Vocabulary Learning and Instruction
Volume 3, Issue 2, September 2014
http://vli-journal.org

Article

A New General Service List: The Better Mousetrap We've Been Looking for?

Charles Browne
Meiji Gakuin University
doi: http://dx.doi.org/10.7820/vli.v03.2.browne

Abstract

This brief paper introduces the New General Service List (NGSL), a major update of Michael West's 1953 General Service List (GSL) of core vocabulary for second language learners. After describing the rationale behind the NGSL, the specific steps taken to create it and a discussion of the latest 1.01 version of the list, the paper moves on to comparing the text coverage offered by the NGSL against both the GSL as well as another recent GSL published by Brezina and Gablasova (referred to as Other New General Service List [ONGSL] in this paper). Results indicate that while the original GSL offers slightly better coverage for texts of classic literature (about 0.8% better than the NGSL and 4.5% more than the ONGSL), the NGSL offers 5–6% more coverage than either list for more modern corpora such as *Scientific American* or *The Economist*.

In the precomputer era of the early 20th century, Michael West and his colleagues undertook a very ambitious corpus linguistics project that gathered and analyzed more than 2.5 million words of text and culminated in the publication of what is now known as The General Service List (GSL; West, 1953), a list of about 2000 high-frequency words that were deemed important for second language learners. However, as useful and helpful as this list has been to us over the decades, it has also been criticized for being based on a corpus that is considered to be both dated (most of the texts were published before 1930), as well as too small by modern standards, and for not clearly defining what constitutes a "word."

In February 2013, on the 60th anniversary of West's publication of the GSL, my colleagues (Brent Culligan and Joseph Phillips of Aoyama Gakuin Women's Junior College), and I put up a website (www.newgeneralservicelist.org) that released a major update of West's GSL known as the New General Service List (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 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 2800 high-frequency words which met the following goals:

Please cite this article as: Browne, C. (2014). A new general service list: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction*, *3* (2), 1–10. doi: 10.7820/vli.v03.2.browne

(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 generalizability and validity of the list
(2) to create a 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 a 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 6 months after we released the 1.0 version of the NGSL, another GSL was put out by Brezina and Gablasova (August, 2013), which, for the purpose of this article, I will hereafter refer to as the "Other New General Service List (ONGSL)" in order to avoid confusion. Although the ONGSL looks to be a very impressive piece of research, the purpose of their list and the way it was developed seems to be a bit different than the purpose and development process we undertook for the NGSL presented here. The authors 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 Lancaster-Oslo-Bergen Corpus (LOB), and The BE06 Corpus of British English (2006), 100 million for the British National Corpus [BNC], and 12 billion words for the En Ten Ten 12 corpora), and resulted in the identification of 2494 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 had 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. We are following many of the principles West and his colleagues used both for developing as well as for improving the list over time, and are thus referring to the latest version of the NGSL in this article as NGSL 1.01.

The purpose of this brief paper is to explain a bit more about the NGSL as well as to give some initial comparisons in text coverage between the GSL, NGSL, and ONGSL for a range of different text types.

1 The NGSL: a Word List Based on a Large, Modern Corpus

One of the obvious axioms of corpus linguistics is that any word frequency lists that are 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, while arguably useful for second language learners of the time, 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 a NGSL that offered better coverage with fewer words. We were lucky enough to be given full unrestricted access to the CEC, a multibillion word corpus that contains both written and spoken text data for British and American English as well as to the Cambridge Learner Corpus, a 40 million word corpus made up of English exam responses written by English language learners, and promised by Cambridge University Press 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, as well as a wide range of other tools including publicly available tools such as Lawrence Anthony's very useful AntConc (http://www.antlab.sci.waseda.ac.jp/software.html) program as well as several specialized bits of software that we developed specifically for the purposes 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.

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

Table 1. CEC Corpora Used for Preliminary Analysis of 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
Nonfiction	35,443,408
Radio	28,882,717
Spoken	27,934,806
Documents	19,017,236
TV	11,515,296
Total	1,282,909,322

Table 2. CEC Corpora Included in Final Analysis for NGSL

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

(a potential problem with the ONGSL since the variance between the largest and smallest corpus is 12 billion words). Second, both of these sub-corpora did not fit the profile of general English text types we were looking for with the Newspaper sub-corpus showing a marked bias toward 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.

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 be re-added in the 2.0 version of the list which is due out in early summer of 2014).

Then we 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 Standard Frequency Index (Carroll, Davies, & Richman, 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 Corpus of Contemporary American English (COCA) to make sure important words were include or excluded as necessary.

2 NGSL Version 1.01

Though we were as careful and systematic as possibly in the process of developing the original NGSL, like West and his colleagues before us, 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 we could do in relative isolation, but with the next very important step being 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 (www.newgeneraservicelist.org), publishing and presenting about the list at more than a dozen conferences around the world and creating a NGSL social media presence on websites such as on Facebook. Through these efforts and the excellent feedback and suggestions 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 2818 to 2801 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 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

3 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 separately. 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 laid out by the generalizability and productivity of the affixes.

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.

4 Text Coverage: Covering Your Bets with the NGSL

One of the most important goals of this project was to try to develop a 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 that the way of counting the number of words in each list 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, nonsystematic 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 1964 word families in the GSL and 2368 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 2368 word families in the NGSL provides 90.34% coverage while the 1964 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.

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

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	1964	3623	84.24
NGSL	2368	2818	90.34

ONGSL in the analysis. All calculations were conducted using Lawrence Anthony's excellent AntWordProfiler, which easily allows for the uploading of vocabulary word lists 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, e.g., 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/Academic Word List (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 Guttenberg (2014). As a collection of classic literature texts (the newest texts available for download in Project Guttenberg 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* magazine, spanning issues from 2001 to 2010 (Culligan, 2013a). The third corpus, too, was also quite modern, a 13 million word sample taken from the *Scientific American* magazine covering issues published between 1993 and 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, 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 close to 0.8% coverage of the GSL despite using 700 fewer lemmas.

Table 4. Comparison of GSL, NGSL, and ONGSL Coverage Figures for Three Different Genres

Word list	Number of headwords	Number of unique headwords	Number of types	Number of lemmas	Number of BNC/COCA word families	Classic	*Scientific American*	*The Economist*
GSL (Nation level 6)	1986	1927	9293	3553	2245	86.17%	65.87%	76.55%
ONGSL	2228	2189	6365	2130	1929	82.76%	68.68%	78.30%
NGSL 1.01	2801	2801	8481	2801	2483	85.35%	71.34%	81.75%
						12,377,844	13,047,726	27,337,358

Table 5. Coverage Figures for Two Well-known Novels

	Number of headwords	Number of unique headwords	Number of lemmas	Coverage of *The Count of Monte Cristo*	Coverage of *Dracula*
GSL range	1986	1927	3553	85.6	90.6
NGSL 1.1	2801	2801	2801	84.8	89.9
ONGSL	2228	2189	2130	82.2	87.4

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 Cristo* 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.

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.

5 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 was as useful as possible for students, teachers, and researchers around the world and have thus created a dedicated website (www.newgeneralservicelist.org) to gather all associated NGSL resources in one place. Here, you can download the 1.01 (and 1.0) versions 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, we 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 and listening 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.antlab.sci.waseda.ac.jp/antwordprofiler_index.html).

References

Bauer, L., & Nation, P. (1993). Word Families. *International Journal of Lexicography*, *6*, 253–279. doi:10.1093/ijl/6.4.253

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics*. doi:10.1093/applin/amt018

Browne, C. (2013, July). The New General Service List: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *7 34*, 13–16. Retrieved from http://jalt-publications.org/tlt/issues/2013-07_37.4

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.

Carroll, J. B., Davies, P., & Richman, B. (eds) (1971). *The American Heritage Word Frequency Book*. Boston, MA: Houghton Mifflin.

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

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

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 Guttenberg. (2014). Retrieved from http://www.gutenberg.org

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