FREQUENCY 2.0: Incorporating homoforms and multiword units in pedagogical frequency lists

Thomas Cobb
Université du Québec à Montréal

The importance of frequency as a principle for organizing language learning, while long promoted in principle (Palmer, 1941; West, 1953), has recently become feasible in practice with three new developments: theoretical support from acquisition theorists (Ellis, 2002); the assembly of truly enormous, representative and accessible language corpora (Davies, 2011; Leech, Rayson & Wilson, 2001); and the extraction of pedagogically relevant lexical information (Nation, 2006) and grammatical information (Biber et al., 1999) from them. Since about 1990, this frequency information has regularly been deployed in the development of language courses and learning resources, particularly lexical resources such as dictionaries and tutorial computer programs for learning vocabulary. Now, however, at least in the area of lexis, the frequency approach must face two consequences of its own success: larger corpora and stronger tools of analysis have revealed not just useful ranked lists of word forms, but (1) the extent of homonymy and homography hidden within them, and (2) the extent of multiword units with meanings independent of their component words. The present paper makes the case for including both types of information in pedagogically oriented frequency lists. It shows firstly why this should be done, then reviews some new research that is making it possible, and finally develops and pilot-tests a way of doing it. The underlying theme is that the technologies that raised the problems of homoforms and multiword units can also be used to solve them.

1. Introduction

Applying corpus insights to language learning is slow work with roughly one or two interesting advances per decade. In terms of lexis and frequency: Tim John's corpus and concordance package MicroConcord became available in 1986, enabling language practitioners to build concordances and calculate word frequencies in their own texts and compare these to more general word frequencies in the small corpora bundled with the program. In the 1990's, Heatley and Nation's (1994) Vocabprofile, a computational deployment of West's (1953) General Service List (GSL) integrated with a series of academic lists, allowed

EUROSLA MONOGRAPHS SERIES 2
L2 vocabulary acquisition, knowledge and use, 79-108

practitioners to perform MicroConcord's two functions together: analyzing texts in terms of the frequency of their individual words both in a particular text and in the English language as a whole. The 2000's have been largely devoted to exploiting the 100-million word British National Corpus (BNC; Aston & Burnard, 1998) and the frequency lists derived from it (Leech et al., 2001). Some important exploitations have been the pedagogical adaptation of these lists (Nation, unpublished), and then their incorporation in a vocabulary test (Beglar & Nation, 2007), deployment in a Vocabprofile update (Nation, 2006), use in a variety of research enterprises (discussed below), and dissemination to researchers, teachers and learners on the World Wide Web (partly via the *Compleat Lexical Tutor* Website, or *Lextutor,* www.lextutor.ca). A likely near-term development will be the incorporation of US English into the scheme from the COCA, or Corpus of Contemporary American English (Davies & Gardner, 2010).

A key element in the pedagogical adaptation of the BNC lists is the expansion of the grouping unit from the lemma (headword and inflections) to the word family (lemma and transparent derivations; Bauer & Nation, 1993). For example, the lemma for the noun *cup* would be *cup* and *cups*, but the family would be these plus the derived verb *to cup* (one's hands), which involves a changed part of speech but not a change in the basic meaning. The development of the family concept is based on learning principles rather than linguistics or computational principles: a learner who understands *cup* will have no problem understanding *cup your hands*.

The appeal of pedagogically oriented lexical frequency information in the language teaching industry appears to be large, an impression that can find quantitative support in Lextutor's user statistics. Since coming on line in 2005, Lextutor's user base has doubled every year and currently generates more than 10,000 concordances, frequency lists, or lexical profiles daily. Lextutor's most utilized resource is Web Vocabprofile, an online adaptation of both Heatley and Nation's original Vocabprofile (1994) and Laufer and Nation's (1995) Lexical Frequency Profiler (LFP), which categorizes every word of any text in terms of both family membership as well as the overall rank of the family in either the GSL or the BNC, calculating a profile by percentage. For example, five of the six words in this sentence, *The cat sat on the mat,* are very frequent (from the BNC's first 1,000 word families by frequency), but one, *mat,* is less frequent (from the fourth 1,000). One can thus state that the text comprises 83% first thousand items, and go on to predict that this text could probably be handled by an intermediate learner who could be predicted to know five of its six words leaving just one to work out from context or look up.

Teachers and learners use this type of analysis to determine and modify the difficulty level of texts. Frequency profiling thus connects the rough-and-ready instructional design end of language learning with the frequency-based learning principles of acquisition researchers like Ellis and Larsen-Freeman (e.g., 2009) at the other. Vocabprofile analysis is fairly simple in both concept and function, and has received empirical validation in both English (Laufer & Nation, 1995; Morris & Cobb, 2004) and French (Ovtcharov, Cobb & Halter, 2006; Lindqvist, 2010) and is a mainstay in the ongoing text coverage and comprehension research (Nation, 2006; Schmitt, Jiang & Grabe, 2011; van Zeeland & Schmitt, in press).

Taking Vocabprofile as an example of how frequency information is being used in the language learning field, we can continue with a finer grained account of the slow but steady evolution roughed out above. As already mentioned, the original frequency list at the heart of Vocabprofiling (West's, 1953, two thousand-item General Service List) has now been replaced by the BNC list (Leech et al., 2001) as adapted and divided by Nation (unpublished) into 14 family thousand-lists. The increase in the number of lists from two to 14 allows much finer grained profiles of texts, clearer distinctions between texts, and a substantial reduction in the percentage of words that cannot be categorized. Other developments in the concept and software are mainly modifications suggested by practitioners, including colour coding of frequency zones, automated treatment of proper nouns, and the sequential re-analysis of evolving text modifications (Cobb, 2010). However, these and related developments have not involved a rethinking of the basic idea, which is to match text words to static frequency information straight out of a computer program whose knowledge of language is limited to counting up the items between empty spaces and judging where they are the same or different to each other and to words in a database.

While it has been possible to do a good deal of frequency work using this simple definition of *word,* the definition was based on two assumptions known to be incorrect but believed to pose relatively minor problems. It was assumed that homoforms (an umbrella term for homonyms, like *river banks* and *money banks,* and homographs, like *to read* and *well read*) could be provisionally ignored. It was also assumed that multiword units (MWUs, phrases with meanings independent of their individual words, like *up to you* and *a lot*) could be overlooked, at least for a while. But larger corpora and growing familiarity with their contents has now revealed the extent of the homoforms and MWUs that lie hidden in between-the-spaces frequency lists. That is, many single words are really two words, and many phrases are really single words. These arguably merit separate entries in a pedagogical frequency list, as well as revamped frequency ratings and pedagogical emphases. It may be that *a_lot* (of anything) should be taught without reference to *a lot* (to build a house on), and *banks* (for money) should be introduced to beginners and *banks* (of rivers) reserved for later, rather than mixing everything together, as happens at present and is tacitly supported by existing Vocabprofile software. Without accounting for this information within and beyond the space-defined word form, existing frequency profiles are almost certainly inaccurate to some unknown degree. Or to put it another way, frequency profiling could be even more useful than it is now. Fortunately, much of the human and computational spade work has already been done to achieve this.

2. FREQUENCY 2.0: Why it is needed

West's hand-made General Service List (1953) of 2,000 high-value lexical items for English teaching made careful distinctions not only between homoforms, which are clearly different words (*money banks* and *river banks*), but also between main senses of words (*cloud banks* and *river banks*). The limitations of this list are that it is small (2,000 word families), intuitive (with only rudimentary frequency information), narrowly pedagogical (no vulgarities allowed), and largely inapplicable to text creation or modification except through handwork with small texts. These shortcomings have now been more than compensated for by lists based not only on huge corpora like the BNC, but also by the systematic inclusion of *range* (the distribution of items across the BNC's 100 subdivisions) as a second consideration in their construction. And yet it is ironic that in the newer lists, the old distinctions have temporarily been lost between both word senses and homoforms. Distinguishing word senses may not be crucial to such an enterprise, if, as Beretta, Fiorentino and Poeppel (2005) argue, these are normally computed in real time from a single entry in the mental lexicon. Nation (e.g., 2001) has long argued for a pedagogy focusing on the "monosemic" concept underlying the polysemes. Nonetheless, homoforms do pose a problem.

The BNC frequency list produced by Leech et al. (2001), while lemmatized for part of speech, does not distinguish between different words that are merely linked by a common word form. A trip to the Web version of the BNC (at http://bncweb.lancs.ac.uk/) reveals that the program is able to output lemmas (related morphologies of the same word form) but not distinguish homoforms. Nor does the newer list by Davies and Gardner (2010) drawing on the even larger Corpus of Contemporary American English (COCA, 425 million words, see Figure 1).

The combined meanings of *bank* shown in Fig. 1 place the word-form at rank 701 in the frequency list, hence in the first 1,000 words by frequency. But this placement is almost certainly an artifact of lumping the two *banks* together, as shown by the collocates *account, loan,* and *river* in line 3. $Bank_1$ and $bank_2$ are clearly distinct words linked mainly by a resemblance of form (and possibly a common etymology that few language users would be aware of). The reason

Figure 1. Homoform lumping in Davies & Gardner (2010)

701 bank *n*
adj central, national, federal, large, commercial, international, reserve, foreign, major, outer
noun world•, •account, loan, river, money, invest < credit, development, fund, saving *verb* rob•, buy, •lend, own, •fail, finance, charge, issue, borrow, operate
52366 | 0.93

for failure to distinguish between the two *banks* is, of course, clear. The amount of textual information that is summarized in a small compilation like Figure 1 is vast (the figure 52,366 at the bottom refers to the number of instances of *bank* in the COCA corpus), such that there is no easy way to insert human judgment into the process. A human investigation of the context for each of these entries, followed by a count-up, is presumably the only way to tell the different *banks* apart, and this is an arduous task.

However, with some quick and dirty human-computer cooperation based on random sampling, this prising apart can be done for many practical purposes. For example, here is a mini-experiment for the word-form *bank* based on the 50 random non-lemmatized samples offered for free by the BNC website at http://www.natcorp.ox.ac.uk/. Entering a search for *bank* reveals that the BNC contains 17,603 lemmatized instances of this item (all noun forms combined). Then, eyeballing and counting up the separate meanings from the available 50 random concordance lines over 10 runs, we find a remarkably consistent 43 to 50 lines of money or blood *bank* and only 5 to 7 of river or cloud *bank*. Thus a rough 86% to 96% of the 17,603 uses pertain to money *bank*, or minimally 15,138 occurrences, so it is probably safe in its first-1,000 position (see Figure 1 for BNC cut-offs). But *river bank* is instead a medium frequency item (7 uses in 50, or 14% of the BNC's 17,603 total occurrences amounts to 2,465 occurrences, placing it near the end of the third 1,000 by frequency).

The recent large-corpus based lists also fail to distinguish between MWUs that are compositional, like *a+lot* (to build a house on), and ones that are non-compositional, like *a_lot* (of money), in the sense that the individual words do not add up to the accepted meaning of the unit (as suggested in the notation of an underscore rather than a plus sign). But once again the corpora make it possible to do so. Passing large corpora through computer programs identifies a wealth of information about all the ways that words co-occur in more than random sequences and the extent to which they do so (Sinclair, 1991). In Figure 1, we see

COCA's main collocates of *bank*, with bullet signs indicating whether each falls consistently before or after the key word (*world*• = World Bank, •*account* = bank account). What the computer output does not show is that not all collocates are created equal. In some, the node word and collocate retain their independence (*an international bank*), while in others they do not (*World Bank*, *Left Bank*, *West Bank*). Degree of connectedness can to some extent be predicted by frequency of found versus predicted co-occurrence in such measures as mutual information or log-likelihood, as calculated by programs like BNC-Web (which gives *international bank* a mutual information (MI) value of 3.04 and *West Bank* a value of 5.82 or almost double).

In two BNC-based studies, both again involving computational analysis with human follow-up, Shin and Nation (2007) and Martinez and Schmitt (2012) identified unexpectedly large numbers of recurring word strings in the highest frequency zone of the language. Shin and Nation's co-occurrences (*you know, I think, a bit*) were for the most part compositional items which, if incorporated into the existing frequency scheme, would count as first 2,000 items. There was no proposal actually to incorporate these items into standard frequency lists, but merely to argue for their importance to language learners. Martinez and Schmitt's focus, on the other hand, was specifically on high-frequency co-occurrences that they judged to be non-compositional, or idiomatic, i.e. which have, in specific environments, independent meanings and hence deserve their own places within standard frequency lists. Using a methodology to be described below, these researchers identified 505 such MWUs in the first five thousand-lists of the BNC (or just over 10%), distributed over these lists in the manner shown in Table 1.

Table 1. Distribution of Martinez and Schmitt's MWUs by 1000-group

Number of MWUs	Zone (by 1000)	Proportion of zone (%)
32	1k	3.2
75	2k	7.5
127	3k	12.7
156	4k	15.6
97	5k	9.7

Incorporating homoform and MWU information into frequency lists could cause quite extensive changes in their composition. If a word form like *arm,* a first thousand item, were found to be about equally implicated in weaponry and anatomy, it is doubtful that either of these would remain a first 1,000 item: one or both might be bumped down to second thousand or beyond. If Martinez and

Schmitt's 505 MWUs were given their rightful places and added to the current frequency lists, then quite a number of existing items would be displaced from zone to zone (which are arbitrary divisions in any case). The result would be a set of lists something like the one imagined in Table 2.

Table 2. The type of frequency list needed

1000 List	3000 List
bank_1	bank_2
of_course	course
something	something_of_a

Incorporating these two kinds of information would also have strong effects on the deployment of frequency information in the profiling of novel texts. Profiling would no longer be a simple matter of matching a word in a text to its family headword and thence to its counterpart in a frequency list. Rather, the profiler would have to interpret both homoforms and MWUs in context, in order to determine which meaning of a homoform was applicable (*bank_1 or bank_2*), and in the case of MWUs whether a particular string was compositional or non-compositional ('look *at all* the bugs', or 'I don't like bugs *at all*'). It is this incorporation of context that is the qualitative transformation implied in the term Frequency 2.0.

3. The feasibility of reworked frequency lists

Frequency profiling up to present has been based on single word forms. It has relied on matching stable word frequencies to equivalent word forms in a given text. The modification proposed here involves not only extensive modification of the lists, but also a real-time contextual analysis of each potential homoform or MWU to determine its true identity in a particular text. These are dealt with in turn.

3.1. Multiwords

Whether for homoforms or MWUs, the first task is to identify the item involved, assign it to a category ('money *bank*' or 'river *bank*'; '*a lot* of money' or 'build on *a lot*'), calculate the frequency of each in a large corpus, and give each a place in the standardized vocabulary lists used by course developers, test writers, and computer programs like Vocabprofile. A methodology for doing this work is under development in a new crop of student research projects in vocabulary.

Table 3. The highest frequency MWUs from Martinez and Schmitt (2012)

Integrated List Rank	MWU	Frequency (per 100 million)	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	I'm **going to** think about it.
483	OF COURSE	26966	He said he'd come **of course**.
489	A FEW	26451	After **a few** drinks, she started to dance.
518	AT LEAST	25034	Well, you could email me **at least**.
551	SUCH A(N)	23894	She had **such a** strange sense of humor.
556	I MEAN	23616	It's fine, but, **I mean**, is it worth the price?
598	A LOT	22332	They go camping **a lot** in the summer.
631	RATHER THAN	21085	Children, **rather than** adults, tend to learn quickly.
635	SO THAT	20966	Park it **so that** the wheels are curbed.
655	A LITTLE	20296	I like to work out **a little** before dinner.
674	A BIT (OF)	19618	There was **a bit** of drama today at the office.
717	AS WELL AS	18041	She jogs **as well as** swims.
803	IN FACT	15983	The researchers tried several approaches, **in fact**.
807	BE LIKELY TO	15854	To be honest, I'm **likely to** forget.
825	GO ON	15610	He **went on** for a while before stopping for lunch.
845	IS TO	15232	Obama **is to** address the media this afternoon.
854	A NUMBER OF	15090	**A number of** concerns were raised.
879	AT ALL	14650	Do you have any kids **at all**?
888	AS IF	14470	They walked together **as if** no time had passed.
892	USED TO (PAST)	14411	It **used to** snow much more often.
894	WAS TO	14366	The message **was to** be transmitted worldwide.
908	NOT ONLY	14110	**Not only** was it cheap, it was delicious.
913	THOSE WHO	13951	He would defend **those who** had no voice.
934	DEAL WITH	13634	The police had several issues to **deal with**.
939	LEAD TO ('CAUSE')	13555	Excessive smoking can **lead to** heart disease.
951	SORT OF	13361	It's **sort of** why I'm here.
974	THE FOLLOWING	12963	He made **the following** remarks.
984	IN ORDER TO	12762	We shared a room **in order to** reduce costs
988	HAVE GOT (+NP)	12734	I don't know what he **has got** planned.

The largest investigation into non-compositional MWUs to date was performed by Ron Martinez and his PhD supervisor Norbert Schmitt (Martinez & Schmitt, 2012). These researchers set Scott's text analysis program Wordsmith Tools 6.0 the task of generating a list of all the recurring 4, 3, and 2-word strings, or n-grams, in the 100-million word BNC, a computer run of just under four days. Lemmas rather than word forms or families were used for this stage of the analysis, so that for example all forms of a verb are included in the analysis (*have to* as well as *had to*) as is occasionally but not consistently marked in Table 3 (in the form of *is/are* and *a/an*). From this massive output, those items with fewer than 787 occurrences were eliminated (787 is the cut-off for inclusion in the first 5,000 headwords of the existing BNC-based Vocabprofile scheme, the number 5,000 being chosen for pedagogical relevance as the words most language learners are likely to be concerned with). The surviving 15,000 items were then hand-sorted in a double randomization procedure. For each candidate MWU, Wordsmith was asked to generate two random 100-word listings, which were then hand sorted into compositional vs. non-compositional meanings of the MWU. For example, in the case of the phrase *at first*, this process yielded 16 compositional uses like 'attack *at first* light' in a single iteration of this process and also 16 in the other. Non-compositional uses such as '*at first* I wasn't sure' were more frequent; there were 84 non-compositionals in one round and 85 in the other. In cases such as this, where there was a discrepancy, the lower of the two numbers was used. The original raw frequency per 100 million was then multiplied by (in this case) .84 to produce the frequency for the non-compositional meaning of the phrase (for *at first,* 5177 x .84=4275, placing it in the third thousand-list according to the cut-offs shown in Table 5). Following this method, instances of the non-compositional *at all* extrapolated to 14,650 occurrences, and thus it was placed at position 879 in the full BNC list, in other words in the first 1000 group (Table 2). In total, 505 MWUs were thus determined and situated throughout the first five lists. The 35 provisional first thousand level items are shown in Table 3, with BNC frequency and computed list rank.

It is almost certain that these rankings are not final. Some of the examples chosen suggest uncertainty in the groupings (such as the last item in Table 3 – the NP is present only with a transformation). But more broadly, compositionality, as Martinez and Schmitt propose, is a cline or continuum, such that different researchers could have selected different non-compositional units from the computer's offering. Research by Grant and Nation (2006), working with a different idea of compositionality, would suggest a less extensive list than the one proposed by Martinez and Schmitt. They feel that most of the proposed non-compositional MWUs are merely metaphorical extensions of the compositional (if *a lot* with a house on it is a large space, and *a lot* of money is a large

Table 4. Eighteen homoforms where most common meaning < 90% of 500 concordance lines

MISS	fail to get or have	50.00%	title	50.00%				
YARD	land	56.60%	36 inches	43.40%				
NET	web	59.36%	total	40.64%				
REST	remainder	62.20%	recuperate	37.80%				
RING	(to produce the) sound of a bell	67.47%	circle	32.53%				
WAKE	become awake	75.80%	a track left behind	23.00%	vigil	1.20%		
SPELL	letter-by-letter/ incantation	75.95%	interval of time	24.05%				
LIKE	resembling	76.20%	opposite of dislike	23.80%				
RIGHT	not left	77.40%	legal rights	22.60%				
POOL	water	78.62%	combine resources	21.38%				
LEAVE	part from	78.96%	direction	17.03%	permission	0.80%	leaf	3.0%
BAND	group of people	79.00%	ring	21.00%				
FIRM	business	80.12%	strong/solid	19.88%				
SET	to place/to be firm	80.40%	a collection	19.60%				
ARM	body part	83.00%	weapon	17.00%				
DEAL	an amount	84.00%	to distribute	16.00%				
HOST	of a party	85.28%	multitude	13.91%	consecrated wafer	0.81%		
WEAVE	interlace threads	87.80%	move from side to side	12.20%				

amount of money, then there is a clear similarity between the two, such that they can be seen as members of a single 'monoseme'). Thus the exact MWUs eventually to be integrated into standard frequency schemes remain to be determined. Nonetheless it seems likely that at least some of Martinez and Schmitt's selections are not very controversial (*at all, as well as* from the first 1,000 list, and *as far as* and *as long as* from the second, clearly have both compositional and non-compositional meanings). It also seems clear that Martinez and Schmitt's basic methodology for determining such items, a large-scale crunching of matched corpus samples followed by a principled selection by humans and the calculation of a frequency rating, is likely to prove the best means of working toward a standard set of MWUs. Following that, the question will be how to deploy this information in live Vocabprofiles of novel texts, and this is a question that can be tackled while the exact target items are not yet settled.

3.2. Homoforms

The work on homoforms was performed by Kevin Parent in the context of doctoral work with Nation. Parent took West's GSL list of 2,000 high frequency items as a starting point, on the grounds that most homoforms are found in the highest frequency zones and also that these would be of greatest pedagogical relevance. Wang and Nation (2004) had already shown that there were only a handful of such items (about 10) in the 570-word Academic Word List (AWL; Coxhead, 2000; a compendium of third to sixth thousand level items). In the GSL, Parent identified 75 items with two or more headwords in the *Shorter Oxford English Dictionary (SOED),* a dictionary which marks homoforms explicitly with separate headwords. For each of these 75 items, he generated 500 random concordance lines from the BNC, and hand-sorted them according to the SOED's headwords. He found that for 54 of the 75 items, the commonest meaning accounted for 90% or more of the 500 lines (surprisingly *bank* itself falls into this category, along with *bear* and *bit*; the others can be seen in Table 1 in the Appendix). Some of the remaining items whose homoformy is less skewed are shown in Table 4. Thus, we see in the first row that half of the uses of *miss* pertained to loss, or failing to have or to get something, while the other half occurred in titles (such as *Miss Marple*).

Some points about Table 4 are in order. First, the items are not lemmatized, or divided into parts of speech (POS), but are simple counts of word forms. This is because while the different meanings of a homoform sometimes correspond to a difference in POS (to *like* somebody vs. look *like* somebody), sometimes they do not ('I broke my *arms*' vs. 'I left the *arms* outside the house'). In the absence of knowing which of these two types of homoform is predominant in English, Parent's decision was to begin the analysis with word forms. Second, Parent's analysis was confined to true homoforms. This meant that he did not include words with plausible etymological relationships (gold *bar* and drink at a *bar*) and words that while undifferentiated in writing are nonetheless differentiated in speech ('*close* [shut] the door' and '*close* [near] to dawn'). The analysis is now being expanded to include all effective homoforms, roughly 100 items in the highest frequency zones. Third, as shown in Table 4, Parent's list was also confined to cases where the least important meaning of a homoform set was greater than 10% in the BNC. It has often been argued that there is no point in handling items where one meaning is vastly predominant (e.g., Wang & Nation, 2004) since the labour to do so would be great and the differences minor. However, once a methodology for assigning differential frequencies is developed, it is arguably feasible to deal with a larger number of homographs and take less frequently used members into account. For example, as already mentioned the 10% criterion leaves 'river *bank*' lumped with 'money *bank*', which intuitively seems an inaccuracy, and one that can easily be avoided once this analysis and technology is in place. A useful target is probably all the homoforms in the first 5,000 word families where the less frequent member or members account for more than 5% of cases.

Following the calculation of proportions from the 500-word samples, each item would be tagged (possibly as *miss_1* and *miss_2*) and assigned by extrapolation its two (or sometimes more) new places in the frequency lists. The evenly divided *miss* is currently a first-1,000 item, with 19,010 lemmatized occurrences in the BNC (raw information available from BNC-Web, http://bncweb.lancs.ac.uk/). But if half of these (about 9,505) are apportioned to each meaning of *miss*, then neither *miss_1* nor *miss_2* belongs in this first 1,000 category. As the first row of Table 5 shows, only lemmas occurring 12,696 times or more in the BNC qualify as first 1,000 items. Rather, both would feature in the second 1,000 zone (between 4,858 and 12,638 occurrences). In cases where a meaning distinction corresponds to a POS distinction, as with *miss,* then the POS-tagged BNC could provide even more precise information (in this case that the verb is 10,348 occurrences and the noun 8,662, both still in the second 1,000). Counts could be refined and cutoffs change as the proposed amendments are made and items shifted up and down the scale. List building would ideally be left to an expert in developing and applying inclusion criteria, with Paul Nation as the obvious candidate since he has already developed a principled method of balancing frequency and range, spoken and written data, and corpus as well as pedagogical validity, into the existing BNC lists.

Table 5. BNC's first five 1000-list cut-offs by token count (for lemmas)

K1	>12639
K2	4858 - 12638
K3	2430 - 4857
K4	1478 - 2429
K5	980 - 1477

Source: R. Martinez (2009)

Table 6 gives a sense of what this new arrangement would look like. Parent's proportions have been multiplied against BNC frequency sums and sorted according to Martinez' cut-offs in order to give a provisional look at the thousand-level re-assignments that could flow from Parent's data in Table 3. The thousand (or *k*) levels in the first column on the left are the current composite k-levels from the BNC; those in the third and subsequent columns are provisional new k-levels for the independent meanings of the homoform. (These are even *highly* provisional since they merely result from multiplying Parent's percentages from 500 lines against BNC word-form totals from 100 million words). The goal in presenting this data at this point is merely to give a flavour of the changes being proposed. Also of interest may be any compatibility issues arising from combining data from several analyses.

Note that the original 1,000-level ratings as presented in Table 6 may not be identical to those in Nation's current fourteen 1,000 lists in all cases (*spell* is shown as 2k in Table 6, but in Vocabprofile output it is 1k). That is because Nation's first two 1,000 levels (1k and 2k) are derived from the spoken part of the BNC corpus (10 million words, or 10 percent of the full corpus), in order to ensure for pedagogical reasons that words like *hello* will appear in the first 1,000 word families. All ratings in Table 6 are based on information from the unmodified BNC, in an attempt to employ a common scale to think about moving items between levels.

Table 6 shows provisional list assignments for the 18 items of Parent's analysis that would be most likely to affect frequency ratings, in that the less dominant meaning is nonetheless substantial (between 10% and 50%). As is shown, only seven items (the top six plus *pool*) would require shifting the dominant member to a lower frequency zone (e.g., from first thousand to second). Similarly, in the remainder of the homoforms identified by Parent, the reanalysis proposed here will most often leave the dominant member of a homoform at its existing level. (The remainder of Parent's analysis is shown in Table 1 in the Appendix [further analysis under way, January, 2013)]). So is this reanalysis worth the trouble?

Table 6. Provisional adjustments to frequency ratings for homoforms

MISS **19,010** **(currently)** **1k**	fail to get or have	50.00% 9,505 **(provisionally)** **2k**	title	50.00% 9,505 **(provisionally)** **2k**		
YARD **6,627** **2k**	land	56.60% 3,751 **3k**	36 inches	43.40% 2,876 **3k**		
NET **7,578** **2k**	web	59.36% 4,494 **3k**	total	40.64% 3,076 **3k**		
REST **18,368** **1k**	remainder	62.20% 11,425 **2k**	recuperate	37.80% 6,943 **2k**		
RING **12,114** **2k**	sound of a bell	67.47% 4322 **3k**	circle	32.53% 2161 **3k**		
WAKE **4,981** **2k**	become awake	75.80% 3,776 **3k**	a track left behind	23.00% 1,146 **5k**	vigil	1.20% 60 **>14k**
SPELL **3,806** **3k**	letter-by-letter/ incantation	75.95% 2,889 **3k**	interval of time	24.05% 913 **6k**		
LIKE **155,813** **1k**	resembling	76.20% 118,729 **1k**	opposite of dislike	23.80% 37,083 **1k**		
RIGHT **103,410** **1k**	not left	77.40% 80,039 **1k**	legal rights	22.60% 23,370 **1k**		

>>>

>>>

POOL **5,818** **2k**	water	78.62% **4,573** **3k**	combine resources	21.38% 1,244 **5k**				
LEAVE **63,807** **1k**	part from	78.96% 50,343 **1k**	direction	17.03% 10,847 **K2**	permission	0.80% 510 **8k**	Tree leaves	3.01% 191 **13k**
BAND **9,005** **2k**	group of people	79.00% 7114 **2k**	ring	21.00% 1891 **4k**				
FIRM **19,890** **1k**	business	80.12% 15,912 **1k**	strong/solid	19.88% 3,938 **3k**				
SET **53,544** **1K**	to place/ to be firm	80.40% 42,835 **1k**	a collection	19.60% 10,495 **2K**				
ARM **20,051** **1K**	body part	83.00% 16,725 **1K**	weapon	17.00% 3,426 **3k**				
DEAL **28,065** **1k**	an amount	84.00% 23,575 **1k**	to distribute	16.00% 4,490 **3K**				
HOST **4,327** **3K**	of a party	85.28% 3,678 **3K**	multitude	13.91% 601 **7K**	consecrated 0 wafer	.81% 34.6 **>14K**		
WEAVE **1,213** **5K**	interlace threads	87.80% 1,065 **5k**	move from side to side	12.20% 148 **>14K**				

Bumping the minor member down a zone could yield rather different text profiles from those at present. If teachers are looking for texts at a particular level, say one matched to their learners as a means of building fluency, or ahead of their learners to build intensive reading skills, then just a few items (*band_2* or *host_2*) can push a short text above or below the 95% (Laufer, 1989) or 98% known-word comprehension threshold (Nation, 2006). Given the air time given in the recent research literature to the 95 vs. 98% difference as a factor in comprehension (Schmitt et al., 2011), small differences are clearly important. Similarly when Vocabprofiles are used to assess the lexical richness of student writing (Laufer & Nation, 1995) or speech (Ovtcharov et al., 2006; Lindqvist, 2010), a small number of lower frequency items can make a large difference to the lexical richness scores of short texts.

To summarize, the resources, methodologies, and motivation for a significant upgrade of the Frequency 1.0 scheme are largely in place. These include a methodology for identifying the main homoforms and MWUs for the pedagogically relevant zones of the BNC, a means of assigning them frequency ratings, and a first application of this methodology. There is clearly much more to do in this phase of the project, yet even when this is accomplished there will still be the matter of deploying this information in the real-time profiling of particular texts.

4. Deployment of new lists in profiles of novel texts

A theme in this chapter is that the pedagogical application of a relatively simple frequency analysis of a large corpus has now necessitated a more sophisticated frequency analysis. The presence and then the extent of multiword units was first noticed and eventually tallied over the 2,000s, and now there is really no choice but to incorporate this information into the analysis. Similarly homoforms: the difference between 'the *rest* of the day' and 'a *rest* for a day' may seem a fairly minor phenomenon in a 1-million word corpus, where many minor partners in homograph pairs probably did not feature at all owing to the flukes of a small sample, but in the BNC's 100-million there is no denying its importance. A second theme in this paper, however, is that while large corpora pose new problems, they also contain within them the solutions to these problems, as will be shown in the plan for deploying updated frequency information.

The goal is to reconfigure Vocabprofiling computer programs so that each *rest* or *bank* is tagged and assigned its own frequency level. In this way, two texts, like "Pound a stake into the bank to hold the dog" and "Stake out the bank for a hold up with a dog from the pound," would be assigned quite different profiles. In considering how software can be programmed to make such distinctions, it is useful to ask how humans distinguish *bank*$_1$ from *bank*$_2$ and *at_all* from *at* + *all.* Clearly, they do it through an implicit analysis of the linguistic and situational context of the utterance, something a computer program cannot fully do at present, or maybe ever. However, a large part of a homoform's context is its particular lexical associates, which a computer program can easily identify.

The lexical associates in question are the frequent collocations that, while occurring with most words, are not so bound together that they form MWUs. In other words, these are collocates that maintain their independent or compositional meanings, as for example *fast* often collocates with *car,* and yet *fast car* is not normally viewed as a unit. In Davies and Gardner's list above (Fig. 1), the top noun collocations for 'money *bank*' are *account* and *loan,* and while no collocates are offered for 'river *bank*', these could include *grassy*, *steep*, *fishing,* or *Thames*. The discovery that large corpora have made available is, first, the great extent of these collocations, but second the fact that they are largely non-overlapping in character, at least in the case of homoforms and MWUs. We do not have *steep* money banks or *accounts* at river banks. We buy, look at, or covet *a lot* on which to build a house, but for this we need to pay or borrow *quite a lot* or *a whole lot* of money. Stubbs (2009) and Hoey (2005) both argue for systematic collocation as the means by which the mind distinguishes both polysemes and homoforms (Stubbs, p. 19, suggests this "can be done automatically" but with no reference to a running example). A test of this assertion begins with obtaining an adequate listing of collocations for a sample collection of homoforms and MWUs. A preliminary set of collocations for such a sample is explored in the next section by way of illustration.

5. A database of collocates

A listing of collocates for any single-word lemma can be generated at Sharp-Europe's BNC-based *Just-The-Word* online collocational database (at http://www.just-the-word.com/). The database supplies all collocates for an entered item if there are five or more instances of the item in the corpus; it looks within a span of five words on either side. Thus for Parent's collection of 178 homoforms, a collection of collocates down to a frequency of 10 was straightforward to produce. These collocations are, of course, not counted according to which meaning of a homoform they refer to (*between bank,* for example, is simply presented as a collocation having a frequency of 42), so once again the computer analysis has to be followed by a human sorting. This sorting is under way, but will be tested here on the first 10 items of Table 4, those most likely to cause a change in frequency rating. Table 2 in the Appendix shows the entire collocation listings for the two meanings of *bank* as generated by *Just-The-Word.*

Figure 2. BNC-Web's first 15 collocates for *at all* sorted by Mutual Information

Collocation parameters:			
Information:	collocations	Statistics:	Mutual information
Collocation window span:	3 Left - 3 Right	Basis:	whole BNC
Freq(node, collocate) at least:	50	Freq(collocate) at least:	50
Filter results by:	Specific collocate:	and/or tag: no restrictions	Submit changed parameters Go!

There are 5592 different types in your collocation database for "[word="at"%c] [word="all"%c]". (Your query "[word="at"%c] [word="all"%c]" returned 16941 hits in 2930 different texts, thinned with method *random selection* to 5000 hits)

No.	Word	Total No. in whole BNC	Expected collocate frequency	Observed collocate frequency	In No. of texts	Mutual information value
1	levels	12,047	2.733	172	144	5.9758
2	hardly	8,397	1.905	69	63	5.1788
3	none	8,979	2.037	62	60	4.9278
4	times	29,194	6.623	195	167	4.8799
5	nothing	32,216	7.308	184	161	4.654
6	costs	15,161	3.439	65	63	4.2402
7	anything	27,431	6.223	114	103	4.1953
8	no	226,707	51.430	666	518	3.6948
9	any	121,761	27.622	291	255	3.3971
10	not	451,291	102.378	848	617	3.0502
11	look	51,972	11.790	83	77	2.8155
12	n't	316,187	71.729	438	330	2.6103
13	if	253,205	57.441	219	202	1.9308
14	like	147,567	33.476	97	89	1.5348
15	did	135,699	30.784	82	74	1.4134

A listing of collocates for MWUs is unfortunately not so simple to obtain, since *Just The Word* as presently configured does not perform searches for strings longer than one word (e.g., does not offer the typical collocates for a two-word string like *at all*). Fortunately, however, BNC-Web does handle multi-words, outputting a collocate list tagged by frequency and mutual information value (the degree of connectedness between headword and collocate). A small selection of high frequency MWUs from Martinez and Schmitt's collection (Table 3) was chosen for which there seemed to be little doubt of the existence of both a compositional and non-compositional version (*at all, as well as,* and *a lot* from the first 1,000, and *as far as* and *as long as* from the second).

The working hypothesis here is that the members of both homoforms and MWUs can be distinguished by collocations, but there are nevertheless some differences between the two. One is that some MWUs do not have a compositional meaning at all, or else it is extremely unlikely, and hence there is no point performing the collocational part of the analysis. For instance, it is hard to think of a compositional way to use *in order to* or *by and large* ('Zebras thundered *by and large* vultures flew overhead'?) so these can be tagged as MWUs and assigned their frequency rank without deliberation.

BNC-Web can generate lists of lemmatized collocates for the 505 MWUs in question, and provide both raw frequency and mutual information values for each one, which allows for trimming of the list to a manageable human task. The program's output for the most connected 15 collocates of *at all* (sorted by mutual information value) is shown for illustration in Figure 2. For *at all* as a compositional phrase, the frequent collocates mostly involve words like *levels*, *times,* and *costs* (thus *at all* levels, etc.) and as a non-compositional phrase they largely involve negative quantifiers like *none*, *hardly*, and *nothing* (thus nothing *at all,* etc.) and this once again must be hand sorted. A compilation of the most frequent 50 collocates of *at all*, sorted into compositional and non-compositional lists that an updated Vocabprofile can use to do its sorting is shown in Table 3 in the Appendix.

From these diverse sources, a database of collocates for both homoforms and MWUs can be fashioned.

6. Program function

The goal is for a modified Vocabprofile program to be able to assign homoforms and MWUs to their correct identities through an analysis of the high frequency collocates in the context (in this case choosing a span of four words on either side, following Sinclair's, 1991, suggestion). The program's job is to go through a text, and for any word or phrase it recognizes as a potential MWU or homoform (from an existing list), inspect the context for items from the two collocate sets from its database, and use this information to categorize the item as, e.g., *bank_1* or *bank_2*, or as *at_all* (non-compositional unit) or *at all* (compositional separate words).

This procedure is intended to simulate a much reduced version of what humans do when they encounter ambiguous words or phrases. Further human-like functions of the program include (1) a coherent information assumption and (2) a competition procedure for conflicting information. For the first, once for instance *bank* has shown itself to be *bank_2* (river bank) in a particular text, then in the absence of further information the next occurrence is also assumed to be this same kind of *bank* on the grounds that it is uncommon for the two senses of a homograph to appear in the same text (money *banks* located on river *banks*?). Where this does happen, however, by the second assumption collocates are simply counted up on a competition basis (most collocates wins) in an elemental version of the "cue summation model" proposed by MacWhinney (1989, p. 200) for similar language choices. In future, this calculation could be refined by inclusion of strength-of-relationship information from a corpus, such as mutual information value.

The way this procedure would work in a Frequency 2.0 Vocabprofile is as follows: The user enters a text for analysis. The Familizer subroutine (lextutor.ca/familizer) translates every word form in the text into a family headword (e.g., every *had* is changed to *have*) based on Nation's (2006) pedagogical rendering of the BNC frequency list. The disambiguator routine (living in prototype form at lextutor.ca/concordancers/text_concord/) then reads through the text-as-families, first in three-word, then two-word n-grams (to pick up any *at all*-like items) and then in singles. Every n-gram and single is weighed against the program's stop list of potential homoforms. In the singles phase, for example, the program comes across the headword *miss,* finds the item to be in its stop list, and thus opens its collocational database for this item (an abbreviated version of this database, coded for reading by a PERL routine, is shown in Fig. 3). The program inspects the eight words surrounding *miss* in the text (four to the left, four to the right). If it finds *bare*, *boat,* or *bus,* it parses the word as the 'loss' type of miss, *miss_1*. If it finds *girl, young, pretty*, or other similar titles like *mister,* or a following word with a capital letter (miss Smith), it parses the word as *miss_2*. If there are multiple occurrences of *miss* and the program finds collocates supporting both interpretations, the majority association wins. In the event of a tie or a lack of any match, any previous parsing is repeated, following the reasoning already mentioned. In the rare event (except at the very beginning of a text) of no collocate matches and no previous parsing, then the parsing assigned is miss_0.

Figure 3. Database with collocates for two members of the homograph miss

MISS	+ miss missed unmissed misses missing
loss	| (i | you | he | she | they | we) miss | aim | bad | bare | beat | boat | bus | (can | can_1 | cannt | cannot) miss | chance | date | deadline | dreadful | fail | family | foot | forget | heart | hit | lack | lose | lot | mark | match | mile | moment | much | narrow | near | never | opportunity | plane | point | putt | race | really | target | terrible | thing | train | trick | tube | want to miss |
title	| daughter | dress | girl | hair | kiss | lady | little | marry | master | mister | mistress | mrs | niece | pretty | sister | spinster | universe | victorian | young | Miss ([^the])([A-Z][a-z]+[ ]*) | she |

In the n-gram phase of the analysis, if an instance of *at all,* for example, is found, it is tested against the non-compositional collocates for this entry (Fig. 4), and if none is found in the environment, then the individual components are returned to the analysis as single words (where *at* and *all* will both be classed 1k items). The collocational criteria for the two meanings of *at all* are shown in Fig 4. The prepositional meaning is nearly always followed by *the*; the quantity meaning of *at all* is almost always preceded by a negating term like *never*, plus optional intervening other words (like '*never* <u>*saw him*</u> *at all,* which can be picked up by the regular expression [*a-z**].

Figure 4. Distinguishing collocates for a multi-word unit

AT_ALL	+
quantity	| (no | not | nothing | none | any | never | any | anything | few) [a-z]* at_all | if [a-z]* at_all |
preposition	| at_all the |

7. How well do collocates do their work? A Mini-Experiment

7.1. Research question

Can homoforms including MWUs with a compositional and non-compositional meaning be reliably distinguished by the collocational resources currently available?

7.2. Context

It is frequently claimed that there are few true synonyms in a language owing to differences in contexts of use and especially the distinct collocations that different senses of words typically enter into (Sinclair, 1991). This claim should be even more applicable to forms which are not just synonyms but have no related meaning whatever. However, to date many examples but few proofs are offered for this claim, which therefore remains intuitive. The proof of the claim would be if the collocations that appear to distinguish the meanings of a homoform in a particular corpus could predict the same distinctions in a novel text or corpus.

7.3. Procedure

The BNC was mined for all collocations with a frequency > 10 for the first three items from Parent's selection in Table 6 (*miss, yard,* and *net*) and two selections from Martinez and Schmitt's selection in Table 3 (*a lot* and *at all*) in the manner of the information in Table 2 in the Appendix for *bank*. For each item, roughly 200 collocations, with some variability in the number, were hand sorted into those corresponding to each meaning, which in the case of *miss* was tagged as miss_1 or miss_2. The collocations were coded in the PERL scripting language to match text strings within ten words on either side of each test item, including strings with an unpredicted intervening word (*miss train* would also match *miss<u>ed</u> <u>their</u> train*). Novel contexts for the five items were obtained by searching a corpus of simplified stories for texts containing both meanings of each of the homoforms. For example, Wilde's *The Picture of Dorian Gray* (Oxford Bookworms Series; 10,500 running words; 1,000 headwords) bears three instances of *miss* with both parsings represented. All instances were extracted as concordance lines of roughly 30 words (80 characters on either side of the keyword). These concordance lines served as a greatly truncated 'text' that would test the program's ability to use context information to disambiguate the homoforms. The next step was to feed this test text into a computer program that accesses the collocational database. The program breaks a text (in this case, the set of concordance lines with homographs) into family headwords, identifies the current search term, and looks for pattern matches in its collocation set. Each time it makes a match it records the fact and awards a point to the relevant meaning.

7.4. Results

The collocational information is clearly able to distinguish the two meanings of the homoform *miss*. Figure 5 shows the Dorian Gray output for *miss,* followed by the record of the decision process.

Figure 5. "*miss*" in simplified *The Picture of Dorian Gray* - Bookworm Level 4

Parsed concordance

034. omething to say to you.' That would be lovely. But wont you MISS_1 your train?' said Dorian Gray, as he went up the step

035. , You look like a prince. I must call you Prince Charming.' MISS_2 Sibyl knows how to flatter you.' You dont understand

036. g, Harry. I apologize to you both.' My dear Dorian, perhaps MISS_2 Vane is ill,' said Hallward. We will come some other

Program's reasoning

34. 2 0 miss_1
to you' That would be love But wont you MISS you train' say DORIAN Gray as he go up
— miss 'you MISS'
— miss 'train'

35. 0 1 miss_2
like a prince I must call you Prince Charming' MISS Sibyl know how to FLATTER you' You dont understand
— miss 'MISS Sibyl' (CAP)

36. 0 1 miss_2
I apology to you both' My dear Dorian perhaps MISS Vane be ill' SAY Hallward We will come some
— miss 'MISS Vane' (CAP)

The program's reasoning as shown in the output is thus: Before starting, the algorithm reduces all words to familized headwords (e.g., *go* not *went* in line 34). To parse the instance at concordance line 34, a pronoun subject (*I|you|he*, etc) before the keyword, and the presence of the high frequency collocate *train* anywhere in the string, give a score of 2-0 for miss_1 (loss). The challenge point in this and the many other runs of this experiment is where the meaning of the homoform changes. This happens in line 35, where there is no match suggesting miss_1 (loss), and one piece of evidence for miss_2 (title), namely *miss* followed by a word with a capital letter, giving a score of 0-1 and a verdict of miss_2. In line 36, a capital letter is once again the decider, now backed up by the coherent information assumption. A score of 0-0 would have led to a continuation of the previous parsing and that would have been correct.

Similarly, the Bookworms version of Conan Doyle's *Tales of Mystery and Imagination* was found to bear both meanings of *at all,* and once again the collocations were able to distinguish these (Fig. 6), largely through discovering various quantifiers like *few, none, any* and *if* for the non-compositionals and a following *the* for the compositional (these are underlined in the concordance output for emphasis).

Figure 6. "*at all*" in simplified *Tales of Mystery & Imagination* – Bookworm Level 3

020.	sons of the richest families of England. There was <u>nothing</u> **at_all_1** to stop me now. I spent my money wildly, and passed
021.	n and the strange fears I had felt. <u>If</u> I thought about them **at_all_1**, I used to laugh at myself. My life at Eton lasted f
022.	htening, and <u>few</u> people were brave enough to enter the room **at_all_1**. In this room, against the farthest wall, stood a hu
023.	nd held it there for many minutes. There was <u>no</u> life in him **at_all_1**. Now his eye would not trouble me again. Perhaps you
024.	lantern was closed_2, and so no light came out of it, <u>none</u> **at_all_1**. Then slowly, very slowly, I put my head inside the
025.	d it. I started walking around the streets at night looking **at_all_2** <u>the</u> cats, to see if I can_1 find another one like Pl

In the five test cases, all significantly longer than the ones shown here, the collocation database was able to correctly identify the relevant meaning of the single word or multiword homoform in at least 95% of cases. Accuracy can be increased by expanding the size of the database (Fig. 4 is far from an exhaustive list of at all the collocates Web-BNC offers for *at all*), but at the expense of slowing the program down and making it less useful for practitioners.

7.5. Discussion

There is thus evidence that collocations can indeed simulate the function of human judgment in this task and hence that the full database of collocates for the high frequency homoforms and MWUs is worth building.

Further, it should be noted that the task set to the computer program in the mini-experiment is unrealistically difficult. As mentioned, few natural/normal/real texts contain both meanings of a homoform in as close proximity as in the special texts used here to test the program, which were chosen precisely for the presence of both meanings of the homoform. In a natural text, one meaning is normally established and then the algorithm's default procedure ("use previous") almost invariably leads to a correct assignment – and the success rate over the many trials performed by the author is more like 98%.

8. Conclusion

The pieces of Frequency 2.0 are at hand and, although hailing from quite disparate quarters, merely require assembly. The most frequent and most pedagogically relevant homoforms have been identified, separated, and assigned initial frequency ratings, and a methodology is in place to move the analysis down the scale to the vast number of homoform items in English where the minor member represents fewer than 5% of occurrences. Refinements there will certainly be, and the question of what makes an MWU non-compositional will need further thinking, but the methodology is likely to be something similar to the one proposed here. Further, while the first round of this work had to be accomplished by humans, prizing apart the *banks* and *at all's* by inspecting samplings of concordance lines, for subsequent rounds a means is available to automate this task using a computer program in conjunction with a collocational database such that sampling should not be necessary: within a year or two, the collocational database should be completed for both the Parent and Martinez items, or principled sub-sets thereof, and it should be possible to assemble the pieces and create a complete set of trial lists, incorporating both types of homoforms, as hypothesized in Table 2.

When that happens, an important task will be to establish new cut-offs – that is, new frequency counts. The alert reader will have noticed that in several of the analyses above, the original word-form cut-offs were used for proposed new frequency assignments, whereas in fact, every re-assignment will shift all the cut-offs. For example, if the first thousand list is defined as every BNC lemma represented by more than 12,369 occurrences (Table 5), and the non-compositional meaning of *a lot* is found to have more occurrences than this, then it should be included as a first thousand item – and the current last item will be bumped to the second thousand list.

Also on the to-do list will be to establish a coding format for the different meanings of homographs (*bank_1* and *bank_2,* or *bank_money* and *bank_river*? and *at_all* for non-compositional MWUs but plain *at* and *all* for compositional?); to settle on the exact list of MWUs to include; to settle on the percentage of main-meaning occurrences (90% or 95%) that makes handling separate meanings worth program time; and to decide whether to limit the single word analysis to the first five thousand-word families or to proceed further. Benefits to be realized will be more accurate Vocabprofiling (extent to be determined), greater credibility for this methodology within the scientific community, and more effective language instruction.

References

Aston, G., & Burnard, L. (1998). *The BNC handbook: exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press.

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

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

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research, 24,* 57-65.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Harlow, UK: Pearson Education.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*(1), 181-200.

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

Davies, M., & Gardner, D. (2010). *Frequency dictionary of contemporary American English: Word sketches, collocates, and thematic lists.* New York: Routledge.

Davies, M. (2011). Word frequency data from the Corpus of Contemporary American English (COCA). [Downloaded from http://www.wordfrequency.info on 2012-07-02.]

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*(02), 143-188.

Ellis, N. C., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning, 59*, 90-125.

Grant, L., & Nation, P. (2006). How many idioms are there in English? *International Journal of Applied Linguistics, 151,* 1-14.

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

Hoey, M. (2005). *Lexical priming: A new theory of words and language.* London: Taylor and Francis.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System, 14*(2), 151-162.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

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

Lindqvist, C. (2010). La richesse lexicale dans la production orale de l'apprenant avancé de français. *La Revue canadienne des langues vivantes, 66*(3), 393-420.

Martinez, R. (2009). *The development of a corpus-informed list of formulaic expressions and its applications to language assessment and test validity.* PhD thesis, University of Nottingham.

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

MacWhinney, B. (1989). Competition and lexical categorization. In R. Corrigan, F. Eckman, & M. Noonan (Eds.), *Linguistic categorization* (pp. 195-242). Amsterdam: Benjamins.

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of TESL student performance. *System, 32*(1), 75-87.

Nation, P. (2001). *Learning vocabulary in another language*. London: Cambridge.

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

Nation, P. (Unpublished). *The frequency ordered 1,000 word family lists based on the British National Corpus.*

Ovtcharov, V., Cobb, T., & Halter, R. (2006). La richesse lexicale des productions orales: mesure fiable du niveau de compétence langagière. *Revue Canadienne des Langues Vivantes, 63*(1), 107-125.

Oxford Bookworms Library. London: Oxford University Press.

Palmer, H. E. (1941). *A grammar of English words: One thousand English words and their pronunciation, together with information concerning the several meanings of each word, its inflections and derivatives, and the collocations and phrases into which it enters.* London: Longman.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *Modern Language Journal, 95*(1), 26-43.

Sharp, Europe. Just The Word: Collocational Database. [Website http://www.just-the-word.com/, accessed 20 November 2011.]

Shin, D., & Nation, P. (2007). Beyond single words: The most frequent collocations in Spoken English. *ELT Journal, 62*(4), 339-348.

Sinclair, J. (1991). *Corpus, concordance, collocation.* London: Oxford University Press.

Stubbs, M. (2009). Technology and phraseology: With notes on the history of corpus linguistics. In U. Romer & R. Schulze (Eds.), *Exploring the lexis-grammar interface* (pp. 15-32). Amsterdam: Benjamins.

van Zeeland, H. and Schmitt, N. (in press). Lexical coverage and L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics.*

Wang, K., & Nation, P. (2004). Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics, 25,* 291-314.

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

APPENDIX

Table 1. Full list of Parent's GSL homoforms

air	coast	faint	late	page	rest	step
arm	company	fall	lay	pan	right	stick
article	concentrate	fast	lead	park	ring	still
ball	contract	fence	leave	passage	rock	stir
band	count	figure	left	patient	roll	stone
bank	country	file	letter	pen	row	story
bar	course	fine	lie	pick	scale	strike
bear	court	fire	light	plant	school	swallow
bear	cross	firm	like	pole	season	table
belt	crush	flat	line	policy	second	tend
bill	cry	fold	live	pool	sense	train
bit	culture	foot	lock	port	sentence	trip
boil	cure	formal	lot	pot	set	type
boot	curl	forward	love	pound	shoot	wake
bowl	current	game	match	present	shoot	watch
box	date	general	mean	present	shower	weave
bridge	deal	go	metre	press	slip	well
brush	degree	habit	might	pretty	sock	whip
camp	die	hand	minute	private	sound	wind
can	down	hide	miss	produce	spell	wound
case	drag	host	mouse	pupil	spirit	yard
cell	draw	how	nature	race	spring	
charge	drive	just	net	rail	staff	
chest	duty	kind	nut	rank	stage	
close	ear	knot	order	realize	state	
club	egg	last	organ	repair	steep	
	even					

Table 2. Collocates for two banks, from Just-The-Word database, frequency >10, span=5 word-forms either side, hand-sorted into independent meanings

Money banks

world bank	714	development bank	86	director of bank	51
central bank	690	bank on	84	bank announce	50
bank account	422	bank balance	78	bank credit	50
bank holiday	409	swiss bank	76	bank provide	49
bank manager	298	bank rate	74	private bank	49
national bank	272	major bank	73	money in bank	49
commercial bank	226	bank lend	71	clearing bank	48
european bank	215	state bank	67	international bank	48
merchant bank	201	bank clerk	64	president of bank	48
royal bank	191	bank and company	62	bank offer	47
bank loan	189	British bank	61	bank statement	47
investment bank	165	american bank	57	french bank	45
between bank	142	bank and institution	57	bank official	45
go to bank	117	borrow from bank	55	leave bank	44
midland bank	113	include bank	55	german bank	43
big bank	104	branch of bank	55	reserve bank	43
governor of bank	97	bank or building society	55	clearing bank	40
bank deposit	95	bank hold	53	creditor bank	40
foreign bank	91	bank note	53	bank strip	40
bank and building society	90	japanese bank	52	bank lending	39
large bank	87	data bank	51	bank agree	38

>>>

>>>

bank pay	38
chairman of bank	38
work in bank	37
join bank	37
bank buy	37
leading bank	37
bank governor	37
break bank	36
bank lending	36
overseas bank	35
bank charge	35
bank debt	35
allow bank	34
have in bank	33
rob bank	33
issue by bank	33
bank issue	33
bank sell	32
bank able	32
land bank	32
bank branch	32
loan from bank	32
way to bank	32
northern bank	31
be bank	30
bottle bank	30
street bank	30
bank robbery	30
bank base rate	30
memory bank	29
put in bank	28
bank cut	28
bank staff	28
manager of bank	28
force bank	26
provide by bank	26
Independent bank	26
bank report	26
pay into bank	25
street bank	25
union bank	25
bank robber	25
account at bank	25
customer of bank	25
fund and bank	25
bank and fund	25
regional bank	24
bank act	22
bank refuse	22
bank seek	22
irish bank	22
issuing bank	22
bank interest	22
head of bank	22
group of bank	22
Western bank	21
role of bank	21
clear bank	20
enable bank	20
close bank	20
bank operate	20
bank raid	20
line bank	19
sponsor by bank	19
bank charge	19
bank require	19
trust bank	19
bank borrowing	19
bank corporation	19
bank vault	19
subsidiary of bank	19
establishment of bank	19
take to bank	18
bank create	18
asian bank	18
account with bank	18
Government and bank	18
eastern bank	17
piggy bank	17
state-owned bank	17
city bank	17
bank card	17
debt to bank	17
oblige bank	16
approach bank	16
bank publish	16
bank deal	16
bank overdraft	16
agreement with bank	16
name of bank	16
available from bank	16
bank and house	16
bank up	16
own by bank	15
work for bank	15
persuade bank	15
bank president	15
bank show	15
accept by bank	14
deposit in bank	14
make by bank	14
set up bank	14
offer by bank	14
owe to bank	14
shanghai bank	14
write to bank	14
bank step	14
retail bank	14
jeff bank	14
bank employee	14
bank finance	14
bank funding	14
bank customer	14
bank estimate	14
consortium of bank	14
building society and bank	14
bank and government	14
receive from bank	13
draw on bank	13
sell to bank	13
co-op bank	13
deposit with bank	13
bank to bank	13
get in bank	12
hold by bank	12
pay to bank	12
take by bank	12
bank assistant	12
bank guarantee	12
bank creditor	12
Balance at bank	12
currency and bank	12
Building society or bank	12
bank and credit	12
bank or company	12
deposit with bank	11
bank grant	11
bank intervene	11
failed bank	11
gene bank	11
bank post	11
bank operating	11
bank interest rate	11
chair of bank	11
money from bank	11
company and bank	11

River banks

west bank	240	steep bank	45	left bank	28
river bank	210	opposite bank	42	east bank	27
along bank	194	west bank	42	left bank	26
south bank	166	top of bank	42	stand on bank	15
far bank	94	grassy bank	41	occupied bank	14
its banks	85	north bank	41	shingle bank	12
down bank	73	sit on bank	30	situate on bank	11
up bank	53	swain bank	30	walk along bank	11
south bank	48	burst bank	28		

Table 3. Collocates for *at all* (57 idiomatic or non-compositional, 11 compositional) selected from the BNCWeb's most frequent and most connected 100 (by log-likelihood of co-occurrence) as the basis for database entry (Fig. 6)

Non-Compositional

(anything) at all wrong	(no) interest at all	at all — (phrase end)
(didn't) notice at all	(no) problem at all	at all' (phrase end)
(didn't) seem at all	(no) reason at all	at all possible
(didn't) sleep at all	(no) sense at all	at all! (sentence end)
(doesn't) bother (me) at all	(no) sound at all	at all. (sentence end)
(doesn't) exist at all	(no) trouble at all	at all? (sentence end)
(doesn't) look at all	(not) aimed at all	did (not) at all
(don't care) at all about	(not) at all actually	hardly at all
(don't care) at all except	(not) at all clear	if at all
(don't care) at all really	(not) at all easy	mention at all
(don't see it) at all	(not) at all sure	never (did it) at all
(don't) like at all	(not) at all surprised	no … at all
(don't) mind at all	(not) changed at all	nobody at all
(don't) remember at all	(not) doubt (it) at all	none at all
(don't) see at all	(not) pleased at all	not at all
(no) good at all	(not) worried at all	nothing at all
(no) harm at all	any at all	n't … at all
(no) help at all	anything at all	scarcely at all
(no) idea at all	anywhere at all	without (any) at all

Compositional

avoided at all (costs)	at all sites	at all events
avoid at all (costs)	at All Saints	at all costs
at all times	at all levels	at all ages
at all stages	at all hours	