learner's lexical repertoire as part of their knowledge of language has gained special attention in linguistic research. As a result of the analysis of data produced by language speakers both in written and spoken forms, studying different aspects of vocabulary knowledge has become established as an important area in language research. Within this context, researchers often felt the need to cite Wilkins's (1972, p. 111) aptly phrased statement that "While without grammar very little can be conveyed, without vocabulary nothing can be conveyed".

In spite of recognizing the value of words as building blocks of language, no significant progress was made in vocabulary research until recently thanks to the use of advanced computer technology. This has allowed linguists to collect, part-of-speech tag and analyze large amounts of vocabulary data. These technological advances led to the thriving of the discipline of corpus linguistics. Because traditional dictionaries do not provide enough information about the actual, real-life, use of words in skills such as writing and speaking nor in genres such as story-telling, academic articles or other written media, relying on corpus analysis using technology has become necessary. This has made it possible to collect and store large computer-readable corpora representing language use in a certain period of time, and to search those corpora with little effort and in very little time. This progress allowed for the identification of patterns in language use. Without the analysis of huge language corpora and by relying on intuition alone, many aspects of language use would have escaped researchers' attention. For instance, many words that seemed perfect synonyms turned out, after close investigation using corpus linguistics techniques, to be used differently, especially in relation to the words they collocate with. Actually, the difference may not be limited to collocations, but may also involve differences in meaning. Similarly, corpus analysis has also allowed

https://doi.org/10.34874/PRSM.ijal-vol10.48103

linguists to get a thorough understanding of the distribution and dispersion of words over different registers and genres. The analysis of large amounts of vocabulary as used by actual speakers has also allowed applied linguists to highlight the role of vocabulary in the overall development of language proficiency and also in the development of specific language skills such as reading, listening and writing. To cite only one example in this general introduction, Nation (2001) reports that vocabulary tends to correlate significantly with reading, writing, listening and speaking, and in some cases, vocabulary accounts for about 70% of the variance in the scores of the language skill. This should come as no surprise, given that words are the building stones of language, and without them, communication cannot happen.

One of the major contributions of corpus linguistics to vocabulary research, if not *the* major one, is the calculation of frequency. Although native speakers can develop sharp intuitions about how frequent a word or an expression is, their intuitions are rarely sharp enough to determine exactly how frequent an item is or whether one item is more or less frequent than another. Computer software is able to count not only the number of tokens or types in corpora of millions of running words, but it can also classify types into lemmas or word families and count their frequency in those corpora. If a corpus is compiled carefully in such a way that is highly representative of the usage of a language in a particular period, the use of a specialized computer program can provide a fairly accurate picture of how frequent any word or expression is in the language. Similarly, advances in computer technology and corpus linguistic research has made it possible not only to gain a clear understanding of the frequency and dispersion of single words, but also the frequency and strength of association between word combinations (i.e. collocations) and the frequency and strength of association of words within and syntactic constructions (i.e. collostructions). Thus, probably for the first time in history, linguists have at their disposal sufficient data about usage to describe language structure in a more realistic manner than was possible before.

Corpus linguistics methods and techniques are even more useful to applied linguists. In language education, for example, it has long been noted that formal descriptions of a language can be of little use to curriculum or textbook designers, teachers, or learners. The mastery of words and grammar rules alone may enable learners to produce intelligible messages in a target language, but it can hardly enable them to communicate in a native-like way. Learners need to know how native speakers express themselves and how they formulate their messages in order to be able to use language in a more fluent way, while respecting native norms as to usage and contextual constraints. Obviously, the best way to acquire that kind of knowledge is to be exposed to as much native input as possible, but many foreign language learners often rely on formal classes where they do not have access to that kind of input. Therefore, formal education is called to provide explicit instruction on how a target language is used natively. Thanks to the analysis of large data, curriculum designers have become able to sort language vocabulary into frequency categories which are, in turn, targeted at different levels of instruction. This facilitates language learning and provides learners with the most frequent and urgent items that would allow them to express their basic communicative needs. The Common European Framework of Reference, for instance, has identified different lev-

els of language proficiency and also the type of linguistic items which are expected in the performance of the learners at each level.

Researchers are also interested in understanding the development of proficiency among foreign languages learners. For a long time, especially after the advent of formal schools of linguistics like Generative Grammar, the focus in second language acquisition (SLA) was on the acquisition of the system or the structure of language. Even after the introduction of Hymes's (1972) notion of 'communicative competence', research in SLA continued to favor the formal aspects of language, probably on the assumption that linguistic competence is determined by a universal, allegedly innate, faculty rather than by the circumstances of language use. By using information about performance gathered through corpus linguistics techniques, researchers are able to gain insight into the acquisition of language through language use rather than by relying only on the competence of a 'perfect' speaker as was the case before. As a case in point, a lot of studies have tracked different tacit areas of language development, including the influence of the frequency of word occurrence and co-occurrence on the acquisition of different aspects of language such as vocabulary and grammatical constructions (e.g, Davies, 2009; Durrant and Schmitt, 2009; Evert, 2009; Granger, 2018, 2021; Laufer, 1989; Laufer and Nation, 1995; Milton, 2009; Nation, 2001, 2006; Schmitt, 2014; Schmitt and Schmitt, 2020).

Having highlighted these achievements in English corpus linguistics, it is important to note that not all languages have witnessed similar developments. While sophisticated software programs have been developed to analyze English and a few other languages, Arabic is still lagging behind in this respect. Among the factors that could explain this state of affairs one can cite the structure of Arabic itself, the low interest in the acquisition of Arabic as a foreign language, the assumption that standard Arabic is a first language in the Arabic-speaking world, etc. We, therefore, underscore the deficiency in the studies of Arabic vocabulary which are mainly corpus-based due to lack of sophisticated computer tools which can deal with such a complex system. The current issue of the *International Journal of Arabic Linguistics* (IJAL) is an attempt to contribute to bridging this gap between vocabulary studies in Arabic and other languages.

This special issue includes a set of empirical studies which deal with different aspects of the vocabulary of Arabic. In the first paper on 'Frequency and text coverage in SA based on Arabic Internet Corpus', Ahmed Ech-Charfi tackles the issue of word frequency lists in Arabic. By addressing the limitations of Sawalha's list extracted from Arabic Internet Corpus (Sawalha and Atwell, 2011), he attempts to develop a lemmatized frequency list of Arabic vocabulary based on a consistent and explicit definition of the Arabic word lemma. The paper also uses the extracted word list to calculate lexical coverage in Arabic. While frequency lists are abundant in English, Arabic suffers from shortage in this area. This makes researching other aspects of Arabic vocabulary and also the relationship between vocabulary and other language skills quite difficult to conduct. Therefore, one of the merits of this study is that it helps researchers to conduct further studies on Arabic corpora and also on the teaching of

Arabic using the revised frequency list.

In the second paper, the same author exploits the frequency list described in the first paper to develop an Arabic version of Nation and Beglar's (2007) Vocabulary Size Test. The article presents the results of using a test of 140 words selected from the first 14 frequency bands of the revised list. The results of the study indicate that the test can significantly differentiate between the three proficiency levels of participants used in the study. This shows that the frequency levels it is based on reflect real language proficiency. The results of this test piloting further support the accuracy of the methodology adopted in the compilation of the frequency list on which the test is based as well as the ability of the test to draw lexical profiles of learners with different proficiency levels.

In the third paper, Brahim Ait Hammou deals with the issue of vocabulary coverage and its relationship with reading comprehension in Arabic. While the topic of vocabulary coverage has gained paramount importance in studies conducted on languages such as English, the topic has not gained its due importance in Arabic. This article presents interesting findings about the connection between knowledge of the vocabulary in a text and overall comprehension of the text. Similarly, this study also examines the relationship between vocabulary coverage and learners' performance in specific reading skills. While international studies such as PISA (e.g, 2022) continue to show that Moroccan learners are facing severe difficulties in reading comprehension, there is very little research that tries to understand the reasons behind this poor performance. Ait Hammou's paper tries to provide a tentative answer to this question.

The fourth article by Imad Messouab focuses on the relationship between learners' receptive vocabulary knowledge and their performance in reading comprehension in Arabic as a foreign language (AFL). With a total of 125 Arabic L2 participants, the study uses two tests: a Yes/No (Y/N) test and a reading comprehension test (RCT). The results indicate that vocabulary knowledge is a fundamental factor in achieving reading comprehension. Like the study by Ait Hammou, this study also shows that knowledge of vocabulary is an important pre-requisite for performance in reading comprehension.

The fifth article by Abdelaziz Agrram and Ahmed Ech-Charfi tries to measure the receptive collocational knowledge of Moroccan learners of Arabic. Using the Rasch model, this paper examines the receptive collocational knowledge of high frequency items in Modern Standard Arabic. The results of this article shows that the receptive mastery of collocations among Moroccan learners is to some extent underdeveloped. The study also lists some implications for language teaching and research.

The last paper by Taifi Bernoussi investigates morphological awareness in Moroccan learners of Arabic. The test she designed targets mostly inflectional affixes in regular and irregular forms. The results of the piloting stage indicate that learners still struggle with this aspect of Arabic even at relatively advanced levels of study. Given that Arabic morphology relies mostly on roots and templates, the development of morphological knowledge in native and non-native learners is still in need of further investigation, as the difference between concati-

native and non-concatinative languages can turn out to be significant in language acquisition.

Through these studies, we hope that this issue will be a good contribution to applied Arabic linguistics, in general, and to the area of Arabic vocabulary, in particular. Researchers in this area are invited to address the issues tackled in these papers with more depth in the future.

References

Davies, Mark (2009). "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights". In: *International Journal of Corpus Linguistics* 14.2, pp. 159–190. DOI: http://doi10.1075/ijcl.14.2.02dav (cit. on p. 3).

Durrant, Philip and Norbert Schmitt (2009). "To what extent do native and non-native writers make use of collocations?" In: *IRAL - International Review of Applied Linguistics in Language Teaching* 47.2, pp. 157–177. DOI: 10.1515/iral.2009.007 (cit. on p. 3).

Evert, S. (2009). "Corpora and collocations". In: *Corpus linguistics. An international handbook*. Ed. by A. Lüdeling and M. Kytö. Vol. 2, pp. 1212–1248 (cit. on p. 3).

Granger, S. (2018). "Formulaic sequences in learner corpora: Collocations and Lexical Bundles". In: *Understanding formulaic language*. Ed. by S. Granger. Routledge, pp. 228–247 (cit. on p. 3).

— (2021). "Phraseology, corpora and L2 research". In: *Perspectives on the L2 Phrasicon: The View from Learner Corpora*. Ed. by S. Granger, pp. 3–21. DOI: 10.21832/9781788924863-002 (cit. on p. 3).

Hymes, D. H. (1972). "On Communicative Competence". In: *Sociolinguistics: Selected Readings*. Ed. by J. B. Pride and J. Holmes. Penguin, pp. 269–293 (cit. on p. 3).

Laufer, B. (1989). "What percentage of text lexis is essential for comprehension?" In: *Special language: From humans thinking to thinking machines*. Ed. by C. Lauren and M. Nordman. Multilingual Matters, pp. 316–323 (cit. on p. 3).

Laufer, B. and P. Nation (1995). "Vocabulary size and use: Lexical richness in L2 written production". In: *Applied Linguistics* 16.3, pp. 307–322. DOI: 10.1093/applin/16.3.307 (cit. on p. 3).

Milton, J. (2009). *Measuring Second Language Vocabulary acquisition*. Multilingual Matters (cit. on p. 3).

Nation, I. S.P. (2001). *Learning Vocabulary in Another Language*. Second edi. Cambridge University Press (cit. on pp. 2, 3).

— (2006). "How large a vocabulary is needed for reading and listening?" In: *Canadian Modern Language Review* 63, pp. 59–82. DOI: 10.3138/cmlr.63.1.59 (cit. on p. 3).

Nation, P. and D. Beglar (2007). "A vocabulary size test". In: *The Language Teacher* 31.7, pp. 9–12 (cit. on p. 4).

Sawalha, M. and E. Atwell (2011). "Accelerating the processing of large corpora: Using grid computing technologies for lemmatizing 176 million words Arabic Internet corpus". In:

Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2, pp. 1–3 (cit. on p. 3).

Schmitt, N. (2014). "Size and Depth of Vocabulary Knowledge: What the Research Shows". In: *Language Learning* 64.4, pp. 913–951. DOI: 10.1111/lang.12077 (cit. on p. 3).

Schmitt, N. and D. Schmitt (2020). *Vocabulary in Language Teaching*. Cambridge University Press. DOI: 10.1017/9781108569057 (cit. on p. 3).

Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold (cit. on p. 1).