*Reading in a Foreign Language* April 2011, Volume 23, No. 1
ISSN 1539-0578 pp. 65–83

# A primer on the General Service List

Leah Gilner
Bunkyo Gakuin University
Japan

## Abstract

This paper aims to be an introduction to the General Service List (GSL) that brings together descriptive data with material otherwise dispersed throughout the literature. The discussion first provides an historical overview of the work that scholars, researchers, and educators used as foundations for the manufacturing of the GSL. Following, a collection of modern studies is presented in an effort to critically assess the contents and intent of the GSL. In this manner, the paper attempts to provide comprehensive information on the manufacture, content, characteristics, and analyses of the GSL that can serve to inform those interested in the GSL, in particular, and the compilation and assessment of new word-lists, in general.

***Keywords***: word learning, vocabulary teaching, corpus based word lists, token coverage, high frequency words

Over 100 years of analyses of English corpora unequivocally agree on the fact that relatively few words amount for most of the vocabulary used. Approximately 2,000 words account for 70% to 95% of all running words regardless of the source of the text. The most frequent words in English are effectively ubiquitous; there is no escaping them. The relevance of this fact is of utmost importance in all aspects of language development, be that the acquisition of listening, reading, speaking, or writing skills. Instruction that focuses on the most frequent words of English provides students with the largest vocabulary gains possible. Instruction that ignores the most frequent words fosters significant and crippling vocabulary gaps. For let us be clear in this respect: The most frequent words in English are the lexical backbone, or foundation, upon which language use—and, therefore, communication—takes place.

The literature on lexical distributions shows that frequency-range-based word-lists compiled from modern corpora are alike the General Service List (GSL), both in content and token coverage. Divergences exist, but these are not unequal to those found among frequency-range-based word-lists compiled from different modern corpora. Despite the decades passed—and, perhaps, contrary to intuition—analyses of corpora keep returning the same lexical set over and over. Interestingly, our understanding of what to do with this information has not evolved significantly. The GSL remains the best researched of all frequency-range-based word-lists, and the amount of existing research still does not match the relevance of the subject. We are in need of additional—if not new—criteria with which to create, augment, and evaluate word-lists as well as their application in pedagogy. By means of an introduction to the GSL, this paper aims to recognize the value of historical research on the matter of frequency-range-based word-lists, to explicitly state the selection criteria used then and now, and to bring attention back to the importance of language description in language teaching and learning.

It is no accident that the scholar and educator considered by some to be the father of Applied Linguistics—Harold Palmer (1877-1949)—devoted a great deal of his efforts to language selection and description. Language selection is, after all, at the core of language instruction. Moreover, the decision as to what to teach and why, is bound to a description of language that is suitable to the purposes of language instruction.

It is precisely because of the existence of manuscripts containing selections of language that we can trace English instruction (as a foreign language) back to descriptions of grammar by early educators such as John Wallis, Ben Jonson, Christopher Cooper, and Elisha Coles among others in the 1600s. Since then and over the course of several hundreds of years, we continue to find selections and descriptions of different significance, scope, depth, and applicability, some more influential than others, on the exploits and insights of subsequent scholars (for an overview, see Darian, 1972; Howatt, 2004). It is not until the early decades of the 20th century, however, that a concerted, systematic, and formalized approach to selection first took place in the so-called Vocabulary Control Movement.

In retrospect, the chief material contribution of the Vocabulary Control Movement was what is nowadays commonly referred to as the GSL, a subset of the English lexicon selected by means of objective and subjective descriptions of the English language (Carter, 1998; Faucett, Palmer, Thorndike, & West, 1936; Palmer, 1931; Schmitt, 2000). The original publication (West, 1953) is no longer in print, but the word list itself it widely available (e.g., http://www.sequencepublishing.com/academic.html). This paper aims to be an introduction to the GSL that brings together descriptive data with material otherwise dispersed throughout the literature. We begin with an historical overview of the work that scholars, researchers, and educators used as foundations for the manufacture of the GSL. These efforts will then be contrasted against modern studies, seeking to critically assess the contents and intent of the GSL. In this manner, the intention is to provide comprehensive information on the manufacture, content, characteristics, and analyses of the GSL that can serve to inform those interested in a data-driven assessment of the GSL as well as the mechanisms behind the compilation of word-lists—past, present, or future.

## Historical Background

What is commonly known as the General Service List (West, 1953) is actually a reissue of the *Interim Report on Vocabulary Selection* (Faucett et al., 1936). The Interim Report—an annotated vocabulary list of about 2,000 words—was the result of two conferences held in the mid-1930s sponsored by the Carnegie Corporation with the purpose of examining "the part played by word-lists in the teaching of English as a foreign language" (Faucett et al., 1936, p. 1). The Carnegie conferences provided a forum where complementary and competing perspectives in vocabulary description and selection were debated, and a consensus could be reached. In order to properly understand the origins and inception of the GSL, therefore, it is necessary to look back to the early decades of the 20th century and the work of researchers such as Edward L. Thorndike, Ernest Horn, and Itsu Maki among others as well as pioneers in the field such as Lawrence Faucett, Harold Palmer, and Michael West.

Because of its impact on the description of language use, Thorndike's *The Teacher's Word Book* (1921) and *A Teacher's Word Book of 20,000 Words* (1931) merit first mention. In 1921, Thorndike presented frequency information for the 10,000 most frequently occurring words in a corpus of 4.5 million running words from 41 sources including the Bible, children's story books, textbooks, trade manuals, and periodicals. In 1931, Thorndike revised previous results by conducting "extensive additional counts from over 200 other sources including about 5,000,000 words" (Thorndike, 1932, p. iii) and expanded his frequency list to 20,000 words. He was aware that strict frequency counts were not necessarily informative and palliated this by including the parameter of range (occurrence across texts) and by cross-checking against frequency lists compiled by others. The result was a solid revision and validation of his ranking of the 10,000 most frequently occurring words in the English language. A famed psychologist with an interest in learning in general, Thorndike's work in lexical frequency had a significant influence on the discussions that took place at the Carnegie conferences. Together with Horn's work, it provided the scientific basis for objective criteria in the compilation of the GSL.

Horn's work in lexical frequency was fueled by interest in curriculum design and, in particular, the identification of spelling vocabularies. In 1926, Horn published *A Basic Writing Vocabulary* where he provided the 10,000 most frequent words in a collection of texts from 65 sources (business and personal letters, minutes, newspapers and magazines) that, together, added up to a computed estimate of 5,136,816 running words. Horn was keenly aware of "the lack of uniformity among the various investigations in the method of tabulating words" (Horn, 1926, p. 17) that plagued previous studies and he labored to solve the many discrepancies in methodology by providing a "critical review of twelve investigations" (Horn, 1926, p. 8) and implementing a "credit system" (Horn, 1926, p. 50). Like Thorndike, Horn took into consideration the distribution of words across texts and it is a variable of selection addressed by his credit system. Unlike Thorndike, however, Horn's list reported all inflections and derived forms separately.

The possibilities offered by Thorndike's and Horn's work were not missed by Faucett as he focused his efforts on the design of adequate materials for learners of different levels. While teaching in Japan in the early 1930s, Faucett had become interested in the elucidation of objective criteria to apply to the selection of word-lists. In collaboration with Itsu Maki, Faucett combined the two most extensive and credited frequency lists of the time, Thorndike's (1921, 1931) and Horn's (1926). The rationale was "the necessity for combining the two lists on a common statistical plan in order that teachers may readily get the total ratings of words on a 10,000,000 word-occurrence basis" (Faucett & Maki, 1932, p. 2). The Faucett-Maki word-list was published as *A Study of English-Word Values Statistically Determined from the Latest Extensive Word Counts* (Faucett & Maki, 1932), and it would later be the starting point in the selection process at the 1934 Carnegie conference.

Parallel to these efforts, Palmer pursued a different approach to selection based on "definite principles as distinguished from lists […] compiled by objective methods alone" (Palmer, 1931,

p. ii) while serving as director of the Institute for Research in English Teaching (IRET) in Tokyo. His principles included aim of selection, mode of listing, degree of utility and equability, and consistency of word categorization. Palmer was gradually persuaded of the merits of the objectivity afforded by word counts and his 3,000 word-list acknowledged Thorndike's data (Palmer, 1931). It is worth noting that, in Palmer's word-list, entries were presented together with a collection "of their commonest derivatives and compounds" (Palmer, 1931, p. 6). The approach was innovative. It presented a 'word' as a 'headword' (a term coined by Palmer), thereby ensuring a degree of coherence of organization and selection. Palmer's conceptual precision and persuasive argumentation were to have a profound impact on the selection process at the Carnegie conferences despite the fact that any implementation of his principles could not be anything other than largely subjective. Palmer's principles would complement the statistical data derived from other sources.

The defining vocabulary devised by West and used in the creation of the *New Methods Readers* Series (1927) was based on experience and intuition. While serving as an officer of the Indian Educational Service in the early 1920s, West undertook the task of making reading materials more accessible to his students through vocabulary control. He felt that a proper distribution of lexical items that decreased the density of new words would, in turn, increase the readability of a text (Schmitt, 2000; Howatt, 2004). Arising from the extensive work invested in the rewriting and paraphrasing of instructional materials that deliberately discriminated known from unknown words, West eventually isolated a list of about 1,800 words. Along with Palmer's list, West's contribution would constitute the brunt of the subjective word choices brought to the conferences.

In this manner, the Carnegie conferences brought together leading researchers in the field and set out to reach a consensus regarding a vocabulary standard that would be of general service to learners of English as a foreign language. The first conference, "The Use of English as a World Language" held October 15th-19th 1934 in New York, appointed a committee consisting of Faucett, Palmer, Thorndike, and West that would frame, classify, and itemize a tentative word-list. The process of word selection drew on the expertise of the committee members and involved both objective and subjective criteria. The objective selection was primarily based on the Faucett and Maki (1932) frequency list while the subjective criteria was embodied in Palmer's principles, on the one hand, and in Palmer's word-list and West's definition vocabulary, on the other.

The committee recorded the process of selection by means of a "plan of work" (Faucett et al., 1936, p. 11). First, the 1,500 highest ranking words in the Faucett-Maki-Thorndike-Horn list were considered, and "no word was definitely included in the main vocabulary save on a unanimous vote" while "doubtful words were marked and discussed individually" (Faucett et al., 1936, p. 11). Next, they appraised the contents of Palmer's and West's word-lists. A second inspection was then made of the Faucett-Maki-Thorndike-Horn list that took into consideration up to the 5,000 most frequent words. Ten other unidentified "selected lists" (Faucett et al., 1936, p. 12) were consulted. Lastly, a review and reevaluation was made of words that had previously failed to receive the support of all participants.

The committee articulated the subjective criteria for inclusion/exclusion of words as consisting of: structural value, universality, subject range, definition words, word-building potential, and style (Faucett et al., 1936). Words of structural value, namely function words such as prepositions, pronouns, auxiliary verbs, conjunctions, and determiners, were included. Also included were words whose use was not limited to a certain time or place (universality), while those used in restricted domains (religion, moral concepts, proper names) were excluded. Words that were deemed useful for the personal and professional development of students between the ages of 12 and 18 were considered to be of wide subject range and were included. Words perceived as necessary for students to express themselves or for a teacher to define an unknown word were included. In the presence of other criteria, the root of a word was sometimes taken into consideration for its word-building potential as were words that were thought to afford learners more precise expression (style).

Held eight months later on June 11, 1935 in London, the second conference revised the list and issued a formal publication. Arriving at a consensus had been considered essential to the process and was to remain so. The *Interim Report on Vocabulary Selection* (1936) expressly stated the tentative nature of the list and emphasized the desire for feedback by including a questionnaire at the back. The intention was to continue the process of evaluation by encouraging criticism and experimentation. “By 1939 arrangements had been made for its revision, and, with the agreement of the Carnegie Corporation, Dr. Michael West was invited to carry out the work” (West, 1953, p. vi). However, a number of factors, most notably World War II, delayed the work and it was not until 1953 that *A General Service List of English Words* was published by West.

The intervening years between the last of the conferences and the publication of the GSL saw two related works that would make a significant contribution to West’s revision, Lorge and Thorndike’s *A Semantic Count of English Words* (1938) and Lorge’s *The Semantic Count of the 570 commonest words* (1949). As was the case with the word counts in Thorndike (1921, 1931) and Horn (1926), the semantic counts were conducted manually. The monumental effort still boggles the mind. Funded by the Rockefeller Foundation, the 1938 study employed a corpus of 2.5 million words that was later expanded to 5 million. A small army of specially-trained personnel read through the corpus, assaying the context in which a word occurred and making note of its corresponding meaning if it fell within the 32 pages assigned to each researcher from the *Oxford English Dictionary* (specifically, the 13-volume 1933 edition). The 1949 supplementary study was undertaken in order to reorganize the data presented in Lorge and Thorndike (1938) for use in the revision of the *Interim Report*.

As mentioned, the publication of the *Interim Report* word-list explicitly sought feedback. While we know that Faucett coordinated the effort of collecting and analyzing the questionnaires received from around the world (Warwick, CELTE), West’s introduction to the *A General Service List of English Words* does not indicate that these were taken into account. Rather, West avers: “In this reissue of the Carnegie *Report on Vocabulary Selection* the various meanings and uses of the selected words have been rearranged in the light of the Semantic Count [by Thorndike and Lorge]” (West, 1953, p. vii). Indeed, there is no mention of any change whatsoever to the contents of the list and we can only assume that, even if West had new information or insights concerning alternative content, he acknowledged and respected the consensus reached at the Carnegie conferences.

Modifications were made in regards to the senses and derived forms, and these were explicitly marked in the text (for an example, see the entry for the word *line* on page 280 of West, 1953). When contrasted against Lorge's data, it was found that the Interim Report list included some less frequently used senses and excluded some more frequently used ones. According to Lorge: "These facts became, in part, the basis for revising the selection of meanings" (in West, 1953, p. xiii). West stated that the major contribution of the GSL is the isolation of that which is "really essential" (1953, p. viii) for the learner while, at the same time, showing the relative importance of each item. Thus, the GSL presents about 2,000 headwords, each with a choice of meanings and derivations based on (and including) frequency information. The resulting compilation was innovative and groundbreaking, a "major advance on previous word lists and [the breakdown by senses] greatly increased the pedagogical value of the work" (Howatt, 2004, p. 289).

Although available in many university (and otherwise) libraries, West's (1953) *A General Service List of English Words* is out of print. Table 1 shows an example entry. It should be noted that the data in Table 1 is arranged to resemble as closely as possible the printing format of the book. The table borders and header are not in the original although the tabular arrangement of the entry is the same.

Table 1. *An example entry from West's (1953) "A General Service List of English Words*."

| Lemma + POS | Freq. rank | Sense # | Senses |
|---|---|---|---|
| POOR, adj. | 1096e | -1 | *(having little money)*<br>Rich and poor<br>Poor people's children |

Summing up, contemporary researchers identify the following as the most relevant characteristics of the GSL: Frequency, universality (words used in all countries), utility (words used to talk about a wide range of topics), and usefulness (words that can be used to describe or define other words) (Carter, 1998; Carter & McCarthy, 1988; Nation, 1990). Altogether, over three decades of work by an international group of leading researchers resulted in the GSL as we now know it. This list of words (headwords) represented a turning point for language selection and description in the context of language instruction. Its grounding in corpus analysis attested to this. It does not imply, however, that the GSL is either perfectly adequate or superior to all other word-lists. Rather, the GSL was conceived in order to target a specific need, namely, the selection of a core vocabulary of general application in foreign language instruction (Faucett et al., 1936). This being the intention of the authors of the GSL, the following sections assess the degree of success achieved.

## Explicit criticism of the GSL

Unfortunately, study-based sound criticism of the GSL is wanting and the literature is confined to discussions regarding limitations of the utility of the GSL in terms of range (Engels, 1968), age (Richards, 1974), and expandability (Gilner & Morales, 2008a). Before discussing those papers, it is worth mentioning that a wider survey of studies (either using or analyzing the GSL) quickly reveals that there seems to be disagreement as to the actual number of words on the list.

Engels (1968) reported using a list of 3,372 words, Nation and Hwang (1995) a list of 2,147 word families, Nation (2004) a list of 1,986 word families, and Gilner and Morales (2008a, 2008b) a list of 2,284 headwords. Inspection of *A General Service List of English Words* (West, 1953) may explain this divergence. Several inconsistencies appear in, for example, what constitutes a headword (as in the prefix 'mis-') or what constitutes a word family member (as in the use of 'mother-', 'drinking-', or 'district-' in an unspecified number of compounds). Strictly speaking, *A General Service List of English Words* contains 1,907 main entries and 3,751 orthographically different words (in principle, common derivatives and compounds).

Engels (1968) criticized the GSL for not having sufficient "range-possibility" (p. 226). The paper reports on a study that considered the GSL to have 3,372 words and that examined the coverage these words provided for a set of ten randomly selected texts of 1,000 words each. He concluded that the most frequent 1,000 words of the GSL "are most useful words for all didactic purposes" (p. 221) since the analysis found that a word belonging to this subset would appear in up to 9 out of 10 texts. Engels then proposed that "the last 2,000 words cannot be called *general service-words* [italics in the original]" (p. 226) as these failed to appear enough times across texts (from 5 to 2 or less, often none).

Although Engels acknowledges that the size of his collection of texts was small (10,000 words), there is an unmistakable failure in the methodology of the study. Even if one were to think up ten 1,000-word texts made exclusively from the 3,372 GSL words and having these words uniformly distributed across the texts, each word could not appear in more than three texts. In other words, it is simply unrealistic to expect any set of 3,372 words to yield an informative measure of range under these conditions. Moreover, since we know that frequency distributions are significantly sloped among frequent words, it is not surprising that many of the 3,372 GSL words appeared in few (if any) texts. Engels reports that the first 1,000 words on the GSL accounted for 7,312 of the 10,000 words in his collection of texts, leaving the remaining 2,372 GSL words to be represented by an average of 266.8 words per text. The methodology of Engels's 1968 study (entitled "The Fallacy of Word-counts") is clearly defective and the paper's conclusion that the GSL lacks range is evidently coerced by its own methodology (and agenda?). Again, it is numerically impossible to find 3,372 words in a 1,000 word text.

Richards (1974) criticized the GSL for being dated. He observed that language use and instructional needs have changed since the creation of the GSL in the first decades of the 20th century. A consequence of this, he posited, is that "one is immediately struck by the fact that [the GSL] contains a great number of words of limited utility" (p. 71), and he cited, among others, *apologize, express, fear, lump, loyal, mannerism, mild, motion, rain,* and *scent*. In contrast, Richards proposed a number of "words common in the 1970s" (p. 71) and missing from the GSL, specifically, *astronaut, helicopter, pilot, rocket,* and *television.*

Regrettably, the 100,000,000 word British National Corpus (hereafter BNC) was not available at the time, and Richards could not know that the words he considered to be of limited utility have, in general, range and dispersion values comparable to those he proposed as being more useful. Consulting the BNC alphabetical frequency 794,771 word-list provided by Leech, Rayson, and Wilson (2001), we find (frequency-range-dispersion): *apologize* (11-96-89), *express* (121-99-93), *fear* (53-100-94), *lump* (15-98-93), *loyal* (20-96-93), *mannerism* (2-65-89), *mild* (18-100-

93), *motion* (53-99-87), *rain* (64-100-86), and *scent* (11-88-88), as compared to *astronaut* (2-50-81), *helicopter* (16-96-88), *pilot* (43-100-88), *rocket* (9-92-88), and *television* (102-100-93). It should be noted that Richards (1974) explicitly mentioned 17 additional GSL words that he considered questionable. These words have range and dispersion values in the BNC similar to those just cited.

Gilner and Morales (2008a) compiled an English Language Teaching (ELT) corpus of 1,157,493 running words and utilized Nation's BNC-based 14,000 word families as well as the GSL to elicit the lexical distributions in the corpus. The ELT corpus contained eight collections of authentic texts (interview scripts, children's stories, adult novels, movie scripts, technical descriptions, and newspaper articles from three sources) that are often used in the university classroom. The collections deliberately represented a natural grading in terms of linguistic difficulty and expository complexity. Both the GSL and Nation's BNC-based word-lists were used to profile the ELT corpus. Gilner and Morales' study showed that the GSL is neither dated nor lacking sufficient "range-possibility." In fact, the GSL performed in a manner remarkably similar to Nation's list which is based on modern corpora. This is unsurprising since Gilner and Morales also inspected the contents (actual words) of both lists and found a great deal of agreement. Nonetheless, Gilner and Morales' criticism of the GSL identified expandability as a problem for two reasons. First, expanding the GSL is impossible if one tries to remain faithful to the objective/subjective criteria used in its creation. New word-lists can and have been created to complement the GSL, but none can add new word families to the GSL itself without violating the original objective/subjective criteria. Second, research shows that, for example, topic words are found in the 4,000 to 6,000 frequency bands (Nation, 2006), forcing us to look beyond the GSL, that is, beyond a word-list locked in content by its own selection criteria.

As mentioned, criticism is scarce and, with the exception of Gilner and Morales' study (that stated the relatively obvious), it is uninformative. However, additional information about the GSL can be elicited from the studies in the following section.

## Content, coverage, and range of the GSL versus modern word-lists

Nation and Hwang (1995) compared the GSL with word-lists extracted from the Lancaster-Oslo-Bergen (LOB) corpus (Johansson, 1978) and Brown corpus (Francis & Kucera, 1978). The LOB (British English) and Brown (American English) corpora are made up of about 1,000,000 running words each. Both corpora are equally divided into 15 subsections by genre such as reporting, religion, general fiction, and science. The LOB and Brown word-lists were obtained by taking the most frequent words that appeared in 10 or more subsections (range) in each corpus. The final LOB word-list consisted of 1,810 items and the Brown word-list of 2,410 items. Note that, in this study, the GSL contained 2,147 items.

The three word-lists (GSL, LOB, and Brown) were compared and the overlapping items were examined: 1,331 words were found to be shared by all three lists; an additional 614 words were shared by any two of the three lists (Brown/LOB, GSL/Brown, or GSL/LOB). The Brown/LOB overlap added up to 250 items, the GSL/Brown overlap totaled 226 items, and the GSL/LOB overlap came to 138 items. That means 452 GSL words were not found on either of the other two lists as compared to 91 words on the LOB list and 333 words on the Brown list. A comparison of relative inclusion between word-lists is shown in Table 2. It should be noted that the table is not part of Nation and Hwang (1995) but derived from the data provided therein.

Table 2. *Relative measure of inclusion among word-lists.*

| | *Shared* | |
|---|---|---|
| GSL in LOB | 1469 | 68.42% |
| GSL in Brown | 1557 | 72.51% |
| Brown in LOB | 1581 | 65.60% |
| Brown in GSL | 1557 | 64.60% |
| LOB in GSL | 1469 | 81.16% |
| LOB in Brown | 1581 | 87.84% |

We can see that the LOB word-list has the largest amount of shared items. This degree of inclusion is diminished for both the GSL and Brown lists in approximate measures. On average, the LOB list shares 82.5% of its words, the Brown list shares 65.1% of its words, and the GSL shares 70.5% of its words.

While we cannot explain the divergence between the LOB and Brown lists (229 words and 829 words unique to each list, respectively) other than by questioning the size of the corpora and dialectal differences, the results coincide with what we know of the GSL, namely, that it contains the 1,500 most frequent words in the English language (according to the Faucett-Maki-Thorndike-Horn list). Thus, it is posited that the comparison of the GSL with word-lists obtained by frequency and range alone should indeed correlate to a large degree.

The LOB corpus was also used by Nation and Hwang to measure the text coverage provided by the GSL as well as word-lists composed of the overlapping portions of the three lists. The 1,331 word families shared by all lists provided 78.3% coverage of the LOB corpus. Adding the 614 word families shared by any two lists (total 1,945 word families) brought coverage up to 83.4%. The 2,147 word families of the GSL were found to provide 82.3% coverage of the same corpus. Thus, replacing the items that only occur on the GSL with words shared by more recently compiled lists resulted in a difference of 1.1% in coverage. Nation and Hwang observed that the difference was significant but not great.

It is of relevance to note that the corpora used for the manufacturing of the GSL were in excess of five times larger than either the LOB or Brown corpora. Nation (2004) further explores the GSL composition and coverage by comparing the GSL with the BNC, a corpus of much larger size consisting of, as mentioned, 100,000,000 words of spoken (10%) and written (90%) discourse.

Nation extracted three lists containing 1,000 words each from the BNC. The BNC first 1,000 word-list was made by identifying those words of a rank list of 6,500 lemmas that occurred 10,000 times or more in the corpus. Those words that appeared in 98 out of 100 one-million-word sub-corpora were retained and reanalyzed based on their distribution across texts (range).

Those words with a dispersion value (statistical measure of evenness of distribution) of 80 or more were retained. After sorting the retained words by frequency, the first 1,000 words became the BNC 1,000 and were expanded into word families (in accordance with Bauer and Nation 1993 level 6). The BNC second 1,000 and BNC third 1,000 word-lists were made in the same way but used what was left of the pool of words obtained from the initial sorting. The BNC second 1,000 words occurred from 27 to 89 times throughout the whole corpus (note the drop in frequency), appeared in 97 or more sub-corpora, and had a dispersion value of 80 or more. The BNC third 1,000 occurred 10 times or more in 95 or more sub-corpora with a dispersion value of 80 or more. Nation explained that five word families (hesitations, interjections, *alright*, *pardon*, and *fuck*) that occurred frequently in the spoken part of the BNC were included among the BNC third 1,000 even though they did not meet the range and dispersion criteria in the entire corpus. As for the GSL, numbers, days of the week, and months of the year were added to the list for the purposes of the study, resulting in 1,986 word families.

Nation considered the combination of the GSL and the Academic Word List (Coxhead, 2000) of relevance because of the complementary coverage they provide and because the BNC is largely made up from written sources that include vocabulary likely to have been deliberately excluded from the GSL by its authors. The Academic Word List (AWL) is composed of 570 word families "that are not in the GSL and that are frequent and of wide range" (Nation, 2004, p. 7). Regarding its origins, Coxhead explained that the AWL "was compiled from a corpus of 3.5 million running words of written academic text" (Coxhead, 2000, p. 213). The AWL is further described below.

One aspect of Nation's study compared text coverage provided by the GSL, the GSL+AWL, and the three BNC word-lists. Four corpora were employed: a 3.5 million token (running words) written academic corpus; a 300,000 token technical corpus; the 500,000 token Lund corpus of spoken English; and a 3.5 million token fiction corpus of texts from Project Gutenberg. Results indicate that the GSL covered 75.5% of the academic corpus, 82.5% of the technical corpus, 89.6% of the spoken corpus, and 87.1% of the fiction corpus. The combined BNC first 1,000 and second 1,000 word-lists covered 83.9%, 89.8%, 91.1%, and 86.6% of each corpus, respectively. When the coverage of the 2,556 word families of the GSL + AWL were compared with an equal amount of BNC word families (from the combined three BNC lists), text coverage was similar for three of the four corpora. Always in favor of the BNC lists, the biggest difference, 2.0%, was found for the technical corpus; a difference of about 1.0% was observed for the other corpora.

Nation then compared the contents of the GSL+AWL against the contents of the combined (three) BNC lists and found that the two sets contained largely the same vocabulary. In fact, all but four of the first 1,000 GSL words (*hurrah*, *ounce*, *scarce*, *shave*) were found among the three BNC lists. Almost all of the first 1,000 GSL words (97%) were found among the BNC first 1,000 and second 1,000 word-lists while 80% of the second 1,000 GSL words and 80% of the AWL were found among the three BNC lists. All together, 88% of the GSL+AWL is in the three BNC lists, leaving 301 of 2,556 word families unaccounted for.

Based on these findings, Nation's observation regarding the composition of the GSL is that "though the GSL was compiled long before the BNC, when supplemented by AWL, most of it can be found in [the 3,000 most frequent word families in the BNC]" (p. 9). His conclusion is that given the differences in distribution, coverage, and content - possibly a result of the age of the GSL, Nation speculates—a replacement might merit consideration although it is not entirely clear how it should be formulated. Of note, subsequent refinement of Nation's BNC list has lessened the differences in distribution, coverage, and content with the GSL. These findings are further corroborated in the previously mentioned Gilner and Morales (2008a) study.

As part of a study that utilized Nation's BNC-based 14,000 word families to profile and characterize an ELT corpus of 1,157,493 running words (described previously), Gilner and Morales (2008a) provided a comparison between the 2,000 most frequent words in Nation's BNC-based list and the GSL in terms of content, coverage, and range. Results showed that the first 1,000 BNC word families provided 80.43% token coverage of the corpus while the first 1,000 GSL word families provided 80.02% token coverage. The second 1,000 word families from the BNC and GSL provided 7.65% and 6.71% token coverage, respectively. In terms of range, the GSL and the 2,000 most frequent word families in the BNC showed similar metrics across the eight subcorpora that made up the corpus: 44.95% of the BNC and 42.55% of the GSL appear in all eight subcorpora, 24.70% and 21.90% in seven, 15.25% and 12.99% in six, 7.45% and 10.32% in five. The agreement of results comes as no surprise since an item per item comparison between the first 2,000 word families in the BNC and those in the GSL reveals a corresponding degree of agreement in content. In other words, as most of the word families are common to both the BNC list and the GSL, it follows that similar metrics are obtained.

Two other studies are relevant due to the information they provide on the coverage that the GSL affords particular genres. Sutarsyah, Nation, and Kennedy (1994) investigated text coverage of academic and technical discourse and found that the GSL provided comprehensive coverage of both. The study compared an EAP (English for Academic Purposes) corpus with an ESP (English for Specific Purposes) corpus. The EAP corpus contained 160 texts, each 2,000 words long, taken from a variety of academic fields while the ESP corpus was comprised of an Economics textbook that was 300,000 words in length. Findings indicated that the GSL provided coverage of 78.4% of the EAP corpus and 82.5% of the ESP corpus.

It is interesting to elaborate here on Coxhead's (2000) investigation into lexical coverage of academic texts mentioned earlier. Coxhead used an academic corpus of 3.5 million running words which was divided into four subcorpora representing arts, commerce, law, and science. Results indicated that the GSL accounted for, on average, 76.0% of the entire corpus: 77.4% of the arts subcorpus, 76.8% of the commerce subcorpus, 79.1% of the law subcorpus, and 70.7% of the science subcorpus. Coxhead used range and frequency to compile a word-list of statistically relevant words in the corpus yet not in the GSL. This word-list is known as the Academic Word List (AWL) and is the best known example of a companion extension of the GSL. A follow-up pilot study conducted by Coxhead and Hirsh (2007) found that the GSL accounted for approximately seven out of every 10 words in a 1.76 million-word academic corpus.

Hirsh and Nation (1992) provided data on GSL coverage of fiction texts. The investigation examined the range of vocabulary used in three short novels (on average 30,000 tokens) written for young first language (L1) readers and found that the GSL covered 90-92% of each novel. It is interesting to note that the first 1,000 GSL word families alone accounted for about 75% of both the EAP and ESP corpora (Coxhead, 2000; Sutarsyah, Nation, & Kennedy, 1994) as well as about 85% of the three short novels (Hirsh & Nation, 1992).

Summing up, modern studies show that GSL words are among the most frequent in the English language irrespective of the selection criteria (objective or subjective) that merited their original inclusion in the list. This, however, does not imply that the GSL exactly equates to the most frequent words in the English language. Studies that contrast text coverage between frequency-range-based word-lists and the GSL show that differences exist although they are not great. This is unsurprising since frequency distributions are very, very markedly sloped so that the most frequent words will dominate coverage statistics regardless of what other words are included in any particular word-list. The next section will elaborate on this point as it is important to understand it fully and properly.

## The big picture

The reader unfamiliar with the lexical distributions exhibited by English language use might wonder why so much emphasis is placed on frequent words. Naturally, words that are more common may deserve a special place in language instruction, but to what extent can frequency be used to determine if a particular word merits inclusion in language program curricula?

The answer lays in the indisputable fact that lexical distributions are immoderately biased towards a few types (words) at the expense of the entire lexicon. The numbers are staggering. According to the data provided by the analysis of the 100,000,000 word British National Corpus carried out by Leech et al. (2001), 397,041 of the 757,087 unique unlemmatized words (52.44%) in the corpus occur only once in the entire corpus, while the 100 most frequent of the same 757,087 unique unlemmatized words (0.0132%) account for almost 46 million of the running words in the corpus. Table 3 shows a breakdown of the most frequent words in the British National Corpus and their degree of reoccurrence.

Table 3. *Breakdown of the most frequent words in English according to the BNC.*

| Types | % of BNC types | Tokens | Token coverage | |
|---|---|---|---|---|
| 100 | 0.01% | 45,878,600 | Difference | Cumulative |
| 200 | 0.03% | 52,252,200 | 6.37% | 6.37% |
| 500 | 0.07% | 60,610,600 | 8.36% | 14.73% |
| 1,000 | 0.13% | 67,569,500 | 6.96% | 21.69% |
| 1,500 | 0.20% | 71,864,900 | 4.30% | 25.99% |
| 2,000 | 0.26% | 74,950,900 | 3.09% | 29.07% |
| 2,500 | 0.33% | 77,332,500 | 2.38% | 31.45% |
| 3,000 | 0.40% | 79,255,000 | 1.92% | 33.38% |
| 3,500 | 0.46% | 80,828,900 | 1.57% | 34.95% |
| 4,000 | 0.53% | 82,144,700 | 1.32% | 36.27% |
| 4,500 | 0.59% | 83,254,100 | 1.11% | 37.38% |
| 5,000 | 0.66% | 84,214,800 | 0.96% | 38.34% |
| 5,500 | 0.73% | 85,060,900 | 0.85% | 39.18% |
| 6,000 | 0.79% | 85,809,200 | 0.75% | 39.93% |
| 6,500 | 0.86% | 86,480,400 | 0.67% | 40.60% |
| 7,000 | 0.92% | 87,088,700 | 0.61% | 41.21% |

It is evident that extremely few types account for the vast majority of tokens. While the 10 most frequent words occur over 21 million times, there are 397,041 words (52.44%) that occur only once in the entire corpus (Leech et al., 2001). In other words, language users have markedly strong preferences regarding word choice so that words exhibit a "range of frequencies from 60,000 per million down to 1 per million and below" (Ellis, 2002, p. 167). The column labeled *Difference* shows the decreasing percentage of tokens provided by progressively less frequent types, demonstrating how "vertical" the slope of distribution actually is. For example, 500 additional types account for 6.95% (4th row–1,000 types) and yet only 0.96% (12th row–5,000 types) of the tokens in the corpus according to their relative frequency each set of types. The column labeled *Cumulative* shows in a different manner the same exceptional distributional bias so that, for example, the 100 most frequent types account for more tokens than the following 6,900 types together (45.8786% and 41.2101%, respectively).

The observed frequency distributions are not limited to a corpus of large size but can be elicited from any text although, naturally, the approximation of values diminishes according to size and varies according to the type of discourse used in each text. Table 4 presents the results of a quick analysis carried out for this paper in order to illustrate the phenomenon across a sampling of diverse texts. All texts were profiled using the most frequent words in the language as obtained from the BNC analysis just mentioned.

Table 4. *Frequency counts of a variety of texts.*

| | Tokens | Types | 100<br>(%) | 500<br>(%) | 1,000<br>(%) | 2,000<br>(%) | 3,000<br>(%) |
|---|---|---|---|---|---|---|---|
| BNC | 100,000,000 | 757,087 | 45.89 | 60.61 | 67.57 | 74.95 | 79.25 |
| All texts | 240,001 | 10,648 | 48.89 | 62.33 | 68.20 | 76.04 | 79.65 |
| Darwin | 208,985 | 8,925 | 48.46 | 61.45 | 67.38 | 75.58 | 79.26 |
| Carroll | 26,693 | 2,634 | 53.29 | 69.67 | 75.14 | 80.32 | 83.30 |
| NYT | 2,688 | 934 | 41.22 | 57.89 | 63.69 | 70.83 | 74.67 |
| BBC | 883 | 395 | 47.45 | 65.35 | 71.35 | 76.44 | 81.65 |
| VOA | 752 | 359 | 42.95 | 59.04 | 64.36 | 70.88 | 74.87 |

Values fluctuate although trends are uniform and correlate. Approximately half of all tokens of any text are confined to the 100 most frequent words (column labeled *100*) in the language while approximately three-fourths of all tokens of any text are confined to the 2,000 most frequent words (column labeled *2,000*). Darwin's *The Origins of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life* (6th edition, 1872) displays the closest agreement to the BNC numbers despite its age. A second work, also from the 19th century, Carroll's *Alice's Adventures in Wonderland* (Millennium Fulcrum Edition although originally published in 1865) skews towards better coverage quite possibly as the result of being intended for general rather than domain-specific audience. Two political articles randomly chosen in August 2007 from the New York Times and the BBC World News internet sites as well as a randomly chosen article from the VOA Special English internet site illustrate that coverage trends are similar to those observed elsewhere even when the size of the text under consideration is only a minuscule fraction of the size of the BNC.

The substantial coverage that frequent words provide of the language has long been recognized. As early as 1915, Ayres combined "the results of the four most extensive [frequency] studies" (Ayres, 1915, p. 6) available at the time, stating that:

> "There is one salient characteristic common to all these studies. This is the cumulative evidence that a few words do most of our work when we write. In every one of the studies it was found that about nine words occur so frequently that they constitute in the aggregate one-fourth of the whole number of words written, while about 50 words constitute with their repetitions one-half of all the words we write." (Ayres, 1915, p. 8)

Being in possession of Thorndike's and Horn's data, the phenomenon was not missed by the Carnegie committee in the 1930's. Palmer remarked:

> "A carefully selected vocabulary consisting of some 3,000 words (together with their commonest derivatives) constitutes a little over 95% of the contents of all ordinary English texts. […] On the other hand a carelessly selected vocabulary of even double the number of word-units will constitute a considerable lower percentage." (Palmer, 1931, p. 3)

The implications for language instruction are clear. "Learners could [be taught and therefore] use less frequent words, use them well, and communicate effectively with them, but they would be doing so with words different from those used, anticipated, and preferred by fluent speakers" (Gilner & Morales, 2008b).

This section has served to further the understanding that the long standing validity of the GSL is simply the result of it being, first and foremost, a frequency-range-based word-list. As it seems that language use has not changed significantly in regards to its most frequent words over at least the last century, the studies presented in this paper show that (properly compiled) frequency-range-based word-lists are bound to be largely similar regardless from which corpora they originate. Differences will exist but these, as shown, are not sufficient to conclusively favor any frequency-range-based word-list over another. Considering that the GSL is comparatively well established, it is therefore natural that experienced researchers use it in current state-of-the-art studies (Coxhead, 2000; Simpson-Vlach & Ellis, 2010).

## Selection criteria, new word-lists, and closing remarks

Language change (or the lack of it in this respect) aside, there is a self-evident reason why frequency-range-based word-lists are bound to resemble each other: the criteria for selection and validation have not changed since the GSL was first compiled. We still use frequency and range in order to determine not only how common a word is but also how widespread its use is. It does remain unclear where to establish the compromise between the two (frequency and range), that is, when a word is found in many texts yet infrequently while another word is found in fewer texts yet rather frequently. The matter only becomes more complicated when we acknowledge that our sole tool for the validation of word-lists also remains the same as 100 years ago, namely, token coverage (which is, after all, no more than a straight-forward frequency measure). It is important to note that no formalization has been advanced regarding the frequency-range compromise in the criteria for word selection.

Along with frequency and range, there is a third and last criterion that is also part of the GSL heritage, namely, the headword and its family of inflections and derivations. In this regard, however, Bauer and Nation (1993) have advanced our understanding of this form of lexical organization by providing a rationale for structured formalization based on potential learnability through the "transparency" of affixation. Their contribution makes the concept of word family far more amenable to and useful in instruction.

And so, frequency rank, range measure, word family structure, and token coverage remain the entire toolkit with which to build word-lists. The logical next step is to discriminate which senses of the words in a list are the most frequent and have the widest range. This task was undertaken for the GSL by Lorge and West more than half a century ago and has not been repeated since due to the enormous resources required. In fact, it is important to mention that, to date, vocabulary profiler software still cannot be used to distinguish senses (although it can distinguish part-of-speech) and, thus, the information provided in West's 1953 GSL cannot be used to fully quantify the token coverage of a text.

Since the criteria for selection has not changed and the method for validation cannot be programmed into a computer, instructors and researchers seeking new and improved word-lists have two basic choices. The first is to create word-lists that are complementary to the GSL. As mentioned, this has been the approach followed by, for example, Coxhead (2000). The second is to ignore the GSL and create original word-lists from scratch. As mentioned, this has been the approach of, for example, Nation (2006). Regardless of which of these two approaches is adopted, the data has shown that the result will be similar word-lists. This will remain the case as long as we are bound by the frequency-range-headword selection criteria and token coverage validation. The changes from word-list to word-list will be due to the nature of the corpus of origin of each word-list. Thus, the choice of corpus becomes the last criterion for the compilation of word-lists.

Perpendicular to the creation of word-lists are descriptions of pedagogical value of existing word-lists. For instance, Simpson-Vlach and Ellis' (2010) analysis of a 4.2 million word corpus which represented spoken and written academic discourse equally elicited collocational and colligational attributes of interest. Also of relevance, Gilner and Morales (2008b, 2009) examined the phonetic and semantic attributes of the GSL, respectively. In the first case, the 2,284 GSL headwords were transcribed and subsequently analyzed in accord with phonetic parameters such as the frequency of occurrence of the range of sounds in the phonetic inventory, the type and frequency of syllable shapes found as well as the type and frequency of consonant clusters found. Results indicated that headwords in the GSL provide extensive phonetic coverage of the language, that is, the GSL headwords encapsulate the majority of the phonetic characteristics of the English language. Gilner and Morales (2009) introduced the concepts of *cohesion* and *reach* in order to describe the semantic relationships evidenced by the GSL headwords in contrast to randomly compiled control word-lists. Simply put, *cohesion* is used to refer to the relationships that exist between/among words in a list (and the concepts they denote); *reach* is used to refer to the relationships that exist between/among words in a list and the rest of the lexicon. Unlike the randomly compiled control word-lists, the GSL headwords were found to be highly polysemous, to be tightly cohesive, and to possess a great reach. These kinds of studies provide valuable information for materials design and instruction.

Lastly, the GSL is unique in that it is one of a few artifacts to have survived the shifts in paradigms and perspectives that so often influence applied linguistics research and language teaching. Indeed, the GSL is part of the collective consciousness and one would be hard pressed to find someone in the field who has not heard of this word-list. Yet, it is less likely that, having heard of the GSL, the same individuals are aware of how it came about, the considerations underpinning its creation, or its critical assessment. This paper has provided an introduction to the GSL, a summary of its historical background, and an account of the procedures that led to its compilation. Additionally, a survey of modern studies that examine the coverage and composition of the GSL has served to contextualize it within present-day concerns. By bringing together information dispersed throughout the literature (and over the generations), we hope to have added to the shared understanding of one of the legacies in the field.

It is not unreasonable to state that, without the data presented herein, it is not possible to have a discussion regarding the validity and adequacy of the GSL or any other word-list. We cannot be satisfied with dismissing the GSL because it was compiled decades ago when the data tells us that, if there is fault to be found with the GSL, it is not due to its age. If we fail to apply proper critical mechanisms when assessing existing work, how can we do better with new research? One of the reasons why the work on word-lists has largely stalled (in relation to its importance) is precisely because of the lack of an informed discussion on the topic. What is certain is that one cannot accept that there is no merit in the frequency distributions observed in language use (that is, speaker choices) nor that data-driven identification of subsets of the lexicon is without interest to language learning, pedagogy, or the field.

## References

Ayres, L. P. (1915). *Measuring scale for ability in spelling*. Russell Sage Foundation: New York.

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

Bogaards, P., & Laufer, B. (2004). *Vocabulary in a second language.* Amsterdam: John Benjamins.

Carter, R. (1998). *Vocabulary: Applied linguistic perspectives*. New York, NY: Routledge.

Carter, R., & McCarthy, M. (1988). *Vocabulary and language teaching*. New York, NY: Longman.

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

Coxhead, A., & Hirsh, D. (2007). A Pilot Science Word List for EAP. *Revue Française de Linguistique Appliqué*, *XII(2)*, 65–78.

Darian, S. (1972). *English as a foreign language: History, development, and methods of teaching.* Norman, OK: University of Oklahoma Press.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24(2)*, 143–188.

Engels, L. K. (1968). The fallacy of word-counts. *IRAL*, *6(2)*, 213–231.

Faucett, L., & Maki, I. (1932). *A study of English-word values statistically determined from the latest extensive word counts*. Tokyo: Matsumura Sanshodo.

Faucett, L., Palmer, H., Thorndike, E. L., & West, M. (1936). *Interim report on vocabulary selection*. London: P.S. King and Son, Ltd.

Francis, W. N., & Kucera, H. (1978). Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers, from http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

Gilner, L., & Morales, F. (2008a). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguistics*, 41–58.

Gilner, L., & Morales, F. (2008b). Elicitation and application of a phonetic description of the General Service List. *System*, *36(4)*, 517–533.

Gilner, L., & Morales, F. (2009). Lexical and semantic relationships as aids to learnability: Cohesion and reach. *Nagoya University of Foreign Studies, Faculty of Foreign Languages Kiyo, 36*, 73–89.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8(2)*, 689–696.

Horn, E. (1926). *A basic writing vocabulary, 10,000 words most commonly used in writing*. College of Education, University of Iowa.

Howatt, A. (2004). *A history of English language teaching*. New York, NY: Oxford University Press.

Johansson, S. (1978). *Manual of information to accompany the Lancaster-Oslo-Bergen Corpus of British English for use with digital computers*. Oslo: University of Oslo, Department of English.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English*. Harlow: Pearson Education Limited.

Lorge, I. (1949). *The semantic count of the 570 commonest words*. New York, NY: Teachers College, Columbia University.

Lorge, I., & Thorndike, E. L. (1938). *A semantic count of English words*. New York, NY: Teachers College, Columbia University.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Boston, MA: Newbury House.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, *63*(1), 59–81.

Nation, I. S. P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System*, *23*, 35–41.

Palmer, H. (1931). *Second interim report on vocabulary selection submitted to the Eighth Annual Conference of English Teachers under the auspices of the Institute for Research in English Teaching*. Tokyo: IRET.

Richards, J. C. (1974). Word lists: Problems and prospects. *RELC*, *5(2)*, 69–84.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy.* Cambridge: Cambridge University Press.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31(4)*, 487–512.

Sutarsyah, C., Nation, I. S. P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, *25(2)*, 34–50.

Thorndike, E. L. (1921). *The teacher's word book*. New York, NY: Teachers College, Columbia University.

Thorndike, E. L. (1931). *A teacher's word book of 20,000 words*. New York, NY: Teachers College, Columbia University.

Warwick CELTE. Lawrence Faucett's life and career (n.d.). Retrieved from http://www2.warwick.ac.uk/fac/soc/celte/research/elt_archive/halloffame/faucett/life/

West, M. (1926–7). *The new method readers (new series).* Bombay and Calcutta: Longmans, Green.

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

## About the Author

Leah Gilner is an applied linguist and associate professor working in the faculty of Foreign Studies at Bunkyo Gakuin University in Tokyo, Japan. Research interests include: word knowledge acquisition, applied phonetics and phonology, fluency development, and extensive

reading. Email: leahgilner@gmail.com