Applied Linguistics 2015: 36/1: 1–22 © Oxford University Press 2013
doi:10.1093/applin/amt018 Advance Access published on 26 August 2013

Is There a Core General Vocabulary? Introducing the *New General Service List*

[1,*]VACLAV BREZINA and [2]DANA GABLASOVA

[1]Department of Linguistics and English Language, County South, Lancaster University, Lancaster LA1 4YL, UK
[2]Lancaster University, UK
*E-mail: v.brezina@lancaster.ac.uk; d.gablasova@lancaster.ac.uk

The current study presents a *New General Service List* (*new-GSL*), which is a result of robust comparison of four language corpora (*LOB*, *BNC*, *BE06*, and *EnTenTen12*) of the total size of over 12 billion running words. The four corpora were selected to represent a variety of corpus sizes and approaches to representativeness and sampling. In particular, the study investigates the lexical overlap among the corpora in the top 3,000 words based on the *average reduced frequency (ARF)*, which is a measure that takes into consideration both frequency and dispersion of lexical items. The results show that there exists a stable vocabulary core of 2,122 items (70.7%) among the four corpora. Moreover, these vocabulary items occur with comparable ranks in the individual wordlists. In producing the *new-GSL*, the core vocabulary items were combined with new items frequently occurring in the corpora representing current language use (*BE06* and *EnTenTen12*). The final product of the study, the *new-GSL*, consists of 2,494 lemmas and covers between 80.1 and 81.7 per cent of the text in the source corpora.

1. INTRODUCTION

Learning vocabulary is a complex process in which the learner needs to acquire both the form and the variety of meanings of a given lexical item. For beginner learners the main question, of course, is where to start. General vocabulary wordlists can assist in this process by providing common vocabulary items that occur frequently across different texts (Nation and Waring 1997; Nation 2001; Beglar and Hunt 2005; Carter 2012). These lists can be used directly by learners or can aid teachers or textbook writers with the selection of materials appropriate for a particular group of students. Moreover, general vocabulary lists are essential in the development of specialized wordlists (such as academic or technical vocabulary lists) where they serve as the general vocabulary baseline for identification of more specialized vocabulary (Nation and Hwang 1995).

Although there are a number of different lists of frequent lexical items available for English, by far the most influential and widely used wordlist is West's *General Service List* (*GSL*). It has been adopted in both pedagogical practice and vocabulary research (e.g. McCarthy 1990; Hirsh and Nation 1992; Nation 2004; Cobb 2012) and therefore directly influences the way in which essential English vocabulary is conceptualized. In addition, West's *GSL* served also as the non-academic baseline for the creation of the *Academic Word List* (*AWL*) (Coxhead 2000, 2011). It therefore also lies at the heart of the distinction between general and academic vocabulary.

However, a number of problems with West's *GSL* have been pointed out over the years (cf. Gilner 2011). *GSL* was often criticized with respect to the principles on which it was created as well as with respect to its utility. Generally, researchers agree that *GSL* as a guideline for L2 vocabulary learning has long been out of date and requires revision (cf. e.g. Richards 1974; Nation 1990; Carter 2012). It is important to realise that *GSL*, although published in 1953, represents a revised version of the *Interim Report on Vocabulary Selection* from 1936 (West 1953). As a result, *GSL* includes a number of items that are no longer in general use (e.g. *gay* [=happy], *cart*, *shilling*, *servant*, *footman*, *milkmaid*, and *telegraph*) and excludes newer items (e.g. *television*, *computer*, and *Internet*). Richards (1974) has also pointed out a number of inconsistencies in the selection of the words. For instance, *GSL* includes certain words from the semantic field of animals such as *bear*, *elephant*, and *monkey*, but excludes others such as *lion*, *tiger*, and *fox*. Gilner and Morales (2008), on the other hand, argue that the core of the *GSL* overlaps to a large extent with modern corpus-based wordlists (cf. also Nation and Hwang 1995; Nation 2004); however, Gilner and Morales at the same time question the possibility of expanding *GSL* given the combination of objective and subjective criteria on which the original wordlist was based.

In response to the problems identified with the *GSL*, this study offers a bottom-up, quantitative approach to the development of a *New General Service List* (*new-GSL)* by means of examining frequent general words across a variety of language corpora. At the same time, the question of which words should be included in the *new-GSL* is related to a larger issue of existence and stability of general vocabulary, which has been raised on several occasions. Some researchers (e.g. Bongers 1947; Richards 1974; Bogaards 2008) argue that wordlists based on different language corpora show a large variability in the lexical items that they contain. This claim is supported by the evidence from Nation's (2004) study that showed that although the *BNC*-based wordlist included items to a large extent similar to West's *GSL*, the individual items were distributed in a different frequency order in these two wordlists. The question of the existence of a general English vocabulary reflects a similar question asked by Hyland and Tse (2007) in relation to academic vocabulary, namely whether there exists a core academic vocabulary common to a range of academic disciplines. This study, in a similar vein, explores the issue of stability of a core general English vocabulary across a range of written and spoken contexts. The answer is sought by comparing four language corpora *LOB*, *BNC*, *BE06*, and *EnTenTen12* of the total size of over 12 billion running words.

1.1 Wordlists: a quantitative paradigm

Seen from the perspective of current corpus linguistic research (cf. McEnery and Hardie 2011), one of the main problems of West's *GSL* lies in the fact that its compilation involved a number of competing principles that brought a large element of subjectivity into the final product. When reviewing the compilation principles of the *GSL*, we can see that in addition to the quantitative measure of word frequency, West also used a number of 'qualitative' criteria for the selection of individual lexical items. These include (i) the ease of learning, (ii) necessity, (iii) cover, and (iv) stylistic and emotional neutrality (West 1953: ix–x). Let us now briefly discuss these principles.

First, according to the principle of the ease of learning, West included a number of words on the basis of the similarity of the word form, despite the fact that these words did not meet the word frequency requirement. In effect, in addition to the core vocabulary, *GSL* offers learners also a number of relatively infrequent words that can be easily acquired, but are rarely encountered in spoken or written contexts. Secondly, the principles of necessity and cover ensure, according to West, that all 'necessary ideas' are covered in the wordlist and that there are few redundancies. However, with the rapid changes in the modern society, it becomes increasingly more difficult to establish what the range of 'necessary ideas' is. In 1953, West argued that the verb *to preserve* (food), although of relatively low frequency, represents such a necessary idea, as it subsumes *canning*, *bottling*, *salting*, *freezing*, and *jam making*, which cannot be explained without the use of the hyperonym *preserve*. Arguably, for the present-day learner the verb to *preserve* may not be as important, although there is no other word that would express the same idea. The last criterion (stylistic level and emotional neutrality) guarantees that the words in West's *GSL* are stylistically unmarked expressions that focus on communication of ideas rather than expressing emotions. West presupposes that neutral expression of ideas is the primary language function that the learners of English seek and therefore excludes some stylistically marked expressions despite their high frequency. Overall, as can be seen from the examples above, there is a strong tension between the quantitative and the qualitative principles of the wordlist compilation. Moreover, some of the principles (especially the necessity principle) are highly subjective and dependent on the compiler's preferences.

By contrast, this study uses a purely quantitative approach to the compilation of a general vocabulary list that reflects current language use across a large number of contexts. This approach agrees with West in that word frequency alone is not a reliable measure for selecting words important for learners. However, instead of using additional qualitative (subjective) criteria as West did, we chose a combination of three quantitative measures: frequency, dispersion, and distribution across language corpora. These measures guarantee that the words selected for the new vocabulary list are frequently used in a large number of texts and that the wordlist is compiled in a transparent and replicable way.

1.2 Word families and lemmas

At this stage, it is important to briefly discuss the main organizing principle of the new wordlist. While West's *GSL* as well as a number of more recent pedagogical wordlists are organized according to the word-family principle (cf. Bauer and Nation 1993; Nation 2001), the lexical units used in the *new-GSL* are lemmas, that is, groups of lexical forms with the same stem that belong to the same word class (cf. Francis and Kučera 1982). This means that whereas in West's *GSL* the unit broadly understood as a word includes a set of forms related by both inflectional and derivational affixes, the *new-GSL* restricts the scope of each word to the appropriate headword and its inflectional variants. The following example illustrates the difference: a lemma with the headword *develop* (verb) includes also the inflectional forms *develops*, *developed*, and *developing*. A word family with the same headword would in addition include adjectival derivatives *undeveloped* and *underdeveloped* as well as the nominal forms *development*, *developments*, *developer*, and *developers.* Note that word families do not distinguish between word classes and thus in the example above the adjectival and verbal uses of the form *developing* (as in *developing countries* vs. *They are developing a wordlist*) are merged.

It is important to realize that the choice of the lexical unit has major ramifications for the word-selection process and ultimately for the usability of the wordlist. Word families and lemmas are connected with a different approach to lexical generalizing over large language corpora and both carry certain problems (cf. Gardner 2007 for detailed discussion). The word-family principle is based on the conviction that the lexicon can be usefully divided into larger morphologically related units. The principle operates with the underlying assumption that the meaning of a derived word is largely transparent and can be understood on the basis of the knowledge of the individual morphological components. This according to Bauer and Nation (1993: 253) guarantees that 'once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort.' This makes word families a useful tool especially for the research and pedagogy concerned with receptive uses of vocabulary (e.g. recognition of words in reading).

In many cases, however, the assumption of transparency is difficult to maintain because of two inter-connected reasons. First, it is the semantic distance of the words that can be included under one headword in a word family. For example, the semantic relationship between the pairs of words such as *to train* and *trainers* (shoes), *please* and *unpleasantly*, *part* and *particle*, *value* and *invaluable*, etc. may not be immediately apparent. Each of these pairs is included in Nation's version[1] of the *GSL* under the same word family. Secondly, the ability to use word families successfully relies on users' morphological skills, which may not always be at an adequate level. Although psycholinguistic research (Nagy *et al.* 1989) shows a close relationship of morphologically related words in the mental lexicon of L1 speakers, recent evidence on both productive (Schmitt and Zimmerman 2002) and receptive (Ward and Chuenjundaeng

2009) morphological knowledge of learners suggests that students' word-building skills are in many cases overestimated.

The *new-GSL* is conceived of as a list of the most frequent English vocabulary suitable for both receptive and productive use. With respect to pedagogical uses of the list, this wordlist is primarily intended for beginner learners whose morphological awareness and word building skills can be limited. In the creation of the *new-GSL*, the preference was therefore given to lemmas rather than word families. This enables us to restrict the wordlist to the most frequent items with greater precision than if we followed the word-family principle.

1.3 Aim of the study and research questions

The aim of this research is to develop a *new-GSL* that offers a list of lexical items frequently occurring across different texts as well as different language corpora. The study takes a purely quantitative low-inference approach to identifying the *new-GSL* items. It focuses on three main criteria of inclusion: (i) word frequency, (ii) dispersion, and (iii) stability of a lexical item across different corpora. The study addresses five research questions (RQs). The first three RQs are related to identifying the items to appear in the *new-GSL*. The last two RQs focus on the evaluation of the *new-GSL*.

RQ1: Is there a substantial overlap between frequent lexical items in different general language corpora?
RQ2: What is the lexical core common to different language corpora?
RQ3: What lexical items represent a recent development in the English language?
RQ4: What is the difference between the *new-GSL* and West's *GSL* (and the *AWL)*?
RQ5: What percentage of words in text does the *new-GSL* cover (as opposed to West's *GSL*)?

The rationale behind including both West's *GSL* and the *AWL* in the evaluation of the *new-GSL* was based on the fact that these two wordlists are interdependent (the former represents the general vocabulary baseline for the latter). Moreover, West's *GSL* and the *AWL* are often considered together in vocabulary studies as two steps in vocabulary learning (cf. Nation 2004).

The article discusses the development and evaluation of the *new-GSL*. The wordlist itself is provided in full at *Applied Linguistics* online.

2. METHOD

2.1 Data

For the investigation of the core general vocabulary in English, four language corpora were selected to represent a variety of corpus sizes and approaches to sampling and representativeness: *The Lancaster-Oslo-Bergen Corpus (LOB)*, *The*

Table 1: Corpora used: A comparison

Corpora	*LOB*	*BNC*	*BE06*	*EnTenTen12*
Tokens	1 million	100 million	1 million	12 billion
Period	1961	1990s	2005–7	2012
Variety of English	British	British	British	International
Spoken component	No	Yes (10%)	No	No
Sample size	2k words of each text	40–50k words of each text	2k words of each text	whole documents included
No. of texts	500	4,049	500	21.55 million
Sampled text-types	15 genres of writing	Imaginative (20%) and informative (70%) writing + speech (10%)	15 genres of writing	www—a wide range of documents

British National Corpus (BNC), *The BE06 Corpus of British English (BE06)*, and *EnTenTen12*. Table 1 provides basic information about these corpora.

The most notable difference between the corpora is their size, which ranges from one million tokens (*LOB*, *BE06*) to 12 billion running words (*EnTenTen12*). In fact, *EnTenTen12* is one of the largest corpora available for any language today. Three of the four corpora (*LOB*, *BE06*, and *EnTenTen12*) represent the written language only. *BNC*, on the other hand, includes a substantial spoken component of 10 million running words, with ~4.2 million words of informal conversation. The corpora also differ in their sampling policies. While *LOB* and its 'mirror' corpus *BE06* both consist of 500 samples from 15 different genres of writing, each sample comprising 2,000 words (cf. Johansson *et al.* 1986; Baker 2009), the *BNC* samples (>4,000 altogether) are usually longer, ranging between 40 and 50 thousand tokens (cf. Aston and Burnard 1998). *EnTenTen12*, which is a result of extensive web crawling (*Sketch Engine trac: enTenTen* 2012), comprises >21 million texts ranging in formality from very informal language of Internet blogs to very formal language of legal documents. Chronologically, the corpora offer samples of language from the 1960s (*LOB*) to the present (*BE06* and *EnTenTen12*). With the exception of *EnTenTen12*, which reflects the international character of English on the Internet, the corpora represent the British variety of English.[2]

All corpora used in this study were built by independent language sampling and therefore do not have any textual overlap. The detailed reasons for selecting the individual corpora are:

1. Although being relatively small corpora by today's corpus linguistics standards, *LOB* and *BE06* were carefully sampled to reflect a variety of written English genres, including newspapers, fiction, essays, and scientific writing.

They both belong to the Brown family of language corpora and were thus compiled according to the same principles. The main difference between these two corpora (as well as the main reason for selecting them) lies in the fact that *LOB* offers an insight into the use of written English in the 1960s, while *BE06* reflects the use of written English in the late 2000s. Comparing the vocabulary in these two corpora thus enables us to identify lexical changes in the English language during the 40-year period that lies between the corpora.

2. *BNC*, which was compiled in the early 1990s, represents a mid-size corpus that has become a standard tool for investigating different language patterns. It is a balanced sample of British English that includes a substantial spoken part (10 per cent). It has been a basis for the frequency dictionary of written and spoken English (Leech *et al.* 2001) and a source of frequency lists used in a number of vocabulary studies (e.g. Nation 2004; Gilner and Morales 2008; Webb and Rodgers 2009). In addition, *BNC* is the principal source of the recently published *Phrasal Expressions List* (Martinez and Schmitt 2012).

3. *EnTenTen12* is by far the largest of the four corpora used in this study. It is more than a hundred times larger than the *BNC* and more than 10,000 times the size of the *LOB* and *BE06* corpora. In contrast to the previous three corpora, its representativeness has not been achieved by meticulous proportional sampling, but is an artifact of its enormous size and coverage of a wide variety of online texts. *EnTenTen12* is a result of extensive web-crawling and cleaning of the raw data available online (*Sketch Engine trac: enTenTen* 2012). In this respect, *EnTenTen12* comes closest to Sinclair's (1991, 2004) idea of a monitor corpus, as it reflects the most current developments in the English language.

2.2 Procedure

The procedure included the following steps:

1 Creation of wordlists based on the four corpora (*LOB*, *BNC*, *BE06*, and *EnTenTen12*).
2 Comparison of wordlists pairwise (RQ1).
3 Identification of a common lexical core among the four wordlists and extraction of the shared items (RQ2).
4 Identification of lexical items reflecting recent vocabulary changes in the English language based on *BE06* and *EnTenTen12* (RQ3).
5 Creation of the *new-GSL.*
6 Comparison of West's *GSL* and the *AWL* with the *new-GSL* (RQ4).
7 Investigation of the coverage of text by the *new-GSL* and West's *GSL* (RQ5).

The first step of the procedure was to generate wordlists based on the four corpora described above. For this purpose, we used the *Sketch Engine*, which is a sophisticated web-based corpus-handling tool designed primarily for lexicographic purposes (Lexical computing 2012). The main reason for choosing the

Sketch Engine was the fact that it provides access to all four corpora used in this study (*LOB*, *BNC*, *BE06*, and *EnTenTen12*).[3] In the *Sketch Engine*, we created four lemmatized wordlists (*LOB*, *BNC*, *BE06*, and *EnTenTen12*) that included the information about the word class[4] as well as the absolute and the *average reduced frequency (ARF)* of each lemma. ARF is a measure that takes into account both the absolute frequency of a lexical item and its distribution in the corpus (Savický and Hlaváčová 2002; Hlaváčová 2006). Thus if a word occurs with a relatively high absolute frequency only in a small number of texts, the ARF will be small (cf. Čermák and Křen 2005; Kilgarriff 2009). All four wordlists were then sorted according to the ARF that ensured that only words that are frequent in a large variety of texts appeared in the top positions in the wordlists.

The frequency lists ordered according to the ARF were further manually processed to ensure that only words with valid non-specific meanings were included. In this procedure, all proper nouns, letters, and erroneous entries were deleted. In addition, spellings of all words were standardized according to the British norm. Finally, the first 3,000 words (according to the ARF ranking) from each of the lists were extracted and compared. Table 2 offers an overview of the wordlists used in the comparison.

As can be seen from Table 2, all four wordlists included similar proportions of the individual word classes. Almost half of the items were nouns followed by verbs (22 per cent), adjectives (16–17 per cent), and adverbs (7–9 per cent). Grammatical words (conjunctions, prepositions, pronouns, determiners, and quantifiers) represented between 5 and 7 per cent of the items.

To address the first RQ (*Is there a substantial overlap between frequent lexical items in different general language corpora?*), a series of pairwise comparisons between the vocabulary items on the four wordlists (sorted by ARF) were carried out. The aim of the comparison was to establish the percentage of common lemmas between pairs of wordlists. In addition, for each comparison, the ranks of the common lemmas in the two wordlists in question were correlated using Spearman's rho (Oakes 1998) to ascertain whether the positions

Table 2: Distribution of word classes in the wordlists

Wordlists	Word class						
	Nouns	Verbs (+ modals)	Adjectives	Adverbs	Conjunctions and prepositions	Pronouns	Other (gram. words)
LOB-3000	1470 (49%)	637 + 10 (22%)	490 (16%)	221 (7%)	73 (2%)	28 (1%)	71 (2%)
BNC-3000	1369 (46%)	635 + 10 (22%)	516 (17%)	271 (9%)	101 (3%)	23 (1%)	75 (3%)
BE06-3000	1489 (50%)	656 + 10 (22%)	465 (16%)	212 (7%)	77 (3%)	26 (1%)	65 (2%)
EnTenTen12-3000	1479 (49%)	656 + 10 (22%)	478 (16%)	211 (7%)	75 (3%)	27 (1%)	64 (2%)

of the items in the individual wordlists were similar. Overall, six pairwise comparisons were carried out.

To address the second RQ (*What is the lexical core common to different language corpora?*), all four wordlists were compared, and the shared items were extracted. In addition, the ranks of the individual items in the different lists were put together and the median rank was calculated. With four wordlists, the median rank was a useful measure of the central tendency in the data, as it was calculated as a mean rank between two central values.

To address the third RQ (*What lexical items represent a recent development in the English language?*), the wordlists based on the two most recent corpora *BE06-3000* and *EnTenTen12-3000* were compared, and the shared lexical items were extracted. These were in turn compared against the common lexical core extracted as part of RQ2 (see above) and a list of current words was created. For each of the current words, a mean rank was calculated based on the rank in *BE06-3000* and *EnTenTen12-3000.*

The compilation of the *new-GSL* involved combining the wordlists created in the previous two steps: the common lexical core items were put together with current words shared by the two corpora reflecting the present-day use of the English language (*BE06* and *EnTenTen12*). The items in the *new-GSL* were then marked according to the presentation conventions in Table 3 and the list was alphabetized. Finally, all entries were manually checked for consistency.

Once created, the *new-GSL* was compared with West's *GSL* and the *AWL* (to address the fourth RQ). The headwords of the *new-GSL* were traced in West's *GSL* first and second 1000 word families as well as in the *AWL* to find out not only the general overlap but also the differences in distribution of the vocabulary items. For this procedure, VocabProfile in Cobb's (2012) *Compleat Lexical Tutor* was used.

To address the last RQ (*What percentage of words in the four corpora does the new-GSL cover?*), four new lemmatized wordlists with absolute frequencies of occurrences of words were created. The items in the *new-GSL* were then traced in the new wordlists based on the absolute frequencies. The overall percentage of coverage of the *new-GSL* items was calculated for each of the corpora. For comparison, the coverage of West's *GSL* was also calculated for all four corpora.

Table 3: Presentation conventions of the new-GSL

FIRST 500 WORDS (BOLD, RED, and CAPITALS) **500–1000 words (bold type)** 1001–2500 (plain type)	***new-GSL:* base**
Current words (italics)	***new-GSL:* current**

3. RESULTS AND DISCUSSION

3.1 Comparison of wordlists pairwise (RQ1)

To test the stability of lexical items across different corpora, the four wordlists (*LOB—3000*, *BNC—3000*, *BE06—3000*, and *EnTenTen12—3000*) were compared. The items in these lists were sorted by the ARF to take into consideration both the frequencies of the words and their dispersions in text. Table 4 provides an overview of the pairwise comparison of the individual wordlists. For each comparison, the number of shared items (together with the percentage of overlap) and the correlation between the ranks of the individual vocabulary items are provided.

As can be seen, for all six pairwise comparisons, the number of shared items is very high with a 78 to 84 per cent overlap. Moreover, the overall high correlations between the ranks of the shared vocabulary items ($r_s = .762$ to $r_s = .870$, all $p < .001$) indicate that the shared words are distributed in a comparable way in the wordlists. This result is a first indication of the fact that there is a strong stable core of common vocabulary.

Let us look at some individual comparisons in detail. Tables 5–7 show comparison of selected pairs of corpora that provide a strong contrast with respect to (i) their date of compilation, (ii) size, and (iii) the sampling technique. Each of the tables provides a breakdown of the overall comparison score into individual word classes establishing that the pattern of overlap is stable across all types of words.

Table 5 displays a comparison between the wordlists based on the *LOB* corpus and its current counterpart *BE06*. We can see that although the two corpora were compiled with a 40-year gap between them, they show a remarkably large lexical overlap both in total and across different word classes. In particular, grammatical words (conjunctions, prepositions, pronouns, determiners, quantifiers, etc.), which occur mostly within the first 500 words in the lists, have almost identical rank distributions as the nearly perfect correlations (.971, $p < .001$ and above) indicate. Lexical words (nouns,

Table 4: Comparison of the first 3,000 vocabulary items

Corpora	*LOB-3000*	*BNC-3000*	*BE06-3000*	*EnTenTen12-3000*
LOB-3000	x	2,497 (83.2%) $r_s = .870$, $p < .001$	2,458 (81.9%) $r_s = .832$, $p < .001$	2,352 (78.4%) $r_s = .762$, $p < .001$
BNC-3000	x	x	2,514 (83.8%) $r_s = .870$, $p < .001$	2,428 (80.9%) $r_s = .819$, $p < .001$
BE06-3000	x	x	x	2,518 (83.9%) $r_s = .826$, $p < .001$
EnTenTen12-3000	x	x	x	x

Table 5: LOB and BE06: comparison of the first 3,000 vocabulary items

Word class	*LOB-3000*	*BE06-3000*	Overlap	Correlation
Nouns	1,470	1,489	1,178	$r_s = .770$, $p < .001$
Verbs (+modals)	637 + 10	656 + 10	543 + 10	$r_s = .854$, $p < .001$
Adjectives	490	465	377	$r_s = .802$, $p < .001$
Adverbs	221	212	188	$r_s = .898$, $p < .001$
Conjunction and prepositions	73	77	72	$r_s = .971$, $p < .001$
Pronouns	28	26	26	$r_s = .988$, $p < .001$
Other (gram. words)	71	65	64	$r_s = .975$, $p < .001$
Total	3,000	3,000	2,458 (81.9%)	$r_s = .832$, $p < .001$

Table 6: BE06 and EnTenTen12: comparison of the first 3,000 vocabulary items

Word class	*BE06-3000*	*EnTenTen 12-3000*	Overlap	Correlation
Nouns	1,489	1,479	1,229	$r_s = .767$, $p < .001$
Verbs (+ modals)	656 + 10	656 + 10	551 + 10	$r_s = .847$, $p < .001$
Adjectives	465	478	384	$r_s = .808$, $p < .001$
Adverbs	212	211	184	$r_s = .886$, $p < .001$
Conjunction and prepositions	77	75	73	$r_s = .970$, $p < .001$
Pronouns	26	27	26	$r_s = .847$, $\mathrm{p} < .001$
Other (gram. words)	65	64	61	$r_s = .970$, $p < .001$
Total	3,000	3,000	2,518 (83.9%)	$r_s = .826$, $p < .001$

Table 7: BE06 and BNC: comparison of the first 3,000 vocabulary items

Word class	*BE06-3000*	*BNC-3000*	Overlap	Correlation
Nouns	1,489	1,369	1,208	$r_s = .841$, $p < .001$
Verbs (+modals)	656 + 10	635 + 10	570 + 10	$r_s = .918$, $p < .001$
Adjectives	465	516	386	$r_s = .844$, $p < .001$
Adverbs	212	271	199	$r_s = .842$, $p < .001$
Conjunction and prepositions	77	101	69	$r_s = .950$, $p < .001$
Pronouns	26	23	22	$r_s = .962$, $p < .001$
Other (gram. words)	65	75	50	$r_s = .916$, $p < .001$
Total	3,000	3,000	2,514 (83.8%)	$r_s = .870$, $p < .001$

verbs, adjectives, and adverbs) show also a large proportion of overlap with the correlations ranging from .770 to .898 ($p < .001$). The main differences between the wordlists lie in the inclusion/exclusion of a variety of nouns, verbs, and adjectives.

LOB-3000 includes a number of words from the domains of

- religion and aristocracy: *blessing, castle, catholic, chamber, chapel, cross, crown, devil, devotion, kingdom, knight, parish, priest, saint, spiritual, worship.*
- culture: *architecture, artistic, composer, concert, conductor, dancing, exhibition, opera, orchestra, organ, painter, poet, poetry, portrait, verse.*

On the other hand, in *BE06-3000* we can notice words from the following areas:

- current technology: *CD, computer, email, Internet, mobile, online, video, web, website.*
- current political and social issues: *environmental, gay, immigrant, immigration, obesity, organic, poverty, smoking, smoker, terrorism.*

Table 6 offers a detailed comparison between the wordlists based on two corpora of current English—*BE06* and *EnTenTen12*. The corpora differ in many respects. They were compiled using different strategies and, most importantly, they differ in their size. Despite this fact, there is a large overlap (83.9 per cent) in the first 3,000 items on the frequency lists. In fact, the overlap is two per cent larger than that between *BE06* and *LOB* (see Table 5). This can be explained by the fact that *BE06* and *EnTenTen12* (unlike *BE06* and *LOB*) both reflect the current language use.

Interestingly, the main differences in the vocabulary between *BE06* and *EnTenTen12* can be found in the areas of (i) Internet and computer technology and (ii) Business and advertising. It is not surprising that both of these areas are strongly (over)represented in the Internet-based corpus *EnTenTen12*, which reflects the dominant online registers. The following examples show the lexical items occurring solely in *EnTenTen12-3000*:

> Internet and the computer technology: *blog, browse, browser, database, download, file, hardware, host, info, install, installation, laptop, menu, net, PC, print, program, scan, server, software, tablet, tag, technical, upgrade, virtual*
>
> Business and advertising: *advertise, advertising, affordable, banking, CEO, competitor, corporation, dealer, developer, discount, enterprise, exclusive, executive, hire, insurance, investor, liability, logo, manufacture, manufacturer, manufacturing, market, marketing, marketplace, purchase, retail, retailer, sponsor, trade, transaction*

Finally, Table 7 compares the wordlists based on the one-million-word corpus *BE06*, which represents different genres of written language, and the 100-million-word *BNC*, which includes a substantial (10 per cent) spoken component. The overall overlap between the two wordlists (83.8 per cent) is almost as high as the one between the two corpora of current English (see Table 6). Also noticeable is the very large overlap in verbs with 570 lexical and 10 modal verbs in common and almost a perfect correlation ($r_s = .918$, $p < .001$).

As expected, the main lexical difference between the corpora can be observed in the occurrence of words reflecting technological development after the early 1990s in *BE06*, e.g. *CD*, *Internet*, *mobile*, *online*, *web*, and *website*. In addition, *BE06-3000* contains a number of informal words and Americanisms such as *guy*, *hit*, *mate*, *movie*, *mum*, and *sexy* which do not appear in the *BNC* wordlist.

3.2 Common lexical core among four wordlists (RQ2)

To identify the common lexical core, all four wordlists were compared. In the previous section, we have seen that when the individual wordlists were compared pairwise their overlap was between 78 and 84 per cent. The lexical core common to all four wordlists consists of 2,122 items, which represents almost 71 per cent overlap. This is a high number considering the diversity of corpora on which the wordlists are based. Table 8 shows the breakdown of the overlap according to individual word classes.

As can be seen from Table 8, it is not only the closed (grammatical) word classes that show a high degree of overlap, but, interestingly, a high overlap was also found among open (lexical) word classes such as nouns, adjectives, and verbs. This is to some extent a surprising result, given the variety of textual sources in the four corpora.

3.3 New lexical development (RQ3)

Three hundred and seventy-eight vocabulary items were identified as belonging to the lexical overlap specific to the current English corpora *BE06* and *EnTenTen12*; these were lexical items not occurring among the top 3,000

Table 8: All four wordlists compared: A common lexical core

Word class	*LOB-3000*	*BNC-3000*	*BE06-3000*	*EnTenTen 12-3000*	Overlap
Nouns	1,470	1,369	1,489	1,479	1,009
Verbs (+ modals)	637 + 10	635 + 10	656 + 10	656 + 10	488 + 10
Adjectives	490	516	465	478	317
Adverbs	221	271	212	211	166
Conjunction and prepositions	73	101	77	75	63
Pronouns	28	23	26	27	22
Other (gram. words)	71	75	65	64	47
Total	3,000	3,000	3,000	3,000	2,122

lemmas in the other two corpora representing older language use. These items can be categorized into three major groups representing:

- New words (forms): *Internet* (n, 701),[5] *website* (n, 860), *online* (adj, 987), *email* (n, 1696)
- New meanings/functions of old words: *user* (n, 775), *via* (con, 899), *network* (n, 1008), *client* (n, 1,047), *mobile* (adj, 1,348), *file* (n, 1,470), *web* (n, 1,622)
- Old words with recent prominence: *medium* (n, 609), *phone* (n, 612), *key* (adj, 660), *technology* (n, 664), *guy* (n, 696), *kid* (n, 736), *environment* (n, 751), *computer* (n, 861), *movie* (n, 1,336), *definitely* (adv, 1,357)

It is interesting to note that the major sources of lexical innovations are (i) semantic extensions of word meanings and (ii) the increase in the word frequency, which signals a shift towards a more general usability. For example, words such as *user*, *network*, and *file* are now prominently used when referring to the online environment, although they also retain their 'offline' meanings. In addition, words such as *computer*, *technology*, and *media* have relatively stable meanings, but these forms have gained prominence in general use. Based on the corpus evidence, we can safely assume that these items belong to the general lexis rather than to academic or specialized vocabulary, although for instance Coxhead (n.d.) includes these three items in her *AWL*. The last source of lexical innovations is newly created words such as *Internet*, *website*, *online*, and *email*. There are, however, only a handful of these among the 3,000 most frequent words.

3.4 The *new-GSL*

Based on the answers to RQ2 and RQ3, the *new-GSL* was compiled. It is composed of two parts: (i) the common lexical core and (ii) the items representing recent lexical development based on the two current wordlists (*BE06-3000* and *EnTenTen-3000*). Before compiling the final wordlist, all items were manually checked for consistency and erroneous entries were removed[6]. In total, the *new-GSL* consists of 2,494 items, 2,116 of which belong to the vocabulary shared by all four wordlists (base part) and 378 to the current vocabulary items. Table 9 provides detailed information about the structure of the *new-GSL* according to the word class of the items.

The largest proportion of the *new-GSL* (almost a half) is formed by nouns followed by verbs (22 per cent), adjectives (14.7 per cent), and adverbs (7.4 per cent). As expected, closed word classes (conjunctions, prepositions, pronouns, determiners, quantifies, etc.) represent only 7 per cent of the *new-GSL* vocabulary. The whole *new-GSL* is available from *Applied Linguistics* online.

3.5 The *new-GSL* compared with West's *GSL* and the *AWL* (RQ4)

To provide the first evaluation of the *new-GSL*, we compared this list with its predecessor (West's *GSL*) and a *GSL*-derived wordlist—Coxhead's (2000) *AWL*.

Table 9: The new-GSL: word classes

Word class	Items	Per cent
Nouns	1,204	48.16
Verbs	548	21.92
Modals	10	0.40
Adjectives	368	14.72
Adverbs	186	7.44
Adverbial particles in phrasal verbs	9	0.36
Prepositions or conjunctions	78	3.12
Pronouns	41	1.64
Determiners, quantifiers, or particles	49	1.96
abbreviations	5	0.20
To as infinitive marker	1	0.04
Existential *there*	1	0.04

Table 10: The comparison of the new-GSL with West's GSL and the AWL

Wordlist	Number of items		
	Types	Lemmas	Word families
new-GSL	5,115	2,494	
West's *GSL*[a]	7,826	4,114	2,000
AWL	3,099		570

[a] The numbers are based on the most widely used version of *GSL* in Nation's RANGE program (Heatley *et al.* 2002).

In essence, the *AWL* represents an academic extension of West's *GSL* and therefore these two wordlists have been used as the basic point of comparison for the *new-GSL*. Table 10 compares the *new-GSL* with West's *GSL* and the *AWL* according to their organizing principles. As discussed in Section 1.2, the main vocabulary unit of the *new-GSL* is a lemma, which includes a headword together with its variants based on inflectional morphology. West's *GSL* and the *AWL*, on the other hand, are organized around word families, which include not only inflectional, but also derivational morphology (cf. Bauer and Nation 1993). In effect, as can be clearly seen from Table 10, the decision to include only the most frequent lemmas (as opposed to whole word families) considerably reduces the number of types (different forms) in the *new-GSL*. While West's decision to include 2,000 word families resulted in >4,000 lemmas and ultimately almost 8,000 word types appearing in the wordlist, our decision to organize the *new-GSL* around lemmas (2,494 in total) led to a dramatic reduction in the number of word types (5,115). In addition to the 2,494 headwords, the *new-GSL* includes 2,619 morphological variants of the inflected words. Ultimately, by adopting a notion of the word that more appropriately reflects the use of words in text and their distribution in corpora, the *new-GSL* reduces the size of the wordlist by almost 40 per cent. Instead of the original ∼4,100 lemmas, the *new-GSL* includes only 2,494 most frequent and commonly distributed words.

Because West's *GSL* and the *AWL* on the one hand and the *new-GSL* on the other hand are organized according to different principles (word families and lemmas, respectively), it is fairly difficult to draw a direct comparison between these wordlists. Despite this fact, we can trace the occurrence of the *new-GSL* items in West's *GSL* and the *AWL* and thus identify the differences in the distribution of the core vocabulary in these wordlists. Table 11 shows the distribution of the *new-GSL* headwords in West's *GSL* first 1,000 and second 1,000 word families as well as in the *AWL*.

We can see that the majority of items from the first 1,000 words in the *new-GSL* are included in West's first 1,000 word families. Interestingly, however, 97 items from the first 1,000 words in the *new-GSL* occur only in the specialized *AWL* wordlist. These are very common words such as *couple*, *image*, *team*, *computer*, *area*, *individual*, *environment*, and *job*, which arguably belong to the general rather than the academic vocabulary and should therefore be excluded from the *AWL*. Surprisingly, the largest proportion of the second 1,000 words in the *new-GSL* overlaps also with West's first 1,000 word families. This is to some extent a reflection of the fact that West's first 1,000 word families include a number of less frequent words morphologically related to high frequency words. Thus, for instance, West's 1000 word families comprise items such as *appearance (1,263)*,[7] *existence (1,556)*, *industrial (1,582)*, *payment (1,355)*, *reader (1,069)*, *specialist (1,840)*, *unable (1,379)*, and *writing (1,618)* owing to the fact that these are lumped with their more frequent 'relatives' from the same word family, namely *appear (279)*, *exist (708)*, *industry (553)*, *pay (228)*, *read (316)*, *special (377)*, *able (252)*, and *write (213)*.

Table 11: The overlap between the new-GSL and West's GSL + the AWL

Wordlists	*new-GSL 0–1000*	*new-GSL 1001–2000*	*new-GSL 2001–2500*
West 0–1000[a]	816	461	131
West 1000–2000[a]	78	252	133
AWL	97	222	126
Off list	9	65	104
Total	1,000	1,000	494

[a] ∼2000 lemmas.

The last 500 items from the 2000–2500 band in the *new-GSL* are relatively evenly distributed among West's first 1,000, West's second 1,000, and the *AWL*. Again, this shows a major difference in the ranking principles between the *new-GSL* and the two other wordlists as exemplified above. Finally, 178 items from the new *GSL* do not occur either in West's *GSL* or the *AWL*. These are words such as *career*, *guy*, *huge*, *Internet*, *kid*, *online*, and *website*, representing a mixture of informal words and vocabulary referring to new technologies.

3.6 Coverage of text (RQ5)

Finally, Table 12 shows the results of a vocabulary coverage test of West's *GSL* and the *new-GSL*. The former consists of ~4,100 lemmas organized into 2,000 word families, the latter comprises only 2,494 lemmas. Table 12 reports on the proportions of tokens in four different corpora (*LOB*, *BNC*, *BE06*, and *EnTenTen12*) that are covered by the two wordlists.

The results suggest that there is not a large difference in the coverage between the two wordlists. Whereas West's original *GSL* covers a slightly larger proportion of older corpora (*LOB* and *BNC*), the coverage of the current language corpora (*BE06* and *EnTenTen12*) is either almost equal or slightly favours the *new-GSL*. The coverage proportions of West's *GSL* listed in Table 12 are comparable with those reported in the literature (for a discussion cf. Gilner 2011: 73–6).[8] However, the main difference in text coverage between West's *GSL* and the *new-GSL* lies in the fact that the *new-GSL* achieves a similar coverage with substantial reduction (~40 per cent) in the number of lemmas that appear in West's *GSL*.

4. CONCLUSION

This study brought strong evidence about the stability of the core English vocabulary across a variety of language corpora including different written and spoken contexts. We examined the overlap between 3,000 most frequent vocabulary items in four different corpora and identified a substantial correspondence between these corpora in terms of the number of shared items as well as the distribution of the words in the wordlists.

Table 12: Comparison of vocabulary coverage: West's GSL and the new-GSL

Wordlist	Corpora			
	LOB	*BNC*	*BE06*	*EnTenTen12*
West's *GSL*	84.1%	82%	80.6 %	80.1%
new-GSL	81.7%	80.3%	80.1%	80.4%

As mentioned in the introduction, some researchers (e.g. Bongers 1947; Richards 1974; Bogaards 2008) questioned the stability of the core vocabulary in different language corpora. Moreover, Engels (1968: 213) referring to Frumkina's research maintains that 'the minimum number of words for a corpus that might lead to a valid word-count is ten million'. Contrary to these claims, our study shows that if a suitable methodology is used that takes into account not only absolute (raw) frequencies of words, but also their dispersions, similar results can be obtained from a one-million-word corpus (e.g. *LOB* or *BE06)* and a 12-billion-word corpus (*EnTenTen12)*.

Further, the high correlations between the 3,000 most frequent words in the four corpora indicate stability of the relative frequencies of individual items across the four wordlists, that is, there is a strong tendency for individual words to occur with similar ranks in all four wordlists. This finding not only provides further evidence about the existence of the core vocabulary, but may also serve as a basis for research exploring the changes in the general vocabulary of the English language. Thus for example, against the background of the stable lexical core, this study identified 378 vocabulary items common to the current language corpora (*BE06* and *EnTenTen12*), which do not occur in the older corpora (*LOB* and *BNC*) with the same high frequencies.

The *new-GSL* consists of a total of 2,494 words. It can be divided into the base part (2,116 items) and the current vocabulary part (378 items). The *new-GSL* is compiled according to the lemma principle. While losing some advantages that word families bring (cf. Bauer and Nation 1993), opting for an alternative to the traditional word-family approach allowed us to narrow down the wordlist to significantly fewer forms than included in West's *GSL* and at the same time retain comparable coverage of text. This methodological choice also plays an important role in the ranking and the frequency-bands organization of words in the *new-GSL*, which reflects more precisely the actual occurrence of words in text. Pedagogically, this feature of the *new-GSL* is important for creating lexically appropriate teaching materials for different groups of learners (cf. Nation 2003) (e.g. simplified versions of texts).

The analyses that evaluated the practical usefulness of the *new-GSL* showed its effectiveness in covering about 80 per cent of the texts in the corpora with only 2,494 lemmas. This is a significant reduction compared with the ~4,100 lemmas that West's *GSL* needs to reach similar coverage. Moreover, as can be seen from the comparison of our general wordlist with the *AWL*, the *new-GSL* can also help narrow down the number of academic words more accurately than West's *GSL* by excluding general words such as *computer*, *technology*, and *media* from the academic vocabulary. Thus, the *new-GSL* can be used as an effective research tool in the development of specialized wordlists (cf. Nation and Hwang 1995).

The *new-GSL*, moreover, represents a wordlist based on a transparent compilation procedure. This is an important dimension of the new wordlist that makes it more readily usable for pedagogic as well as research purposes. Because the reasons behind the selection of the words in the wordlist are stated explicitly, the study can be replicated with other corpora following the same methodology. For practical purposes, the transparent methodology enables further extensions of the *new-GSL*, a feature not readily available for West's *GSL* (cf. Gilner and Morales 2008; Gilner 2011).

Several limitations to the wordlist presented in this article should, however, be acknowledged. The selection of the words from the four corpora was based on the judgements connected with automatic word class identification (see Section 2). Although generally reliable, the tagging procedure results in a small percentage of inaccuracies that can lead to slightly imprecise word counts. Every effort was, however, made to check the consistency of the word classes in the final wordlist. In addition, it is important to note that the *new-GSL* is based primarily on the British variety of English with a particular focus on written language. Although the selection process included also an International variety of English (*EnTenTen12*) as well as a spoken component in the *BNC*, further research is desirable to establish the stability of the core vocabulary in spoken and non-British contexts. With respect to regional varieties, preliminary findings of a study that replicated the research with American English corpora suggest that there is surprisingly small variation between the two varieties in the most frequent vocabulary (Gablasova and Brezina [in preparation]). However, a list of items that systematically appear across corpora of American English can be used as an extension of the present wordlist. Likewise, an extension with the most frequent items found across different spoken corpora could be used to complement the present wordlist. Finally, because the *new-GSL* was designed as a list of single words, further research is needed to account for multi-word lexical items such as the expressions on the PHRASE list (Martinez and Schmitt 2012).

In sum, the *new-GSL* was designed with transparency and flexibility in mind both as a research and a teaching tool. With respect to the diversity of ESL/EFL contexts, it is deemed more useful to envision the use of our wordlist as a vocabulary base with the possibility of further additions, rather than a wordlist that strives to cater to a mixed cluster of heterogeneous expectations and needs.

SUPPLEMENTARY DATA

The whole *new-GSL* is available at *Applied Linguistics* online.

ACKNOWLEDGEMENT

The authors would like to thank the three anonymous reviewers for their helpful comments on an earlier draft of the article.

FUNDING

ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference (ES/K002155/1).

Conflict of interest statement. None declared.

NOTES

1 There are different versions of West's *GSL*. Gilner (2011) reports that the original *GSL* (West 1953) includes 1,907 main entries for individual word families and 3,751 orthographically different words. Nation (e.g. 2004) extended the list for the use in the RANGE program. His version includes 1,986 word families and >4,000 lemmas.

2 The focus on a single variety of English enabled us to limit possible variation connected with regional varieties of English and thus create a baseline against which other varieties can be compared (cf. Gablasova and Brezina [in preparation]).

3 *EnTenTen12* is a corpus developed for the *Sketch Engine* and is available only through the *Sketch Engine* interface (Lexical computing 2012).

4 In morphologically annotated corpora, the word class is determined automatically by a tagging program. The *Sketch Engine* uses *Tree tagger* with the average accuracy of >96 per cent for English (Schmid 1994).

5 The letter in the parentheses indicates the word class of the item (n = noun, adj = adjective, con = preposition or conjunction, adv = adverb); the number indicates the rank of the item in the *new-GSL.*

6 The manual processing involved checking for inconsistencies both in the headwords and word-class ascriptions. As a result of this process a handful of entries were deleted and some word-class ascriptions were corrected.

7 The number in parentheses shows the rank in the *new-GSL*.

8 Previous studies show that the coverage of West's *GSL* largely depends on the genre ranging from 75 per cent for academic texts to 90 per cent for speech and fiction (Hirsh and Nation 1992; Nation 2004). The remaining portion of the texts in language corpora is usually attributed to academic and lower frequency lexical items.

REFERENCES

Aston, G. and **L. Burnard.** 1998. *The BNC Handbook: Exploring the British National Corpus with Sara*. Edinburgh University Press.

Baker, P. 2009. 'The BE06 corpus of British English and recent language change,' *International Journal of Corpus Linguistics* 14/3: 312–37.

Bauer, L. and **P. Nation.** 1993. 'Word families,' *International Journal of Lexicography* 6/4: 253–79.

Beglar, D. and **A. Hunt.** 2005. 'Six principles for teaching foreign language vocabulary: A commentary on Laufer, Meara, and Nation's 'ten best ideas',' *The Language Teacher* 29/7: 7–10.

Bogaards, P. 2008. 'Frequency in learners' dictionaries' in E. Bernal and J. DeCesaris (eds). Proceedings of the XIII EURALEX International Congress, Barcelona, 15–19 July, pp. 1231–6.

Bongers, H. 1947. *The History and Principles of Vocabulary Control*. Wocopi.

Carter, R. 2012. *Vocabulary: Applied Linguistic Perspectives*. Routledge.

Čermák, F. and **M. Křen.** 2005. 'Large corpora, lexical frequencies and coverage of texts,'. Proceedings from the Corpus Linguistics Conference, Birmingham, 14–17 July.

Cobb, T. 2012. 'Compleat lexical tutor,' available at http://www.lextutor.ca/. Accessed 15 July 2012.

Coxhead, A. 2000. 'A new academic word list,' *TESOL Quarterly* 34/2: 213–38.

Coxhead, A. 2011. 'The Academic Word List 10 years on: Research and teaching implications',' *TESOL Quarterly* 45/2: 355–62.

Coxhead, A. n.d. 'The academic wordlist,' available at http://www.victoria.ac.nz/lals/resources/academicwordlist. Accessed 12 May 2012.

Engels, L. K. 1968. 'The fallacy of word-counts,' *International Review of Applied Linguistics in Language Teaching* 6/3: 213–31.

Francis, W. N. and **H. Kučera.** 1982. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin.

Gablasova, D. and **V. Brezina.** in preparation. 'American English supplement to the *New General Service List*'.

Gardner, D. 2007. 'Validating the construct of word in applied corpus-based vocabulary research: A critical survey,' *Applied Linguistics* 28/2: 241–65.

Gilner, L. 2011. 'A primer on the general service list,' *Reading in a Foreign Language* 23/1: 1.

Gilner, L. and **F. Morales.** 2008. 'Corpus-based frequency profiling: Migration to a word list based on the British National Corpus,' *The Buckingham Journal of Language and Linguistics* 1: 41–58.

Heatley, A., P. Nation, and **A. Coxhead.** 2002. '*RANGE and FREQUENCY programs*. Victoria University of Wellington,' available at http://www.victoria.ac.nz/lals/about/staff/paul-nation. Accessed 30 July 2012.

Hirsh, D. and **P. Nation.** 1992. 'What vocabulary size is needed to read unsimplified texts for pleasure?' *Reading in a Foreign Language* 8: 689–96

Hlaváčová, J. 2006. 'New approach to frequency dictionaries—Czech example'. Paper presented at the 5th International Conference on Language Resources and Evaluation, Genoa, 24–26 May, available at http://www.lrec-conf.org/proceedings/lrec2006/pdf/11_pdf.pdf.

Hyland, K. and **P. Tse.** 2007. 'Is there an 'academic vocabulary'?' *TESOL Quarterly* 41/2: 235–53

Johansson, S., E. Atwell, R. Garside, and **G. Leech.** 1986. 'The tagged LOB corpus: User's manual. Norwegian Computing Centre for the Humanities Bergen, available at http://khnt.hit.uib.no/icame/manuals/lobman/index.htm. Accessed 7 October 2012.

Kilgarriff, A. 2009. 'Simple maths for keywords,'. 'Paper presented at the Corpus Linguistics Conference at Lancaster', available at http://ucrel.lancs.ac.uk/publications/CL2009/171_FullPaper.doc.

Leech, G., P. Rayson, and **A. Wilson.** 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Longman.

Lexical computing. 2012. ''The sketch engine,' available at http://www.sketchengine.co.uk. Accessed 30 July 2012.

Martinez, R. and **N. Schmitt.** 2012. 'A phrasal expressions list,' *Applied Linguistics* 33/3: 299–320.

McCarthy, M. 1990. *Vocabulary*. Oxford University Press.

McEnery, T. and **A. Hardie.** 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Nagy, W., R. C. Anderson, M. Schommer, J. A. Scott, and **A. C. Stallman.** 1989. 'Morphological families in the internal lexicon,' *Reading Research Quarterly* 24/3: 262–82.

Nation, P. 1990. *Teaching and Learning Vocabulary*. Heinle & Heinle.

Nation, P. 2001. *Learning Vocabulary in Another Language*. Cambridge University Press.

Nation, P. 2003. 'Materials for teaching vocabulary' in B. Tomlinson (ed.): *Developing Materials for Language Teaching*. Continuum, pp. 394–405.

Nation, P. 2004. 'A study of the most frequent word families in the British National Corpus' in P. Bogaards and B. Laufer (eds): *Vocabulary in a Second Language*. John Benjamins, pp. 3–14.

Nation, P. and **K. Hwang.** 1995. 'Where would general service vocabulary stop and special purposes vocabulary begin?' *System* 23/1: 35–41

Nation, P. and **R. Waring.** 1997. 'Vocabulary size, text coverage and word lists' in N. Schmitt and M. McCarthy (eds): *Vocabulary: Description, Acquisition and Pedagogy*. Cambridge University Press, pp. 6–19.

Oakes, M. P. 1998. *Statistics for Corpus Linguistics*. Edinburgh University Press.

Richards, J. C. 1974. 'Word lists: Problems and prospects,' *RELC Journal* 5/2: 69–84.

Savický, P. and **J. Hlaváčová.** 2002. 'Measures of word commonness,' *Journal of Quantitative Linguistics* 9/3: 215–31.

Schmid, H. 1994. 'Probabilistic part-of-speech tagging using decision trees,' available at http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf. Accessed 30 July 2012.

Schmitt, N. and **C. B. Zimmerman.** 2002. 'Derivative word forms: What do learners know?' *TESOL Quarterly* 36/2: 145–71

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse*. Routledge.

Sketch Engine trac: enTenTen. 2012. Available at http://trac.sketchengine.co.uk/wiki/Corpora/enTenTen. Accessed 30 July 2012.

Ward, J. and **J. Chuenjundaeng.** 2009. 'Suffix knowledge: Acquisition and applications,' *System* 37/3: 461–9.

Webb, S. and **M. P. H. Rodgers.** 2009. 'The lexical coverage of movies,' *Applied Linguistics* 30/3: 407–27.

West, M. 1953. *A General Service List of English Words: with Semantic Frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology*. Longman.