### *In the CUP Handbook of English Corpus Linguistics*

**Vocabulary**

Ron Martinez

San Francisco State University

Norbert Schmitt

University of Nottingham

*In both L1 and L2 pedagogy, the question of which vocabulary should be taught and in which order of priority has long been asked (Nation, 2001), with researchers increasingly turning to corpora for the answer. As reviewed in Leech (2011), one reason is the longstanding perception that the most common words should be taught first, and indeed there is a good body of research to support that notion (e.g. Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Read, 2004; Stæhr, 2009). As this frequency-usefulness lexical relationship has both notionally and empirically endured the test of time, so too have researchers’ attempts to identify and list vocabulary using frequency information informed by corpus evidence. This chapter will take a critical look at the development of some key corpus-informed vocabulary lists in detail, examining the lessons that can be taken away from each attempt. We will then present a detailed account of the development of our own corpus-derived list – the PHRASE List (Martinez & Schmitt, 2012) – reflecting on and building from the lessons learned from earlier vocabulary lists.*

**1. The enduring relevance of early corpus-informed vocabulary list research**

Although the use of computerized corpora to enhance the analysis of lexis seems a rather recent phenomenon, the perceived need to use a corpus to help put some order into the apparent lexical vastness of English for pedagogical purposes has actually existed for some time. One of the earliest and most famous efforts was initiated by Edward Thorndike of Teachers College, Columbia University, starting as early as the 1910s. Realizing that “even expert teachers have very inadequate and inaccurate notions of the relative frequency and importance of words” (Thorndike, 1921: 360), Thorndike undertook to compile a list of the words that schoolchildren should know (or be taught), informed by empirically-derived frequency data. Therefore, in much the same fashion as many corpus linguists might proceed today, Thorndike sampled from texts that American children were generally required or likely to read at that time, including a variety of literary works and standard textbooks, arriving at a corpus of approximately 5 million words. Taking “about ten years” (Thorndike, 1921: 346), Thorndike (with the help of assistants) eventually combed through the entire corpus by hand, tallying and tabulating each occurrence, and eventually published the influential *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading for Children and Young People* (Thorndike, 1932).

Obviously, with computerized corpora, what took Thorndike some ten years to complete might today only take ten days, or ten hours, or even ten seconds. Nonetheless, on a qualitative level, the corpus research and concomitant issues taken on by Thorndike and colleagues over a century ago remain largely unchanged. Much like today, the researchers had to make decisions about what corpus size was needed (or adequate), how and why they would choose the texts that comprised the corpus, and, perhaps of greatest relevance to the present chapter, what would count as a *word*: “Are we to count ‘Apply for a job’ and ‘Apply a cup to the lips’ as one word or two? … We shall certainly teach ‘large’ (‘a large box’). Shall we teach ‘at large’?” (West, 1937: 436). Such issues are still discussed today in corpus-related vocabulary research:

I would argue that suggested applications of corpus research based on frequency of word forms, without considerations of word meanings, will invariably suffer from one of three problems – or combinations of the three: (a) they will overestimate the true coverage of the word forms; (b) they will underestimate the actual user knowledge required to negotiate the word forms; and/or (c) they will underestimate the actual number of meanings inherent in the word forms. (Gardner, 2007: 253)

These potential problems with the Thorndike (1932) list were indeed eventually recognized by Thorndike himself, who in 1934 participated in a special Carnegie Corporation-funded conference on the selection of vocabulary, which in turn culminated in an influential special report entitled the *Interim Report on Vocabulary Selection* (Palmer, West & Faucett, 1936). Mentioned in the report are the problems that Gardner (2007) raises, that is, a list that only takes frequency data into account without attention to semantics may be of limited pedagogical value. Thorndike therefore teamed up with colleague Irving Lorge and published the seminal *A Semantic Count of English Word*s (Lorge & Thorndike, 1938). The researchers went back to the original Thorndike (1932) data and cross tabulated each word with individual senses in *The Oxford English Dictionary*, so that the word *game*, for example, was listed with a raw frequency of 638 (in the corpus of 5 million words), but was “semanticized” (West, 1953: xii) into 29 separate units, each unit with its own token frequency.

While such semantic granularity certainly represented a great improvement over the raw frequency-driven original lists, it also meant that a list that was already considered somewhat unwieldy for the average teacher due to its pure size (30,000 words), was now perhaps even more off-putting, thus defeating the purpose for creating a list in the first place.

However, the *Interim Report on Vocabulary Selection*, in an attempt to make the list more user-friendly, also contained a special section that listed the 2,000 words judged as most essential by the Carnegie committee (mostly due to corpus frequency, but also more subjective judgments related to usefulness, difficulty, and style) , called the *General Service List* (GSL). In 1939, arrangements were made to have Michael West, a teacher of English working especially in India, revise the GSL using the semantic information now published by Lorge and Thorndike. It is this version of the list, published in 1953 under the title *A General Service List of English Words* (West, 1953), that has proved to be the one of the most enduring and influential corpus-informed lists of English vocabulary, widely employed and referenced in language teaching research still to this day (e.g. Lubliner & Hiebert, 2011; Nation & Beglar, 2007; Matsuoka & Hirsh, 2010; Parent, 2012; Smith, 2011). A possible reason for this lasting popularity is the way in which West made the GSL more pedagogy-friendly. Using his experience as a language teaching professional, West not only applied the Lorge and Thorndike semantic frequency data to the original 2,000-word GSL, but further enhanced it by dividing the individual senses “more coarsely” (West, 1953: vii).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **GAME** | 638e | (1) | (*amusement, children’s play*)  Fun and games  It’s not serious; it’s just a game | **9%** |
|  |  | (2) | (*with the idea of competition, e.g., cards, football, etc*.)  A game of football  Indoor games; out-door games | **38%** |
|  |  | (3) | (*a particular contest*)  We won, six games to three  I played a poor game  Play a losing game (10-5%) | **23%** |
|  |  | (4) | (*games = athletic contest*)  Olympic Games | **8%** |
|  |  | ? | [= *animals*, 11%; game-/, game-birds, etc., 5%]  [=*fun*, Make game of, 0-5%] |  |

**Figure 1** *Entry for ‘game’ in the West (1953) General Service List*

As can be seen in Figure 1, the 29 semantic divisions in the Lorge and Thorndike (1938) list are broken down into an arguably much more digestible 4 senses. Moreover, the list is made even more accessible by the addition of percentages, facilitating the prioritization of which sense to teach/target first.

Hence, what perhaps can be learned from early corpus-informed English vocabulary studies is that while corpus size matters, as do frequency counts, pure quantitative data alone are perhaps not easily adopted by those who would apply such information (e.g. in teaching and syllabus design). If we are to take the West GSL as a model, it is important to consider the data qualitatively as well. This may be especially important when considering frequency data, as was eventually recognized by Lorge and colleagues: “In general, the more frequent the occurrence of a word, the greater is the variety of meanings in which it is used” (Lorge, in West, 1953: xi). To ignore or not provide adequate semantic information about the most frequent words in English is to be essentially disingenuous about the true weight of the learning burden of that vocabulary.

Furthermore, the West GSL also teaches us that it is important to present such data in ways that are user-friendly and relatively easily interpretable by non-specialists. While the publication of the most frequent 30,000 schoolbook words – each with meticulous semantic counts – was surely a momentous achievement, what ultimately emerged as most practical and useful was the far less detailed qualitative breakdown of those, in which a practicing EFL teacher was instrumental.

*1.1. Lessons learned (and seemingly forgotten) in the advent of computerized corpora*

Interestingly, at least for a while, some of the important innovations that emerged over the course of the development of the West (1953) GSL fell by the wayside amid the excitement of the potential that computers brought to corpus linguistic study, in some ways regressing to the Thorndike work of the early 1900s. Henry Kučera and Nelson Francis of Brown University in 1961 developed the *Brown Corpus*, at just 1 million words substantially smaller than the corpora used by Thorndike and colleagues, but generally considered the first ‘modern’ corpus of English since the text was digitized and inputted into a computer. Kučera and Francis published wordlists based on computerized analyses of the corpus (Kučera & Francis, 1967), and included valuable dispersion data in the lists. They also made use of part-of-speech (POS) tagging, but it was very crude and required manual proofreading. Most notably, unlike the West (1953) GSL, semantics were not taken into account.

Another list that perhaps could have benefitted from the lessons learnt in the development of the West (1953) GSL was Avril Coxhead’s *Academic Word List* (Coxhead, 2000). Not unlike Thorndike and colleagues’ motivations for their lists, Coxhead wanted to identify words that students needed to know in order to function in school; in the case of the *Academic Word List* (AWL), the interest was more centered on success in higher education. Prior to the AWL, there were other lists that had been published and which did not use computerized corpora, such as the *University Word List* (Xue & Nation, 1984), which in turn was informed by the Campion and Elley (1971) *Academic Vocabulary List*[[1]](#footnote-1) and the *American University Word List*[[2]](#footnote-2) (Praninskas, 1972). However, Coxhead’s corpus was collected in electronic form, and was much larger than most predecessors at 3.5 million words, made up of four balanced sub-corpora (arts, commerce, law, and science). Words were included in the corpus if they were shown to not be on the GSL (West, 1953), and occurred at least 10 times in each of the sub-corpora (and in at least 15 of the 28 subject areas).

Much like Thorndike and those that followed him, the unit of counting was not word forms but ‘base’ forms that included any inflected and/or derived forms, or what Bauer and Nation (1993) called a ‘word family’. Coxhead’s analysis ultimately rendered a total of 570 word families that met the criteria for inclusion in the AWL. However, as mentioned earlier in this chapter, arguments can be made against the assumption that if one understands a lemmatized word it is reasonable to assume that changes in morphology should not present too much difficulty. This is true for the general English corpora from which the Thorndike, Lorge and West vocabulary lists emerged, and is probably even truer for vocabulary for specific genres of English, as is the case of the AWL. Consider, for example, the case of the word *drama* in the AWL: it is clear that it features in the list because of the words *dramatic* and *dramatically* when discussing trends and statistics in academic texts (e.g. ‘dramatic change’, ‘increased dramatically’); less clear is the extent to which the word *drama* alone is semantically related to those forms.

Furthermore, Coxhead’s wording of her first research question is also revealing (Coxhead, 2000: 218):

*Which lexical items occur frequently and uniformly across a wide range of academic material but are not among the first 2,000 words of English as given in the GSL (West, 1953)?*

Implicit in that question is the assumption that knowledge of the GSL obviates the need to repeat those words in any specialized list for English for Academic Purposes. However, consider the following extract from a business academic journal (from Bhuian, Menguc & Bell, 2005: 9):

**Abstract**

Within the **literature** of marketing and management, researchers have explored different **models** that examine the **relationships** between market orientation, entrepreneurship, and performance. In this **paper**, we offer a new model that includes curvilinearity in the moderating effect of entrepreneurship on the relationship between market orientation and performance.

The four words in boldface are those which are in the GSL but were not included in the AWL. They are also words that are employed in this academic genre in a more specialized way that deviates from their more canonical and dominant senses in the GSL. It has also been noted that the GSL items excluded from the AWL also ignore the importance of academic collocations, such as the omission of *address* (‘address an issue’) and *control* (‘control group’) (Durrant, 2009).

We therefore again see the dangers of focusing on quantitative measures alone. Without a doubt, the Coxhead (2000) AWL has proved an invaluable tool, and because of the principled way in which Coxhead compiled her corpora, to this day represents one of the better – if not the best – lists of academic vocabulary in English. However, it is unfortunate that the value of the more qualitative analyses that made the GSL such an important tool for so many years, especially with respect to consideration of semantic information, was not retained when the AWL was compiled. It would seem that the same question asked by West and colleagues in early part of last century – *we shall include ‘large’, but shall we include ‘at large’?* – is one that should always be asked when analysing corpora (computerized or otherwise) for vocabulary.

*1.2 Using new technology to address old questions*

Leech, Rayson, and Wilson’s (2001) *Word Frequencies in Written and Spoken English* (henceforth, *WFWSE*) is noteworthy since it was informed by the large 100 million word British National Corpus (BNC), but that is not why it is included in the present chapter. Although it did not include semantic information, WFWSE is the first computer-corpora-derived list to include items beyond single orthographic words, counting *a great deal*, for example, as a separate lexical item from *great* and *deal*[[3]](#footnote-3) (Figure 2).

|  |  |  |  |
| --- | --- | --- | --- |
| **Word** | **POS** | **Derivations** | **Frequency (p/million)** |
| A / a | Lett | : | 268 |
| A | NoP |  | 38 |
|  |  | A | 10 |
|  |  | A. | 28 |
| a bit | Adv | : | 119 |
| a great deal | Adv | : | 14 |
| a little | Adv | : | 104 |
| a lot | Adv | : | 40 |
| abandon | Verb |  | 44 |
|  |  | abandon | 12 |
|  |  | abandoned | 26 |
|  |  | abandoning | 5 |
|  |  | abandons | 1 |
| abbey | NoC |  | 20 |
|  |  | abbey | 19 |
|  |  | abbeys | 1 |
| Aberdeen | NoP |  | 14 |
|  |  | Aberdeen | 14 |
| ability | NoC |  | 105 |
|  |  | abilities | 13 |
|  |  | ability | 91 |
| able | Adj | : | 304 |

**Figure 2** Sample from the alphabetical list portion of WFWSE (Leech et al., 2001)

Leech et al. called such items ‘multiword units’, identified through strict formal criteria:

Multiword units are items which are treated as a single word token, even though they are spelt as a sequence of orthographic words. Because they function grammatically as single words (e.g. the conjunction *so that*, the preposition *in spite of*, *at least* as an adverb), they are treated as entries in their own right. (p. 8)

Interestingly, it was the computer itself (or the corpus tagging system) that compelled the researchers to include these multiword units. The *Constituent-Likelihood Automatic Word-Tagging System* (CLAWS) (Garside, 1987), used to analyze the corpus, was devised to automatically identify the word class of orthographic words (i.e. noun, adjective, adverb, and so on) in a corpus, but it was found that certain words, such as those in the expression *to and fro*, deviated from their normal tags elsewhere in the corpus (Blackwell, 1987) when the tagging system was still in its development and training phase. Therefore, a mechanism called the ‘ditto-tag’[[4]](#footnote-4) (Blackwell, 1987: 111) was created, allowing CLAWS to search a corpus for “specific sequences of words … whose syntactic role in combination differs from the syntactic role played by the same words in other contexts” (Blackwell, ibid.):

For example, *so that* is made up of two word strings but functions in the same way as a one-word conjunction: it simply does not make sense to analyze it (say) as an adverb preceding a conjunction. (Leech et al., 2001: 14)

Clearly, the empirically-validated inclusion of lexical items beyond single orthographic words in a frequency-driven vocabulary list represented an important step forward. However, this automated method of identification had a number of limitations, including being limited solely to phrases with immutable forms (e.g. *in order to*, *in accordance with*, *with respect to*) for which CLAWS could reliably assign a grammatical function (e.g. ‘preposition’). This means that important lexical items that do vary (e.g. by inflection, separation), such as phrasal verbs (e.g. *set up* 🡪 *setting up*, *set* something *up*), could not receive ditto tags automatically and are therefore not represented at all in WFWSE. There also was a complete reliance on statistical data, and unlike the lists that were produced before the era of widespread use of computerized corpora, there was no attention in WFSWSE to the individual senses of the words or phrases listed. Further, even when items would qualify as multiword expressions under the automatic ditto-tag CLAWS system, the mechanism does not have the same degree of accuracy as its other, single-word tagging counterpart[[5]](#footnote-5). On the other hand, the lists in WFWSE do provide some valuable information, such as frequencies for both written and spoken corpora within the BNC, and the differences between them for each word.

In the next section we will describe the development of our own frequency list, one that tries to incorporate some of the important strides made over the years in corpus-informed vocabulary studies.

## 2. The *PHRASE List*

So far, we have reviewed selected previous corpus-informed vocabulary list projects that we judge noteworthy for the valuable lessons that each has added over the years, including 1) the importance of not taking raw frequency counts at face value (i.e. the importance of also considering semantics), and 2) how the answer to the early question of whether multiword lexical units should be counted in vocabulary lists in addition to the conventional, single orthographic words, seems to be ‘yes’. Finally, it has been our assessment that West’s (1953) General Service List represents an excellent early model because it not only incorporated both those attributes, but it also was a list not intended to be comprehensive in nature; the list took 30,000 words with thousands more semantic frequencies and broke those down into what was subjectively deemed the ‘nitty gritty’ for the purposes of English language teaching, presenting it all in a format that was accessible by the non-specialist user. It is therefore in a similar vein that we strived to design our own list of selected multiword expressions; expressions carefully picked, not unlike the items in the GSL, from a list of thousands of other candidates for their pedagogical relevance. The final product, the *PHRASal Expressions List* (or *PHRASE* *List*), and greater explication of its rationale are presented in full in Martinez and Schmitt (2012). This section provides greater detail on how the actual list was compiled, reflecting on the issues raised earlier in this chapter.

### *2.1 Revisiting issues of frequency and semantics*

Our main objective was to create a list which would have pedagogic utility, mirroring purposes similar to the GSL and AWL lists, but for multiword expressions, to be ultimately juxtaposed and even integrated with such lists. Indeed, the proposal itself is not really a new one:

Some  items  larger  than  a  word  behave  like  high  frequency  words.  That  is,  they  occur frequently  as  multiword  units  (*good  morning*,  *never  mind*),  and  their  meaning  is  often  not  clear  from  the  meaning  of  the  parts  (*at  once*,  *set out*).  If  the  frequency  of  such  items  is  high  enough  to  get  them  into  a  general  service  list  in  direct  competition  with  single  words,  then  perhaps  they  should  be  included (Nation & Waring, 1997: 18).

Nation and Waring seem to suggest that the two most important criteria for inclusion of multiword expressions in a GSL-like list is (high) frequency, and semantics – semantic opacity, in particular (“meaning … not clear from the meaning of the parts”). We concur with both these suggestions.

However, while 2,000 entries seemed “a large vocabulary” (West, 1937: 433) when the GSL was written, today we know that 5,000 words is a better estimate of a pedagogically-relevant (e.g. ‘adequate reading comprehension’) functional threshold (Laufer & Ravenhorst-Kalovski, 2010; cf. Capel, 2010: 5; Davies & Gardner, 2010; Hindmarsh, 1980; Milton, 2009: 180). We therefore decided that multiword expressions that came within the same frequency range as the first 5,000 words in a frequency list might be a sensible target.

Regardless, as argued throughout the present chapter, frequency should not be considered to the exclusion of semantics. As already discussed, while the inclusion of multiword expressions did occur in the Leech et al. (2001) WFWSE, the items were limited to those that could be identified through the CLAWS tagging system. Using the examples in the Nation and Waring (1997) quote above, while *at once* was included in WFWSE, *good morning*, *never mind*, and *set out* were not. Nation and Waring suggest that multiword expressions lacking semantic transparency might be particularly worth including in a vocabulary list like the GSL; arguably, *good morning* is the only one of those examples that does not clearly meet that criterion. So why not devise software to cast the phraseological net wider than the one used for WFWSE? Unfortunately, as Leech and colleagues are no doubt fully aware, while surely a more inclusive list along the lines Nation and Waring suggest is desirable, a more comprehensive identification of such items basically requires reverting to Thorndikian methods of manual counting: asking a computer to produce a list of multiword expressions is still somewhat akin to writing ‘buy fruit’ on one’s shopping list – and sending someone else to do the shopping.

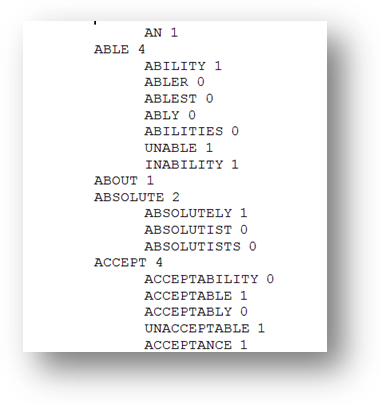
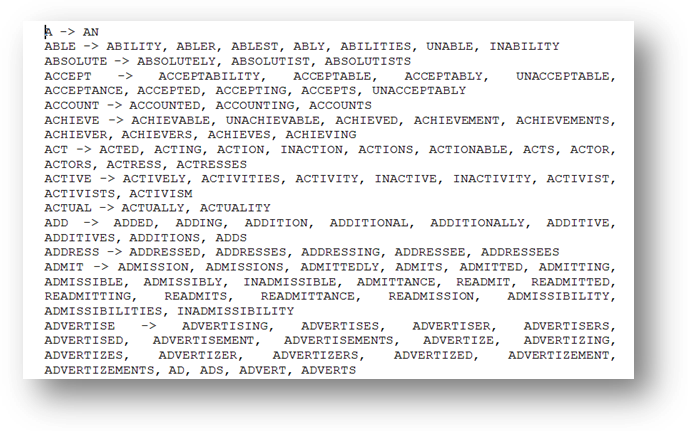
### *2.2 ‘Shopping for fruit’*

The next step was to decide on the corpus source of the language data for the list. After careful consideration, the full 100 million word BNC was deemed the best choice from among the publicly available large corpora for a number of reasons, including its size, diversity and reputation. As seen since the days of Thorndike’s early works, size does matter when it comes to corpus studies, especially when conducting qualitative research to explore patterns – as is the case in the present study. The BNC offers one of the largest publicly available corpora, which can be purchased on DVD-ROM. Not only is the BNC sufficiently large in size for conducting quantitative analyses, it is also derived from fairly diverse sources[[6]](#footnote-6), which the GSL and AWL also taught is important. Moreover, the BNC is the corpus that has most recently been used in the design of vocabulary research, lists and tests (e.g. Leech et al., 2001; Nation, 2006; Nation & Beglar, 2007; and the BNC-20 *Vocabulary Profiler* available on the *Lextutor* website) – instruments into which the PHRASE List may be usefully incorporated.

Of course, a corpus of the size of the BNC cannot be easily analyzed without the use of some kind of specialized software to be able to observe patterns using all the data contained in it. *WordSmith Tools* (version 5.0) was chosen, among other reasons, due to its compatibility with the latest BNC XML edition (used in our study) and ability to generate both word frequency lists and lists on recurring strings of words (or ‘n-grams’).

Since the text files are unreadable in the XML format they are encoded in for the DVD, *WordSmith Tools* was used to convert all the files into a plain unformatted text. The ‘Wordlist’ application embedded in *WordSmith Tools* was then used to upload all the texts (over 4000) and construct what is called an ‘index’ of all the words in the corpus. An index analysis collects vital information about each word in the corpus (e.g. dispersion, collocation, and so on), and is necessary if one wishes to run an analysis of recurrent word strings.

Once the index was complete (a file of over 1.5 gigabytes), the list was further analyzed and restructured in a process of lemmatization and grouping into word families. There were two reasons for this. First, as reviewed in the chapter, lists like the GSL and AWL do not present individual word forms (e.g. ‘go’, ‘going’, ‘went’), but instead their lemmatized, base forms. Further, popular tests of vocabulary knowledge informed by such lists (e.g. the *Vocabulary Levels Test* (Nation, 1990)) assume that a learner who can recognize *interest*, for example, will also readily discern the meaning of derived forms such as *interested*, *uninterested*, *interesting* and *interestingly* (i.e. the word family of ‘interest’). We did not wish to deviate from such a construct since we aimed to have our list be ultimately incorporated into such tests. However, *WordSmith* (like all corpus analysis software) simply recognizes word forms when compiling frequency lists, so this process must be done manually. Therefore, a full list of all the word families (14,000 basewords, or around 62,000 word forms) used by Nation and others for the Vocabulary Levels Test (VLT) was downloaded from the Victoria University of Wellington website (http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx), then carefully reformatted to be readable by *WordSmith Tools*, Figure 3.



**Figure 3**Two versions of word family formatting

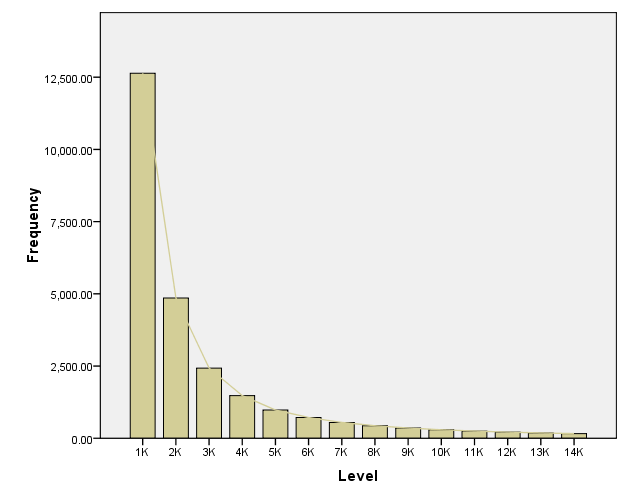
This sorting into word families allowed us to arrive at the crucial frequency band cut-off points with the 5,000 word frequency level, mentioned earlier, already established as a target rage (Table 1).

*Table 1* 1000-level frequency cut-offs (BNC)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Frequency band** | Token frequency cut-off\* |  | **Frequency band** | Token frequency cut-off |
| **1,000** | 12,271 + |  | **8,000** | 434 + |
| **2,000** | 4,455 + |  | **9,000** | 356 + |
| **3,000** | 2,089 + |  | **10,000** | 295 + |
| **4,000** | 1,217 + |  | **11,000** | 249 + |
| **5,000** | 787 + |  | **12,000** | 213 + |
| **6,000** | 620 + |  | **13,000** | 184 + |
| **7,000** | 547 + |  | **14,000** | 162 + |

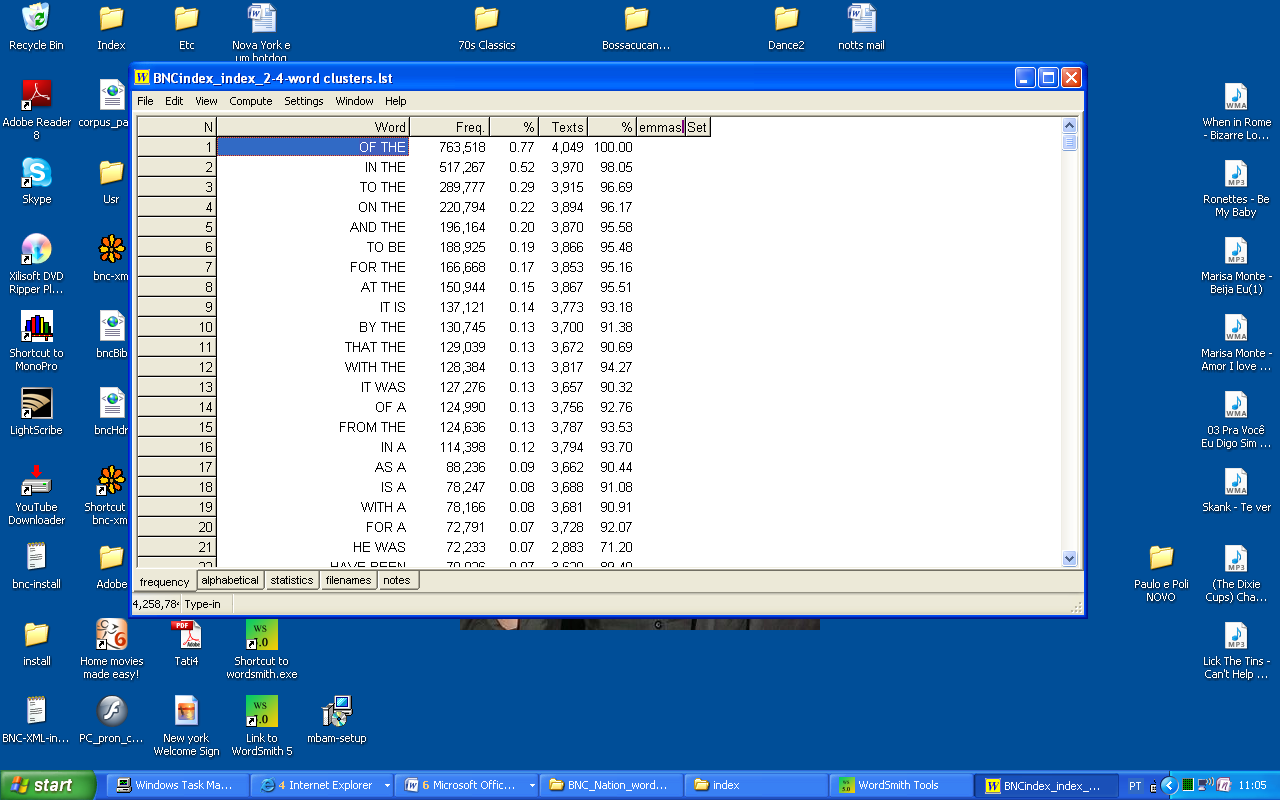
\* Per 100 million, including all tokens within a word family

The frequency band cut-off data in Table 1 might also be visualized graphically as in Figure 4, further evidence of Zipf’s Law (Zipf, 1949).



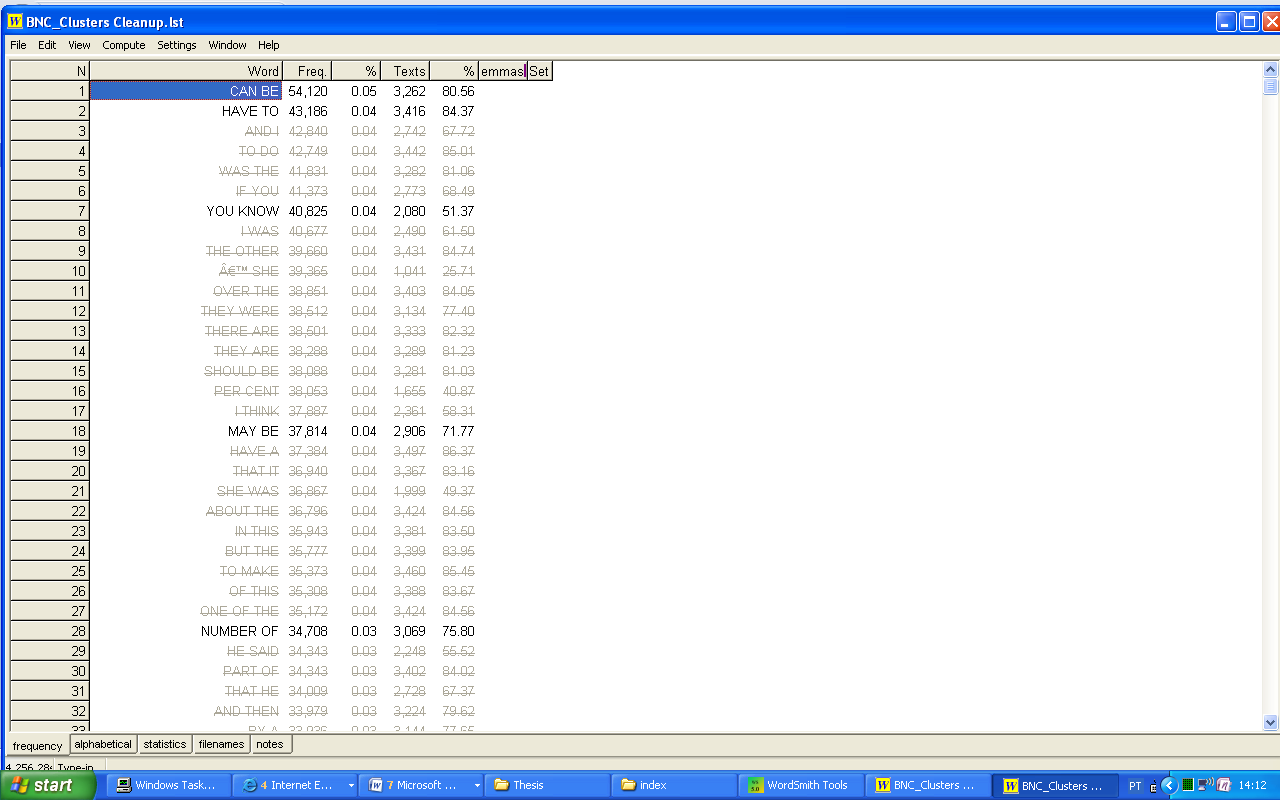
**Figure 4**Diminishing returns in frequency bands

With the frequency cut-off bands now established, we began the actual extraction process by using *WordSmith Tools* to interrogate the original indexed list for any and all n-grams between two and four words long repeated in the corpus at least five times. This search rendered a list of over 4.2 million n-grams. It is interesting to contrast this figure with the single-word index list, which reached a limit of just over 750,000 total word forms (a sample is provided in Figure 5). As our frequency cut-off exercise indicated that we only needed to search for items recurring at least 787 times in the corpus, our candidate pool was limited to 14,500 n-grams – which, as can be seen in Figure 5, would have to be vetted manually.



**Figure 5**A sample of unedited 2-4 grams list derived from BNC

The time-consuming qualitative stage of analysis then began, involving a line-by-line data deletion phase (Figure 6). The first author meticulously went down the n-gram list item-by-item looking for “plausibly formulaic” multiword items (Wray, 2009: 41), guided by carefully pre-established selection (and exclusion) criteria, explained in full in Martinez and Schmitt (2012), but in essence, as cited earlier (Nation & Waring, 1997: 18), choosing items whose “meaning (is) not clear from the meaning of the parts.”

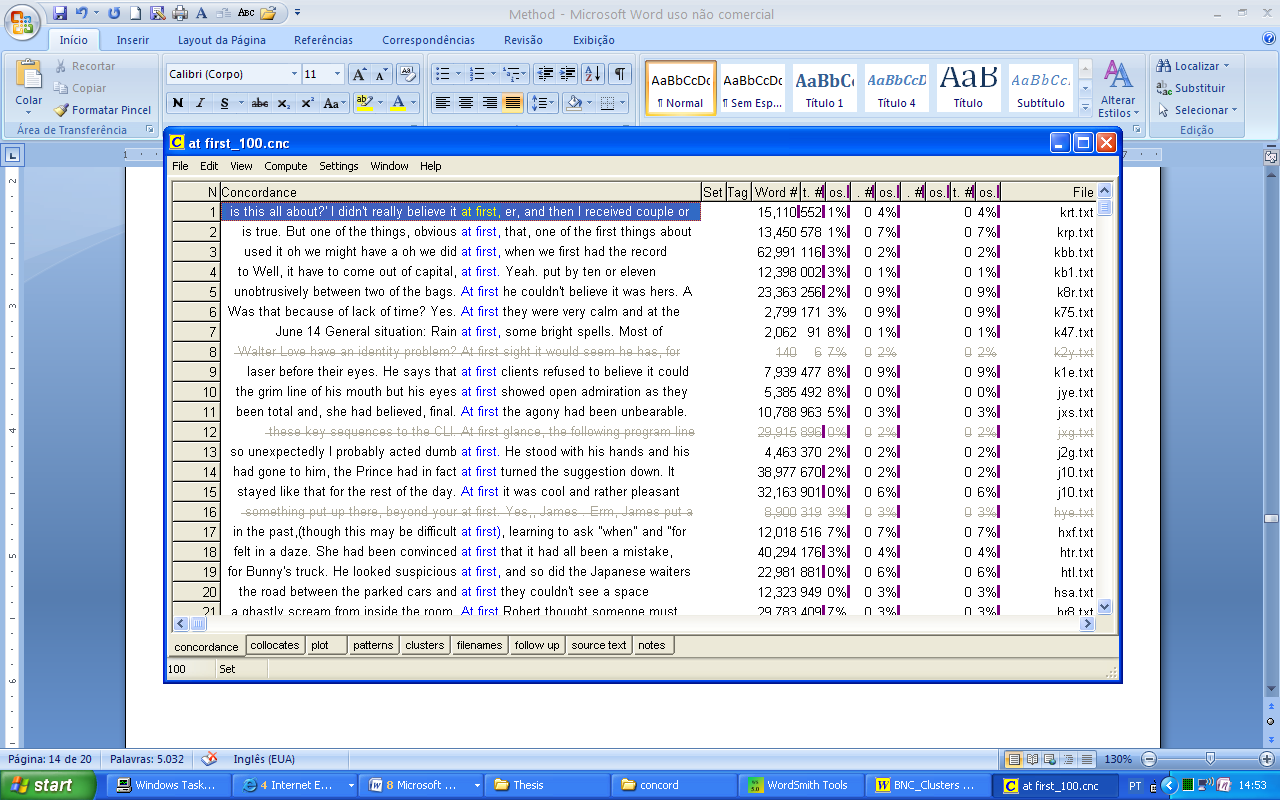


**Figure 6** Example of initial data deletion phase (faded n-grams are deleted ones)

Naturally there were far more lines deleted in the n-gram data than those retained, and even when potential items were identified, it was often just the beginning of a new search; it was quickly discovered that the number of phrasal expressions with unique form-meaning mappings was relatively limited. While there were many expressions of the variety that would receive a ‘ditto-tag’ under CLAWS tagging (e.g. *in spite of*, *rather than*, *as if*) because of their discrete and fixed forms, the vast majority required further investigation – just as Lorge and Thorndike found in 1939 – in order to determine their true frequency in the corpus.

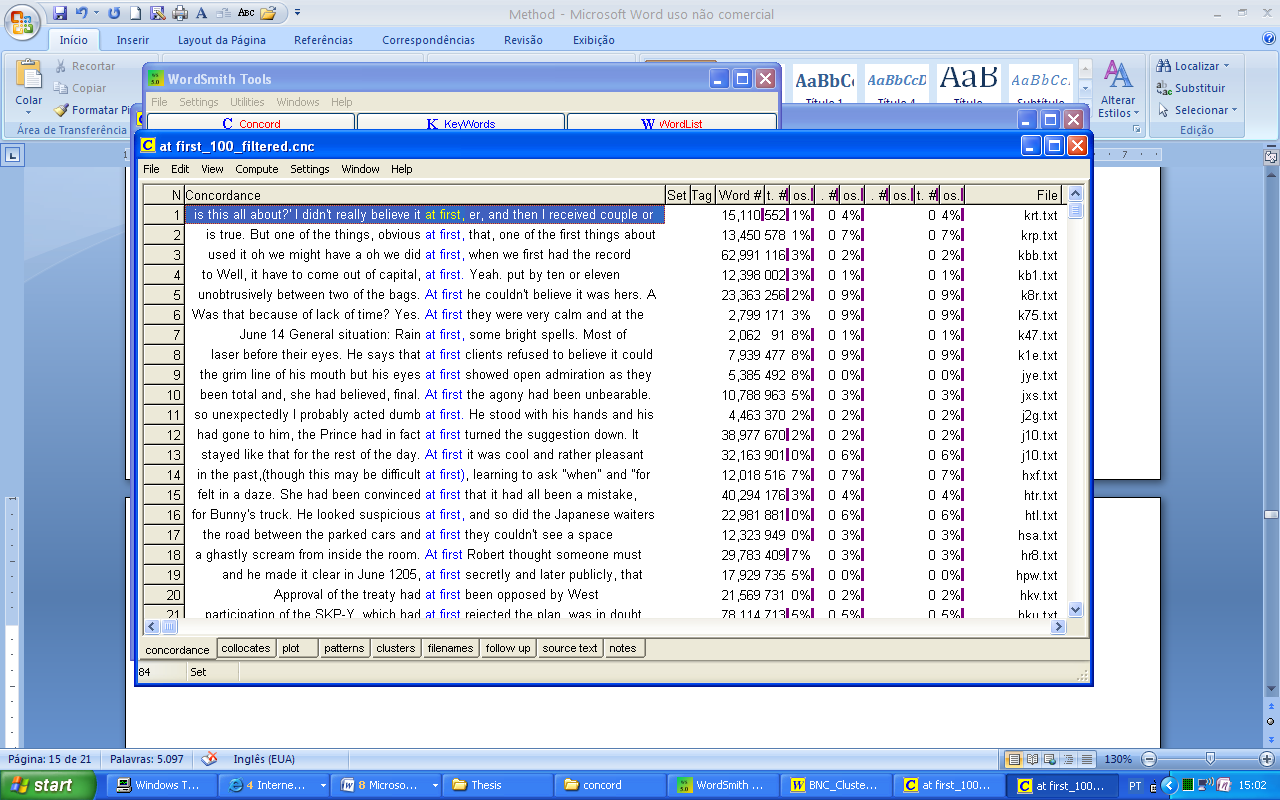
An example is the phrasal expression *at first*. At a glance, it may seem clear that *at first* is an adverbial expression (‘initially’), but with each potential phrasal expression identified an additional concordance of that item was run, and then it would become clear that *at first* also has non-phrasal expression manifestations, as in *love at first sight*. However, since an item like *at first* has a frequency of over 5,000 in the corpus, line-by-line searching was not a viable option. Therefore, a random sampling method was employed instead.

What the random sampling entailed was simply generating a concordance of the potential phrasal expression in question using the entire BNC corpus. Once generated, the concordance was saved and then a special command – ‘delete to N’ – was used to reduce the concordance lines to a random sample of just 100 (Figures 7 and 8).



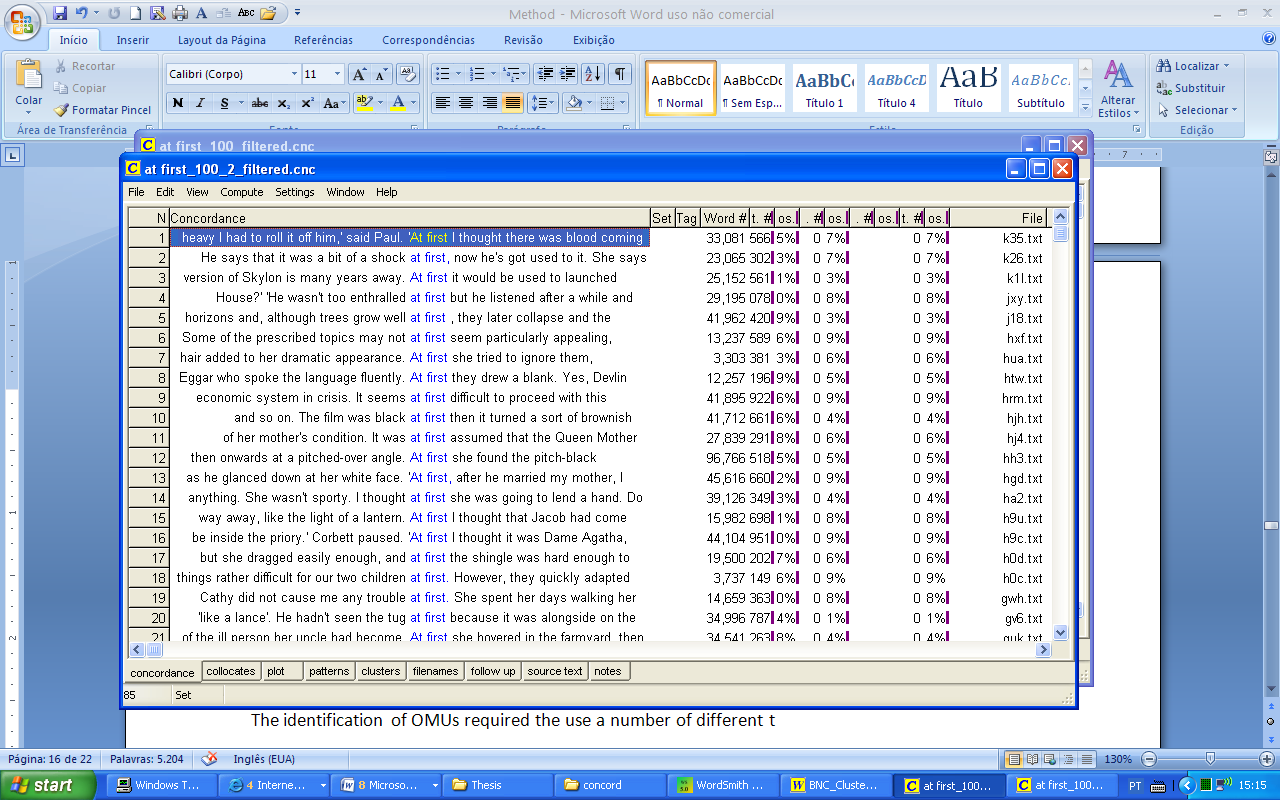
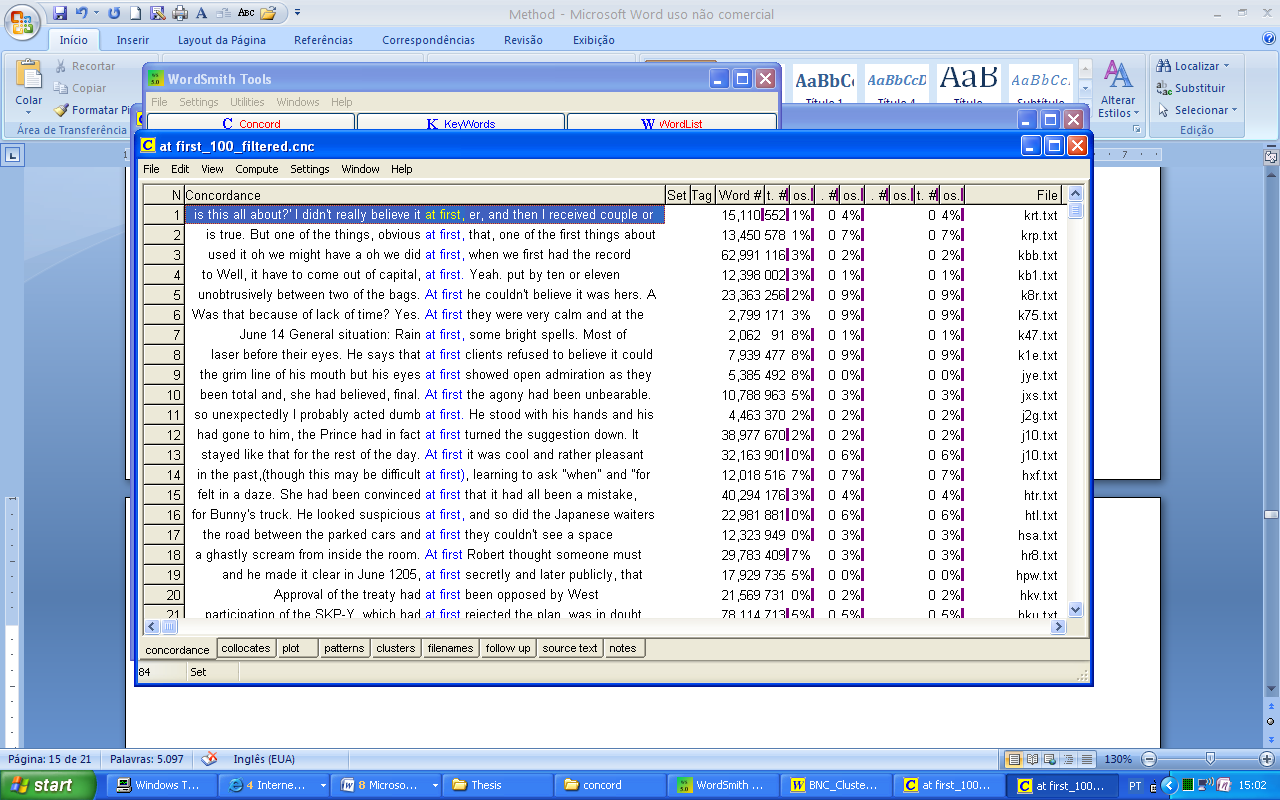
**Figure 7** An example of a random sample for the purposes of data reduction

Once the concordance was reduced to 100, each line was scrutinized and deleted if necessary (Figure 7), until the percentage of lines reflecting the desired sense of the multiword item was arrived at (Figure 8). As seen in the bottom left-hand corner of the window in Figure 8, out of 100 randomly-selected concordance lines, 84 exemplars of *at first* in its phrasal adverbial sense remain – or 84% of the original total.



**Figure 8**Final data deletion example

In order to validate this percentage, a second random sample was generated to check consistency. This method produced stable results, and in cases of minor discrepancies the lower of the two percentages was used (e.g. the two random concordances for *at first* yielded 84% and 85%, so the 84% figure was used – see Figure 9). In the rare cases in which the figures did not match so closely, additional random samples were generated until a reliable percentage figure could be derived. Finally, the frequency figure for each multiword item was calculated by multiplying the total frequency figure by the percentage figure as explained above. For *at first,* this calculation was 5,090 (raw frequency) × .84 (% of desired sense) = 4,275 (adjusted final frequency). The lines in the actual *WordSmith* word list were then edited to reflect the adjustment.



84

85

**Figure 9** Side-by-side comparison of two independent random samples of the same phrasal expression

Also, frequency figures sometimes increased from their original levels. Since the current BNC-derived wordlists are lemmatized and organized into word families, it was decided the same construct should remain in the multiword item list. The expression *take place*, for example, in its uninflected form had a frequency count of just 3,248. However, the form can also be lemmatized:

***take place*** → *takes place, taking place, taken place, took place*

In the case of *take place*, after conflating all of the inflected forms, the count increased from 3,248 to 10,556.

On other occasions, a subtractive method could be employed in order to arrive at a more accurate frequency figure. For example, *opposed to* essentially has two manifestations: *(be) opposed to sth*, and *as opposed to*. The n-gram list is not much help on its own since the program was asked to identify all recurring 2-to-4 word strings, and therefore *opposed to* is subsumed in *as opposed to*. In order to focus on just *opposed to*, it was possible to simply subtract the number of occurrences of the string *as opposed to* (1,615) from the number of times the bigram *opposed to* appears in the corpus (2,674), which rendered a difference of 1,059. In other words, the true frequency of just *opposed to* is 1,059.

Finally, expressions were sometimes encountered that contained variable components. For example, in the BNC, the first exemplar of *shake one’s head* is actually ‘shook his head’ (1,698 occurrences). When a phrase with a variable component such as this one was identified (in this case, mainly the pronoun), a careful follow-up search was conducted in order to identify all variable forms of that expression and arrive at a more accurate frequency count of it. Therefore, after considering *shook his head* (1,698), *shook her head* (1,241), *shook my head* (114), *shake my head* (30), *shaking my head* (17) and so on, the final frequency tally was 3,250.

In all, 505 phrasal expressions were identified that occurred at least 787 times in the BNC – matching the single-word frequency range of the top 5,000 words in that corpus – and which also met the semantic criteria.

## *2.3 Compiling and presenting the list*

Following the analysis presented in Section 2.2, the first version of the PHRASE List was produced. However, the list eventually underwent a number of changes as a result of further analyses and consideration. The first version (Figure 10) lacked the desirable features found in the West (1953) GSL.

IN CONTRAST 2229

THIS STAGE 2223

ALL BUT 2214

ABOVE ALL 2212

RID OF 2212

IN ANY CASE 2159

THANKS TO 2159

GO AWAY 2150

ONCE MORE 2146

OH WELL 2129

**Figure 10** A sample from the first draft of the PHRASE List

For example, colleagues and peer reviewers commented that while certain phrases were readily recognizable as lexical items (e.g. *might as well*, *in the first place*, *take for granted*) with clear, discrete form-meaning mappings, some other phrases eluded immediate interpretation (e.g. *or so*, *all but*, *yet to*). This was obviously a problem, particularly if the list was ultimately intended to be accessible and usable by a wide variety of end-users, including teachers and learners of English. Much in the same way Michael West did for the original (1936) GSL, the decision was taken to contextualize the vocabulary in example sentences (Figure 11).

|  |  |  |  |
| --- | --- | --- | --- |
| Integrated  List  Rank | **Phrase** | **Frequency**  (per 100 million) | **Example** |

|  |  |  |  |
| --- | --- | --- | --- |
| 3149 | **IN CONTRAST (TO)** | 2229 | The inside was amazing **in contrast**. |
| 3152 | **THIS STAGE** | 2223 | We can’t at **this stage**. |
| 3157 | **ALL BUT** | 2214 | She **all but** gave up when she saw her test score. |
| 3160 | **ABOVE ALL** | 2212 | It is **above all** what people care most about. |
| 3162 | **RID OF** | 2212 | She was happy to be **rid of** it. |
| 3197 | **IN ANY CASE** | 2159 | It’s not due till tomorrow **in any case**. |
| 3199 | **THANKS TO** | 2159 | And it’s **thanks to** her research that we know that. |
| 3205 | **GO AWAY** | 2150 | The problem won’t just **go away**. |
| 3207 | **ONCE MORE** | 2146 | I call on you **once more** my fellow citizens. |
| 3220 | **OH WELL** | 2129 | It was due yesterday? **Oh well**. |

**Figure 11** A sample from a later revision of the PHRASE List

As shown in Figure 11 above, the contextualization of the phrases seems to have enhanced the interpretability of the items in all cases, with the meaning of a phrasal expression like *all but* becoming much clearer. What is more, the example sentences also help illustrate the intended sense of phrases. For example, the phrase *oh well* on its own in a list could look like some kind of incomplete stem of a longer sentence (e.g. ‘*Oh well that’s interesting’*), however the contextualization helps to show that it is an item in its own right.

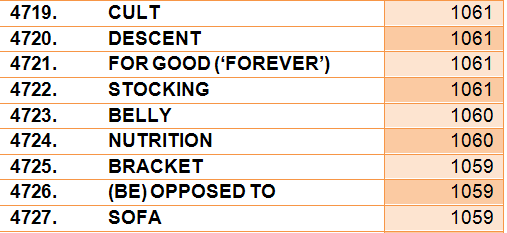
The frequency column remained the same, but in Figure 11 there is the addition of the ‘Integrated List Rank’. What this last figure indicates is where the phrasal expression would rank on a frequency-ranked wordlist derived from the BNC., as exemplified in Figure 12. These data were added as it was felt they helped to conceptualize the relative importance of the phrasal expressions, even more than the raw frequency data.

RANK

FREQUENCY

FREQUENCY

RANK



BEFORE

AFTER

**Figure 12**  Example of integrated list of phrasal expressions and single words

As reviewed earlier in this chapter, we also saw the value of distinguishing genre and modality (written/spoken) information, as done in the lists like the AWL and WFWSE, and wanted to do the same with the PHRASE List. However, the BNC is composed of hundreds of different sub-corpora, and there is no easy way to isolate, say, just general spoken conversation and investigate the frequency of a given phrase in those files. Furthermore, even the individual files in the BNC that are tagged as representing ‘spoken’ English, for example, are actually not what one would immediately think of with respect to that modality of communication, with many BNC files actually containing data of memorized and/or written language that has been read aloud. This problem of genre mislabeling in the BNC has actually long been recognized by users of that corpus (Lee, 2002). Fortunately, an index entitled ‘The BNC World Edition (Bibliographical) Index’, which exists in the form of a publicly-available Excel file, has been developed to address this issue (Lee, 2002: 1):

The BNCW Index spreadsheet was created as one solution to the problem of BNC ‘domain’ categories being overly broad and too inexplicit, to fix classification errors and steer people away from misleading file titles, and to provide a proper navigational map for people wanting to deal with specific ‘genres’ (as generally understood by most people).

Initially, it was hoped that four genres could be isolated and investigated for the PHRASE List: spoken general English (i.e. general conversation), written general English, spoken academic English, and written academic English. However, upon closer examination of the BNCW Index, it was decided that there were not enough corpora of the spoken academic genre present in the corpus to allow for a representative and comparable sample to be generated. Therefore, two million words of the other genres were isolated from the BNC, and merged to form three sub-corpora for further analysis. The 505 phrasal expressions in the PHRASE List were then individually checked for their relative frequency. An example of the data that resulted, and the alterations that took place as a result in the PHRASE List, can be seen in Figure 13.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Integrated  List  Rank | **Phrase** | **Frequency**  (per 100 million) | Spoken general | Written general | Written academic | **Example** |
| 107 | **HAVE TO** | 83092 | **1479** | **502** | **89** | I exercise because I **have to**. |
| 165 | **THERE IS/ARE** | 59833 | **1133** | **997** | **668** | **There are** some problems. |
| 415 | **SUCH AS** | 30857 | **130** | **591** | **620** | We have questions, **such as** how it happened. |
| 463 | **GOING TO (FUTURE)** | 28259 | **587** | **194** | **12** | I’m **going to** think about it. |
| 483 | **OF COURSE** | 26966 | **511** | **327** | **41** | He said he’d come **of course**. |

**Figure 13** Sample of the PHRASE List with numerical genre-sensitive frequency information

However, as presented, these new frequency data pertaining to genre were ultimately found to be more of a hindrance than help, unfortunately. Every person who had the opportunity to look at the new version of the list found the new numbers confusing, and understandably so. First of all, the original data columns containing frequency information are still in the list, so the addition of three new sets of numbers is somewhat daunting. Second, the new sets of frequency information actually are un-interpretable in practical terms. As an example, the phrase ‘such as’ is shown to have 130 attested examples in the spoken general corpora analyzed. However, even when juxtaposed with the subsequent two columns with higher figures, what does that number of 130 mean? To a user picking up the list for the first time, one can imagine it would be very difficult to determine if that figure of 130 means that it is rare, and if it is, how rare it is relative to the other genres.

This led to the development of a new, non-numerical system, which is the one used in the final version of the PHRASE List (sample provided in Figure 14).

\* \* \* = phrase most common in this genre (or as common)

\* \* = phrase less common in this genre

\* = phrase infrequent in this genre

X = phrase rare or non-existent in this genre

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Integrated  List  Rank | **Phrase** | **Frequency**  (per 100 million) | Spoken general | Written general | Written academic | **Example** |
| 107 | **HAVE TO** | 83092 | **\* \* \*** | **\* \*** | **\*** | I exercise because I **have to**. |
| 165 | **THERE IS/ARE** | 59833 | **\* \* \*** | **\* \* \*** | **\* \*** | **There are** some problems. |
| 415 | **SUCH AS** | 30857 | **\*** | **\* \* \*** | **\* \* \*** | We have questions, **such as** how it happened. |
| 463 | **GOING TO (FUTURE)** | 28259 | **\* \* \*** | **\* \*** | **x** | I’m **going to** think about it. |
| 483 | **OF COURSE** | 26966 | **\* \* \*** | **\* \*** | **\*** | He said he’d come **of course**. |

**Figure 14** Genre-sensitive frequency information represented by system of symbols

As seen in Figure 14, a system of four symbols was devised which correspond to the frequencies of each item, previously represented numerically. Each phrasal expression now has at least three stars (‘\*\*\*’) in at least one genre, representing the genre in which that phrase occurs the most. If the frequency of that same phrase in another genre was found to be the same (within 30 percent), the same number of stars was assigned. However, if the token frequency was between 30 to 70 percent less than the highest value at three stars, it was assigned two stars, and if representing only between 5 to 29 percent of the highest value, just one star. Anything less frequent than 5 percent of the highest value was assigned an ‘X’, designating the phrase as rare or non-existent in that genre.

## 

## 3. Conclusion

It has been our aim in this chapter to review what we believe have been important lessons learned (or that should be learned) over various attempts to use linguistic corpora to identify and prioritize vocabulary for English language teaching and learning, so that current and future related research can remain mindful of them. Although it is clear that, with the help of computers, the trail blazed by Thorndike and colleagues a century ago when they painstakingly labored for years through a dense timberland of millions of printed words has since been turned into a veritable superhighway, the lexical forest still needs to be occasionally traversed unhurriedly and observantly in order to be properly appreciated. As evidenced by the enduring influence of Michael West’s 1953 General Service List, vocabulary lists driven by pure quantitative data are more likely to find their way into practical applications when they are complemented by qualitative judgments in a user-friendly presentation. Furthermore, while more exhaustive lists like those compiled by Thorndike and Leech and colleagues have a certain use as works of reference, it seems that the more restricted lists of carefully selected items – such as the highly influential GSL and AWL – are ultimately more widely incorporated into pedagogy. However, the AWL is also an example of the importance of considering semantic as well as frequency data to avoid potentially misleading the end-user of such lists. We believe that our list has benefited from the legacy of successes and caveats that each corpus-informed vocabulary list that precedes our own provides. Nonetheless, we must also recognize that in developing and releasing our own list, that we become a part of that legacy. We do not expect that our research as presented here should dictate the design of future similar studies, but we hope to have contributed to the ever-evolving understanding of how corpora can enhance the study of vocabulary. As West himself said so many years ago,

it is undesirable that there should ever be any one prescriptive list, for that would tend to hamper the liberty of teachers and writers, and do more harm than good. What is needed is a standard form from which infinite divergences may be made, as well as a set of criteria, so that those who diverge may do so with reasoned intention. (West, 1937: 433)

**References**

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

Beglar, D. and Hunt, A. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests*. Language Testing*, 16(1), 131-162.

Bhuian, S.N., Menguc, B., and Bell, S.J. (2005). Just entrepreneurial enough: the moderating effect of entrepreneurship on the relationship between market orientation and performance. *Journal of Business Research*, 58, 9-17.

Blackwell, S. (1987). Syntax versus orthography: Problems in the automatic parsing of idioms. In R. Garside, G. Leech, and G. Sampson, (Eds.), *The Computational Analysis of English* (pp. 110-119). Harlow: Longman.

Campion, M. and Elley, W. (1971). An academic vocabulary list. Wellington, New Zealand: Council for Educational Research.

Capel, A. (2010). A1-B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(e3), 1-11.

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

Davies, M., and Gardner, D. (2010). *A frequency dictionary of contemporary American English*. New York: Routledge.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, 28(3), 157-169.

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

Garside, R. (1987). The CLAWS word-tagging system. In R. Garside, G. Leech and G. Sampson (Eds.), *The Computational Analysis of English* (30-41). London: Longman.

Heatley, A. and Nation, P. (1994). Range. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge: Cambridge University Press.

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

Hu, H.M. and Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403-429.

Kučera, H. and Francis, W.N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Laufer, B. (1991). How much lexis is necessary for reading comprehension? In: Arnaud, P. J. L. and H. Béjoint (Eds.), *Vocabulary and applied linguistics* (126-132). Oxford: Macmillan.

Laufer, B. and Ravenhorst-Kalovski, G.C. (2010). Lexical threshold revisited: Lexical text coverage, learners’ vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15-30.

Lee, D. (2002). Notes to accompany the BNC Word Edition (bibliographical) index. Available at http://www.uow.edu.au/~dlee/home/BNCWIndexNotes.pdf (accessed 18 November 2009).

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, G. Gulquin, and M. Paquot (Eds.), *A Taste for Corpora: In Honour of Silviane Granger* (pp. 7-31). Amsterdam: John Benjamins.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. Harlow: Longman.

Lorge, I. and Thorndike, E.L. (1938). *A Semantic Count of English Words*. New York: Teachers College, Columbia University.

Lubliner, S. and Hiebert, E.H. (2011). An analysis of English-Spanish cognates as a source of general academic language. *Bilingual Research Journal*, 34(1), 76-93.

Martinez, R. and Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, 33(3), 299-320.

Matsuoka, W. and Hirsh, D. (2010). Vocabulary learning through reading: Does an ELT course book provide good opportunities? *Reading in a Foreign Language*, 22(1), 56-70.

Meara, P. and Milton, J. (2003). *X\_Lex, The Swansea Levels Test*. Newbury: Express.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I.S.P. (1983). Testing and teaching vocabulary*. Guidelines*, 5, 12-25.

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

Nation, I.S.P. (2004). A study of the most frequent word families in the British National Corpus. In B. Laufer and P. Bogaards (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Tesing*, ( 3-13). Amsterdam: John Benjamins Publishing Company.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review,* 63(1), 59-82.

Nation, I.S.P. and Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, 31(7), 9-13.

Nation, I.S.P. and Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt and M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (6-19). Cambridge: Cambridge University Press.

Nesselhauf, N. (2004). What are collocations? In D.J. Allerton, N. Nesselhauf, and P. Skandera (Eds.), *Phraseological units: Basic concepts and their application* (1-21). Basel: Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing Company.

O'Keeffe, A., McCarthy, M., and Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Palmer, H.E., West, M., and Faucett, L. (1936). *Interim Report on Vocabulary Selection*. London: PS King and Son, Ltd.

Parent, K. (2012). The most frequent English homonyms*. RELC Journal*, 43(1), 69-81.

Praninskas, J. (1972). *American University Word List*. London: Longman.

Rayson, P. (2008). Software demonstration: Identification of multi-word expressions with WMatrix. Paper presented at the *Formulaic Language Research Network (FLaRN) Third International Postgraduate Conference,* 19 June 2008, Nottingham, UK.

Read, J. (2004). Research in teaching vocabulary. *Annual Review of Applied Linguistics,* 24, 146-161.

Read, J. and Nation, P. (2004). Measurement of formulaic sequences. In Schmitt, N. (Ed.), *Formulaic sequences* (23-36). Amsterdam: John Benjamins Publishing Company.

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

Schmitt, N. (2004). *Formulaic sequences.* Amsterdam: John Benjamins.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88.

Smith, K. (2011). Integrating one hour of in-school weekly SSR: effects on proficiency and spelling. *International Journal of Foreign Language Teaching*.

Stæhr, L. S. (2009). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, 36(2), 139-152.

Taylor, J.R. (2006). Polysemy and the lexicon. In G. Kristiansen, M. Achard, R. Dirven, and F.J. Ruiz de Mendoza Ibanez (Eds.), *Cognitive linguistics: Current applications and future perspectives* (51-80). Berlin: Mouton de Gruyter.

Thorndike, E.L. (1921). Word knowledge in elementary school. *Teachers College Record*, 22(4), 334-370.

Thorndike, E.L. (1932). *A teacher’s word book of the 20,000 words found most frequently and widely in general reading for children and young people*. New York: Teachers College, Columbia University.

Thorndike, E.L. and Lorge, L. (1944). *The teacher’s word book of 30,000 words*. New York: Teachers College, Columbia University.

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

West, M. (1937). The present position in vocabulary selection for foreign language teaching. *The Modern Language Journal*, 21(6), 433-437.

West, M. (1953). A General Service List of English Words. London: Longmans.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics,* 21(4), 463-489.

Wray, A. (2009). Identifying formulaic language: Persistent challenges and new opportunities. In R. Corrigan, E.A. Moravcsik, H. Ouali, and K.M. Wheatley (Eds.), *Formulaic language volume 1: Distribution and historical change* (27-51). Amsterdam: John Benjamins Publishing Company.

Xue, G. and Nation, I.S.P. (1984). A University Word List. *Language Learning and Communication*, 3, 215-229.

Zipf, G.K. (1949). *Human behavior and the principle of least effort*. Oxford: Addison-Wesley.

**OTHER REFERENCES**

British National Corpus (version 3) (BNC XML edition) [CD-ROM]. (2007). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Retrieved from http://www.natcorp.ox.ac.uk/

Cobb, T. Web Vocabprofile. [accessed 17 March, 2008 from http://www.lextutor.ca/vp/], an adaption of A. Heatley and P. Nation, Range (1994).

1. A list designed for overseas students in New Zealand, constructed using a corpus of 301,800 words based on published material (i.e. textbooks, academic journal papers) and a selection of university examination papers, covering the most important academic disciplines at New Zealand universities in the 1970s (Campion & Elley, 1971). [↑](#footnote-ref-1)
2. Compiled using an academic-textbook-derived corpus of 272,466 words, aimed for use by non-native speakers at American universities. The list excluded words from the General Service List (West, 1953). [↑](#footnote-ref-2)
3. There were attempts at including phrases in previously published lists, such as West’s original GSL (1953) and Hindmarsh’s *Cambridge English Lexicon* (1980), but they were always included under headwords and not listed as their own entry, and were based on manual counts of relatively small corpora. [↑](#footnote-ref-3)
4. So called because after CLAWS tagged each word in the corpus, a specially-designed complement to CLAWS (IDIOMTAG) would identify a sequence of two or more words that are serving the function of one grammatical word, and then re-tag just the first word of that sequence and simply add the same tag with ‘dittos’ (“) to the subsequent constituent words (Blackwell, 1987). [↑](#footnote-ref-4)
5. In fact, there are not even any data available for how accurate the ditto-tags are, only that the items tagged as multiwords (‘<mw>’) “should not be included in any assessment of the CLAWS error rate.” (http://www.natcorp.ox.ac.uk/XMLedition/URG/codes.html#defrobs) [↑](#footnote-ref-5)
6. A full listing of all the sources contained in the corpus can be viewed at http://www.natcorp.ox.ac.uk/XMLedition/URG/bibliog.html. [↑](#footnote-ref-6)