*Vocabulary Learning and Instruction*
Volume 8, Issue 1, March 2019
http://vli-journal.org

# An Examination of the New General Service List

Tim Stoeckel
*University of Niigata Prefecture*

## Abstract

The New General Service List (NGSL; Browne, Culligan, & Phillips, 2013b) was published on an interim basis in 2013 as a modern replacement for West's (1953) original General Service List (GSL). This study compared GSL and NGSL coverage of a 6-year, 114-million word section of the Corpus of Contemporary American English (COCA), and used COCA word frequencies as a secondary data source to identify candidates for addition to the NGSL. The NGSL was found to provide 4.32% better coverage of the COCA than the GSL. Moreover, several candidates were identified for inclusion to the NGSL: three are current members of the NGSL's companion list, the New Academic Word List (Browne, Culligan, & Phillips, 2013a); five are words whose usage has increased in recent years; and five are individual types that appear to have been miscategorized during the original development of the NGSL. Because NGSL word selection was based on not only empirical but also subjective criteria, the article calls for the addition of annotations to the NGSL to explain decisions regarding low-frequency NGSL constituents and high-frequency non-constituents.

**Key Words:** New General Service List

## Introduction

The New General Service List (NGSL; Browne, Culligan, & Phillips, 2013b) was published in 2013 as a modern replacement for West's (1953) original General Service List (GSL) for the purposes of both research and pedagogy for second-language learners of English (Browne, 2014). As the NGSL is currently considered an interim list open to examination and debate by interested researchers (Browne, 2014), this article investigates NGSL coverage and word frequencies in a large, modern corpus in hopes of promoting discussion and possible refinement of the list.

### *The General Service List*

The original GSL (West, 1953) was constructed based upon frequencies in a 5-million word corpus (for some words, 2.5 million) as well as subjective criteria on the usefulness of words to learners of English. Generally, word forms with shared roots were grouped together, but there was no single guiding principle for what to include in each group. Under *broad*, for example, are only *breadth* and the compound *broadcast*. The GSL was extensively revised by Paul Nation in the early 1990s so that many inflected and derived forms were added (e.g., *broader* and *broadly* under *broad*) and some compound forms (e.g., *broadcast*) were removed.[1] Numbers, months, days of the week, and letters of the alphabet

Please cite this article as: Stoeckel, T. (2019). An Examination of the New General Service List. *Vocabulary Learning and Instruction, 8* (1), 53–61. https://doi.org/10.7820/vli.v08.1.stoeckel

were also added to the list. This is the most widely used version of the GSL today, as it is utilized in Lextutor VocabProfiler (http://www.lextutor.ca/vp/) and accompanies the Range (Heatley, Nation, & Coxhead, 1994) and AntWordProfiler (Anthony, 2013) programs.

Although the GSL has been valuable in teaching and research, the corpus from which the list was derived is now dated and considered small by modern standards. Additionally, while the inclusion of derivational and compound forms may be useful in raising awareness of related word forms, it is arguably not the most efficient way to help students learn the most important words for second language acquisition. For many learners, derivational knowledge cannot be assumed from knowledge of a headword (McLean, 2017; Mochizuki & Aizawa, 2000; Ward & Chuenjundaeng, 2009), which implies that the learning burden for the GSL entails learning not only headwords but also many separate constituents.

## The New General Service List

The NGSL was developed in part to address these limitations. It was derived from the analysis of a 273-million word section of the more modern Cambridge English Corpus (CEC). As an organizing principle, words in the NGSL are grouped into modified lemmas. A regular lemma consists of a headword plus inflected forms that are of the same part of speech as the headword (e.g., the nominal headword *show* plus the plural *shows*). A *modified* lemma is comprised of a headword in all of its various parts of speech together with the inflected forms for each part of speech. Thus, the modified lemma for *show* includes the noun forms above as well as the verbs *shows*, *showed*, *showing*, and *shown*. This use of the modified lemma is consistent with McLean (2017), who found that most L2 learners with knowledge of a headword also have receptive knowledge of its inflected forms. This word-grouping principle determined what was to be included in each NGSL modified lemma regardless of whether individual word types actually occurred in the CEC (Browne, 2014). Exceptions included cases in which a word form was both the head of one modified lemma and a constituent of another, in which case the word form in question was sometimes considered only as the canonical form of its own modified lemma (B. Culligan, personal communication, March 28, 2018). Thus, *rose* is not found under *rise*, but it is considered the head of the modified lemma comprised of *rose* and *roses*. Other exceptions were the inclusion of pluralized gerunds (e.g., *teachings* under *teach*) for approximately 10% of the gerunds in the list. The NGSL consists of 2801 modified lemmas, and there is a supplementary list of 52 entries comprised of numbers, months, and days of the week.

Word selection for the NGSL was based on both empirical evidence and subjective considerations. To empirically rank words, Carroll's *U* (Carroll, 1971) was used; this statistic expresses an adjusted words-per-million (wpm) that accounts for differences in both dispersion across and size of corpus sections. As shown in Table 1, the most common 1000 words in the CEC occur with an adjusted frequency of at least 93.4 wpm, and the top 2000 words are of at least 36.5 wpm. Because the NGSL has 2801 entries, the modified lemma *utility*, with a *U* value of 21.2 and rank of 2801, can be used as a benchmark for assessing whether a word belongs in the NGSL based on empirical evidence alone (CEC *U* values available at http://www.newgeneralservicelist.org/). Regarding subjective criteria, there was ongoing input from Paul Nation as well as comparisons of the NGSL to other existing lists to "make sure important words were included or excluded as necessary" (Browne, 2014, p. 40).

We get a sense of the balance between these quantitative and subjective considerations from Table 2, which shows the number of entries in the NGSL from different adjusted-frequency levels of the CEC. These follow an expected pattern, with most 1–2K level words present in the NGSL or its supplementary list. Notably however, seven 1K and eight 2K words are absent; with *U* values between 37.4 and 137.5, the adjusted frequencies for these words in the CEC far exceed many NGSL constituents. Because the list is intended to give maximal coverage of English texts with the fewest words, and as explanations regarding the subjective criteria for individual word selections are unavailable, an examination of non-empirically grounded decisions would seem worthwhile. It may also be advisable to further investigate coverage provided by the NGSL. Browne (2014) has reported that the NGSL (90.34%) offers approximately 6% better coverage than the GSL (84.24%) of general English discourse, but this comparison may have favored the NGSL because it used the section of the CEC from which the list itself was derived. Moreover, although the precise composition of this CEC section is unavailable, the complete CEC is comprised mostly of British English sources (65.4%; D. Moser, Cambridge University Press, personal communication, April 16, 2018), suggesting the NGSL may reflect British usage tendencies. There would therefore be merit in cross-checking NGSL membership with word frequencies in corpora composed primarily of North American or other world English sources.

Table 1. Carroll's *U* Values for Words at Selected Frequency Ranks in the CEC

| Modified Lemma Head | Rank | *U* |
|---|---|---|
| *the* | 1 | 60,909.9 |
| *somebody* | 1000 | 93.4 |
| *revolution* | 2000 | 36.5 |
| *mortgage* | 2500 | 26.0 |
| *utility* | 2801 | 21.2 |
| *quit* | 3000 | 19.0 |
| *explicit* | 3500 | 14.7 |

*Note:* Carroll's (1971) *U* values downloaded from http://www.newgeneralservicelist.org/

Table 2. NGSL Entries from Various Adjusted-Frequency Levels of the CEC

| | Adjusted-Frequency Band | | | | | |
|---|---|---|---|---|---|---|
| List | 1K | 2K | 3K | 4K | 5–10K | >10K |
| NGSL | 981 | 987 | 796 | 35 | 2 | 0 |
| NGSL Supplementary | 12 | 5 | 8 | 6 | 1 | 15 |

*Note:* Adjusted frequencies based on Carroll's (1971) *U* values downloaded from http://www.newgeneralservicelist.org/

## Purpose

This study investigated NGSL coverage and word frequencies in the Corpus of Contemporary American English (COCA; Davies, 2008). The research questions (RQ) were:

1. How does the NGSL compare to the GSL in terms of coverage of the COCA and each of its individual sections?
2. Does analysis of the COCA as a secondary data source reveal candidates worthy of consideration for addition to the NGSL?

## Methods

### *Materials*

The study corpus was comprised of word/lemma/part-of-speech files for a 6-year section (2010–2015) of the COCA. Regarding GSL and NGSL word lists, to make comparisons using the most readily available variants today, the version of the GSL that accompanies Range (downloaded from https://www.victoria.ac.nz/lals/about/staff/paul-nation) and version 1.01 of the NGSL (downloaded from http://www.newgeneralservicelist.org/) were used. Because of the additions Paul Nation made to the GSL (described above), the NGSL supplementary list, with letters of the alphabet added by the researcher, was included in order to make fair comparisons between the NGSL and GSL.

### *Corpus Cleaning*

Part-of-speech tags in the complete set of COCA files were used to identify proper nouns, foreign words, marginal content, non-words, and all other discourse. To assess the accuracy of these classifications, a 2000-token sample was also categorized manually, with a 98.3% agreement rate. Tokens categorized as non-words were then removed from the COCA files, and the size of each corpus section was recorded (Table 3).

### *Analyses*

For RQ1, AntWordProfiler (Anthony, 2013) was used to obtain the coverage percentage provided by the GSL and NGSL of the COCA and each of its sections.

Table 3. Number of Tokens Before and After Removal of Non-Words from the 2010–2015 Section of the COCA

| Section | Before Cleaning | After Cleaning |
|---|---|---|
| Fiction | 29,985,804 | 23,705,513 |
| Spoken | 29,621,441 | 23,348,630 |
| Academic | 27,208,629 | 20,913,397 |
| Magazine | 29,122,476 | 23,317,601 |
| Newspaper | 28,965,412 | 22,825,277 |
| Total | 144,903,762 | 114,110,418 |

For RQ2, AntWordProfiler was first used to obtain frequencies for NGSL constituents and for off-list types. A file of modified lemmas was then made for the 1000 most frequent off-list types, and this was used with AntWordProfiler to obtain frequency counts for off-list modified lemmas in each COCA section. Carroll's *U* (Carroll, 1971) was then calculated for all NGSL, NGSL supplementary, and off-list modified lemmas. These *U* values and those derived from the CEC in the original development of the NGSL (downloaded from: http://www.newgeneralservicelist.org/) were examined to identify candidates for addition to the NGSL.

## Results and Discussion

For the 3000 modified lemmas with the highest adjusted frequencies in the CEC, the Pearson's correlation for *U*-values in the CEC and the COCA was 0.991, indicating a very high degree of similarity in the occurrence of high-frequency vocabulary in the two corpora. Regarding coverage, though somewhat lower than Browne's (2014) original comparison, the NGSL plus its supplemental list performed well, providing 4.32% better overall coverage than the GSL (83.66% vs. 79.34%, Table 4). Notably, even though the *academic* and *newspaper* sections of the CEC were excluded from analysis during the development of the NGSL (Browne, 2014), the advantage in coverage provided by the NGSL was greatest in these two COCA sections, +9.07% and +4.87%, respectively.

RQ2 asked whether analysis of the study corpus would reveal candidates for inclusion to the NGSL. The empirical benchmark for NGSL membership in the COCA was *U* = 22.7 for the modified lemma *con* (ranked 2801), which is similar to that of the CEC reported above. There were 282 off-list modified lemmas above this threshold in the COCA. Although this is partially due to differences that exist between any two corpora (see Nation, 2016 p. 99), three categories of words above this threshold will be presented as NGSL candidates.

First are three high-frequency words that currently belong to the NGSL companion list, the New Academic Word List (NAWL; Browne, Culligan, & Phillips, 2013a). Thirty NAWL members have *U* values greater than the empirical threshold for NGSL membership in both the COCA (*U* ≥ 22.7) and the CEC (*U* ≥ 21.2). However, many of these could be considered specialized academic vocabulary on the basis of higher adjusted frequencies in the academic corpus used

Table 4. A Comparison of GSL and NGSL Coverage of the 2010–2015 Section of the COCA

| Section | Proper Nouns | Foreign | Marginal | NGSL | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | NGSL | Supp | Total | GSL | Diff |
| Fiction | 3.62 | 0.04 | 0.25 | 85.38 | 0.66 | **86.03** | **85.11** | +.92 |
| Spoken | 4.65 | 0.01 | 0.65 | 88.08 | 0.70 | **88.78** | **85.59** | +3.18 |
| Academic | 6.58 | 0.10 | 0.04 | 79.57 | 0.84 | **80.41** | **71.34** | +9.07 |
| Magazine | 5.29 | 0.03 | 0.07 | 81.06 | 0.75 | **81.81** | **77.70** | +4.11 |
| Newspaper | 8.42 | 0.06 | 0.05 | 79.57 | 1.25 | **80.82** | **75.95** | +4.87 |
| Total | 5.67 | 0.05 | 0.22 | 82.82 | 0.84 | **83.66** | **79.34** | +4.32 |

*Note:* Coverage totals for each list are in boldface.

Table 5. NAWL Members That Appear More Frequently in Non-Academic Discourse

| Modified Lemma | *U* Values | | | Occurrences per Million in the 2010–2015 Section of the COCA | | | | |
|---|---|---|---|---|---|---|---|---|
| | COCA | CEC | NAWL Corpus | Academic | Fiction | Magazine | Newspaper | Spoken |
| *candidate* | 117.40 | 94.29 | 94 | 109.6 | 13.5 | 72.9 | 212.6 | 268.7 |
| *conference* | 90.81 | 117.04 | 64 | 141.0 | 32.6 | 77.7 | 165.3 | 75.2 |
| *click* | 42.09 | 24.61 | 8 | 27.3 | 60.7 | 70.6 | 30.1 | 28.8 |

*Note:* CEC and NAWL Corpus *U* values downloaded from http://www.newgeneralservicelist.org/

Table 6. COCA and CEC U Values for Words Whose Usage Is Increasing

| Modified Lemma | *U* | | |
|---|---|---|---|
| | COCA | CEC | *M* |
| *website* | 76.07 | 19.13 | 47.60 |
| *blog* | 31.40 | 6.60 | 19.00 |
| *immigration* | 51.69 | 16.91 | 34.30 |
| *solar* | 43.66 | 14.03 | 28.84 |
| *click* | 42.09 | 24.61 | 33.35 |

*Note:* CEC *U* values downloaded from http://www.newgeneralservicelist.org/

Table 7. Changes in Word Frequency between 1990 and 2014 for Five NGSL Candidates

| COCA Section | Occurrences per Million | | | | |
|---|---|---|---|---|---|
| | *website* | *blog* | *immigration* | *solar* | *click* |
| 1990–1994 | 0 | 0 | 28.71 | 23.79 | 6.69 |
| 1995–1999 | 3.34 | 0 | 29.69 | 23.46 | 14.55 |
| 2000–2004 | 13.39 | 1.01 | 34.33 | 35.34 | 20.48 |
| 2005–2009 | 27.90 | 21.26 | 51.05 | 40.91 | 21.23 |
| 2010–2014 | 67.87 | 23.21 | 58.55 | 43.39 | 22.23 |

*Note:* Data from the online version of the COCA at https://corpus.byu.edu/coca/.

to create the NAWL (e.g., *impact* and *authority*; data available at http://www.newgeneralservicelist.org) or because in the COCA they occur most frequently in the academic section (e.g., *aspect* and *distribution*). Setting aside such cases, *candidate*, *conference*, and *click* remain as candidates for inclusion to the NGSL (Table 5).

The second category includes five words whose usage has recently increased (Table 6), probably due to changes in technology (*website*, *blog*, *click*), world events (*immigration*), or both (*solar*). Changes over time were documented with data from the full online version of the COCA (Table 7).

The final category consists of five types (*best*, *better*, *rose*, *born*, and *criteria*) that could have been listed as constituents under current NGSL headwords but were instead classified as heads of their own modified lemmas. Let us consider *best* as an example. Although *best* is the superlative form of the NGSL headword *good*,

Table 8. Constituents of Two Modified Lemmas

| Modified Lemma | *U* | | |
|---|---|---|---|
| | COCA | CEC | *M* |
| *best* | 445.48 | 1.47 | 223.48 |
| *better* | 433.92 | 3.13 | 218.53 |
| *rose* | 78.05 | 12.01 | 45.03 |
| *born* | 102.46 | 0.25 | 51.35 |
| *criteria* | 18.83 | (unlisted) | - |

*Note:* CEC *U* values downloaded from http://www.newgeneralservicelist.org/

Table 9. Occurrences of Modified Lemma Constituents for Good and Best in the COCA

| Type | Noun | Verb | Adjective | Adverb |
|---|---|---|---|---|
| *good* | 329 | 2 | 500,960 | 6,688 |
| *goods* | 19,797 | 1 | 0 | 0 |
| | | | | |
| *best* | 86 | 1795 | 179,702 | 55,846 |
| *bests* | 2 | 131 | 0 | 0 |
| *bested* | 0 | 361 | 10 | 0 |
| *besting* | 0 | 127 | 1 | 0 |

*Note:* Data from the online version of the entire COCA at https://corpus.byu.edu/coca/.

it is not listed in the NGSL because it was treated as the headword for the modified lemma consisting of *best*, *bests*, *besting*, and *bested*, and the *U* value for this modified lemma was too low for NGSL membership. In the COCA, *best* is tagged as a noun, verb, adjective, or adverb depending on the context, and its 445.48 COCA *U* value shown in Table 8 is derived from all of these. It is unclear how the CEC *U* value for *best* was derived, but the very low value in Table 8 (*U* = 1.47) suggests it might only have included occurrences for constituents of the verbal lemma rather than across parts of speech. If so, this would explain the exclusion of this very common word from the NGSL.

For three of the words shown in Table 8, an argument could be made for inspecting frequencies at each part of speech to determine which headword to place it under. Using this approach, *best* would belong under *good* because both *best* and *good* occur predominantly as adjectives (Table 9). Similarly, *better* and *rose* would become constituents of *good* and *rise,* respectively (data omitted due to space limitations). *Born* would be listed under *bear* as it was in the original GSL, but because many modern dictionaries list *born* as an adjective rather than as the past participle of *bear*, and because the connection between *bear* and *born* is probably no longer felt (Online Etymology Dictionary, www.etymonline.com), perhaps *born* could be listed separately. Whichever approach is taken, if the frequency with which *born* occurs is more accurately reflected in the COCA data, it ought to be included in a general service vocabulary list. Finally, the case of the plural form *criteria* appears to be straightforward: it is currently a stand-alone headword in the NAWL, but it belongs under the NGSL headword *criterion*.

## Conclusions

This study corroborates the findings of Browne (2014) regarding the substantially better coverage provided by the NGSL in comparison to the GSL. Using the COCA as a secondary data source, it also found a very high level of agreement between NGSL members and words occurring with high frequency in the COCA. Twelve candidates were tentatively identified for inclusion in the NGSL. Three currently belong to the NAWL but appear commonly across discourse types; several others have seen increased use in recent years; and five others, each of which is both the headword of one modified lemma and a constituent of another, may have been inadvertently excluded from the NGSL due to the use of only partial CEC frequency values.

It is hoped that this study will promote further discussion and debate of the NGSL. Although quantitative evidence was found to support the addition of a small number of words to the NGSL, it would be helpful to know the original rationale behind the decisions to exclude these words. More broadly, similar to the commentary that supplements some entries in West's original GSL, perhaps a next useful step in the development of the NGSL could be to add brief explanations to low-frequency NGSL constituents and high-frequency non-constituents so that interested parties would better understand non-empirically grounded membership decisions.

## Note

[1] This GSL variant is not organized according to Bauer and Nation's (1993) word family levels, as it was made prior to that publication (P. Nation, personal communication, September 27, 2017).

## Acknowledgements

I would like to thank Brent Culligan and Charlie Browne for helping me understand technical aspects of the NGSL. In addition, this article is part of a larger collaborative study with Phil Bennett, whose many valuable insights are reflected here. This research was funded in part by a 2018 JALT Vocabulary SIG grant.

## References

Anthony, L. (2013). *AntWordProfiler (Version 1.4.0) [Computer Software]*. Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253–279. doi:10.1093/ijl/6.4.253.

Browne, C. (2014). The New General Service List 1.01: Getting better all the time. *Korea TESOL Journal*, *11*(1), 35–50. Retrieved from https://koreatesol.org/sites/default/files/pdf_publications/KTJ11-1web.pdf#page=44

Browne, C., Culligan, B. & Phillips, J. (2013a). *The new academic word list.* Retrieved from http://www.newgeneralservicelist.org

Browne, C., Culligan, B., & Phillips, J. (2013b). *The new general service list.* Retrieved from http://www.newgeneralservicelist.org

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

Davies, M. (2008) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present [Corpus].* Retrieved from https://corpus.byu.edu/coca/

Heatley, A., Nation, P., & Coxhead, A. (1994). *Range [Computer software].* Retrieved from https://www.victoria.ac.nz/lals/about/staff/paul-nation

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics.* Advance online publication. doi:10.1093/applin/amw050.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28*(2), 291-304. doi:10.1016/S0346-251X(00)00013-0.

Nation, I. S. P. (2016). *Making and using word lists for language learning.* Amsterdam, Netherlands: John Benjamins Publishing Company.

Ward, J., & Chuenjundaeng, J. (2009). Suffix knowledge: Acquisition and applications. *System, 37*(3), 461–469. doi:10.1016/j.system.2009.01.004.

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