International Journal of Arabic Linguistics

# The Construction and Piloting of an Arabic Vocabulary Size Test

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

**ملخص:**
يقترح هذا المقال اختبارا لتقدير عدد المفردات التي يعرفها متعلم العربية على منوال الاختبار الذي صاغه نايشن وبغلار لتقدير حجم المعجم عند متعلمي اللغة الإنجليزية. وقد اعتمد اختبار العربية على قائمة المفردات الشائعة التي استخلصها مجدي صوالحة من متن عربية الانترنيت بعد تعديلها وتنقيحها، كما بيناه في المقال السابق من هذا العدد. يشتمل هذا الاختبار على 140 مفردة منتقاة بصفة عشوائية من المفردات الأربعة عشر ألف الأكثر شيوعا، مستعملة في سياق لا يساعد إلا على معرفة هل هي اسم أم فعل أم حرف. ويقابل كل كلمة أربع تعاريف على المتعلم أن يختار منها التعريف الصائب فقط. وتشير نتائج التجريب الأول للاختبار أنه قادر على تمييز المتعلمين بحسب مستواهم الدراسي، إلا أن أثر الشيوع لا يظهر جليا، خاصة بعد الألف السادسة، حيث تميل المتوسطات إلى الطلوع والنزول.

## Abstract

This paper introduces an Arabic vocabulary size test, modelled on Nation and Beglar's (2007) Vocabulary Size Test. The Arabic version is based on a word list extracted initially from the Arabic Internet Corpus by Majdi Sawalha and made available at the Leeds University website. This list was further developed by Ech-Charfi (2024) by adopting an explicit definition of the Arabic word lemma and removing undesirable items. The test is composed of 140 items selected from 14 frequency bands and used in contexts that are not conducive to guessing. For each item, four options are provided, all of which are definitions of words selected from the same band as the stem. The piloting indicates that the test can differentiate between learners' proficiency levels. The effect of frequency, however, though noticeable, is not decisive since the group means do not decrease consistently from high frequency to low frequency bands.

**Keywords**: Receptive vocabulary size; measuring vocabulary; Modern Standard Arabic; Arabic vocabulary.

# 1 Introduction

Measurement of vocabulary knowledge is a cornerstone of language learning research in general, and research on vocabulary, in particular. By measuring vocabulary size, both researchers and teachers could predict learners' proficiency and performance in various language tasks, as vocabulary knowledge has been found to correlate highly with different skills (cf. Nation, 2001). Such a measurement also makes possible the tracing of language development in learners at various stages in the learning process. Although these are the major benefits of vocabulary testing, they are by no means the only ones, as the literature on different languages has shown. The issue, however, is that vocabulary size tests are not available for all languages, nor even for the major international languages. One such language is Arabic, a language spoken by around 360 million native speakers and one of the six languages of the United Nations; the language is also learned by millions of non-Arab Muslims around the world. Such a situation can be explained partly by the lack of a reliable word frequency list extracted from a representative corpus (cf. Ech-Charfi, 2023).

In this paper, I will draw on a word list I developed based on the one extracted by Majdi Sawalha[1] from the Arabic Internet Corpus in order to design an Arabic version of Nation and Beglar's (2007) Vocabulary Size Test. (A detailed description of how the list was developed can be found in Ech-Charfi, 2024). The paper is constituted of three main sections. In addition to this introductory section, Section 2 will explain how the test was designed and the respects in which it follows closely in the footsteps of (and those in which it differs from) Nation and Beglar's version. Section 3 discusses the piloting of the test and its results along with their implications for the test itself as well as for the list it was based on. Finally, Section 4 concludes with implications for the measurement of vocabulary knowledge in Arabic.

# 2 Test construction

As explained in Ech-Charfi (2024), Sawalha's list includes 100,000 items and accounts for 74,191,620 word tokens. But because the notion of lemma on which it is based does not seem to be coherent or consistent, it was re-lemmatized in accordance with both the dominant definition of word lemma in the literature and the nature of Arabic morphology. Items in Latin script were also deleted, and so were colloquial words that did not have cognates in Modern Standard Arabic (MSA). The rationale behind this decision was not to construct a pure language, but rather to comply with the expectations of teachers and learners of MSA, who generally do not consider colloquial varieties of Arabic as part of their target language. Proper nouns as well were dropped since this class of words are generally not treated as words of any language and are not translatable. The outcome of this process was a much shorter list than Sawalha's, comprising only around 22,000 lemmas, a figure that accounted for 55,935,369 tokens of the Internet Arabic Corpus.

[1] The list is available at http://corpus.leeds.ac.uk/frqc/i-ar-forms.num

The recompiled list, however, included a number of entries (1742 in total) consisting of two or more homographs. Homographs are words written in the same way but have different pronunciations, given that the Arabic writing system does not generally use vowel diacritics. Even with the use of concordances generated by the Leeds corpora platform, all attempts to disambiguate the homographs failed to yield reliable results. Since some of the homographs were already listed in Sawalha's list with vowel diacritics, either because the diacritics were provided in the texts or because there were clues in their linguistic environment which indicated which meaning was intended, it was decided that an estimation of their frequency could be made on that basis. For instance, the written form بعد could be pronounced as 'baʕd' (after), 'buʕd' (distance), or 'baʕud' (to be far away). The three were listed as separate lemmas with their token frequency, but there was a fourth ambiguous item, with its own frequency. A decision was made to split the frequency of the ambiguous item over the other three in proportion, rather than equally, to their frequency. The reasoning behind such a decision is that the three meanings of the homograph are likely to have the same distribution in the corpus as the disambiguated words. In the other cases where this estimation was not possible, the frequency was divided equally over the potential pronunciations of the word. Obviously, this rather ad hoc measure will affect the quality of the word list to a certain degree.

In addition to disambiguating homographs, functional words were also dropped from the list. Functional words in Arabic include prepositions, particles, pronouns, subordinators, coordinators, question words and a few others. As in other languages, these are high frequency words and are usually learned at early stages of contact with the language. That is why they are often dropped from word lists used in designing vocabulary tests intended even for foreign language learners (cf. Milton, 2009). Given that MSA is used extensively in Arabic-speaking countries, we assumed that learners would develop acquaintance with high frequency words early enough to justify the exclusion of such words right at the outset. They were 134 in total, but they accounted for almost 18 million tokens.

The final list consisted of exactly 19,968 items that represented more than 47 million tokens of the corpus. Following Nation and Beglar (2007), the most frequent 14,000 words were targeted for testing. As the authors explain, this figure accounts for more than 98% of running words in authentic English texts and, consequently, to achieve a reasonable understanding of these texts requires acquaintance with these 14,000 words. For MSA, however, the relation between reading comprehension and vocabulary size is still not well understood, as no study on the topic is visible yet. However, according to an estimation made by Ech-Charfi (2024), the first 15 bands account for almost 95% of the corpus considered; more or less the same finding applies to genres such as children's stories and standardized reading tests like PISA. If we take into account the fact that the first 100 most frequent words (which have been dropped here) account for more than 37% of the running words, according to the same study, we will realize that the 14 bands targeted in the present test should account for a level well beyond 95% of a non-technical authentic text of the kind intended for students in high school. Therefore, 14 bands seems to be a reasonable target, at least for this initial stage of research on vocabulary in MSA.

The design of the Arabic Vocabulary Size Test followed in the footsteps of Nation and Beglar (2007). Specifically, 10 items were randomly selected from each band of 1000 words. These were used in contexts that indicated their part of speech only, without allowing for more possibilities to infer their meaning. Four definitions were provided for each target item. The definitions corresponded to the meaning of items belonging to the same part of speech and drawn from the same band. As to the words used to define them, they were mostly selected from the first three bands or, on a few occasions, they had colloquial cognates. The reasoning behind this decision is that high frequency words are likely to be known by the test takers, and so are words that have colloquial cognates, even when they are not very frequent. Here's an example from the second band:

شابَ: {شابَ} الّرجلُ
أ- وَضَعَ جَبْهَتَه على الأرْض
ب- صار غَنِيّا أكثر من غيره
ت- صار شَعْره أبْيضاً
ث- سَالت دُموعه

As can be seen, the stem is cited first in bold type, and then inserted between braces (also in bold type) when used in a sentence in order to avoid all kind of ambiguity as to which is the target item. In the first definition, the word *ʒabhah* (forehead) is preferred over the more usual *ʒabīn*, first because the former occurs in the second band and the latter in the ninth and, second, because the former also coincides with its equivalent in the Moroccan colloquial Arabic variety. Also, the four options define words from the same frequency band as the target word (viz. second band), and all of those words are verbs that can fit perfectly in the context of the sentence. Good care was taken to ensure that the correct answers were equally distributed over the four positions indicated by the letters ا - ب - ت - ث.

In some respects, however, the design of the Arabic version of the Vocabulary Size Test differs from that of the English version. Specifically, while Nation and Beglar (2007) use word family as a unit of measuring vocabulary, the list on which the Arabic version of the test was based used word lemma instead. The two units in the two languages have not been compared so far, and it is not clear how their difference affects coverage or vocabulary knowledge. This fact explains in part another point of difference between the Arabic and the English versions of the test regarding the words used in the definitions. While Nation and Beglar (2007) drew these words from the first two bands only, we had to extend selection to the third band in the list as well. In English, research has shown that the most frequent 2,000 word families constitute the threshold for any significant use of the language. In comparison, no study has established any threshold for MSA yet and, consequently, we do not know precisely what words constitute the basic vocabulary of the language. In practical terms, however, the first two bands were found to be insufficient for the definition of many items and so selection was extended to the following band.

After this presentation of the test design, we will describe the piloting of the test and its results. Some of the implications on the validity of the test will also be discussed[2].

# 3 Test piloting

## 3.1 Participants

In order to test the extent to which the Arabic Vocabulary Size Test can distinguish between learners of various proficiency levels, three groups were recruited from three different levels of education: one from middle school, one from high school and one from university. All the three groups were in their first year and the test was administered early in the school year to measure the amount of vocabulary students have at the end of their previous learning stage, i.e. primary school for the middle schoolers, middle school for the high schoolers, and high school for the university students. The middle school group was constituted of 31 students, the high school group of 28, and the university group of 12, making a total of 71 students. Most of them finished the task in 30 – 45 minutes, but some of them took about an hour.

The results were entered first in an Excel spreadsheet for each item either as 1 if the answer was correct or as 0 if not. Later, the results of each of the fourteen bands were calculated and tallied and, finally, the total was multiplied by 100 to yield an estimate of the receptive vocabulary size of each participant. The scores of the bands and the estimated vocabulary size were finally exported to an SPSS file to be submitted to some statistical procedures.

## 3.2 Results and Discussion

The first analysis conducted on the results relates to the internal consistency of the test. In this regard, the scores of the fourteen bands were submitted to a Cronbach test to see the extent to which they co-vary. The Alpha coefficient turned out to be .94, a coefficient that is high enough to indicate that the bands are consistent with each other. Indeed, all the fourteen components showed equal coefficients, suggesting that the overall reliability of the test would not be affected if any of them were to be deleted. Obviously, for a test to be validated, more sophisticated analyses are needed. In particular, what is needed is not only whether different groups of items (viz. bands) correlate with each other, but also whether individuals items tend to exhibit patterns of ease or difficulty and the extent to which these patterns can be considered as reflecting difference in the target constructs (i.e. word frequency and vocabulary knowledge). This objective, however, is beyond the scope of the present study, as statistical measures like Item Analysis or Rasch Analysis require considerable data. Regarding the test's ability to distinguish between learners from different proficiency levels, we note that the participants' scores do indicate that the test satisfies some types of validity such as predictive and concurrent validity. Table 1 provides a summary of the results:

[2] A copy of the test and its key can be requested from the author on his personal e-mail address.

**Table 1:** Descriptive statistics

| Group | Mean | S.D | MIN | MAX |
|---|---|---|---|---|
| Middle schoolers | 7654,84 | 1978,35 | 3400 | 11200 |
| High schoolers | 10431,03 | 1395,45 | 7100 | 12500 |
| University Students | 12846,15 | 652,60 | 11500 | 13500 |

As can be observed, the standard deviation values indicate that the three groups vary in terms of homogeneity, with the middle schoolers being the least homogeneous and the university group the most homogeneous. Apart from that, the results show that the higher the study level, the higher the means tend to be, and vice versa. More specifically, the mean score of the university students approaches 13,000 words out of 14,000, that of the high schoolers less than 11,000, while that of the middle schoolers is less than 8,000. Given that the three levels are separated by two grades, the difference between the mean scores is quite expected and, therefore, the scores themselves can only be seen as reflective of the different amounts of vocabulary the participants have. The minimum and the maximum scores also increase along study level. Thus, while the minimum score recorded for the middle schoolers is barely 3400, the high schoolers exceeded that by almost 4000 words and that of the university students surpassed that of the second group by more than that figure. The maximum scores also exhibit the same tendency, although the difference between the scores recorded for the three groups is not as high as that between the minimum scores. In this respect at least, the test seems to fare well.

When these results are submitted to statistical analysis, the difference between the three groups turns out to be highly significant. This is indicated by the results of an independent sample ANOVA test displayed in the following table:

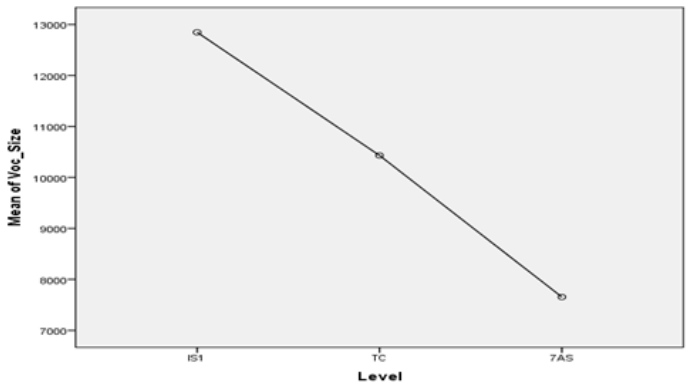
**Table 2:** Results of ANOVA

| | Sum of squares | df | Mean | F | Sig. |
|---|---|---|---|---|---|
| Between | 273815698,5 | 2 | 136907849,2 | 54.1 | .000 |
| Within | 177131150,9 | 70 | 2530445,01 | | |
| Total | 450946849,3 | 72 | | | |

As the significance value indicates, the difference between the three means cannot be the result of chance except at a probability level less than .001, which is highly significant. This fact is presented visually in Figure 1.

The curve exhibits a steep rise as we move from the lowest level, i.e. middle school, to the highest, i.e. university students. A post hoc analysis indicates that all group comparisons are also highly significant. These findings form a clear indication that language proficiency, as defined by study level, does have an impact on vocabulary knowledge and, more importantly, that the Arabic Vocabulary Size Test is capable of reflecting this development.

Figure 1: Comparison of group means




This remark holds not only for the test as a whole, but also for the individual bands. A comparison between the horizontal cells in Table 3 shows that practically all the mean scores increase as the study level increases:

**Table 3:** Group means for the 14 bands

| Band | Middle school | High school | University |
|---|---|---|---|
| K1 | 770,97 | 920,69 | 953,85 |
| K2 | 725,81 | 820,69 | 915,38 |
| K3 | 667,74 | 837,93 | 915,38 |
| K4 | 725,81 | 875,86 | 953,85 |
| K5 | 663,33 | 862,07 | 953,85 |
| K6 | 646,67 | 803,45 | 923,08 |
| K7 | 454,84 | 686,21 | 946,67 |
| K8 | 548,39 | 772,41 | 938,46 |
| K9 | 486,21 | 720,69 | 953,85 |
| K10 | 462,96 | 682,76 | 915,38 |
| K11 | 406,67 | 572,41 | 900,00 |
| K12 | 416,67 | 613,79 | 838,46 |
| K13 | 393,55 | 603,45 | 823,08 |
| K14 | 460,00 | 658,62 | 915,38 |

The mean scores in the first column representing the middle school group are all lower than those in the second column that represent the high school group, and these in turn are lower than those of the university group in the third column. These results are obviously in support of the robustness of the test, at least as far the pilot groups are concerned.

When an independent samples ANOVA test is run, the results turn out to be significant at a level less than .001 for all the 14 bands. An LSD post hoc analysis, however, indicates that not all pairwise comparisons are statistically significant. K1 for example shows that only the lowest group mean is statistically different while the other two are not. In comparison, K2

differentiates between the three groups while K3 again differentiates only between middle schoolers, on one hand, and high school and university students, on the other. Once we go beyond the fifth band, however, all pairwise comparisons become significant. This is an interesting finding since it could be interpreted as a piece of evidence supporting the robustness of the frequency list on which the Arabic Vocabulary Size Test was based.

A comparison of the 14 bands also indicates that their mean scores show some difference. Their overall scores are displayed in the following table:

**Table 4:** Descriptive Statistics

| Band | Mean | Std. Deviation |
|---|---|---|
| K1 | 877,61 | 143,359 |
| K2 | 808,96 | 150,485 |
| K3 | 792,54 | 214,132 |
| K4 | **825,37** | 197,977 |
| K5 | **807,46** | 188,546 |
| K6 | 767,16 | 195,702 |
| K7 | 637,31 | 269,588 |
| K8 | **723,88** | 252,916 |
| K9 | 688,06 | 235,824 |
| K10 | 649,25 | 250,075 |
| K11 | **567,16** | 267,643 |
| K12 | 585,07 | 257,761 |
| K13 | 580,60 | 223,769 |
| K14 | **635,82** | 281,070 |

A repeated measures ANOVA indicates that the differences between means is significant at a level lower than .001. But as the SPSS does not allow for the possibility of pairwise comparisons for this statistical procedure, we cannot know for sure which pairs of means are statistically significant and which are not. By eyeballing the data, however, some pairs seem distant enough to allow for a non-chance interpretation, while others appear to be too close for that kind of reading. But irrespective of statistical significance, the more important issue is how to interpret the difference itself.

One thing for sure is that frequency alone cannot explain all the differences. A quick look shows that, while the means tend generally to decrease from the most frequent to the least frequent bands, there are a few irregularities (indicated by bold type in Table 4 above). Between K3 and K4, for instance, there is a rise of more than 30 points, and between K7 and K8, the rise exceeds 80 points. K14 as well exhibits a rise of more than 50 points compared to the previous band. These kinks are rather unexpected, given the hypothesis that more frequent words tend to be learned before less frequent ones. K14 in particular appears to be odd in that its mean score is closer to that of K7 and higher than the means of K11 – K13. In other words, the test takers recognized almost as many words in K14 as in K7, and more words in

K14 than in K11 – K13. Therefore, although the role of frequency is undeniable, there must be other confounding factors interfering with the test takers' performance on the test.

It should be pointed out that such irregular patterns are not unfamiliar in vocabulary tests developed on the basis of word frequency. Studies on the acquisition of English vocabulary in EFL (English as a foreign language) and ESL (English as a second language) contexts have repeatedly reported that learners as groups sometimes perform better on low frequency bands than in higher frequency bands (for a review, see Milton, 2009). Aizawa (2006), for example, measured the receptive vocabulary size of Japanese university learners of English in Tokyo using a test that targets the most frequent 8000 words and concluded that, while they tend to decrease consistently down to the fourth band, the means become irregular in the other four bands. In fact, even regarding the first bands, many studies that used Meara's Yes/No test have reported an irregular pattern between the second and the third bands, with a rise in the third band significantly higher than the second (Milton, 2009). In the Moroccan context, Ait Hammou (2019) noted a similar pattern among 460 high school EFL learners that occurs consistently among the whole sample as well as among the different study level, program stream or gender sub-groupings. What can be concluded is that extraneous factors can sometimes weaken the effect of frequency but they never override it.

One such extraneous variable extremely important in the case of Arabic is the similarity between MSA and the colloquial. Arabic is well known for being a diglossic language in the sense that two distinct but related varieties of it can be identified: a written standard common to all Arabic-speaking countries and a colloquial that varies from one country to another and from one place to another even within the same country (cf. Ferguson, 1959). Because of widespread literacy among the population, especially those who have benefited from formal education, many speakers are able to make connections between standard forms and their colloquial equivalents and to switch easily and smoothly between the two varieties. In fact, this situation is so complex that some linguists have proposed to analyze the language as forming one continuum of styles ranging between the most classical used in very formal contexts and the most colloquial used in very informal contexts (cf. Badawi, 1973). Consequently, it is extremely difficult to determine word frequency in the language based on a single corpus, if by word it is meant both the standard form and its colloquial equivalent. The reason is that one variety is used basically in writing while the other is used mostly in speaking and only sporadically in the written form (see the papers in Høigilt and Mejdell (2017)). As a result, the standard and its colloquial equivalent can diverge widely in frequency. Given this fact, it is possible that a low frequency word in our list – assumed to represent the standard – can have a high frequency counterpart in the Moroccan Arabic colloquial, which could have stood behind the irregularities noted in Table 4 above. This issue, however, has not received due attention in the literature, but it cannot be overlooked in any future investigation into Arabic vocabulary learning. So far, very little is known about the effect of the colloquial on learning standard Arabic, and whether it is positive or negative.

A final note concerns the frequency list on which the Arabic Vocabulary Size Test is based. As explained in Section 2 above as well as in Ech-Charfi (2024), this list was based on another list extracted by Majdi Sawalha from the Arabic Internet Corpus. Therefore, I had neither the possibility nor the expertise to check the validity of the steps and procedures followed by Sawalha in the extraction process. Some remarks, however, suggest that the original list may not be flawless. More specifically, some items seem to have a higher frequency in the list than is indicated by a simple query for its concordances. For instance, the word ذل has a token frequency of 300891 according to Sawalha's list, but a query for its concordances returns only 269 matches, corresponding to 275.485 instances per million words. In comparison, the quantifier كل, which has a frequency of 289525 and, thus, lower in rank, returns 2193 matches, corresponding to 2245.868 instances per million words. Although less than two dozen such cases have been identified, their number cannot be determined with any certainty. Therefore, the usefulness of the list on which the vocabulary size test is based remains less satisfactory than we would have hoped.

Overall, the Arabic Vocabulary Size Test seems to work well enough to be used with native speakers, given the lack of better alternatives. In fact, it seems that the test can help identify the major factors affecting the development of vocabulary knowledge, many of which are not well understood and have gone unnoticed so far. Thus, the results it yields could be used as a basis for further research in this under-researched area of Arabic learning.

# 4 Conclusion

This paper has provided a description of an Arabic version of Nation and Beglar's (2007) Vocabulary Size test. The test uses a word list developed by Ech-Charfi (2024) on the basis of a lemmatized list extracted by Majdi Sawalha from the Arabic Internet Corpus and made available at the Leeds University website. The piloting of this Arabic version has shown that the test can differentiate between learners with various language proficiency and that the effect of frequency, though not decisive, is nonetheless noticeable.

The piloting of the Arabic Vocabulary Size Test has yielded a few interesting implications. Perhaps the most interesting of these implications is that learners at the primary and middle school levels still have relatively poor vocabulary size which would not enable them to read authentic texts with ease. Another implication is that knowledge of the colloquial apparently affects the use of MSA, as learners try to make connections between MSA words and their colloquial counterparts. While this could prove to be beneficial, it might also hide some harm if the counterparts are only partially equivalent or are false friends. These remarks call for contrastive studies of the two varieties as well as for a detailed investigation of the role of the colloquial in the process of learning MSA. All these and many more issues are still pending serious consideration on the part of researchers of Arabic.

# References

Ait Hammou, B. (2019). “The receptive vocabulary size of high school EFL learners in Morocco”. MA Thesis. Mohamed V University in Rabat (cit. on p. 33).

Aizawa, K. (2006). “Rethinking frequency markers for English-Japanese dictionaries”. In: *English Lexicography in Japan*. Ed. by M. Murata et al. aishukan-shoten, pp. 108–119 (cit. on p. 33).

Badawi, E. (1973). *Mustawayāt al-ʿarabiyya al-muʿāṣira fī miṣr (Levels of Modern Standard Arabic in Egypt)*. Dār al-Maʿārif (cit. on p. 33).

Ech-Charfi, A. (2023). “Word frequency and lexical coverage in English and Arabic”. In: *Journal of Applied Language and Culture Studies* 6.3, pp. 1–19 (cit. on p. 26).

— (2024). “Frequency and text coverage in Standard Arabic based on Arabic Internet Corpus”. In: *International Journal of Arabic Linguistics* 10.1, pp. 7–24 (cit. on pp. 25–27, 34).

Ferguson, C. A. (1959). “Diaglossia”. In: *Word* 15, pp. 325–340 (cit. on p. 33).

Høigilt, J . and G. Mejdell (2017). *The Politics of Written Language in the Arab World: Writing Change*. Brill (cit. on p. 33).

Milton, J. (2009). *Measuring Second Language Vocabulary acquisition*. Multilingual Matters (cit. on pp. 27, 33).

Nation, I. S.P. (2001). *Learning Vocabulary in Another Language*. Second edi. Cambridge University Press (cit. on p. 26).

Nation, P. and D. Beglar (2007). “A vocabulary size test”. In: *The Language Teacher* 31.7, pp. 9–12 (cit. on pp. 25–28, 34).