

Developing a vocabulary size test in Greek as a foreign language

James Milton^{*} and Thomaï Alexiou^{**}

University of Swansea^{}*
*Hellenic Open University^{**}*

Abstract

This paper presents a vocabulary size test in Greek as a foreign language derived from the first 5000 lemmatised words of the Hellenic National Corpus. It is designed as a Greek version of Meara and Milton's (2003) X-Lex checklist test which has test forms in a number of other languages. Such a test should provide useful information to both learners and teachers of Greek as a foreign language in an important area of language knowledge, and should help indicate both student level and measure learning progress. The test has been trialled on learners of Greek at CEFR A1, A2, B1 and B2 levels in the School of Modern Greek, Aristotle University of Thessaloniki. Results suggest that the test works reliably, has good construct validity and provides estimates of vocabulary size which are believable and fit well with estimates of vocabulary size at CEFR levels in other languages.

Keywords: vocabulary size test, Greek, CEFR, X-Lex, frequency

1. Introduction

Gaining a good knowledge of vocabulary is an essential tool for developing proficiency in a foreign language. In the foreword of the book *Modelling and assessing vocabulary knowledge*, Long and Richards (2007: xii) characterise vocabulary as “the core component of all the language skills”. Greece is probably not a country where this needs to be emphasised but in some countries, Britain being one of them, the value of vocabulary is doubted in modern foreign language teaching communities. The reason for the importance of vocabulary in developing proficiency is not hard to imagine. As Wilkins (1972: 111) puts it, “without grammar very little can be conveyed, without vocabulary *nothing* can be conveyed.”

In an important area of language knowledge such as this it would make sense for teachers, and for the learners themselves, to monitor knowledge of vocabulary in order to check their progress toward the goals of communication and to establish whether the level of knowledge required for formal exams has been achieved. Yet, until very recently, we have not had formal tools for measuring vocabulary knowledge capable of being equally used across a wide range of learners from different learning contexts and backgrounds. These tools are now emerging and the methodologies used are capable of being applied to other languages, apart from English. This opens up the useful

possibility of establishing a fairly standard way of estimating a learner's vocabulary knowledge which can work across different languages.

2. The study

The methodologies for estimating vocabulary knowledge in learners tend to be based on frequency counts, and they focus on the most frequently occurring vocabulary. There are good reasons for this, the first being that the most frequent vocabulary in a language tends (but only tends) to be learned earliest. It is good practice to focus a test on what is known rather than on what is not known. A second reason is that the most frequent lexis in a language features disproportionately in text coverage and this – how many words in a text one recognises and understands – can explain a lot about how much a learner will understand and can communicate, to others. A third reason is that frequency counts are textbook-neutral, so regardless of the specific material a learner has been exposed to, a test constructed in this way can identify the words they have learned. These tests appear to work well in practice (e.g., Meara and Milton's X-Lex test (2003)). They are very reliable, and give believable estimates of a learner's knowledge within the limitations they work.

One specific limitation of such tests is that they work on a sample of the most frequent vocabulary, often 5,000 or 10,000 words, and they can only form an estimate of knowledge within this range. Learners are likely to have some knowledge of words outside this range as course materials are thematically selected and cannot reasonably work within strict frequency limitations, so total learner vocabulary knowledge is underestimated.

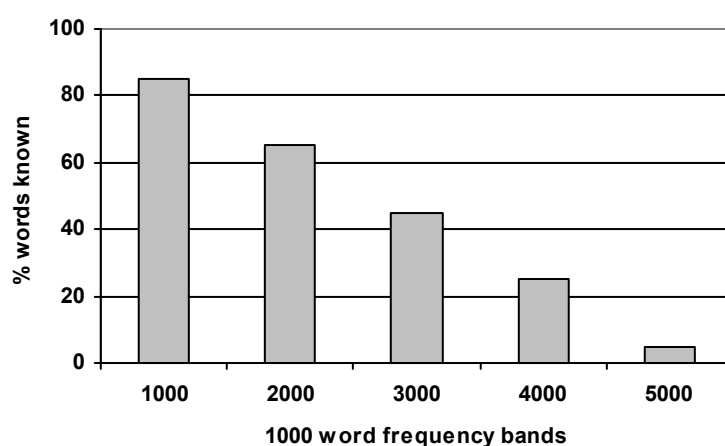
The test form we have used is illustrated in Table 1. A complete version of the Greek test is provided in the Appendix.

Table 1. Illustration of test format

| | | | | | |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|----------------------------------|----------------------------------|---------------------------------|----------------------------------|
| Please look at these words. Some of these words are real French words and some are invented but are made to look like real words. Please tick the words that you know or can use. Here is an example. | | | | | |
| <input checked="" type="checkbox"/> chien | | | | | |
| Thank you for your help. | | | | | |
| <input type="checkbox"/> de | <input type="checkbox"/> distance | <input type="checkbox"/> abattre | <input type="checkbox"/> absurde | <input type="checkbox"/> achevé | <input type="checkbox"/> manchir |

The test is a Yes/No test which asks the learner to identify the words they know from a selection of individually presented words. The words are selected, in this illustration, from across the most frequent 5000 lemmatised words in French. 20 words are taken from the first 1000 most frequent words, 20 from the second 1000 words, 20 words from the third, fourth and fifth band, so that an estimate of knowledge of those 5000 words can be made. In addition to these words there are 20 false words, words which are not real words, but follow all the rules of word formation and sound combinations in the language being tested; in the example French but in our case Greek. The false words allow an estimate to be made of over-estimation on the part of the learners. Remember that it is not always easy to be certain if you know a foreign language word or not, and, due to economy of practice, learners will tend to give themselves the benefit of the doubt. The false words allow for this kind of over-estimation to be calculated and the scores are adjusted accordingly. Real words which are correctly identified score 50, providing a basic score out of 5000. False words which are identified as real, result in a deduction of 250 from the basic score allowing for a more accurate estimate of vocabulary size to be made. These tests allow the way vocabulary is learned to be checked against a model of learning. Meara's 1992 frequency model is presented in Figure 1.

Figure 1. Meara's (1992) frequency model of vocabulary learning

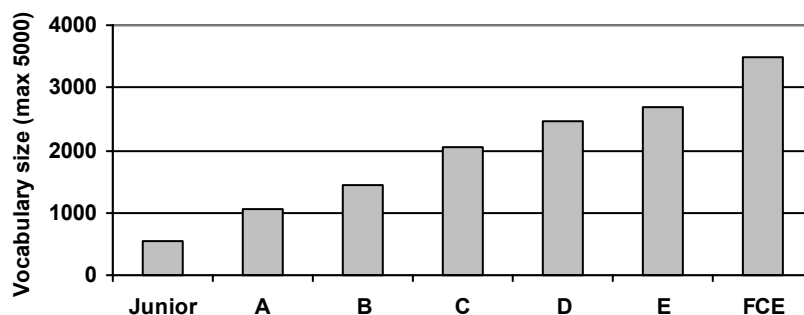


This suggests that learning will tend to be concentrated in the most frequent bands so if a learner knows 80% of the vocabulary in the most frequent 1000 word band, then he or she will know less in the second 1000 words band, still less in the third and fourth 1000-word bands, and so on. This gives a frequency profile, high on the left and sloping

downwards to the right. As learning progresses, this profile moves upwards, until it hits 100%, complete knowledge of a frequency band. And then a plateau emerges and learning begins to concentrate in the infrequent bands, so the profile moves to the right. Empirical evidence shows that this is, indeed, pretty much what happens in groups of learners, and among most individuals (Milton 2007).

Estimates of vocabulary knowledge made this way appear to give very credible results which conform to Meara's model. Figure 2 demonstrates vocabulary uptake, measured using a test of the most frequent 5000 words in English, across all the classes in a language school in Greece (Milton 2006).

Figure 2. Vocabulary growth among Greek learners of EFL (Milton 2006)



This suggests that about 500 words are learned every year in about 100 contact hours, and that learners take Cambridge FCE with about 3500 words word knowledge. Extensive work on the vocabulary content of the course books suggests that 500 words a year is actually a rather good estimate of average knowledge in the bands being tested. And 3500 words for FCE sits well with Hindmarsh's (1980) estimate of 4500 words as a requirement for this level; remember that learners will also have vocabulary knowledge outside the most frequent 5000 words.

Linking vocabulary size estimates with formal exams also allows vocabulary to be built into the Common European Framework of Reference for Languages (The Council of Europe 2003). The vocabulary sizes which emerge (taken from Milton and Alexiou 2009) are shown in Table 2 and are indicative, not being absolute requirements for each level, of course. Learners will vary in their vocabulary knowledge but also in what they can do with their knowledge; but these figures probably reflect a truth. It would be impossible to become, for example a very proficient C2 language user without having a

very sizable vocabulary – you would just be too limited in what you could understand and what you could say.

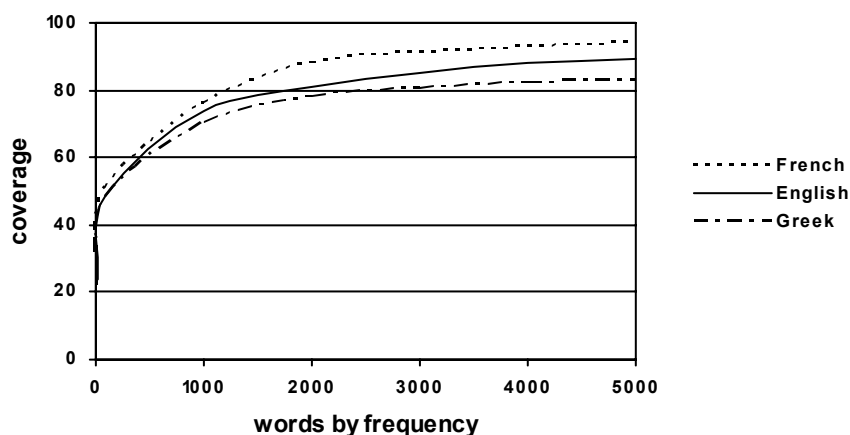
Table 2. Vocabulary size and the CEFR (Milton and Alexiou 2009)

| <i>CEFR level</i> | <i>XLex (5000 max)</i> | |
|-------------------|------------------------|---------------|
| | <i>English</i> | <i>French</i> |
| A1 | <1500 | 1160 |
| A2 | 1500 - 2500 | 1650 |
| B1 | 2750 - 3250 | 2422 |
| B2 | 3250 - 3750 | 2630 |
| C1 | 3750 - 4500 | 3212 |
| C2 | 4500 - 5000 | 3525 |

The vocabulary size associated with each level is slightly different in French, with English requiring greater vocabulary knowledge for the same level of communicative ability. This probably reflects differences between the two languages and the coverage which the most frequent vocabulary gives.

Our intention, therefore, is to create a vocabulary size test in Greek along lines that are equivalent to other tests which appear to work well and provide useful information to learners and teachers. To do this we have drawn on the Hellenic National Corpus which is a 9 million word corpus drawn from the Greek newspaper *Nea*, and more specifically from the cultural, sociological and sports sections.

To give us a workable frequency list to draw test items from, this corpus has had proper names and other items not relevant for teaching stripped out and it has been lemmatised. Inflections and derivations work rather differently in Greek derivations compared to English and French but this process brings the corpus into line with the English and French corpora, actually making them more similar, and allows a selection of frequent words to be made to form the Greek test. A frequency list was produced of the most frequent items and from this list a Greek test has been constructed similar to the English and French vocabulary size tests: 20 words selected from each of the five 1000 word bands, and 20 pseudo-Greek words. The coverage provided by words in the Greek frequency list is plotted up and shown in Figure 3.

Figure 3. Coverage from Greek, French and English frequency lists

It appears that the most frequent lexis of Greek is slightly less frequent than that of English (taken from Carroll, Davies and Richman 1971) or French (taken from Baudot 1992). In English the most frequent words provide large amounts of coverage; the most frequent 150 words give about 50% coverage in a typical text. About 2000 words give about 80% coverage and this seems to be an important figure because below this level of knowledge learners struggle for comprehension of texts and struggle to communicate, while beyond this level they begin to understand the gist of passages they read or hear, and can begin to be independent in communication. In French the line of coverage is higher than the English line which suggests that it is possible to understand and to do more with slightly fewer French words than it is in English. This may account for the lower French vocabulary size estimates at each CEFR level (Table 2). The Greek line is lower than the English line which suggests that more words in Greek might be required to attain the levels of communicability associated with each CEFR than is the case in English. While English might require only 2000 words for A2 level, in Greek maybe 2500 words would be needed. And while B1 in English requires about 3000 words, Greek might require 3500 words. This difference may be a product of the nature of the corpus we are drawing on, there is no spoken Greek element, for example, which would very likely boost the frequency levels of the most frequent words. But it may, also, simply be a feature of Greek.

3. Research questions

Our intention, having created a test for estimating vocabulary size in learners of Greek, is to see how well it works. In particular we intend to examine:

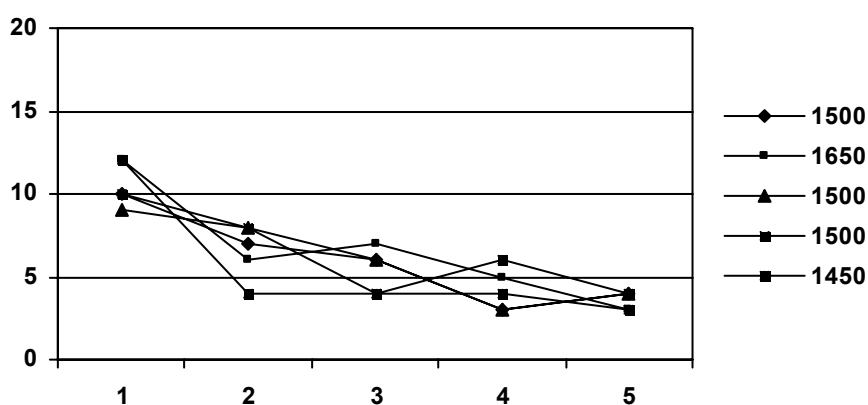
- (a) whether the test is reliable – and produces the same score for the same learner when knowledge has not changed
- (b) whether the frequency effects observable in other languages can be seen in Greek
- (c) whether the frequency profile changes with level and knowledge in the expected manner
- (d) whether the test differentiates between learners of different levels in predictable ways (and suggests vocabulary knowledge required for each CEFR level).

4. Results and discussion

4.1 Test reliability and frequency effects

A preliminary trial suggests that the test is reliable. A single learner took five different forms of the test (different words and false words) in an afternoon. He had little formal Greek training and has learned primarily from books and holidays. A high score was not expected and the results obtained are shown in Figure 4.

Figure 4. Test and retest scores and profiles for an individual test taker



It is clear from Figure 4, where the scores at each 1000 word frequency band level are recorded in addition to the overall vocabulary estimate, that something like the frequency profile emerges and that even where scores on each 1000 word level may vary slightly, the overall scores, on the right, are remarkably consistent. They suggest the learner is of a quite low level.

4.2 Vocabulary profile and level of ability

The study further tested a group of learners in Greece. 64 adult learners at a range of levels from A1 to B2 took the test. They came from different backgrounds and had been learning Greek as a foreign language in the School of Modern Greek, Aristotle University of Thessaloniki. The length of their studies ranged between one month and two years. They were tested around the end of October so they had just started the academic year.

Figure 5. Profiles for learners in CEFR bands A1, A2, B1 and B2

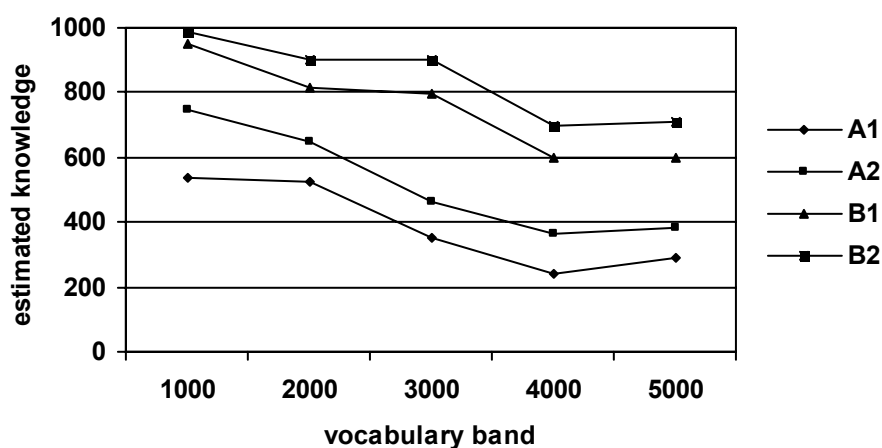
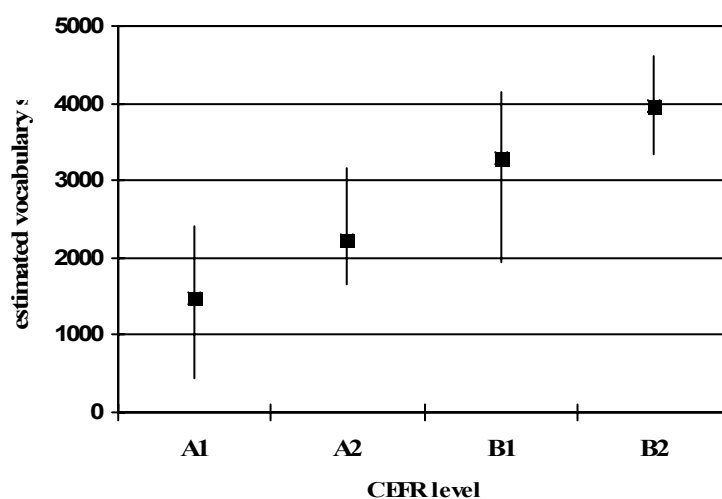


Figure 5 summarises the results drawn from students at each of the four CEFR bands. The average scores at each frequency band suggest that the frequency effect is present, as expected, and that the frequency profiles rise with progress as Meara's theoretical profile suggests. The profile even appears to be flattening out above 90%. The only variation is noticed in the size of the knowledge known in frequency band 5, suggesting knowledge of comparatively infrequent vocabulary. A Friedman test confirms the very strong trend in this data, the score on each successive frequency band being lower than the preceding one, $\chi^2 = 190.595$, $p < .001$. These results, conforming to theory, suggest the test has strong construct validity.

4.3 Vocabulary size and CEFR levels

Figure 6 presents the mean vocabulary size scores for learners at each of the four CEFR levels we have.

Figure 6. Mean, maximum and minimum vocabulary scores for learners in CEFR groups A1, A2, B1 and B2.

The mean scores increase with each level, as might be hoped, with A1 the lowest vocabulary knowledge and B2 the highest. The relationship between CEFR level and vocabulary size is strong and its statistical significance is confirmed by an ANOVA, $F(3) = 63.121$, $p < .001$. A Tukey test provides multiple comparisons between levels and this confirms that the differences between the scores at every level are statistically significant (Table 3).

Table 3. Multiple comparisons between means scores at CEFR levels.

| total | | | | | |
|--------------------------|----|------------------------|-----------|-----------|-----------|
| Tukey HSD ^{a,b} | | | | | |
| group | N | Subset for alpha = .05 | | | |
| | | 1 | 2 | 3 | 4 |
| 1.00 | 19 | 1486.8421 | | | |
| 2.00 | 12 | | 2237.5000 | | |
| 3.00 | 17 | | | 3288.2353 | |
| 4.00 | 16 | | | | 3956.2500 |
| Sig. | | 1.000 | 1.000 | 1.000 | 1.000 |

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.547.

b. The group sizes are unequal. The harmonic mean of the group sizes is used. Type I error levels are not guaranteed.

However, the neat regularity of the growth in mean vocabulary size scores, and the strong statistical relationship with the CEFR level, disguises the variation within each level and the degree to which learners at each level can overlap in their levels. It appears that there are learners at the top of the A1 group who might, in terms of their vocabulary

level, fit into the B1 class, and vice versa. These are teaching groups, after all, which have not been rigorously separated by level.

These figures suggest that in Greek as in English and French, rough vocabulary sizes can be associated with each successive CEFR level. It was suggested earlier that the figures at each level would be higher for English than for French. It has also been suggested that the figures for Greek would be higher than for English. When the vocabulary levels for Greek at each level are added to the chart for CEFR levels (Table 4), this is what appears to happen. 1500 words as a mean score seems to be at the top end of the estimate for A1 in English, as does 2237 at level A2. In B1 and B2 the size estimates exceed the English estimates.

Table 4. Vocabulary size in Greek and the CEFR

| <i>CEFR Levels</i> | <i>XLex (5000 max)</i> | | |
|--------------------|------------------------|---------------|--------------|
| | <i>English</i> | <i>French</i> | <i>Greek</i> |
| A1 | <1500 | 1160 | 1486 |
| A2 | 1500 - 2500 | 1650 | 2237 |
| B1 | 2750 - 3250 | 2422 | 3288 |
| B2 | 3250 - 3750 | 2630 | 3956 |
| C1 | 3750 - 4500 | 3212 | |
| C2 | 4500 - 5000 | 3525 | |

5. Conclusion

These results suggest that the frequency based vocabulary size test in Greek appears to work very successfully. The results confirm a strong frequency effect in Greek L2 vocabulary acquisition as anticipated in theory and as occurs in other foreign languages. The test successfully distinguishes between learners at different levels of the CEFR framework and appears to give credible figures for learners' level of vocabulary knowledge. These figures appear to mesh well with the predictions for vocabulary suggested by the coverage obtained from the frequency data.

This study is a first step in validating this testing tool and in order to confirm its reliability, we intend to carry out the majority of the tests at the end of the academic year. We also have some supporting evidence that by using coverage figures drawn from word frequencies, we can tie the CEFR levels to vocabulary sizes in a whole

variety of languages other than English, French and Greek. This should help make the CEFR system both more robust and more transparent.

Acknowledgements

We are grateful to Dr George Mikros for helping to provide the corpus and for carrying out much of the background work in generating lemmatised frequency lists. We are also grateful to Mrs Martha Vazaka and her colleagues at the School of Modern Greek, Aristotle University of Thessaloniki for kindly providing access to their classes as well as the learners themselves for their willingness to participate in our study.

References

- Baudot J. (1992). *Fréquences d'utilisation des mots en français écrit contemporain*. Montréal: Les Presses de l'université de Montréal.
- Carroll J.B., P. Davies and B. Richman (1971). *The American heritage word frequency book*. Boston: Houghton Mifflin.
- Hindmarsh R. (1980). *Cambridge English lexicon*. Cambridge: Cambridge University Press.
- Long M.H. and J.C. Richards (2007). Series Editors' Preface. In H.Daller, J.Milton and J.Treffers-Daller (eds), *Modelling and assessing vocabulary knowledge*. Cambridge; Cambridge University Press, xii-xiii.
- Meara P. (1992). *EFL vocabulary tests*. University College Swansea: Centre for Applied Language Studies.
- Meara P. and J. Milton (2003). *X_Lex, The Swansea levels test*. Newbury: Express.
- Milton J. (2006). X_Lex: the Swansea Vocabulary Levels Test. In C. Coombe, P. Davidson and D. Lloyd (eds), *Proceedings of the 7th and 8th Current Trends in English Language testing (CTELT) Conference* vol 4. UAE: TESOL Arabia, 29-39.
- Milton J. (2007). Lexical profiles, learning styles and construct validity of lexical size tests. In H.Daller, J.Milton and J.Treffers-Daller (eds), *Modelling and assessing vocabulary knowledge*. Cambridge; Cambridge University press, 45-58.
- Milton J and T. Alexiou (2009). Vocabulary size and the Common European Framework of Reference in Languages. In B.Richards, H. Daller, D. Malvern, P. Meara, J. Milton and J. Treffers-Daller (eds), *Vocabulary studies in first and second language acquisition*. Palgrave: Macmillan, 194-211.
- The Council of Europe (2003). *Common Framework of Reference for Languages*. Cambridge: Cambridge University Press.
- Wilkins D.A. (1972). *Linguistics in language teaching*. London; Arnold.

Appendix

Greek vocabulary size test examples; a note on testing and scoring

The test presented here contains 20 words randomly selected from each of the first five 1000 word frequency bands found in the lemmatised version of the Hellenic National Corpus. It also contains 20 false words which are designed to allow the amount of over-estimation which any forced answer test produces to be calculated and the scores adjusted. To make the contents clear to users, words from the first 1000 word band are presented in column 1 of the test. Words from the second 1000 word band are presented in column 2, and so on. False words are presented in column 6. It is usually good practice to randomise the presentation of test words.

To enable an estimate of the words a learner knows out of the most frequent 5000 lemmatised words in Greek, scoring is as follows. Award 50 for each real word checked by the testee and total these to produce a *raw score*. From the *raw score* deduct 250 for each false word which is checked to produce an *adjusted score* and the estimate of words known.

Greek vocabulary size test 1

Please look at these words. Some of these words are real Greek words and some are invented but are made to look like real words. Please tick the words that you know or can use. Here is an example.

αστυνομία ✓

Thank you for your help.

| | | | | | |
|---------------|---------------|----------------|--------------|---------------|--------------|
| ο | ανάλογος | ρούχο | περιγραφή | αστυνόμος | αλογομένη |
| μόνο | έπειτα | πίνω | υπερβολή | προβλεπόμενος | νερολός |
| μικρός | διεύθυνση | προτεραιότητα | συγκρότηση | ισόβιος | βιλός |
| πόλη | πλησιάζω | βίος | ψωμί | υπότιτλος | κολίμας |
| προσπαθώ | ανά | μωρό | βεβαιότητα | χριστιανικός | απέριος |
| βέβαια | ξεχωρίζω | απόπειρα | σκάζω | διακύμανση | ζόλος |
| αλλάζω | αδυναμία | χτύπημα | μελωδία | εποπτεία | φελί |
| διαθέτω | προοπτική | έξοχος | φράγκο | πρωτοτυπία | τρε |
| εκατομμύριο | προκριματικός | ιδίως | ιθύνων | κατάμεστος | τέτριο |
| όριο | καθορίζω | μήνυση | ξυλένιος | πρωτοπορία | σερό |
| βγάζω | μελετώ | ασφαλής | εμμονή | επιμελητής | στρίμα |
| συγκεκριμένος | μόνιμος | προσόν | αντίληπτός | μεραρχία | σκελικά |
| εις | κάμερα | προορισμός | εορτή | δημοτική | ματριτάκι |
| νιώθω | συμβάλλω | εργάτης | τιμόνι | ξαφνιάζω | αρχεότητος |
| υποβάλλω | πλανήτης | εμπνέω | εφικτός | μετανάστευση | τραπεζόλ |
| κύκλος | ρεύμα | δρομολόγιο | μετανιώνω | κινητήρας | γεραντοπολίο |
| ουσιαστικά | προσωπικά | συντήρηση | αποδοκιμασία | επιλύνω | δενερή |
| σπουδαίος | πλάνο | ικανοποιητικός | υπονομεύω | επιζητώ | τεποταπολίο |
| σωστός | στρατηγική | ψυχρός | λεσχη | ρητορικός | μυχανίο |
| πειθω | σπάνια | χειλος | διγάζω | αφοσιώνομαι | βατορά |