The New General Service List Version 1.01: Getting Better All the Time

Charles Browne
Meiji Gakuin University, Tokyo, Japan

The purpose of this brief paper is to explain a bit further about the New General Service List (NGSL), as well as to give some initial comparisons in text coverage between the General Service List (GSL), the NGSL, and the Other New General Service List (ONGSL) for a range of different text types.

INTRODUCTION

The General Service List (GLS) is a list of about 2,000 of the most commonly used words based on a corpus of written English compiled by West (1953). Although the original GSL was a remarkable, pre-computer era, corpus-derived list of important high-frequency words for second language learners that has been used for more than 60 years, the corpus it was based on is now considered to be quite dated (most words were published in the 1800s to early 1930s), small by modern standards (the original analysis was done with a corpus of only 2.5 million words), and in need of a clearer definition of what constitutes a "word" within the list.

In February 2013, on the 60th anniversary of West's publication of the GSL, my colleagues and I put up a website (www.newgeneralservicelist.org) that released a major update of West's GSL known as the NGSL. This list was derived from a carefully selected 273 million-word subsection of the 2-billion-word Cambridge English Corpus (CEC). The 1.0 version of the NGSL was then published in several journals, including the July issue of *The Language Teacher* (Browne, 2013).

Following many of the same steps that West and his colleagues did (as well as the suggestions of Professor Paul Nation, project advisor and one of the leading figures in modern second language vocabulary

acquisition), we did our best to combine the strong, objective scientific principles of corpus and vocabulary list creation with useful pedagogic insights to create a list of approximately 2,800 high-frequency words that met the following goals:

1. To update and expand the size of the corpus used (273 million words) compared to the limited corpus behind the original GSL (about 2.5 million words) with the hope of increasing the validity and the ability to generalize the list.
2. To create an NGSL of the most important high-frequency words useful for second language learners of English, which gives the highest possible coverage of English texts with the fewest words possible.
3. To make an NGSL that is based on a clearer definition of what constitutes a word.
4. To be a starting point for discussion among interested scholars and teachers around the world with the goal of updating and revising the list based on this input (in much the same way that West did with the original interim version of the GSL)

Unbeknownst to us, about six months after we released the 1.0 version of the NGSL, another General Service List was put out by Brezina and Gablasova (August, 2013), which I will refer to as the Other New General Service List (ONGSL) in order to avoid confusion. Although the ONGSL looks to be a very carefully constructed and impressive piece of research, the purpose of their list and the way it was developed seems to have a slightly different focus than what we undertook for the NGSL presented here. The authors of the ONGSL state that they used a purely quantitative approach to try to identify high-frequency words that were common across several different corpora, two of which were hugely different in size (1 million words for the LOB and BE06 corpora, 100 million for the BNC, and 12 billion words for the En Ten Ten 12 corpora) and resulted in the identification of 2,494 lemmas (according to their way of counting).

Our own NGSL project has been more directly focused on the needs of second language learners and teachers, and started with a selection of sub-corpora that were carefully balanced in size so as to avoid one corpus or type of text dominating the frequencies (which appears to be a real problem in the ONGSL) and, just as with the original GSL, our NGSL project employed both quantitative as well as qualitative methods

to attempt to identify the words that are most useful to the needs of language learners while providing the highest possible coverage.

Like the original GSL, which was released to the public in 1936 as an interim list, one that was revised and refined for more than 17 years before being published as the GSL in 1953, so too, our NGSL list should be seen as one that is still in its interim stages, released to the public in evolving versions (with 1.01 being the latest) and through various venues, including conferences, research papers, the web, and social media, with the hope that the list will be used, discussed, debated, and improved over time.

THE NGSL: A WORD LIST BASED ON A LARGE, MODERN CORPUS

One of the obvious axioms of corpus linguistics is that any word frequency list generated from a corpus will be a direct reflection of the texts in that corpus. In the case of the original GSL, there are many words on the list, which were arguably useful for second language learners of the time, but seem a bit dated for the needs of today's learners. For example, the GSL contains many nautical terms (oar, vessel, merchant, sailor, etc.), agricultural terms (plow, mill, spade, cultivator, etc.), religious terms (devil, mercy, bless, preach, grace, etc.) as well as many other terms that seem less likely to occur frequently in texts that the modern second language learner would likely use in the classroom (telegraph, chimney, coal, gaiety, shilling, etc.). As much as my colleagues and I were in awe of how much West was able to accomplish without the benefits of computers, digital text files, scanning equipment, or powerful corpus analysis software, we felt that the GSL was long overdue for an update and hoped that the application of modern technology to a more modern corpus could result in an NGSL that offered better coverage with fewer words.

Cambridge University Press offered us full unrestricted access to the Cambridge English Corpus (CEC), a multi-billion word corpus that contains both written and spoken text data for British and American English, as well as the Cambridge Learner Corpus, a 40-million-word corpus made up of English exam responses written by English language learners. They furthermore agreed that whatever list we derived from

their corpus could be made available to the public for free. We began our development of the NGSL in early 2010, using both the SketchEngine (2006) tools that Cambridge provided and a wide range of other tools, including publicly available ones such as Lawrence Anthony's very useful AntConc program (http://www.laurenceanthony.net/software.html), along with several specialized bits of software that we developed specifically for the purpose of this project.

The initial corpus we used was created using a subset of the CEC that was queried and analyzed using the SketchEngine corpus query system (http://www.sketchengine.co.uk/). The size of each sub-corpus that was initially included is outlined in Table 1:

TABLE 1. CEC Corpora Used for Preliminary Analysis of the NGSL

Corpus	Tokens
Newspaper	748,391,436
Academic	260,904,352
Learner	38,219,480
Fiction	37,792,168
Journals	37,478,577
Magazines	37,329,846
Non-Fiction	35,443,408
Radio	28,882,717
Spoken	27,934,806
Documents	19,017,236
TV	11,515,296
Total	**1,282,909,322**

The newspaper and academic sub-corpora were quickly eliminated for very similar reasons. First, although statistical procedures can be used to correct for minor differences in the size of sub-corpora, it was clear that the newspaper sub-corpora at 748,391,436 tokens and the academic sub-corpora at 260,904,352 tokens were dominating the frequencies and far too large for this kind of correction (a potential problem with the ONGSL since the variance between the largest and smallest corpus is 2 billion words). Second, both of these sub-corpora did not fit the profile

of general English text types that we were looking for, with the newspaper sub-corpus showing a marked bias towards financial terms and the academic sub-corpus being from a specific genre not directly related to general English. As a result, both corpora were removed from the compilation.

Table 2 shows the sub-corpora that were actually used to generate the final analysis of frequencies. While smaller than the corpus described in Table 1, the corpus is still more than 100 times the size of the corpus used for the original GSL and far more balanced as a result:

TABLE 2. CEC Corpora Included in Final Analysis for the NGSL

Corpus	Tokens
Learner	38,219,480
Fiction	37,792,168
Journals	37,478,577
Magazines	37,329,846
Non-Fiction	35,443,408
Radio	28,882,717
Spoken	27,934,806
Documents	19,017,236
TV	11,515,296
Total	**273,613,534**

The resulting word lists were then cleaned up by removing proper nouns, abbreviations, slang, and other noise, and excluding certain word sets such as days of the week, months of the year, and numbers (this proved to be a controversial decision and these word sets will most likely be re-added in the 2.0 version of the list in early 2015).

We then used a sequence of computations to combine the frequencies from the various sub-corpora while adjusting for differences in their relative sizes. Specifically, we used Carroll's measure of dispersion, (D_2), estimated frequency per million (U_m) and the Standard Frequency Index (SFI; Carroll, Davies, & Richman, 1971; Carroll, 1971) to combine the frequencies from the various sub-corpora while adjusting for differences in their relative sizes.

Finally, based on a series of meetings and discussions with Paul

Nation about how to improve the list, the combined list was then compared to other important lists such as the original GSL, the BNC, and COCA to make sure important words were included or excluded as necessary.

NGSL VERSION 1.01

Though we were as careful and systematic as possible in the process of developing the original NGSL, we view the release of the 1.0 version of the NGSL as no more than an interim list, representing the best research and development that we could do in relative isolation. The next very important step was to release the NGSL publicly so that teachers and researchers around the world could begin to react to it, and give ideas and advice on how to improve it. To this end, most of 2013 was devoted to making the list and a variety of NGSL-related resources available via a dedicated website (http://www.newgeneralservicelist.org/), publishing and presenting about the list at more than a dozen conferences around the world, and creating an NGSL social media presence on websites such as Facebook. Through these efforts and the excellent feedback and suggestions that we have received from many experts, we are now releasing the 1.01 version of the NGSL both here and on the NGSL website. The net result of these changes will decrease the number of NGSL headwords by 17 from 2,818 to 2,801 with the following being the main changes made:

Two Words Added

- Insertion of TOURNAMENT, which was accidentally deleted in the initial analysis.
- YEAH, which was originally counted as a derived form of YES, is now counted under its own headword.

Nineteen Words Deleted

- Four numbers were deleted and moved to the supplemental list:
  ZERO
  BILLION
  FIFTEEN
  FIFTY

- The inflected parts of speech of pronouns were demoted and listed under their canonical objective pronoun:
  HER was listed under SHE.
  HIM and HIS were listed under HE.
  ITS was listed under IT.
  ME and MY were listed under I.
  OUR and US were listed under WE.
  THEIR and THEM were listed under THEY.
  THESE was listed under THIS.
  THOSE was listed under THAT.
  WHOM and WHOSE were listed under WHO.
  YOUR was listed under YOU.

Why Weren't Word Families Like Those in the Original GSL Used?

It is important to remember how the original GSL counted words. The GSL did not amalgamate frequency counts for derived forms, but it did combine the frequencies for word forms regardless of parts of speech. For example, the frequency counts for both the noun and verb forms of CARE are summed, while the frequency counts for the derived forms CAREFUL and CARELESS are listed separately (Figure 1).

Following the publication of Bauer and Nation's *Word Families* (1993), the number of words included under the headword expanded greatly. They stated a word family consisted of a base word and all its derived and inflected forms that can be understood by a learner without having to learn each form separately" (p. 253). For example, CARE under the word family rubric contains, along with the inflections of the verb and noun, the following: CARE, CAREFUL, CAREFULLY, CAREFULNESS, CARELESS, CARELESSLY, CARELESSNESS, CARER, CARERS, UNCARED, and UNCARING. However, the assumption that the form "can be understood by a learner without having to learn each form separately" has been called into question. Research by Schmitt and Zimmerman (2002) "did not support a strong facilitative effect for knowledge of words within a word family" (p. 158).

Another problem with determining which words would be included under the headwords using the word family concept was suggested by Gardner (2007), who wrote "case-by-case assessments of affixed word

forms would be necessary to determine if a prolific derivational affix was acting transparently or not" (p. 247). This of course adds a level of subjectivity to the compilation of the word list and an avenue to list differentiation, resulting in difficulty in interpreting coverage statistics reported for a variant word list going under the same name, such as is the case with the current GSL coverage claims coming from substantially different word lists.

CARE	1134e		
care, n.		? [(1) *(anxiety; rather literary)*	
		Cast all your cares aside	
		The cares of office	14%]
		(2) *(caution)*	
		Take care, or you'll fall	22%
		(3) *(responsibility)*	
		Take care of the baby	
		Take great care of it; it's glass	
		In his care	
		J. Smith, care of Mrs Jones, 22 High St.	25%
care, v.		(1) *(feel anxious for)*	
		He cares only for his own interests	(4%)
		I don't care! I don't care what you do	(7%) 11%
		(2) *(idea of responsibility)*	
		The child has been well cared for	8%
		(3) *(wish—usually in questions, negatives, or hypothetical)*	
		Would you care to read this?	
		Do you care to come out for a walk?	
		I don't much care for dancing	14%
		? [(4) *(love)*	
		Does she really care for him?	2.4%]
careful, adj.	244e	A careful person; careful work; be careful not to break it	97%
careless, adj.	102e	A careless worker; careless work	94%

FIGURE 1. An example of how the original GSL counted words.

What Constitutes a "Word" in the NGSL?

There are many ways to define a word for the purpose of counting frequencies. The simplest is to look at "types," where each form is counted as a different word regardless of part of speech. For example, LISTS would include both the third-person singular form of the verb LIST and the plural form of the noun LIST.

The second method is to count "lexemes" where homographs are counted separately, but all the inflected forms of a word are added together. For example, the nouns LIST and LISTS would be counted together but not with the verbs LIST, LISTS, LISTED, and LISTING, which would be counted as a separate item. Inflections for nouns include the plural and the possessive. Verb inflections include the third person, the past, and the participles. Inflections for short adjectives include the comparative and the superlative.

The third method of counting words is called "word families" and was proposed by Bauer and Nation (1993). Word families include the inflected forms and certain derived forms. The NGSL uses a modified lexeme approach, where we count the headword in all its various parts of speech and include all inflected forms. Unlike the traditional definition of a lexeme, it includes all the inflected forms from the different parts of speech. For example, LIST would include LISTS, LISTED, LISTING, and LISTINGS. It does not include any of the derived forms using non-inflection suffixes. Variations such as the difference between US and UK spelling are also grouped within the same lexeme.

Why Are Unusual Lemmas Like WINDOWING and WHILES Included as Part of the Headwords WINDOW and WHILE?

Word lists are created in different ways and for different purposes, and what is or is not included in a list really depends on the final purpose. Although the version of the NGSL that you will see on either the free Quizlet flashcard program, or the free NGSL with definitions in the easy English file, contains only the headword since the purpose is teaching, you may notice that the main NGSL list includes not only the headword but also a wide range of its associated lemmas, including several that may seem strange or unusual. This is because another

purpose of the NGSL was to be useful to researchers who are analyzing real world texts to identify the frequency of words in order to predict the probability of the reader encountering the lemma. When faced with making the word set for a given headword, one can use evidence or arbitrarily imposed rules. For example, when making the revised version of the GSL in 1995, Bauman and Culligan chose an evidence-based approach. If the derived form did not appear in the Brown Corpus, they did not include it. This resulted in the exclusion of many legitimate derived forms.

For the NGSL, we wanted to address two primary tasks. First, we wanted to predict the probability of the reader encountering the lemma. To do so, our lists were used to analyze real-world texts to identify the frequency of words. Second, we wanted to identify unique lemmas that were not on our word list. In Probability Theory, there is something called an event space. Basically, it is the set of all possible ways a rare or frequent event can happen.

Once the parameters of the event space are defined, only those words are permissible. It may sound logical to conclude that only high-frequency events be included in the list, but what does a researcher do when a rare event occurs? Do they ignore the event and maintain the event space or do they update the event space? More concretely, what should researchers do when they encounter words that clearly belong to a Level 6 affix family (Bauer & Nation, 1993) but are not on the word list? Should they ignore it and pretend it is a unique occurrence, or add it, thus changing the list? We have chosen the latter, evidence-based approach, including lemmas with even a very low or no occurrence in the main list so that researchers who are doing corpus research with the NGSL using analytical tools such as VocabProfiler and AntWordProfiler can explore questions and issues beyond what the typical EFL learner or teacher might be interested in. English is an incredibly flexible language with words shifting parts of speech with ease, as Susanna Centlivre showed in 1709 with her creative use of the word “but” with the phrase, "But me no buts." We chose rule-based and completeness.

Why Weren't the Numbers, Days of the Week, and Months of the Year Included?

Although these word sets were excluded from the NGSL proper in the same way they were excluded from the original GSL, they are

actually included as an appendix in the main NGSL Excel file. Though pulling these words out had a negative effect on our coverage figures, it seemed to be the right decision from a pedagogic point of view. In the case of days of the week and months of the year, it was consistent with our decision (and most corpus-derived vocabulary lists) not to include proper nouns. Furthermore, keeping them in would have caused another kind of problem since not all items of each lexical set occurred at a high enough frequency to appear on the NGSL list even within the 273 million sample of the CEC corpus used for this project.

Why Weren't Letters of the Alphabet Included in the NGSL?

The alphabet by itself is used as signs or symbols, often as placeholders, like numbers or bullets in a list. They are often used in sequences or stand in as variables in formulas. While they are of interest in the field of semiotics, they cannot be classed as words, but are more often used in the same way as smiley faces or other emoticons.

Text Coverage: Covering Your Bets with the NGSL

One of the most important goals of this project was to try to develop an NGSL that would be more efficient and useful to language learners and teachers by providing more coverage with fewer words than the original GSL. One of the problems with making a comparison between the two lists, indeed between any well-known vocabulary lists, is the way the number of words were counted in each list, which needs to be done according to the same criteria. As innovative as the GSL was at the time of its creation, West's definition of what constituted a word was, by his own admission, non-systematic and arbitrary: "no attempt has been made to be rigidly consistent in the method used for displaying the words: each word has been treated as a separate problem, and the sole aim has been clearness" (West, 1953, p. viii).

This means that for a meaningful comparison between the GSL and NGSL to be done, the words on each list need to be counted in the same way. As was mentioned in the previous section, a comparison of the number of "word families" in the GSL and NGSL reveals that there are 1,964 word families in the GSL and 2,368 in the NGSL (using Level 6 of Bauer and Nation's, 1993, word family taxonomy). Coverage within

the 273 million word CEC is summarized in Table 3, showing that the 2,368 word families in the NGSL provides 90.34% coverage, while the 1,964 word families in the original GSL provides only 84.24%. That the NGSL with approximately 400 more word families provides more coverage than the original GSL may not seem a surprising result, but when these lists are lemmatized, the usefulness of the NGSL becomes more apparent as the more than 800 fewer lemmas in the NGSL provide 6.1% more coverage than is provided by West's original GSL.

TABLE 3. Comparison of Coverage for the CEC by the GSL and NGSL Word Lists

Vocabulary List	Number of "Word Families"	Number of "Lemmas"	Coverage in CEC Corpus
GSL	1,964	3,623	84.24%
NGSL	2,368	2,818	90.34%

After analyzing coverage of the CEC corpus for the GSL and NGSL word lists, the next step taken was to compare coverage figures against other kinds of corpora I had at my disposal. In this round of analysis, I have also included the ONGSL in the analysis. All calculations were conducted using Lawrence Anthony's excellent AntWordProfiler, which easily allows for the uploading of vocabulary wordlists and texts to be analyzed as long as they have been converted to .txt files. For this comparison, all word lists used were first converted to modified lemmas so that word counts would be done in the same way. A modified lemma is one that combines all possible parts of speech into one lemma. For example, the modified lemma for ROUND includes the inflections for the noun, verb, and adjective; for example, ROUND, ROUNDS, ROUNDED, ROUNDING, ROUNDINGS, ROUNDER, and ROUNDEST.

Please note that the slight difference in number of word families and lemmas between the analysis done in early 2013, shown in Table 3, and the results given for this report, in Tables 4 and 5, are due to the fact that the GSL in Table 3 was taken from the GSL/AWL version of the Range program (Heatley, Nation, & Coxhead, 2002). These lists were not specifically cited to have been developed up to Affix Level 6 (Bauer & Nation, 1993) while the lists from the BNC/COCA, shown in Table 4, are. Therefore, the headwords from the GSL/AWL word lists were matched to the derived forms from the BNC/COCA lists.

The first corpus used was a 12 million word corpus of the top 100 most important classic works of English literature as rated by professors of English literature at several top Japanese universities (Browne & Culligan, 2008). All texts selected were ones that were available in the public domain for download and analysis via Project Gutenberg (2014). As a collection of classic literature texts (the newest texts available for download in Project Gutenberg are at least 50 years old), it was hypothesized that the word list, which was based on the oldest corpus, the original GSL, would probably provide the highest coverage.

The second corpus was a more modern corpus of 27 million words taken from *The Economist*, spanning issues from 2001 to 2010 (Culligan, 2013a). The third corpus, too, was also quite modern, a 13-million-word sample taken from *Scientific American*, covering issues published between 1993 to 2000 (Culligan, 2013b). Here it was hypothesized that one of the word lists based on more modern corpora (either the NGSL or the ONGSL) would provide more coverage.

As can be seen from Table 4 below, the GSL provided slightly better coverage (0.8%) than the NGSL for the corpus of classic literature and a more substantial 3.4% higher coverage than the ONGSL. That the GSL, which is based on a corpus with a far older collection of texts, provided the best coverage of a collection of older literary texts is perhaps an expected result, but a more surprising one was that the NGSL, which is based on a more modern corpus, was able to come within 0.8% coverage of the GSL despite using 700 fewer lemmas.

TABLE 4. Comparison of GSL, NGSL, and ONGSL Coverage Figures

Word List	Number of Headwords	Number of Unique Headwords	Number of Types	Number of Lemmas	Number of BNC-COCA Word Families	Classic Literature	Scientific American	The Economist
GSL (Nation, Level 6)	1,986	1,927	9,293	3,553	2,245	86.17	65.87	76.55
ONGSL	2,228	2,189	6,365	2,130	1,929	82.76	68.68	78.30
NGSL 1.01	2,801	2,801	8,481	2,801	2,483	85.35	71.34	81.75
						12,377,844	13,047,726	27,337,358

If we narrow down the results for classic literature to look at coverage for two well-known novels within the corpus, *The Count of Monte Christo* and *Dracula*, Table 5 shows very similar results with the

GSL giving slightly better coverage than the NGSL (0.8% and 0.7% more coverage, respectively), with the NGSL giving 2.5-2.6% more coverage than the ONGSL.

TABLE 5. Coverage Figures for Two Well-Known Novels

	Number of Headwords	Number of Unique Headwords	Number of Lemma	Coverage of *Count of Monte Cristo*	Coverage of *Dracula*
GSL Range	1,986	1,927	3,553	85.6	90.6
NGSL 1.1	2,801	2,801	2,801	84.8	89.9
ONGSL	2,228	2,189	2,130	82.2	87.4

When looking at coverage figures for the two more modern genre-specific corpora, the efficiency of the NGSL becomes more apparent, with the NGSL giving 3.5% more coverage than the ONGSL, and 5.5% more coverage than the GSL for the *Scientific American* corpus and similar figures of 3.5% and 5.2% more coverage for *The Economist* corpus.

Where to Find the NGSL and Associated Resources

From the very beginning, our focus has been less on simply publishing an academic paper on a new list of words than it has been on creating a list of high frequency words that would be as useful as possible for students, teachers, and researchers around the world. One culmination of this effort is our dedicated website (www.newgeneralservicelist.org), which gathers all associated NGSL resources in one place. Here you can download the 1.01 (and 1.0) version of the NGSL in lemmatized or headword form, as well as all papers that have been written on the NGSL, and see a list of past and upcoming conference presentations on the list. Because word lists are only useful to learners and teachers if there are definitions and learning tools, I have already written original definitions for all words in easy English for all NGSL words and uploaded the entire list in 50 word blocks (by frequency) to the free Quizlet vocabulary flashcard learning program (www.quizlet.com). As for analytical tools, the NGSL is already

available on the free Online Graded Text Editor (OGTE) program (http://www.er-central.com/ogte/), which is part of the free Extensive Reading Central website (www.er-central.com) developed by Charles Browne and Rob Waring, as well as on Tom Cobb's wonderful VocabProfile tool (http://www.lextutor.ca/vp/eng/), and will soon also be available via Laurence Anthony's free AntWordProfiler Program (http://www.laurenceanthony.net/antwordprofiler_index.html).

THE AUTHOR

Charles Browne is Professor of Applied Linguistics and head of the EFL teacher training program at Meiji Gakuin University in Japan, and a well-known expert on English education in Asia. He received his Ed.D. from Temple University and is a specialist in CALL (Computer Assisted Language Learning) and second language vocabulary acquisition. Over the past 25 years that he has worked in Japan, Dr. Browne has published dozens of research articles and books, including *New Perspectives in CALL for Second Language Classrooms*. He was the first National Chairman of the JET (Japan Exchange and Teaching) Program, worked for the Japanese ministry of education as a teacher-training specialist and textbook specialist, and has led language learning product development for several software companies.

REFERENCES

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279.

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics, 2013,* 1-23.

Browne, C. (2013). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher, 7*(34), 13–16.

Browne, C., & Culligan B. (2008). [A collection of the top 100 works of classic English literature as rated by Japanese university professors]. Unpublished raw data.

Culligan, B. (2013a). [A corpus of *The Economist* magazine articles from 2001-2010]. Unpublished raw data.

Culligan, B. (2013b). [A corpus of *Scientific American* magazine articles from 1993-2000]. Unpublished raw data.

Carroll, J. B. (1971). Statistical analysis of the corpus. In *The American Heritage word frequency book* (pp. xxi-xl). Boston, MA: Houghton Mifflin.

Carroll, J. B., Davies, P., & Richman, B. (1971). Guide to the alphabetical list. In *The American Heritage word frequency book* (pp. 1-4). Boston, MA: Houghton Mifflin.

Gardner, D. (2007). Validating the construct of *word* in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics*, *28*(2), 241−265.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation

Project Gutenberg. (2014). *Partners, affiliates and resources*. Retrieved from the Project Gutenberg website: http://www.gutenberg.org

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? *TESOL Quarterly*, *36*(2), 145−171.

West, M. (1953). *A general service list of English words.* London, UK: Longman, Green & Co.