The Lexical Characteristics of Specialized Texts

Simon Fraser
Hiroshima University

Reference Data:

Fraser, S. (2005). The Lexical Characteristics of Specialized Texts. In K. Bradford-Watts, C. Ikeguchi, & M. Swanson (Eds.) *JALT2004 Conference Proceedings*. Tokyo: JALT.

In this study, Chung and Nation's (2003) rating scale methodology was used to identify technical terms in pharmacology and applied linguistics textbooks. It was found that around 35% of the running words in the pharmacology text and 15% in the applied linguistics text were technical in nature, figures which imply that technical vocabularies are much larger than had previously been supposed. The finding that more than half of the total word types in the pharmacology text are technical (considerably more than Chung and Nation's figure of 37.6% for anatomy) suggests that there may be substantial differences between sub-disciplines within a particular field. The study also helps us towards a better understanding of the different kinds of words that make up a technical vocabulary, and shows that many of the most frequent technical terms are in fact common words with a "hidden" technical meaning ("cryptotechnical" words).

本研究では薬理学と応用言語学の術語を識別するため、ChungとNation（2003）による4ステップ分類型評価スケールを使用した。薬理学のテキストにおいてはラニングワードの約35%が、応用言語学のテキストでは約15%が実際専門用語とみなされており、この数値は専門的語彙の占める割合が、以前考えられていたものより相当大きいことを示している。薬理学テキストのワードタイプの半分以上が専門用語であるという結果（Chungと　Nationの出した解剖学の場合の37.6%に比べるとかなり大きな割合となる）は、医学分野においても異なる領域でかなりの差が存在しうることを示唆している。本稿では専門的語彙を構成する単語の特徴も解明しており、最も頻度の高い術語は、実はごく一般的な単語に専門的な意味が「隠れている」タイプのものであるという事実を明らかにしている。

Recent years have seen an increasing interest in the role played by the specialized vocabulary needed for academic study (e.g. Chung & Nation, 2003; Coxhead, 1998; Sutarsyah, Nation & Kennedy, 1994; Ward, 1999). A great deal of attention has been paid to academic vocabulary, which comprises those words occurring frequently across a range of academic texts. A general academic vocabulary has been identified by Coxhead (2000), whose 3.5 million word Academic Word List provides good coverage of a wide variety of academic texts. For learners with more specific goals, however, knowledge of the technical terms associated with a specific discipline will also be necessary.

A number of studies have attempted to identify technical vocabulary. One way of doing this is to rely on intuitive judgment and use a semantically-based rating scale (e.g. Baker, 1988, Farell, 1990). Medical science, well-known for its use of words which are incomprehensible to the layperson, has received particular attention. The specialized vocabulary of medicine has been defined by Salager (1985, p.6) as "those high-frequency, context-bound, or topic-dependent, terms particular to a given medical specialty". Chung and Nation (2003) used a rating scale approach to identify this kind of vocabulary in an anatomy textbook, and by way of comparison carried out a similar study with an applied linguistics text. Their results are of special interest, as they suggest that the size and importance of technical vocabulary may have been seriously underestimated in the past.

Chung and Nation's study

The texts selected for analysis by Chung and Nation were *Clinically Oriented Anatomy* (Moore & Dalley, 1999) and *Learning a Second Language through Interaction* (Ellis, 1999). These were chosen because Chung has tertiary qualifications in both fields and was thus able to use her specialist knowledge in the classification of words.

Chung and Nation developed a four-point rating scale for identifying technical terms. They found that technical vocabulary made up a very substantial proportion of both the different words (types) and the total running words (tokens) in the texts. Figure 1 shows the distribution of high frequency words, words found in the academic word list (AWL), technical words, and low frequency words. High frequency words here are those found in the most frequent 2,000 word list, which is based on West's (1953) *A General Service List of English Words.* One in every three running words in the 450,000 word anatomy text, and one in every five in the 93,445 word applied linguistics text was found to be a technical word. For both texts, the proportion of technical vocabulary was far higher than Nation's (2001, p.12) estimate that typically only 5%, or one in twenty words, is technical.

Chung and Nation's study, then, is an important one, implying as it does that technical vocabulary plays a far more significant role in specialized texts than had hitherto been realized. It is, however, an analysis of the words in only two textbooks, and more data is of course required before we can reach any definite conclusions. We therefore need to know whether Chung and Nation's methodology can be successfully applied by other researchers, and if their findings can be replicated. The present study attempts to determine this by using a similar rating scale procedure in an analysis of pharmacology and applied linguistics textbooks. In addition, the study expands on previous research by suggesting new categories for the different kinds of words that make up a technical vocabulary, and by examining some of the characteristics of these words.

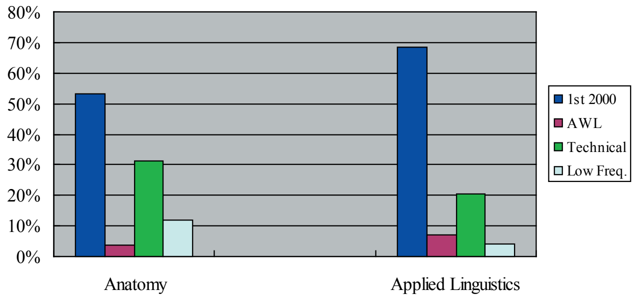

Figure 1. Running words (tokens) in anatomy and applied linguistics texts

The Present Study

Procedure

The following texts were scanned and saved in Text format in preparation for computer analysis:

- *Medical Pharmacology at a Glance* (Neal, 2003) (No. of running words = 58,413)
- *Issues in Applied Linguistics* (McCarthy, 2001) (No. of running words = 56,998)

Nation's RANGE computer program (available at http://www.vuw.ac.nz/lals/) was applied to count word types and tokens. The program could also determine which words are found in the most frequent 1,000 words of English, the second most frequent 1,000 words, and the 570-word family academic word list. A four-point rating scale similar to the one devised by Chung and Nation was used to extract pharmacological terms (see Table 1). Chung and Nation used the term "Step" to label the categories of their scale, but I felt this could be misleading, and to avoid confusion "Category" is used here. The same kind of scale was applied to the applied linguistics text. Words at Categories 3 and 4 were considered to be technical.

Pharmacology was chosen for analysis because this researcher has a degree in the subject, and, like Chung, felt that the specialist knowledge could be of use. Both the pharmacology and applied linguistics texts were considered to be comprehensive introductions to their respective fields, and they were both of similar length, at just under 60,000 running words each.

Table 1. A rating scale for identifying technical words (as applied to the pharmacology text)

Category 1 (C1) Words with a meaning that has no particular relationship to the field of pharmacology (e.g. *probably, differences, breakfast*).
Category 2 (C2) Words with a meaning minimally related to the field of pharmacology (e.g. *water, body, life*). Such words may be related to the body, or used when describing the actions and effects of drugs, and include terms used in the broader scientific/medical field or hospital environment.
Category 3 (C3) Words with a meaning closely related to the field of pharmacology (e.g. *transmitter, malignant, artery*). These refer to body organs, maladies and medical conditions, the actions and effects of drugs, etc., and are also used in general language.
Category 4 (C4) Words with a meaning specific to the field of pharmacology and not likely to be known in general language (e.g. *stenosis, warfarin, presynaptic*).

Inter-rater reliability

An inter-rater reliability check was carried out to ensure consistency in the use of the rating scale and that other researchers could apply it reliably. Sixty randomly selected words, fifteen from each of the four categories, were provided for the rater, an experienced university EFL teacher, to analyze. Table 2 shows that of the 15 words assigned to each of the four categories by the researcher, the rater agreed on the assignment of 15 items out of 15 at C1, 14 out of 15 words at C2, 12 out of 15 words at C3, and 11 out of 15 words at C4. The total agreement score is therefore (15+14+12+11) = 52 out of 60 (i.e. an accuracy score of 0.87). There was disagreement on four of the words at C4, but the rater assigned all of these to C3; that is, they were still categorized as technical words.

Table 2. Inter-rater reliability accuracy score calculated by the number of words assigned to the four categories by the rater and the researcher

Categories chosen by the rater	Categories chosen by the researcher				Total words assigned by the rater
	1	2	3	4	
1	15	1			16
2		14	2		16
3			12	4	16
4			1	11	12
Total words assigned by the researcher	15	15	15	15	Accuracy score =(15+14+12+11)÷60 = 0.87

Results

Comparison of the Applied Linguistics Texts

In Figures 2a and 2b, we see very similar patterns for both applied linguistics texts, with around 15% of the tokens in the McCarthy text and 20% in the Ellis text being technical. (Here, and elsewhere in the study, technical vocabulary has been removed from the GSL and AWL when calculating these.) For word types, the two texts correspond even more closely, with identical figures for technical words (16.3%).

The fact that the results are essentially the same as Chung and Nation's for a text in the same field indicates that the rating scale methodology is reliable, and it is very helpful to establish this before looking at any differences between the pharmacology and anatomy texts.

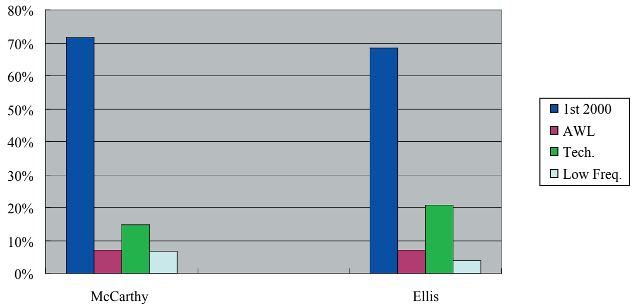

Figure 2a. Applied linguistics (tokens)

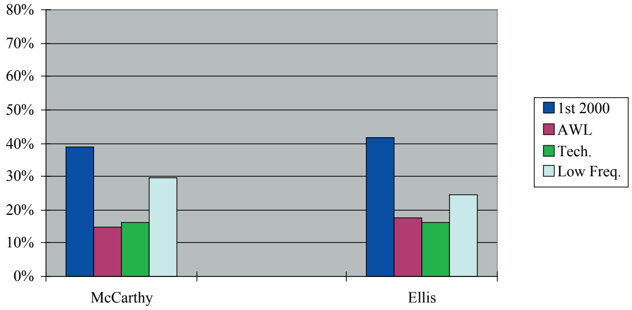

Figure 2b. Applied linguistics (types)

Comparison of Pharmacology and Anatomy texts

Figures 3a and 3b reveal similarities between the two medical texts, but it is the differences that are particularly striking. We see that 35.9% of the tokens in the pharmacology text and a similar 31.2% of the tokens in the anatomy text are technical words. However, although these figures correspond quite closely, it is apparent that pharmacology has a far greater proportion of technical word types than anatomy (55.4% v. 37.6%). A possible reason for this is the length of the texts: the pharmacology text (58,000 words) is much shorter than the anatomy text (450,000 words), and it aims to give a succinct account of the subject, eschewing the case studies and discussion found in the anatomy text. We might expect a more concise text to have less elaboration and repetition, resulting in a higher ratio of technical to non-technical types. Even so, it is surprising that the proportion of technical word types should be so much higher in one medical text than another, particularly as text length did not appear to be a factor in our comparison of the applied linguistics texts. Next, we look at the different categories of technical words (C3 and C4), to see if we can shed any further light on this difference between the pharmacology and anatomy texts.

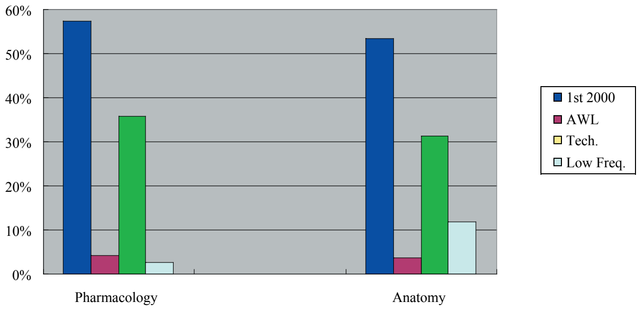

Figure 3a. Pharmacology v. anatomy (tokens)

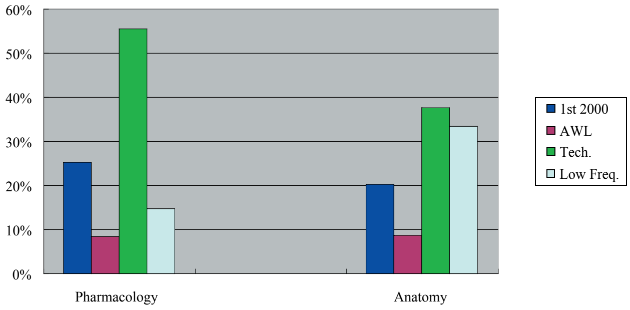

Figure 3b. Pharmacology v. anatomy (types)

Comparison of C3 and C4 Technical Vocabulary Types (Pharmacology and Anatomy)

Figure 4 shows that although the figures for C3 types are very similar for both subjects, the proportion of C4 types is much higher in pharmacology (42.5%) than anatomy (24.2%). It was initially thought that misspelled words (and words with alternative spellings) missed in the initial editing of the text might be artificially inflating the number of technical terms. However, a close examination of the frequency lists showed that there were very few instances of these.

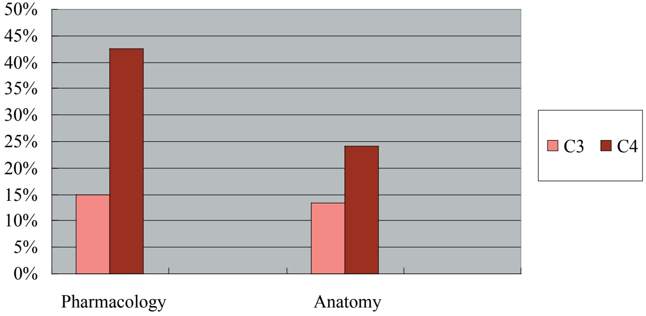

Figure 4. Pharmacology and anatomy C3 v. C4 technical types

Pharmacology C3 and C4 Types and Tokens

Figure 5 shows the percentages of types and tokens for both C3 and C4 technical words. Immediately apparent from the graph is the high proportion of C4 types, although coverage by C3 and C4 words is very similar. This would suggest that there is a good deal more repetition of C3 words than C4 words, and prompts us to investigate the frequency patterns of the two different categories of technical vocabulary.

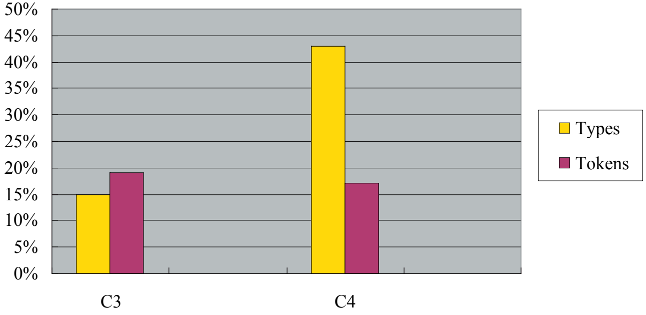

Figure 5. Pharmacology C3 and C4 types and tokens

A Closer Look at Pharmacology C3 and C4 Words

In order to examine more closely the characteristics of the different categories of technical vocabulary, twenty words were chosen at random from each of the C3 and C4 lists obtained from the corpus (see Table 3). What is most striking is that the C3 words occur on average with a much higher frequency than the C4 words, many of which are the names of drugs and are found only once or twice in the corpus. In the pharmacology text we find quite extensive lists of alternative drugs used in a particular treatment, and this is obviously contributing to the high proportion of technical types. The other observation that we can make from the table is that there are two clear sub-categories of C3 words: words such as *bacteria, fever*, and *heart*, which are clearly medical terms but are likely to be known by the layperson; and words like *dependence*, *transmitter*, and *inhibits* which occur in general language but are used in a pharmacological sense here. We might label the former category "lay-technical", and the latter, with their technical meaning in a sense hidden, "cryptotechnical". These differences are summarized in Table 4.

Table 3. A list of randomly chosen pharmacology C3 and C4 words

Typical C3 Words		Typical C4 Words	
	Frequency		Frequency
CELLS	104	HEPATOTOXIC	6
HEART	99	TRINITRATE	6
CHANNELS	79	RIBOSOMES	5
ACTIONS	77	LIPIDS	4
BLOCK	54	OOCYTE	4
FAILURE	51	PALLIDUS	4
INHIBITS	46	VINCRISTINE	4
SYMPATHETIC	40	GONADOTROPHINS	3
LOCAL	39	NORADRENALINE	3
TRANSMITTER	37	FLUCONAZOLE	2
DEPENDENCE	34	METHADONE	2
ELIMINATION	29	SULPHYDRIL	2
NERVOUS	27	ACETYLSALYCYLIC	1
WITHDRAWAL	25	BISPROLOL	1
BACTERIA	22	CYCLOPROPANE	1
COMPOUNDS	16	DISOPYRAMIDE	1
FEVER	15	EURYTHROID	1
STIMULATED	10	IONISATION	1
MESSENGERS	7	KETONE	1
RELAXATION	4	POLYPEPTIDE	1

Table 4. Some characteristics of pharmacology C3 and C4 words

C3 Words	C4 Words
Make up only 15% of the total word types	Make up almost half of the total word types
Typically occur with a relatively high frequency	Typically occur with a relatively low frequency
Types: tokens = approx. 1:11 (c.f. applied linguistics = 1:10)	Types: tokens = approx. 1:3 (c.f. applied linguistics = 1:5)
Can be divided into two categories: 1) CRYPTOTECHNICAL: words with a "hidden" technical meaning 2) LAY-TECHNICAL: words which can be considered technical, but are known in general language	Are often "hapax legomena" (occurring only once in the text), many of which are the names of drugs

Technical words from the high frequency word lists and the AWL

If a typical C3 word is one which is likely to be known in ordinary English, then we would expect many words in this category to come from the high frequency word lists. Tables 5 and 6 show that this is the case, with a large number of C3 technical words being found in the first three lists (the most frequent 1,000 words, the second most frequent 1,000 words, and the AWL). In the pharmacology text, the proportion of C3 types found in these three lists is 23.9%; the proportion of C3 tokens is 38.5%. For applied linguistics, the corresponding figures are 47.4% and 86.1%. Clearly, many of the technical words in both disciplines are common in ordinary English, with a particularly large proportion of the specialist vocabulary in applied linguistics consisting of words familiar to the layperson.

Table 5. Pharmacology: Technical vocabulary from the first three lists

Word list	% of C3 types	% of C3 tokens
Most frequent 1,000 words	4.4%	15.5%
2nd 1,000 most frequent words	8.7%	10.6%
Academic Word List	10.8%	12.4%

Table 6. Applied linguistics: Technical vocabulary from the first 3 lists

Word list	% of C3 types	% of C3 tokens
Most frequent 1,000 words	12.4%	42.6%
2nd 1,000 most frequent words	7.1%	16.1%
Academic Word List	27.9%	27.4%

Discussion

This study supports Chung and Nation's finding that a technical vocabulary can be very large. Similar results to theirs were obtained for coverage by technical words in a medical text, with around one third of the running words being technical. However, there was a clear difference between anatomy and pharmacology regarding word types, with technical words making up a substantially higher proportion of total word types in the pharmacology text. We mentioned that this may be an effect of the type and length of the texts used; nevertheless, it does suggest that there may be considerable variation not only between different technical fields but also between sub-disciplines within a particular field. Of course, categorization by means of a rating scale is a necessarily subjective process, but the fact that the results for the applied linguistics text were very similar to Chung and Nation's suggests that the methodology is reliable and reproducible.

The rating scale method is not without its drawbacks, however, and probably the most problematic of these is that it is extremely time-consuming. In this study, each word-type in the lists produced by the RANGE program had to be categorized, and there are well over 6,000 word-types in both texts. Needless to say, it was impossible to look at all of these words in context. In most cases, the researcher could be reasonably certain of how the words were being used in the text, but there were probably some technical uses of common words that were missed. Also, it may be a mistake to assume that cryptotechnical words are always used in a technical sense. In the pharmacology text, for instance, the word "buffer" did not refer to a kind of chemical solution, as might be assumed; rather, it was being used with its more familiar meaning "to cushion against". Multiword units, too, pose problems: the words "action" and "potential" often occur as the compound "action potential", but this is not evident from the word lists.

One of the main findings is that medical texts contain a far greater proportion of technical words than applied linguistics (prompting the doubters amongst us, perhaps, to question the status of applied linguistics as a science!). However, when considering the "difficulty" of a text, the situation is not as straightforward as it might first appear, as it is apparent that the applied linguistics text contains a sizeable proportion of low frequency vocabulary (almost one third of total types), which will certainly present learners with difficulties.

This study has also shed some more light on the characteristics of the different kinds of words that make up a technical vocabulary. The type-token ratios of C3 and C4 words in the pharmacology text, for instance, show that in general cryptotechnical and lay-technical words occur with a much higher frequency than those technical words which are specific to the field. Because of the coverage that C3 words provide, allocating more classroom time to this category of words would surely be justified.

It has been suggested (e.g. Laufer, 1989) that knowledge of 95% of the word tokens in a text is sufficient to allow reasonable comprehension. However, this may be misleading, as we have seen that many of the common words that learners apparently "know" are used quite differently in a medical text. Other suggestions concerning vocabulary size should also probably be treated with caution: Nation (2001, p.147) for instance, asserts that 95% coverage in an academic text can probably be achieved by knowledge of the most frequent words, the AWL, and just over 1,000 technical and low frequency words. What this study has shown, however, is that learners need a much larger vocabulary than this.

Conclusion

This research has demonstrated the usefulness of the categories proposed in Chung and Nation's (2003) study, while offering some suggestions for alternative categories which more clearly represent the different kinds of technical vocabulary. The findings have helped to clarify what is meant by "technical" vocabulary in relation to words at other frequency levels and given us a better idea of the size and importance of such a vocabulary. The study has also highlighted the need for both intuition and subject knowledge in the interpretation of the results of computer-generated data.

It must be borne in mind, of course, that we are drawing conclusions from studies of a very small number of texts. Much larger corpora (and an infinite amount of time!) will be needed before we can state with certainty the differences between disciplines and arrive at a definitive technical vocabulary for a particular field. What is clear from the limited data, however, is the size of the challenge that EAP learners face in learning vocabulary over and above the high-frequency and academic categories.

In addition to working with larger corpora, future studies might also usefully look in more detail at the nature of technical vocabulary. For example, we already know that many academic words are of Latin or Greek origin, but what are the physical characteristics (word length, number of syllables, and so on) of the different categories of technical words? We have seen that cryptotechnical words have a "hidden" technical meaning in addition to their commonly-known, "core" meaning. Does the relationship between the core and the technical meaning of a cryptotechnical word relate to the difficulty of learning that word? And what are the differences in the way that native and non-native speakers respond to and use these words? Answers to questions such as these would go some way towards furthering our understanding of how technical vocabulary is learned and how it might best be taught.

References

Baker, M. (1988). Sub-technical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles. *Reading in a Foreign Language 1* (1), 54-64.

Chung, T.M., & Nation, I.S.P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language 15* (2), 102-116.

Coxhead, A. (1998). *An academic word list.* English Language Institute Occasional Publication Number 18. Wellington: Victoria University of Wellington.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34* (2), 213-238.

Ellis, R. (1999). *Learning a Second Language through Interaction.* Amsterdam: John Benjamins.

Farell, P. (1990). *Vocabulary in ESP: a Lexical Analysis of the English of Electronics and a Study of Semi-technical Vocabulary*. CLCS Occasional Paper 25, Dublin: Trinity College.

Laufer, B. (1989). What percentage of text-lexis is necessary for comprehension? In C. Lauren & M. Nordman (Eds.), *Special Language: From Humans Thinking to Thinking Machines*. (pp. 316-323). Clevedon: Multilingual Matters.

McCarthy, M. (2001). *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.

Moore, K.L., & Dalley, A.F. (1999). *Clinically Oriented Anatomy* (4th ed.). Philadelphia: Lipincott, Williams & Wilkins.

Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Neal, M.J. (2003). *Medical Pharmacology at a Glance*. Oxford: Blackwell Science.

Salager, F. (1985). *Specialist Medical English Lexis: Classificatory Framework and Rhetorical Function – a Statistical Approach. EMP Newsletter 2* (2), 5-17.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus-based study. *RELC Journal,* 25, 34-50.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language 12*, 309-323.

West, M. (1953). *A General Service List of English Words*. London: Longman, Green & Co.