In Bogaards, P. and Laufer, B.
Vocabulary in a Second Language
John Benjamins, Amsterdam
(2004)

CHAPTER 1

A study of the most frequent word families in the British National Corpus

Paul Nation
Victoria University of Wellington

Abstract

This study compares the General Service List (West 1953) and the Academic Word List (Coxhead 2000) with three 1000 word lists from the British National Corpus. Even though these two sets of lists were developed from quite different corpora and at widely different times, overall they contain much the same vocabulary. This vocabulary however is not distributed in the same way in each set of lists, with the AWL words occurring across the three BNC lists. The BNC lists provided slightly better coverage of a variety of texts and corpora. The BNC lists reflected the adult, British, formal nature of the BNC.

1. Word lists

Making word lists in the field of L2 learning and teaching is usually done for the purpose of designing syllabuses and in particular it is an attempt to find one way of determining necessities (what needs to be learned) as a part of needs analysis. In any needs analysis it is important to decide *whose* needs are being investigated, and then to ensure that the investigation draws on data that is relevant to the people whose needs are being investigated (Nation 2001).

This paper looks at high frequency word lists developed from a very recent analysis of the British National Corpus and shows that it is not appropriate to use these lists unchanged as the basis for syllabus design for learners of English as a second or foreign language in primary or secondary school systems. The reason is that the British National Corpus (BNC) is predominantly a corpus of British, adult, formal, informative language, and most English learners in

primary and secondary school systems are not British, are children, and need both formal and informal language for both social and informative purposes. That is, if the BNC lists were used as a basis for school curriculum design, there would be a mismatch between the nature and goals of the learners, and the nature of the corpus that the lists are drawn from.

This paper will estimate the size and nature of the mismatch and then provide evidence for the mismatch. The procedure used to do these two things involves comparing the high frequency word lists from the BNC with the General Service List and the Academic Word List (GSL+AWL). Let us start by looking at how the BNC high frequency lists were made.

The British National Corpus consists of 100,000,000 running words of English with 10% of the total running words drawn from spoken sources and 90% from written sources (see Figure 1). Leech, Rayson and Wilson (2001) rearranged the corpus into 100 one million running word sub-divisions keeping similar texts together in each sub-division. They then created a list of lemmas occurring 1000 times or more in the corpus. A lemma consists of a headword and its inflected forms where the headword and its inflected forms are all the same part of speech. For example, *diminish, diminished, diminishes, diminishing*. For each lemma they provided frequency data (how often the lemma and each of its members occurred in the 100,000,000 word corpus), range data (how many of the 100 subdivisions the lemma and each of its members occurred in), and dispersion data (how evenly the word occurred across the 100 subdivisions, that is, how similar the frequencies are across the subdivisions of the corpus). If the frequencies were very similar in the different subdivisions, the dispersion figure is close to 1, like 0.89. If they are very different, the dispersion figure is much less than 1. This list is available in written form in a book (Leech *et.*) and in electronic form http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html. Table 1 presents two sample entries from the electronic list.

Table 1. Two samples of entries from the electronic list.

Headword	Part of Speech	Members	Freq	Range	Dispersion
assault	NoC	%	26	98	0.89
@	@	assault	22	98	0.89
@	@	assaults	4	84	0.87
assemble	Verb	%	17	99	0.94
@	@	assemble	4	96	0.92
@	@	assembled	10	97	0.94
@	@	assembles	0	33	0.83
@	@	assembling	2	83	0.92

The first line of each entry is the headword with the total figures for the lemma. The first column gives the headword, the second the part of speech, and the members of the lemma are in the third column. % indicates that this is the headword of the family. @ indicates family members. Note that the headword occurs twice, once representing the whole lemma and once as a member of the lemma. The second column gives the part of speech of the lemma. The fourth column gives the frequency. Note that for the sake of saving space the frequency is given out of 1,000,000 not 100,000,000, although the count is based on 100,000,000 running words. The fifth column gives the range of occurrence with the highest possible being 100. So *assaults* occurs in 84 of the 100 subdivisions of the corpus. The sixth column gives the dispersion with the highest possible being 100 but in practice 99. *assaults* has a dispersion of .87 which is reasonably high. Dispersion is calculated by a formula involving range and frequency in the one hundred subdivisions of the corpus. If the list is downloaded from the web site and put in a word processing programme, it can be sorted on any of the columns.

2. Making the BNC high frequency lists

The first 1000 word list of the British National Corpus was made by taking the just over 6500 entries in the rank list of lemmas with a frequency of 10,000 or higher for the whole 100,000,000 running word corpus from the web site and sorting these by range and removing all lemmas with a range of less than 98 out of the 100 one million word sub-corpora. Then the remaining list was sorted by dispersion, and all lemmas with a dispersion of less than 80 were removed. The list was then sorted by frequency. The first 1000 families were made starting with the items at the top of the list. That is, the first 1000 lemmas were expanded into families. A full list of days of the week, months, numbers, and letters of the alphabet were included even though several of these did not meet the frequency, range, or dispersion criteria. The items *goodbye*, *OK*, and *Oh* were also included even though they did not meet the criteria. The frequency of the items was from 89 per million up (from *ball*).

The second 1000 list was constructed in the same way using what was left of the 6500 lemmas after the first 1000 word families had been made. There were no additions ignoring the criteria. The frequency of the second 1000 was from 27 (*request*) up to 89 (*message*), with a range of 97 up, and a dispersion of 80 up.

The third 1000 list contained words with a frequency of 10 up, a range of 95 up and a dispersion of 80 up. Five word families, which were very frequent in the spoken part of the corpus but which did not meet the range and dispersion criteria, were added to this list. These items were

(1) hesitation procedure (*er, erm, mm, mhm*)
(2) interjections (*ooh, aye, eh, aha, ha*)
(3) *alright*
(4) *pardon*
(5) *fuck.*

These were included in the third 1000 rather than the first because they were low frequency and narrow range in the corpus as a whole

A high range minimum was chosen to make sure that the words were of wide range (general service) and to ensure they occurred in both speech and writing. 10 of the 100 sub-sections of the corpus were spoken English, so a range of 95 ensured that at least 5 of those 95 sub-sections were spoken.

The three word lists are of families not lemmas. Word families include both closely related inflected and derived forms even if the part of speech is not the same. Here are some examples.

ADD
ADDED
ADDING
ADDITION
ADDITIONAL
ADDITIVE
ADDITIONS
ADDS
ADMIT
ADMISSION
ADMITTEDLY
ADMITS
ADMITTED
ADMITTING
ADVANTAGE
ADVANTAGES
DISADVANTAGE
DISADVANTAGES
ADVANTAGING
ADVANTAGED
DISADVANTAGED

In the following discussion, BNC 2000 consists of 2000 word families. The BNC 2nd 1000 consists of the second set of 1000 word families within the BNC 2000. Similarly, BNC 3000 contains 3000 families and the BNC 3rd 1000 contains the third set of 1000 families.

3. The GSL and the AWL

The General Service List (West 1953) is a list of around 2000 headwords (families) largely but not completely chosen on the basis of frequency. The frequency data used in the GSL came from the Thorndike and Lorge counts carried out in the early twentieth century. Frequency was not the only criterion used in making the GSL, but it was the most important. The original GSL did not list numbers, days of the week and months of the year, but in the study described in this paper they were added to the list. When the definition of a word family using Bauer &Nation (1993) level 6 is used, the GSL contains 1,986 word families — a little less than 2000. The GSL has been used as the basis for the early graded reader schemes.

The Academic Word List (Coxhead 2000) was made by looking at the frequency and range of words across the university divisions of Humanities, Science, Commerce and Law. It contains 570 word families that are not in the GSL and that are frequent and of wide range in a wide variety of academic texts. The AWL contains important vocabulary for learners in senior high school and university.

4. Does the BNC 3000 provide better coverage than the GSL plus AWL?

Coverage refers to the percentage of tokens in a text which are accounted for (covered by) particular word lists. The corpora used in the comparison are

(1) a 3,500,000 token written academic corpus with a balance of texts from Science, Arts, Law and Commerce (Coxhead 2000)
(2) a 300,000 token economics text written by one author — M. Parkin *Macroeconomics* (Addison-Wesley, Mass. 1990).
(3) the 500,000 token Lund corpus of spoken English (Svartvik & Quirk 1980)

(4) a 3,500,000 word fiction corpus of texts from Project Guthenburg (Coxhead 2000).

These corpora include written, spoken, academic and fiction texts. In the comparison, it must be remembered that the BNC 3000 contains 444 more word families than the GSL plus AWL, and so should have better coverage because of this.

In Table 2 we can see that the 1st 1000 of the GSL covers 70.9% of the 3,500,000 token academic corpus, the 2nd 1000 words another 4.6% totalling 75.5% with the 1st 1000, and adding the AWL results in a total coverage of 85.5%. In other words 14.5% of the 3,500,000 tokens in the academic corpus are not covered by the GSL plus AWL. The BNC lists provide 1% better coverage.

The BNC provides slightly better coverage of all the corpora. If the coverage by the BNC 3rd 1000 is reduced by 33% to account for the 444 extra words it contains compared to the GSL plus AWL, the BNC extra coverage is less than 1% or in the case of the Academic corpus the advantage goes to the GSL plus AWL. The GSL provides slightly better coverage of the fiction corpus than the BNC 2000.

The BNC 3000 does not provide strikingly better coverage than the GSL plus AWL. The range as shown in the Difference row in Table 2 is from 0.9% to 2.0% with most around 1%.

The BNC 2000 provides much better (7.3% better) coverage of written formal text than the GSL alone. This is probably because the most frequent AWL words are in the BNC 2000 (63% of AWL is in the BNC 2000). Seventy percent of the BNC consists of informative text (see Figure 1) which is the type of text where the AWL is most frequent.

Table 2. Cumulative percentage coverage of a range of corpora by the lists from the BNC and GSL plus AWL.

Corpus	Academic		Parkin		LUND		Fiction	
Levels	GSL+	BNC	GSL+	BNC	GSL+	BNC	GSL+	BNC
1000	70.9	75.5	77.7	80.8	85.6	86.5	81.7	79.8
2000	75.5	83.9	82.5	89.8	89.6	91.1	87.1	86.6
AWL/ BNC 3000	85.5	86.5	91.2	93.2	91.4	92.6	88.5	89.6
Difference		1.0		2.0		1.2		0.9

5. Do most of the words in the lists occur in a range of texts?

Table 3 is based on the same texts as Table 2 but looks to see if *all* the words in the lists are working. That is, does every word family in the lists occur in the various corpora? There could be words in the lists which seem useful but do not occur. For example, the word *chimney* is in the GSL but did not occur at all in the Academic Corpus. In Table 3 we can see that every word family in the BNC 1st 1000 and 2nd 1000 occurred in the Academic Corpus, and 99.2% of the words in the 3rd 1000 of the BNC occurred in the Academic Corpus. In other words, only 8 words did not occur. The GSL plus AWL consists of 2,556 word families. Only 10 word families (0.4%) did not occur in the Academic Corpus. Table 3 shows that the BNC lists are fractionally better than the GSL+AWL but the difference is very small, half a per cent or less which means that less than fifteen out of 3000 word families are affected in each comparison.

6. Does the BNC 3000 contain most of the GSL plus AWL?

The GSL is an old list and the AWL is one with a narrow focus. In spite of this, virtually all the GSL 1st 1000 is in the BNC 3000 (except four words: *hurrah, ounce, scarce, shave*). Most (97%) of the GSL 1st 1000 is in the BNC 2000. At the slightly lower frequency levels, 80% of the GSL 2nd 1000 is in the BNC 3000, and 80% of the AWL is in the BNC 3000. However,107 out of 570 word families are not (18.7%). In total, 88% of the GSL plus the AWL is in the BNC 3000. Only 12% (301 word families out of 2556) is not. Thus, though the GSL was compiled long before the BNC, when supplemented by AWL, most of it can be found in the BNC 3000.

Table 3. Percentage of word families in the lists occurring in various corpora

Corpus	Academic		Parkin		LUND		Fiction	
Lists	GSL+	BNC	GSL+	BNC	GSL+	BNC	GSL+	BNC
1000	99.9	100	94.9	96.6	99.7	100	100	99.5
2000	98.9	100	62.7	85.3	94.9	98.9	99	99.4
AWL/ 3000	100	99.2	93.2	56.1	94.6	91.7	94.7	95.9
Average	99.6	99.7	83.6	79.3	96.4	96.9	97.9	98.2
Difference		0.1	4.3			0.5		0.3

Table 4. Spread of the 10 sublists of the AWL across the BNC

BNC	1	2	3	4	5	6	7	8	9	10	Total (%)
1st 1000	48	29	13	11	9	2	2	1	4	0	119 (21)
2nd 1000	9	28	36	38	33	31	30	21	9	6	239 (42)
3rd 1000	1	1	4	7	9	10	13	17	26	15	105 (18)
Not in BNC	2	2	7	4	9	17	15	21	21	9	107 (19)

7. What happens to the AWL?

The AWL is divided into 10 sub lists (9 with 60 word families, sub list 10 with 30 word families). Sub list 1 contains the 60 most frequent, widest range words, sub list 2 the next 60 and so on. Table 4 shows for example that for sub list 1 of the AWL, 48 of the 60 word families are in the 1st 1000 of the BNC, 9 AWL sub list 1 word families are in the 2nd 1000 of the BNC, 1 is in the BNC 3rd 1000 and only 2 word families in AWL sub list 1 are not in the BNC 3000.

Most of the AWL (81.3%) is in the BNC 3000, and 63% is in BNC 2000. This boosts the BNC 2000 coverage of formal text. Note the bold numbers in the sub lists, showing that many of the word families in the higher AWL sub lists tend to be in the BNC 1st 1000 and 2nd 1000, while many of the word families in the lower AWL sub lists tend to be in the BNC 2nd or 3rd 1000 or not in the BNC. In the BNC data the AWL does not stand out as a separate list but is spread across the BNC lists. This is a result of the nature of the BNC. We will look at this as it is reflected in the vocabulary in the corpus and in the composition of the BNC.

8. The nature and composition of the BNC

The following twenty words are all the words in the BNC 1st 1000 which are not in the GSL or AWL.

> American, announce, appeal, British, budget, campaign, career, client, county, drug, Europe, executive, French, German, okay, Parliament, reference, Scottish, species, television.

Table 5 classifies some of these twenty words and adds example words from the GSL which are not in the BNC 3000. In Table 5 in the row *Young learners vs adults*, *chalk*, *aunt* and *wicked* are considered as words more likely to be useful for younger learners. *Budget*, *campaign*, *client* and *executive*, are considered to

Table 5. Possible reasons for non-overlapping words in the GSL and BNC

Factors	In GSL, not in BNC	Not in GSL, in BNC
Old vs modern	shilling	television, drug
US vs British	republic, gallon, quart	county, Parliament
Young learners vs adults	chalk, aunt, wicked	budget, campaign, client, executive
Proper nouns	–	American, British, Europe, French, German, Scottish

be words more likely to be useful for adult learners. It is likely that West included *chalk* in the GSL not because of its frequency but because of its usefulness in the classroom.

Figure 1 tries to show the proportional make-up of the British National Corpus. The conversation (4%) and imaginative (20%) parts are largely informal text. The remainder is largely formal, informative text (spoken 6% plus written 70%). In order to get some idea of the size of the BNC, 100,000,000 running words has been estimated as being equivalent to approximately 10 years quantity of a person's language experience (Aston & Burnard 1998: 28). The BNC consists largely of informative text

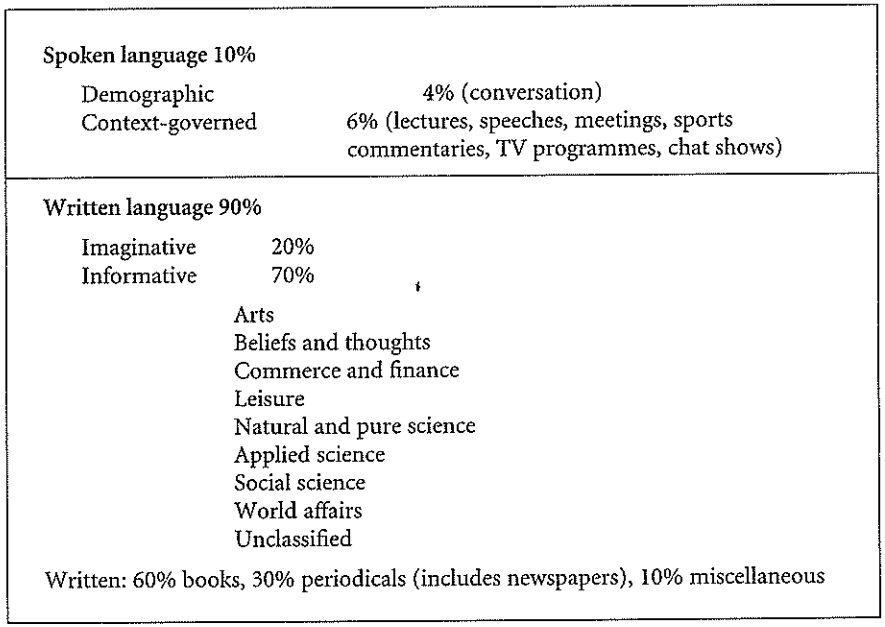

Figure 1. Composition of the British National Corpus

9. Conclusion

The major difference between the BNC 3000 and the GSL+AWL does not lie in the performance of the lists in terms of coverage, but in the way the vocabulary is divided between the three one thousand levels of the BNC, and the two one thousand levels of the GSL and the AWL. The BNC 1st 2000 contains many words from the AWL, whereas in the GSL+AWL, the GSL is largely non-academic and the AWL is wholly academic.

Learners of English as a second language in primary and secondary school systems may be better off using materials based on a replacement for the GSL, with the AWL getting attention at senior high school and university levels. Beginning learners at tertiary level would be better off using materials based on the BNC lists, because of the slightly better BNC coverage.

It is not easy to decide how the GSL could be replaced. As Coxhead (2000) showed, some words could go from the AWL to the GSL, for example *job, sex, percent, area, final.* Some important proper nouns describing countries, people, and languages could be added. A corpus needs to be devised to represent the learning goals of young L2 learners. That is, the corpus would need to contain written and spoken texts that more closely match the uses they would make of their English. At least half of the corpus should be spoken language. Should it be the spoken language of advanced L2 learners or should it be the spoken language of young native speakers? An obvious candidate for the written part of the corpus would be graded readers (also included in the BNC), but many of the graded reading schemes are probably based on the GSL so this could be backward-looking rather than forward-looking. It should probably include school texts as in the Carroll, Davies & Richman count (1971), as this would make the list useful for ESL learners in an English medium school system. Almost one third, 236, out of the top 900 word families in the Carroll et al list do not occur in the BNC 3000. These include *adjective, alphabet, ant, arithmetic, astronaut, aunt, axis,* which are words that may be more immediately useful for school children. It could also include books written for young native speakers. A large number of words (681) in the BNC 3000 are not in the top 5300 types of the New Zealand School Journal corpus. Clearly there is a substantial number of words in texts written for young native speakers which are not in the BNC 3000. Perhaps the language of chat rooms and e-mails should be part of a corpus for making a new GSL.

The main motivation to replace the General Service List is because of its age. There are clearly a few words like *computer, drugs, television,* which are not in it and should be. Perhaps we should respect age rather than see it as an excuse for retirement. However, whatever choices are made, the choices need to represent credible language goals for young learners of English.

References

Aston, L. and Burnard, G. 1998. *The BNC Handbook.* Edinburgh: Edinburgh University Press.

Bauer, L. and Nation, I. S. P. 1993. "Word families". *International Journal of Lexicography* 6: 253–279.

Carroll, J. B., Davies, P. and Richman, B. 1971. *The American Heritage Word Frequency Book.* New York: Houghton Mifflin, Boston American Heritage.

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

Leech, G., Rayson, P. and Wilson, A. 2001. *Word Frequencies in Written and Spoken English.* Harlow: Longman. http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html.

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

Svartvik, J. and Quirk, R. (eds) (1980) *A Corpus of English Conversation.* C. W. K. Gleerup, Lund.

West, M. 1953. *A General Service List of English Words.* London: Longman, Green & Co.