How Large a Vocabulary Is Needed For Reading and Listening?

I.S.P. Nation

Abstract: This article has two goals: to report on the trialling of fourteen 1,000 word-family lists made from the British National Corpus, and to use these lists to see what vocabulary size is needed for unassisted comprehension of written and spoken English. The trialling showed that the lists were properly sequenced and there were no glaring omissions from the lists. If 98% coverage of a text is needed for unassisted comprehension, then a 8,000 to 9,000 word-family vocabulary is needed for comprehension of written text and a vocabulary of 6,000 to 7,000 for spoken text.

Résumé : L'article a pour objectif de parler des essais menés sur quatorze listes de 1 000 familles de mots tirées du *British National Corpus* et de l'emploi de ces listes pour évaluer la taille du vocabulaire nécessaire afin de comprendre sans aide l'anglais oral et écrit. Les essais ont révélé que les listes sont adéquatement triées et ne contiennent aucune omission manifeste. Si on doit connaître 98 % des mots d'un texte pour le comprendre sans aide, il faut un vocabulaire de 8 000 à 9 000 familles de mots pour comprendre un texte écrit et un vocabulaire de 6 000 à 7 000 mots pour un texte oral.

How much vocabulary?

This article sets out to see how large a receptive vocabulary is needed for typical language use like reading a novel, reading a newspaper, watching a movie, and taking part in a conversation.

There are several ways of deciding how many words a learner of English as a second or foreign language needs to know to read without external support. The most ambitious is to try to work out how many words there are in English and to see that as a learning goal. Studies that have tried to do this have come up with figures of 114,000 word-families (Goulden, Nation, & Read, 1990) and 88,500 (Nagy & Anderson, 1984).

63,1 (September/septembre), 59–82

Putting methodological issues aside, the two major objections to this approach are that native speakers do not know all of the words in their first language, and these figures are too large to be sensible learning goals for second language (L2) learners.

A second way of deciding vocabulary learning goals is to look at what a native speaker knows and to see that as the goal. There is a long history of research in this area, but the majority of it is methodologically faulty (Nation, 1993), leading to wildly inflated figures. Reasonably conservative estimates from studies that have attempted to use a sound methodology (Goulden, Nation, & Read, 1990; Zechmeister, Chronis, Cull, D'Anna, & Healy, 1995) indicate that well-educated native speakers know around 20,000 word-families (excluding proper names and transparently derived forms). As a rule of thumb, one year of life equals 1,000 word-families up to the age of 20 or so. There is a lack of well-conducted research in this area. Once again these figures are very ambitious goals for a learning program. Recent unpublished research by the author trialling a test of vocabulary size with highly educated non-native speakers of English who are studying advanced degrees through the medium of English indicate that their receptive English vocabulary size is around 8,000 to 9,000 word-families.

A third way of deciding vocabulary learning goals is to find how much vocabulary you need to know in order to make certain uses of English like read a newspaper, read a novel, watch a movie, or take part in a conversation. Hirsh and Nation (1992), for example, tried to find out how many words you would need to know to read a novel written for teenagers who were native speakers of English. Such novels were chosen because they were considered likely to be among the most accessible texts for native speakers. Hirsh and Nation's estimate was that a vocabulary of around 5,000 words would be needed. In addition to this kind of research, researchers have developed or suggested the development of specialized vocabulary lists (Coxhead, 2000; Ward, 1999) to make certain kinds of texts more accessible. A weakness of the Hirsh and Nation study was that the vocabulary lists that were available at the time were limited to the first 2,000 words of English (West, 1953) and the University Word List (Xue & Nation, 1984). The old Thorndike and Lorge (1944) list had to be used to estimate beyond the first 2,000 word-families. The present study hopes to overcome this difficulty by using lemma lists from the British National Corpus to develop a substantial number of word-family lists that will provide more accurate estimates of the number of word-families needed to read and listen to English intended for native speakers.

Text coverage and comprehension

An important issue in studies of how much vocabulary is needed to read a text or listen to a movie is what amount of text coverage is needed for adequate comprehension to be likely to occur. Putting it another way, how much unknown vocabulary can be tolerated in a text before it interferes with comprehension?

Hu and Nation (2000) examined the relationship between text coverage and reading comprehension for non-native speakers of English with a fiction text. Text coverage refers to the percentage of running words in the text known by the readers. This figure was determined by replacing various proportions of low-frequency words in the text with nonsense words to ensure they were unknown. Reading comprehension was measured in two ways: by a multiple-choice reading comprehension test, and by a written cued recall of the text. These measures were trialled with native speakers before they were used in the study with non-native speakers. With a text coverage of 80% (that is, 20 out of every 100 words [1 in 5] were nonsense words), no one gained adequate comprehension. With a text coverage of 90%, a small minority gained adequate comprehension. With a text coverage of 95% (1 unknown word in 20), a few more gained adequate comprehension, but they were still a small minority. At 100% coverage, most gained adequate comprehension. When a regression model was applied to the data, a reasonable fit was found. It was calculated that 98% text coverage (1 unknown word in 50) would be needed for most learners to gain adequate comprehension. This figure fits with Carver's (1994) findings with native speakers:

> When the material being read is relatively easy, then close to 0% of the words will be unknown, ... when the material is relatively hard then around 2% or more of the words will be unknown, ... and when the difficulty level of the material is approximately equal to the ability level of the individual, then around 1% of the words will be unknown. (p. 432)

As Carver indicates, even 98% coverage does not make comprehension easy. Kurnia (2003), working with a non-fiction text, found that few L2 learners gained adequate comprehension with 98% coverage.

The aim of the present study is twofold. First, it aims to trial word-family lists recently developed from data from the British National Corpus (BNC). Second, it aims to use these lists to see what vocabulary size may be needed to reach a 98% coverage level of a variety of written and spoken texts.

In a partly similar study, Adolphs and Schmitt (2003, 2004) examined the coverage of word types and word-families in spoken corpora (CANCODE and spoken sections of the BNC). CANCODE is the Cambridge and Nottingham Corpus of Discourse in English, consisting of five million words of spontaneous speech. Adolphs and Schmitt's methodology was substantially different from the present study. In the Adolphs and Schmitt studies, percentage coverage figures were found by counting the words that actually occurred in the corpus. Thus the most frequent 1,000 words in their study were the 1,000 words that occurred most frequently in their corpus. In the present study, the word-frequency levels were not determined by of the corpus used. That is, the BNC was used to determine the frequency levels (using range, frequency, and dispersion), and then these frequency levels were applied to other corpora. The reason for doing so was that I wanted the frequency levels to represent the vocabulary size of a typical language user. Such a user would not know only the words in a spoken corpus such as CANCODE but would know other words as well.

We can look at this in another way. Adolphs and Schmitt's research question was as follows: What percentage coverage do various numbers of word-families in that corpus provide? The research question for my study was, How big a vocabulary do you need to get adequate coverage of various kinds of texts?

Adolphs and Schmitt's approach will always result in a higher coverage for the same number of words than in my study, because some words in my frequency lists may not occur in a particular corpus, and frequency of words in a particular corpus might not be the same as their frequency ranking in my lists. This of course reinforces the point that Adolphs and Schmitt make in their studies: 'More vocabulary is necessary in order to engage in everyday spoken discourse than was previously thought' (Adolphs and Schmitt, 2003, p. 425).

Development of the lists

The first part of this study involved the development of fourteen 1,000-word-family lists, using data from the BNC. The BNC is a 100 million–token corpus consisting of 90% written text and 10% spoken text. Word type and lemma lists from the BNC containing frequency, range, and dispersion information are available from http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html and are also published in Leech, Rayson, and Wilson (2001). Detailed information on the development of the lists is available from Paul Nation's Web site, http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx.

63, 1 (September/septembre)

The idea behind developing the lists was that they should represent the higher frequency end of a learner's vocabulary. That is, it is assumed that both native- and non-native-speaking learners acquire vocabulary largely in the order of its range and frequency. High-frequency and wide-range words are generally learned before lower-frequency and narrower-range words. There is evidence that this is so. Read (1988) and Laufer, Elder, Hill, and Congdon (2004) found that learners' scores dropped on the Vocabulary Levels Test and related tests as students moved from higher to lower frequency levels. However, there are problems with using frequency lists in making this kind of test.

As described in Nation (2004), the BNC is largely written, British, formal, and adult, and thus affects the distribution of the words in the lists. For example, in the first 1,000 we have words like *commission, committee, invest,* and *labour*, and in the second 1,000 have words like *crown, chamber, parliament, party*, and *Victorian*, which strongly reflect the nature of the corpus. Words like *hullo, goodbye, pal*, and *damn*, which are very common in spoken language, occur in the fourth 1,000 word-families because spoken language makes up only 10% of the BNC. The first 2,000 word-families contain a reasonable number of words that would not appear in courses for young learners of English, and several words that are known by very young native speakers occur late in the lists. The 1,000 word-family lists were made from a list of lemmas made from the BNC. The range, frequency, and dispersion data that were used for the division of the words into lists is thus based on lemmas and not on word-families. For example, the word-family of *abbreviate* contains the following members: *abbreviate, abbreviates, abbreviated, abbreviating, abbreviation, abbreviations.* This family consists of two lemmas: the *abbreviate* lemma with four members and the *abbreviation* lemma with two members. Word-families include several lemmas and so the frequency, range, and dispersion figures for the lemmas are underestimates of what the figures would be for word-families. One way of adjusting the ordering of items would be to run the word-family lists through the BNC and gather new range, frequency, and dispersion data. This undertaking was beyond the scope of the present study and may not be the best solution. It may be more appropriate to run the lists over separate written and spoken corpora to arrive at two orderings for the items in the lists. There are, however, ways of checking whether the word-family lists are properly ordered. From the first 1,000 to the fourteenth 1,000, the number of tokens, types, and families found in an independent corpus should decrease. That is, when the lists are run over a corpus different from the BNC, the first 1,000-word-family list should account for more tokens, types, and families than the second 1,000

63, 1 (September/septembre)

family list does. Similarly, the second 1,000 word-family list should account for more tokens, types, and families than the third 1,000 family list does, and so on. While this approach does not show that each word-family is in the right list, it does show that the lists are properly ordered. To check this, the fourteen lists were run over a corpus made up of the LOB, FLOB, Brown, Frown, Kohlapur, Macquarie, Wellington written, Wellington spoken, and LUND corpora, which are all available from the International Computer Archive of Modern and Medieval English at http://gandalf.aksis.uib.no/icame.html. LOB and FLOB are 1,000,000-token corpora of written British English; LUND is a 500,000-token corpus of spoken British English

Table 1 contains the data from the LOB corpus as an example. Word list 15 is a large list of proper nouns taken from the BNC and other sources.

The only small inconsistency in the data is evident in the second column, where it can be seen that the tenth 1,000 accounts for slightly more tokens (3,328) than the ninth 1,000 (3,217). Otherwise the figures for tokens, types, and families drop consistently from one thousand to the next. A very similar pattern was found in all the other written corpora. There were two similar small inconsistencies in the tokens of

TABLE 1
Tokens, types, and families at each of the 14 BNC word-family levels in the LOB corpus

Word list (1,000)	Token (%)	Types (%)	Families
1	78,944 (77.86)	4,487 (10.1)	998
2	83,477 (8.23)	4,131 (9.34)	998
3	37,511 (3.70)	3,239 (7.32)	998
4	18,198 (1.79)	2,683 (6.07)	998
5	10,495 (1.04)	2,226 (5.03)	969
6	7,080 (0.70)	1,789 (4.04)	928
7	6,633 (0.65)	1,542 (3.49)	887
8	4,096 (0.40)	1,382 (3.12)	836
9	3,217 (0.32)	1,118 (2.53)	734
10	3,228 (0.32)	1,025 (2.32)	719
11	1,609 (0.16)	753 (1.70)	587
12	1,434 (0.14)	646 (1.46)	498
13	1,211 (0.12)	529 (1.20)	441
14	973 (0.10)	339 (0.77)	288
15	18,519 (1.83)	2,878 (6.51)	2,878
Not in the lists	26,821 (2.65)	15,463 (34.96)	?????*
Total	1,013,9469	44,230	13,747

* The RANGE program is not able to calculate families for words not in the lists.

the spoken corpora (LUND and Wellington spoken), but not in the types and families. The lists are clearly properly ordered.

A second way of checking the validity of the lists is to look at the total number of types in each list. Low-frequency words tend to have fewer family members than high-frequency words, so even though the number of families in each list is the same, the number of types should decline. Table 2 shows the number of types (family members) and families in each of the fourteen 1,000-word-family lists. As can be seen in the second column, the data confirm the expected pattern of decrease. (In the last column, it can be seen that the list for BASEWRD3 contains four extra families [1004]. These are exclamations, hesitations, interjections, etc., that are common in spoken English, but marginal as words.)

A third way of checking the validity of the lists is to make sure that no wide-range, high-frequency words are missing from the lists. To check for error, the lists were run over the nine corpora mentioned above, and the words occurring in three or more of the nine corpora were looked at to see if they should be in the lists. This exercise resulted in the addition of several family members, for example *takings* being added to *take,* and *reds* to *red.* However, no word-families needed to be added to the higher-frequency word lists, although a few replaced gaps in the lists beyond the tenth 1,000. At these levels the nature of the corpus has a very strong effect on what occurs, resulting in some gaps.

It thus seems that the lists may be a reasonably sequenced representation of at least part of a native speaker's vocabulary, and certainly a good representation of the commonly used vocabulary.

TABLE 2
Number of types (family members) and families in each 1,000 word-family list

BASEWRD type	Number	BASEWRD family	Number
1.txt	6,019	1.txt	1,000
2.txt	5,527	2.txt	1,000
3.txt	4,591	3.txt	1,004
4.txt	4,308	4.txt	1,000
5.txt	3,988	5.txt	1,000
6.txt	3,582	6.txt	1,000
7.txt	3,421	7.txt	1,000
8.txt	3,224	8.txt	1,000
9.txt	3,053	9.txt	1,000
10.txt	2,876	10.txt	1,000
11.txt	2,808	11.txt	1,000
12.txt	2,676	12.txt	1,000
13.txt	2,391	13.txt	1,000
14.txt	2,080	14.txt	1,000

The computer program that uses the lists is called RANGE and is freely available from Paul Nation's Web site (Nation & Heatley, 2002). This program cannot distinguish homographs. So RANGE cannot distinguish between homonyms like *Smith* (the family name) and *smith* (blacksmith), and *March* (the month) and *march* (as soldiers do). Thus when the program runs, these uses are not distinguished and would be counted in the same family and as the same type. There was an attempt to deal with this matter wherever possible. For example, *marched, marching, marches, marcher, marchers*, etc., were put in one family and *March* into another. This does not completely distinguish the homonyms, but it is a step towards doing so.

Research on the Academic Word List (Wang & Nation, 2004) suggests that in most cases of homographs, one member of the pair of homographs (for example, *panel* meaning 'committee,' and *panel* meaning 'thin flat sheet') is much less frequent than the other. In the 570 word-family Academic Word List there were 60 families that contained potential homographs. Thirty-nine of these did not have both members occurring in the 3.6 million–word Academic Corpus or had a member that accounted for less than 5% of the total frequency of occurrence of the pair. Being able to distinguish homographs would add to the accuracy of the present study, but it is hoped that not doing so has not weakened the study too much.

RANGE cannot count multi-word units. Thus, the word lists contain compound words but they do not contain phrases. *According to* or *au fait*, for example, might be best counted as units, but in the lists the unit is the single word. Such phrases of course are not ignored. The items that make them up are simply counted as separate words. There is evidence (Grant, 2003; Grant & Nation, 2006) that the number of truly opaque phrases (core idioms) in English is small, and they are infrequent. Although transparent phrases need to be learned for productive purposes, for the receptive purposes of reading and listening they are not a major issue.

There is one further problem with the lists used in this study. The unit of counting used in the lists is the word-family, and the level of the word-family has been set at Level 6 of Bauer and Nation's (1993) scheme for defining word-families. Level 6 includes inflections and over 80 derivational affixes including *-able, -less, -age, -ant, -ward, circum-, neo, -ify, -ist*, and *-y*. Because such a large number of affixes are permitted at this level, they result in some large word-families, especially among the high-frequency words. It appears that higher-frequency stems generally can take a greater range of affixes than lower-frequency words. For example, the high-frequency word-family *nation* at Level 6 has the

following members: *national, nationally, nationwide, nations, nationalism, nationalisms, internationalism, internationalisms, internationalisation, nationalist, nationalists, nationalistic, nationalistically, internationalist, internationalists, nationalise, nationalised, nationalising, nationalisation, nationalisations, nationalize, nationalized, nationalizing, nationalization, nationhood,* and *nationhoods.* The word-family lists group items together that would be perceived as the same words for the receptive skills of listening and reading, and so Level 6 is an acceptable level for advanced learners. If word lists were made for productive purposes, for speaking and writing, the lemma would be the largest sensible unit to use, because each lemma takes different collocates and different grammatical patterns. Some researchers argue for using the word type (Chung, 2003).

The problem faced when deciding on the level of word-family was that the lists were going to be used to represent both a native speaker's and a non-native speaker's vocabulary levels. Ideally there should be several sets of lists ordered by range and frequency, with one set based on word types, where each type is counted as a separate word, the next set based on lemmas, where a word consists of a stem form and its inflected forms of the same part of speech, and so on, up to Level 6 or higher of Bauer and Nation (1993). Making such lists does not mean simply adding or deleting word-family members. Each deleted member and its appropriately related forms would have to appear as a separate word-family in its range and frequency determined place in the sequence of families that make up the lists. In the present study, the decision was made to go with large word-families. This will give a low assessment of how many word-families are needed to read newspapers, novels, etc. That is, if learners' word-families are smaller, a larger number of word-families will be needed to do these tasks.

The assumption that lies behind the idea of word-families is that when reading and listening, a learner who knows at least one of the members of a family well could understand other family members by using knowledge of the most common and regular of the English word-building devices. There is research evidence from native speakers (Bertram, Baayen, & Schreuder, 2000; Bertram, Laine, & Virkkala, 2000; Nagy, Anderson, Schommer, Scott, & Stallman, 1989) that the word-family is a psychologically real unit. Most L2 learners, however, will not have word-families as inclusive as those of native speakers. It should also be noted that adult native speakers will have much larger word-families than the ones used in the present study.

With these cautions in mind, let us now look at how many words are needed to do certain things. We will begin by looking at one text in detail to exemplify how the analysis is done.

How many word-families do you need to know to be familiar with most words in *Lady Chatterley's Lover*?

The novel *Lady Chatterley's Lover* is just over 121,000 tokens long, and it uses a total of just over 5,000 Level 6 word-families. As Table 3 shows, the words are spread over the 14 most frequent 1,000 word-families of the BNC and beyond. The first row of Table 3 shows that the first 1,000 word-families from the BNC account for 97,944 of the running words (tokens) in the novel. This makes up 80.88% of the total running words; 2,258 different word forms (types) are the source of these tokens. These 2,258 types make up 898 word-families. *A* and *an* are counted as two different types, making up one family. The first 1,000 words account for most of the tokens, types, and families. The sixth 1,000 words in contrast accounted for 832 of the tokens, 364 of the types, and 263 of the families. These figures for the sixth 1,000 word-families from the BNC show that most of the types at this level occurred only once in the novel.

Note that from about the sixth 1,000 onwards, each additional 1,000 word-family provides only a small increase in coverage but still involves a reasonable number of word-families.

Here is an extract from *Lady Chatterley's Lover* with the list levels marked. Unmarked words are in the first 1,000 word-families. Those marked with {2} are in the second 1,000 families, with {3} are in the third 1,000, and so on. Those marked with {15} are proper names.

> {15}Constance, his wife, was a {10}ruddy, country-looking girl with {2}soft {2}brown {2}hair and {5}sturdy body, and {2}slow movements, full of {2}unusual {2}energy. She had big, wondering {3}eyes, and a {2}soft {3}mild {2}voice, and seemed just to have come from her {3}native village. It was not so at all. Her father was the once well-known R.A., old Sir {15}Malcolm {15}Reid. Her mother had been one of the {4}cultivated {!}Fabians in the {!}palmy, rather {8}pre-{!}Raphaelite days. Between {3}artists and cultured {4}socialists, {15}Constance and her sister {15}Hilda had had what might be called an {5}aesthetically {2}unconventional {5}upbringing. They had been taken to {7}Paris and {15}Florence and {2}Rome to {2}breathe in {3}art, and they had been taken also in the other direction, to the {15}Hague and {15}Berlin, to great {4}socialist {2}conventions, where the speakers spoke in every {5}civilized {2}tongue, and no one was {!}abashed.

Table 4 lists the headwords of some of the frequently occurring families found at the fourth 1,000 level and beyond in the novel. Note the topic words like *handsome, bitch, thrill*, etc.

TABLE 3
Tokens, types, and families at each word level in *Lady Chatterley's Lover*

Word list (1,000)	Tokens (%)	Types (%)	Families
1	97,944 (80.88)	2,258 (25.70)	898
2	873 (7.21)	1,617 (18.47)	785
3	3,804 (3.14)	1,002 (11.44)	580
4	2,160 (1.78)	717 (8.19)	449
5	1,291 (1.07)	550 (6.28)	370
6	832 (0.69)	36 (4.16)	263
7	733 (0.61)	326 (3.72)	236
8	572 (0.47)	249 (2.84)	189
9	392 (0.32)	202 (2.31)	169
10	290 (0.24)	158 (1.80)	140
11	250 (0.21)	127 (1.45)	111
12	243 (0.20)	99 (1.13)	84
13	134 (0.10)	71 (0.81)	61
14	34 (0.00)	21 (0.24)	18
15	252 (2.08)	212 (2.42)	212
Not in the lists	1,167 (0.96)	784 (8.95)	?????*
Total	121,099	8,757	4,565

* The RANGE program is not able to calculate families for words not in the lists.

TABLE 4
Repeated headwords from *Lady Chatterley's Lover* at low-frequency word levels

Word list (1,000)	Examples of repeated headwords
4	hut, sun, handsome, breast, thrill
5	London, grin, dread, ghastly, dialect
6	gentleman, bitch, womb, inert
7	spite, forlorn, lass
8	queer, flint, lagoon
9	nay, quiver, conceit
10	ruddy, navel, potency
11	hazel, knoll, scullery
12	ay, coop, nowt
13	Bolshevism, gondola, afore
14	bile, crocus

Frequent words not in the lists included *ter, mun, wi, yo,* and *impudence.* These are words that typically bear a close relationship to the topic or genre of the text, the first four representing the dialect of the characters in the book. The words not in the lists are marked by {!} in the text above.

Let us now return to the question of how big a vocabulary you need to be familiar with most words in *Lady Chatterley's Lover*. Table 5 gives

TABLE 5
Cumulative percentage coverage figures for *Lady Chatterley's Lover* by the fourteen 1,000 word-families from the BNC, with and without proper nouns

Word list (1,000)	Coverage without proper nouns (%)	Coverage including proper nouns (%)
1	80.88	82.93
2	88.09	90.14
3	91.23	93.28
4	93.01	95.06
5	94.08	96.13
6	94.77	96.88
7	95.38	97.43
8	95.85	97.90
9	96.17	98.22
10	96.41	98.46
11	96.62	98.67
12	96.82	98.87
13	96.93	98.98
14	96.96	99.01
Not in the lists	97.92	100.00

cumulative percentage coverage figures for the tokens in *Lady Chatterley's Lover.* With a vocabulary of 4,000 word-families and assuming that proper nouns are easily understood, 95.06% of the tokens would be familiar. This means that there would be 1 unknown word in about every 20 running words. With a vocabulary of 9,000 words plus proper nouns, 98.22% of the tokens would be familiar. This means there would be 1 unknown word in about every 50 running words. According to Hu and Nation (2000), this is the minimum desired level for comprehending written narrative.

Assuming that proper nouns can be counted as having a minimal learning burden, a vocabulary of 9,000 words would be needed to read *Lady Chatterley's Lover* without encountering an overwhelming amount of known vocabulary. Let us now see if *Lady Chatterley's Lover* is typical of other novels.

How many words do you need to read a novel?

The novels looked at were *Lord Jim* by Joseph Conrad, *Lady Chatterley's Lover* by D.H. Lawrence, *The Turn of the Screw* by Henry James, *The Great Gatsby* by F. Scott Fitzgerald, and *Tono-Bungay* by H.G. Wells. The texts were taken from the Project Gutenburg site (http://promo.net/pg/). Table 6 summarizes the data.

TABLE 6
Text coverage in several novels

Word list	*Lord Jim* (%)	*Lady Ch.* (%)	*Screw* (%)	*Gatsby* (%)	*Tono-Bungay* (%)
2,000	87.29	88.09	91.71	87.71	86.95
4,000 + proper nouns	94.24	95.06	96.08	95.02	94.36
9,000 + proper nouns	98.06	98.22	98.52	98.47	98.00
Proper nouns	1.04	2.05	0.50	2.12	1.55

The Turn of the Screw reaches 98% with 7,000 word-families plus proper nouns, and *The Great Gatsby* gets there with 8,000. *The Turn of the Screw* has a very small number of proper nouns because there are only four major characters in the novel.

Combining the novels into one corpus gives very similar figures: 2,000 provides coverage of 87.83%, 4,000 plus proper nouns – 94.8%, 9,000 plus proper nouns – 98.24%, proper nouns 1.53%. A vocabulary of 8,000 to 9,000 words is needed to read a novel, and even then, 1 word in 50 will be unfamiliar. A few of these will be repeated topic words, but most will occur only once or twice.

How many word-families do you need to read newspapers?

Studies with the Academic Word List have shown that reading newspapers can be a good way of encountering the vocabulary that is important for academic study, probably because newspaper writing is largely formal and serious and is marked by the Latinate vocabulary found in a range of academic texts. Over 90% of the words in the Academic Word List come from French, Latin, or Greek (Coxhead, 2000).

The newspaper corpora used in this study consisted of Section A of the parallel LOB, FLOB, Brown, Frown, and Kolaphur corpora. Section A of these corpora is entitled Reportage, and each corpus consists of forty-four 2,000-token collections of news articles. Table 7 gives the coverage figures. The coverage figures are very similar for the five corpora.

Here is an extract from the Frown corpus with the word-list levels marked.

> Despite {3}intense White House {4}lobbying, {7}Congress has {2}voted to {5}override the {7}veto of a {3}cable television {2}regulation {3}bill, dealing {7}President {3}Bush the first {7}veto {2}defeat of his {9}presidency just four weeks before the {5}election. Monday night, the {9}Senate {5}overrode

TABLE 7
Percentage text coverage of five newspaper corpora by the BNC word-family lists

Word list	LOB (%)	FLOB (%)	Brown (%)	Frown (%)	Kolaphur (%)
2,000	84.33	83.07	81.54	81.79	84.15
4,000 + proper nouns	95.39	95.10	94.14	93.93	94.64
8,000 + proper nouns	98.31	98.03	97.60	97.28	98.05
Proper nouns	5.29	5.66	6.12	5.43	4.55

the {7}veto 74–25, the same {2}margin by which the {2}upper house {3}approved the {3}bill {2}last month and {2}comfortably above the two-thirds majority needed. Not one {7}senator changed sides, a {2}blow to {3}Bush's {3}prestige after he had heavily {4}lobbied {6}Republican {7}senators, {2}urging them not to {3}embarrass him this close to the {5}election.

Note that *Bush* is a proper noun, but RANGE treats it as an occurrence of the noun *bush*, which is in the third 1,000.

Proper nouns account for 4.55% to 6.12% of the running words in newspapers. This high coverage is not surprising, because newspapers are about people, places, and events. The most common 2,000 words in the BNC account for about 83% of the running words. The most frequent 4,000 words from the BNC plus proper nouns account for around 95% of the running words, and to get to the 98% coverage level a vocabulary of at least 8,000 words plus proper nouns is needed. Thus although the first 2,000 words provide greater coverage of novels (88% compared to 83%), novels have fewer proper nouns and thus a similar vocabulary size of around 8,000 to 9,000 words is needed to read newspapers.

How many word-families do you need to read graded readers?

By way of contrast, let us look at the vocabulary knowledge needed to read a simplified text. *The Picture of Dorian Gray* is in the Oxford Bookworms Series at Level 3, which keeps within a vocabulary of 1,000 words. Proper nouns make up 5.55% of the running words in the text. It has a total vocabulary (including proper nouns) of 682 word-families, and is 10,578 words long. Table 8 shows the vocabulary needed to reach 98% coverage. It is spread across the first 3,000 of the BNC because the ordering of the words in the Oxford Bookworms list is not the same as the ordering of the words in the BNC. The Oxford Bookworms list is a more suitable ordering for learners of English.

TABLE 8
Word levels and text coverage in *The Picture of Dorian Gray*

Word list	Coverage (%)
2,000	91.20
2,000 + proper nouns	96.75
3,000 + proper nouns	98.86
Proper nouns	5.55

There are no words in the novel from the ninth 1,000 onwards and none outside the lists (excluding proper nouns). There are only 20 words from the fourth to ninth 1,000, including *opium, ache, Paris, disease, gentleman, gallery, lazy, drip,* and *London.*

Only a small vocabulary is needed to read this book, and there are few words outside the very high-frequency words to burden the reader.

So far we have looked only at written text. Let us now look at some kinds of spoken text to see what vocabulary size would be needed to cope with those. A weakness of this analysis, however, is that the word lists are made from a corpus that is 90% written. While all the words that are common in spoken English are certainly in the lists, they may not be at the higher-frequency levels of the lists. Lists based solely on spoken corpora would put several of the words in the higher-frequency lists.

How many word-families do you need to know to be familiar with most words in a children's movie?

The popular children's movie *Shrek* was chosen for analysis. The script, excluding stage directions, is almost 10,000 tokens long, and uses a total of almost 1,100 word-families. As Table 9 shows, the words are spread over the 14 most frequent 1,000 word-families of the BNC and beyond. Table 9 shows that the first 1,000 word-families from the BNC account for 8,141 of the running words (tokens) in the movie, comprising 81.54% of the total running words. The first row shows that 708 different word forms (types) are the source of these tokens. These 708 types reduce to 479 word-families. The first 1,000 words account for most of the tokens, types, and families. The sixth 1,000 words in contrast accounted for only 33 of the tokens, 27 of the types, and 23 of the families. These figures for the sixth 1,000 word-families from the BNC show that most of the types occurred only once. Twenty-seven types accounted for 33 tokens.

The coverage of the third 1,000 is high because fillers like *um* and *er* and interjections *oh, uh,* and signs of astonishment *aah* and *ah* are included at that level.

TABLE 9
Tokens, types, and families in *Shrek*

Word list	Tokens (%)	Types (%)	Families
1	8,141 (81.54)	708 (49.65)	479
2	489 (4.90)	252 (17.67)	210
3	63 (6.37)	164 (11.50)	128
4	246 (2.46)	92 (6.45)	80
5	88 (0.88)	49 (3.44)	47
6	33 (0.33)	27 (1.89)	23
7	13 (0.13)	12 (0.84)	11
8	13 (0.13)	12 (0.84)	12
9	16 (0.16)	14 (0.98)	14
10	8 (0.08)	7 (0.49)	7
11	9 (0.09)	9 (0.63)	8
12	7 (0.07)	7 (0.49)	7
13	31 (0.31)	9 (0.63)	9
14	33 (0.33)	3 (0.21)	2
Proper nouns	147 (1.47)	15 (1.05)	15
Not in the lists	74 (0.74)	46 (3.23)	44
Total	9,984	1,426	1,097

Here is an extract from *Shrek* with the list levels marked. Unmarked words are in the first 1,000 word-families.

– This {7}cage is too small.
– Please, don't turn me in. I'll never be {5}stubborn again. I can change. Please! Give me another chance!
– {3}Oh, {2}shut up.
– {3}Oh!
– {2}Next!
– What have you got?
– This little {2}wooden {5}puppet.
– I'm not a {5}puppet. I'm a real boy.
– Five {4}shillings for the {2}possessed {3}toy. Take it away.
– Father, please! Don't let them do this!
– Help me!
– {2}Next! What have you got?
– Well, I've got a talking {4}donkey.

Table 10 lists the headwords of the frequently occurring families found at the fourth 1,000 level and beyond. The proper nouns have been listed together. Note the topic words like *donkey*, *dragon*, *ogre*, etc.

TABLE 10
Repeated headwords from *Shrek* at low-frequency word levels

Word list (1,000)	Examples of repeated headwords
4	donkey (42), dragon (15), damn, knight (11), quest, guy, noble, spell, sun
5	swamp (20), hideous, butt, fiery, kidding, puppet, stubborn, tournament
6	ass, boulder, groom, hum, magnitude, witch, ballad
7	freak
8	firewood
9	coward, dumb
10	lava
11	yank
13	whoa (11), muffin, steed, yonder
14	ogre (25)

It might be argued that the words occurring beyond the fifth 1,000 are words typical of any children's movies. To check this possibility, the vocabulary of *Shrek* was compared to the vocabulary in *Toy Story,* another children's movie. Table 11 presents the data.

As Table 11 shows, beyond the fifth 1,000 level, there are only eight words that occur in both movies. For comparison with Table 10, which lists words from the lower-frequency levels in *Shrek,* here are some of the words from the lower-frequency lists in *Toy Story*: *alloy, atrocity, blink* (seventh 1,000), *buddy, eyeball, jettison* (eighth 1,000), *annihilate, podium* (ninth 1,000), *alpha, buzz, dinosaur* (fourteenth 1,000).

Clearly each movie will bring its own vocabulary from the whole range of levels. As is typical of most texts and collections of texts, a very large proportion of the families will occur only once. In *Shrek* 578 of the 1097 families (53%) occurred only once. It would be impossible even from a brief plot summary of the movie to predict what words from the low-frequency levels would occur.

Let us now return to the question of how big a vocabulary you need in order to be familiar with most words in *Shrek.* Table 12 gives cumulative percentage coverage figures for the tokens in *Shrek.* Proper nouns account for 1.47% of the running words in *Shrek.* With a vocabulary of 4,000 word-families, and assuming that proper nouns are easily understood, 96.70% of the tokens would be familiar to children watching the movie. This means that there would be 1 unknown word in about every 30 running words. With a vocabulary of 7,000 words plus proper nouns, 98.08% of the tokens would be familiar to children watching the movie. This means there would be 1 unknown word in about every 50 running words.

TABLE 11
Number of word families from the fourth 1,000 level on, in *Shrek* and *Toy Story*

Word list (1,000)	Total word families	Word families occurring in both movies
4	125	17
5	83	4 farewell, glow, kid (v.), shave
6	50	1 hug
7	37	1 freak
8	27	–
9	35	2 heck, slime
10	24	2 karate, trash
11	23	–
12	18	–
13	18	1 whoa
14	5	–
Not in the lists	111	1 hallelujah

TABLE 12
Cumulative percentage coverage figures for *Shrek* by word families from the BNC

Word list (1,000)	Coverage without proper nouns (%)	Coverage including proper nouns (%)
1	81.54	83.01
2	86.44	87.91
3	92.81	94.28
4	95.27	96.74
5	96.15	97.62
6	96.48	97.95
7	96.61	98.08
8	96.74	98.21
9	96.90	98.37
10	96.98	98.45
11	97.07	98.54
12	97.14	98.61
13	97.45	98.92
14	97.78	99.25
Not in the lists	98.53	100.00

These vocabulary sizes are not essential for watching and enjoying *Shrek.* Two-year-olds watch *Shrek* with pleasure and get absorbed in the movie. Eight- and nine-year-olds memorize the script from having watched it so many times. A movie has the advantage of providing strong visual support. It has the disadvantage of using spoken language, which is heard and then is gone.

Beyond the sixth 1,000 level, excluding 15 proper nouns, *Shrek* contains 134 word-families (see Table 9). A few of these like *hum, bedtime, dumb,* and *gingerbread* may already be known to children, but

the rest are potential useful additions to their vocabulary. Watching movies could be very good for vocabulary growth.

How many words do you need to cope with unscripted spoken English?

Movies are scripted spoken language, which may differ from spontaneous unscripted spoken language. Two parts of the Wellington Corpus of Spoken English were used to look at the vocabulary of unscripted spoken English. Each part was around 100,000 words long. Two parts were used rather than the whole corpus so that proper nouns could be properly dealt with. Each time a new text is used, there are usually proper nouns not in the proper noun list used by RANGE, and so quite a number of additions are needed to the proper noun list to make sure that proper nouns are separated from other words not in the 14 word lists. One section involved talk-back radio and interviews. In talk-back radio, listeners phone in with their spontaneous comments on the issue being discussed. The other section involved friendly conversation between family members and friends.

Proper nouns account for around 1% of the running words in the spoken selections. Table 13 shows that 3,000 word-families plus proper nouns give over 95% coverage, and 6,000 to 7,000 word-families are needed to get 98% coverage.

Clearly, spoken language makes slightly greater use of the high-frequency words of the language than written language does. Against this we need to consider that greater text coverage than 98% may be needed to cope effectively with the transitory nature of spoken language. A slightly biasing factor in these two samples is that they are both from the Wellington Spoken corpus, and *Zealand* and *Wellington* occur in the ninth 1,000 list, and *Maori* occurs in the fourteenth 1,000. If these were at higher-frequency levels, the coverage of the high-frequency words would have been about 0.4% higher.

TABLE 13
Percentage text coverage of two collections of spoken English by the BNC word-families lists

Word list (1,000)	Talk-back, interview (%)	Conversation (%)
2	89.41	89.35
3 + proper nouns	96.52	96.03
6 + proper nouns	98.26	97.67
7 + proper nouns	98.62	97.95
Proper nouns	1.29	1.03

Will the inclusion of frequently occurring topic words from the low-frequency levels reduce the number of words needed?

When learners read a text or watch a movie or listen to a conversation, there are words that recur because they are closely related to the topic of the text. For example, in *Shrek* the low-frequency words *ogre, swamp, donkey, dragon,* and *knight* occur very often. Similarly in *Lady Chatterley's Lover,* the low-frequency words *ravish, ay, quiver, collier, flint,* and *recoil* occur often. If the learners work out what these words mean early in the text, then their later occurrences are like occurrences of known words, not unfamiliar words. If these topic words are taken out of the low-frequency lists and are considered as known words, does this reduce the size of the vocabulary that learners needed to cope with the texts?

We can investigate this possibility by seeing if words that occur often at the last frequency level needed to reach 98% coverage and the frequency levels beyond that produce a coverage of the text that is greater than the coverage of the last frequency level needed to reach 98%. For example, in *Shrek,* just over 6,000 words are needed to reach 98% coverage. The sixth 1,000 word level covers 0.33% of the running words. Two words occur more than 10 times in the sixth to fourteenth 1,000 levels: *whoa* (11 occurrences) and *ogre* (32 occurrences). These cover 0.43% of the running words, which is greater than the coverage of the sixth 1,000 level – 0.33%. Thus by considering these highly repeated words as known, only 5,000 words would be needed to reach 98% coverage, a saving of a 1,000 word-level.

Similarly in a much longer text, *Lady Chatterley's Lover*, 8,000 words are needed to reach 97.93% coverage. The eighth 1,000 covers 0.47% of the running words. The 24 words occurring 10 times or more from the eighth 1,000 level on cover 0.36% of the running words – not enough to reduce the number of words needed. If 9,000 words were considered as the number needed (they provide 98.25% coverage, and the ninth 1,000 by itself provides 0.32% coverage), the topic words from the ninth 1,000 inclusive on provide 0.21% coverage, which, added to the coverage of the 8,000 words, would give 98.14% coverage. This figure is a little lower than the 98.25% coverage provided by the 9,000 words but is over 98%. This is a saving, 1,000 words less, but as with *Shrek,* it is a negligible saving in that we are dealing with changes in coverage of well under 0.5%.

The 98% target coverage assumes that the learners do not use a dictionary or get help from some other source outside the text. If learners could draw on such help, then a slightly lower coverage figure would be acceptable. However, if the coverage figure were 95%, this

would mean learners would be dealing with 1 unknown word in every two lines of text (1 unknown word in 20), or with 7 unknown words in every minute of speech at 150 words per minute.

Conclusions

If we take 98% as the ideal coverage, a 8,000–9,000 word-family vocabulary is needed for dealing with written text, and 6,000–7,000 families for dealing with spoken text.

Clearly, spoken language makes slightly greater use of the high-frequency words of the language than written language does. In contrast, we need to consider that text coverage greater than 98% may be needed to cope effectively with the transitory nature of spoken language. The data we have looked at in this article suggest the following conclusions.

1. The greatest variation in vocabulary coverage is most likely to occur in the first 1,000 words, and in the proper nouns. The first 1,000 plus proper nouns cover 78%–81% of written text, and around 85% of spoken text.
2. The fourth 1,000 and fifth 1,000 words provide around 3% coverage of most written text, and 1.5%–2% coverage of spoken text.
3. The four levels of the sixth to ninth 1,000 provide around 2% coverage of written text and around 1% coverage of spoken text.
4. The five levels of tenth to fourteenth 1,000 provide coverage of less than 1% of written text and 0.5% of spoken.

Table 14 summarizes these data.

There are fairly stable figures for coverage within a genre such as newspapers, novels, or in the planned corpora of LOB, Brown, etc.

TABLE 14
Average coverage and range of coverage of a series of word levels

Levels	Number of levels	Approximate written coverage (%)	Approximate spoken coverage (%)
1st 1,000	1	78–81	81–84
2nd 1,000	1	8–9	5–6
3rd 1,000	1	3–5	2–3
4th–5th 1,000	2	3	1.5–3
6th–9th 1,000	4	2	0.75–1
10th–14th 1,000	5	< 1	0.5
Proper nouns	1	2–4	1–1.5
Not in the lists	1	1–3	1

Coverage figures for the first 1,000 within a single genre typically vary by no more than 2% or 3%, and variation at the second 1,000 word level is at less than 2%. The raw figures of the range of variation, of course, decreases going down the frequency levels. The few exceptions usually have an obvious explanation, such as the small number of characters in *The Turn of the Screw* or the effect of New Zealand topic words in the New Zealand corpora.

The lists are now available in a revised form at http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx and as part of a Web profiler at http://www.lextutor.ca/vp/bnc/. The words in the revised lists are sequenced largely according to their range and frequency in the 10 million spoken section of the BNC. They are now an alternative to use in the RANGE program instead of the two 1,000 word lists from the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000), which have been used with RANGE and the vocabulary profiling programs up till now. The disadvantage of the BNC lists is that the Academic Word List is not separated out from the frequency levels, and so this important special purposes vocabulary cannot be looked at. In the BNC lists, the Academic Word List vocabulary is spread from the first 1,000 words to the tenth 1,000 words (also see Nation, 2004). On the other hand, the BNC lists cover a very large amount of vocabulary and thus give more detailed estimates of the vocabulary load of texts. The BNC lists are also more recent than the General Service List.

Paul Nation is a professor in applied linguistics at Victoria University of Wellington, New Zealand. He has taught in Indonesia, Thailand, the United States, Finland, and Japan. His books include *Teaching and Learning Vocabulary* (Heinle & Heinle, 1990), *New Ways in Teaching Vocabulary* (TESOL, 1994), and *Learning Vocabulary in Another Language* (Cambridge University Press, 2001). He has also published many articles on teaching and learning vocabulary. His current research project is a computerized test of vocabulary size.

Acknowledgements

This research was done with the support of a Faculty Research Grant from the Faculty of Humanities and Social Sciences, Victoria University of Wellington, New Zealand.

References

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425–438.

Adolphs, S., & Schmitt, N. (2004). Vocabulary coverage according to spoken discourse context. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 39–49). Amsterdam: John Benjamins,

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

Bertram, R., Baayen, R., & Schreuder, R. (2000). Effects of family size for complex words. *Journal of Memory and Language, 42*, 390–405.

Bertram, R., Laine, M., & Virkkala, M. (2000). The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology, 41*(4), 287–296.

Carver, R.P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Reading Behavior 26*(4), 413–437.

Chung, T.M. (2003). A corpus comparison approach for terminology extraction. *Terminology, 9*(2), 221–245.

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

Goulden, R., Nation, P., & Read, J. (1990) How large can a receptive vocabulary be? *Applied Linguistics, 11*(4), 341–363.

Grant, L. (2003). *A corpus-based investigation of idiomatic multi-word units*. Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand.

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

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

Hu, M., & Nation, I.S.P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

ICAME. (2006). International Computer Archive of Modern and Medieval English. Retrieved June 24, 2006 from http://icame.uib.no/

Kurnia, N. (2003). *Retention of multi-word strings and meaning derivation from L2 reading.* Unpublished doctoral dissertation, Victoria University of Wellington, New Zealand.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, *21,* 202–226.

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

Nagy, W.E., & Anderson, R.C. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304–330.

Nagy, W.E., Anderson, R., Schommer, M., Scott, J.A., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly, 24*(3), 263–282.

Nation, I.S.P. (1993). Using dictionaries to estimate vocabulary size: Essential, but rarely followed, procedures. *Language Testing, 10*(1), 27–40.

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

Nation, I.S.P., & Heatley, A. (2002). Range: A program for the analysis of vocabulary in texts [software]. Downloadable from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*(2), 12–25.

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

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

Ward, J. (1999). How large a vocabulary do EAP Engineering students need? *Reading in a Foreign Language, 12*(2) 309–323.

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

Xue, G., & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication, 3*, 215–229.

Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A., & Healy, N.A. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior, 27*(2), 201–212.