CORPUS-BASED FREQUENCY PROFILING: MIGRATION TO A WORD LIST BASED ON THE BRITISH NATIONAL CORPUS

Leah Gilner[*] *and Franc Morales*

ABSTRACT

The use of frequency word lists to profile the vocabulary makeup of a text is one such criterion. It provides a quantifiable characterization and classification of lexical material in terms of corpus-based frequency measures. The process of vocabulary profiling is not without challenges, first among which is the identification of a word list adequate for ELT. The choice will determine the amount of information, if any, that can be derived from a text. This paper provides an appraisal of a frequency word list based on the British National Corpus (BNC) and shows the benefits that can be gained by profiling with this list rather than with the long-established General Service List (West, 1953).

There are two basic appeals to moving on from the GSL. First, the GSL was compiled based on data from corpora tallying up to 10 million tokens while the BNC is ten times larger. The differential in size makes it possible to obtain a more accurate account of the frequency organization of the English lexicon. Second, the GSL leaves uninformative gaps when deployed in profiling, ranging from 10% to 25% depending on the text (Nation and Waring, 1997). Due to the way in which the GSL was manufactured, expanding this word list is nearly impossible if one is to follow the original directives and criteria (Faucett et al, 1936; Lorge, 1949; West, 1953). While necessarily observing different criteria, word lists that supplement the GSL have been proposed (Coxhead, 2000; Xue and Nation, 1984) and investigated (Hyland and Tse, 2007). It can be said that expandability is, perhaps, the weakest point of the GSL and the best reason for seeking a replacement. The BNC affords the possibility of addressing this issue.

However, since the BNC has not been designed to inform ELT, knowledge of its origin and composition is an important factor to take into consideration when seeking to derive information and insight from it. The BNC is a 100-million-word sample, synchronic, general, monolingual, mixed

[*] Nagoya University of Foreign Studies

corpus of present-day British English. The compilation of the BNC was a collaborative undertaking carried out by dictionary publishers (Oxford University Press, Longman, Chambers Harrap) and academic institutions (Oxford University, Lancaster University, the British Library) with financial backing from British government agencies (Leech et al., 2001). About 90% of the corpus is comprised of written language, categorized as imaginative (i.e. fiction) or informative (i.e. non-fiction, expository); most of the written texts date from 1975 or later (20% of imaginative texts date from 1960). The 10-million-word subcorpus of spoken language contains samples recorded between 1991 and 1993; 40% of the samples represent conversational language use, that is to say, "spontaneous interactions engaged in by some 127 adults aged 15 and over" (Leech et al., 2001, p.2); 60% of the samples represent task-oriented speech (lectures, sermons, TV/radio programs, consultations), "those types of […] spoken activity that were unlikely to be recorded by the conversational volunteers during a typical day of their lives (Leech et al., 2001, p. 3).

It stands to reason that not every frequency word list based on the BNC has the potential to be equally informative in ELT. Nation (2004) inspected the 3,000 most frequent word families in the BNC and found that they contain material from the GSL as well as the Academic Word List (Coxhead, 2000). The presence of this academic vocabulary interwoven with words of general service - which are thought to provide the foundation for subsequent learning (Nation, 2001) - made it difficult to "decide how the GSL could be replaced" (Nation, 2004, p.12) when considering, precisely, core vocabulary. Following a different line of inquiry, Nation (2006) produced 14,000 word families, organized according to frequency data from the BNC, in order to asses "how large a receptive vocabulary is needed for typical language use" (p. 59). Of interest, his analyses reveal important information in those gaps uncharacterized by the GSL. For example, 'topic words' were identified in the 4,000 word family and beyond.

The evolution of this list of 14,000 word families (hereafter, BNC-ELT word list) was not limited to expandability (over the 3,000 originally used) but subsequently included the reorganization of the word families according to the spoken subcorpus of the BNC. As Nation explains in the "readme" file of the RANGE software available from his website, "previously the lists had been sequenced using figures from the whole BNC but because of the overwhelming amount of formal written material this resulted in lists that did not satisfactorily represent informal spoken uses of English". It should be noted that range was the main criterion used in the creation of the list and that frequency was the second criterion (I.S.P. Nation, personal communication, February 13, 2008).

Although it is reasonable to assume that further refinement of the BNC-ELT list might take place, we have adopted this latest revision of the list and have used it in the analyses carried out in this investigation in order to characterize a relatively large and varied sample of the kinds of authentic materials that can be used in ELT. As findings will show, this characterization gives powerful reason to migrate from the GSL to the BNC-ELT.

We begin with a closer look at the BNC by inspecting the raw frequency measures it yields and how these metrics characterize the whole corpus, a generalizable insight into lexical choice in language use. Following we present an analysis that illustrates how these same metrics characterize a sample of materials used in ELT. Some structural properties of the BNC-ELT list will then be presented together with a description of the ELT corpus compiled for this investigation. We bring the two together - the BNC-ELT list and the ELT corpus – in a series of analyses that will allow us to explore the extent and manner in which profiling can be used to characterize the lexical content of ELT materials. We close by providing comparative measures between the GSL and the BNC-ELT for referential purposes that may facilitate the transition from one to the other.

The importance of lexical frequency in language use cannot be overstated. Analysis of any reasonable amount of language in use reveals substantial uniformity with regards to lexical content. The numbers are quite impressive. Nearly 50% of all language used is confined to 100 words, 75% to 2,000 words, and 85% to 5,000 words (Leech et al., 2001). These numbers show an extremely sloped curve of distribution where very few items in the vocabulary account for most discourse while the majority of items in the vocabulary occur with severe, even extreme, infrequency (Ellis, 2002). The implications for language instruction are clear. Speakers demonstrate marked preferences when it comes to lexical choice, making it feasible to isolate a vocabulary of objective value.

Table 1 presents data on the lexical coverage that the most frequent words of the English language provide for the British National Corpus. According to the data obtained from Leech et al. (2001), 397,041 of the 757,087 words (52.44%) in the corpus account for only 0.0039% of the occurrences in the corpus, while 100 of the 757,087 words (0.0132%) account for 45.8786% of the occurrences in the corpus. The analysis we provide in Table 1 illustrates this phenomenon in greater detail.

It is evident that exceedingly few types (words) account for the vast majority of tokens (occurrences). While the 10 most frequent types occur over 21 million times, there are 397,041 types (52.44%) that occur only once in the entire corpus (Leech et al., 2001).

Table 1 shows that the first 100 words (types) occur 45,878,600 times (tokens) in the BNC. The next set of 100 types contributes 6,373,600 tokens

(or 6.3736% of the corpus) while the next 300 types add 8,358,400 tokens. From here, the column labeled 'Difference' shows how the amount of tokens contributed by the subsequent addition of types gradually diminishes. The column labeled 'Cumulative' grows as the amount of types increases, yet

Types	Tokens	% of BNC types	% of BNC corpus	Difference	Cumulative
100	45,878,600	0.0132%	45.8786%		
200	52,252,200	0.0264%	52.2522%	6.3736%	6.3736%
500	60,610,600	0.0660%	60.6106%	8.3584%	14.7320%
1,000	67,569,500	0.1321%	67.5695%	6.9589%	21.6909%
1,500	71,864,900	0.1981%	71.8649%	4.2954%	25.9863%
2,000	74,950,900	0.2642%	74.9509%	3.0860%	29.0723%
2,500	77,332,500	0.3302%	77.3325%	2.3816%	31.4539%
3,000	79,255,000	0.3963%	79.2550%	1.9225%	33.3764%
3,500	80,828,900	0.4623%	80.8289%	1.5739%	34.9503%
4,000	82,144,700	0.5283%	82.1447%	1.3158%	36.2661%
4,500	83,254,100	0.5944%	83.2541%	1.1094%	37.3755%
5,000	84,214,800	0.6604%	84.2148%	0.9607%	38.3362%
5,500	85,060,900	0.7265%	85.0609%	0.8461%	39.1823%
6,000	85,809,200	0.7925%	85.8092%	0.7483%	39.9306%
6,500	86,480,400	0.8586%	86.4804%	0.6712%	40.6018%
7,000	87,088,700	0.9246%	87.0887%	0.6083%	41.2101%

Table 1. Breakdown of the most frequent words in English.

showing that the amount of tokens does not quite double even after the inclusion of tokens corresponding to 7,000 types. Summing up, the 100 most frequent types (words) of the language amount to more tokens (occurrences) than the following 6,900 types (words) combined and, together, the 7,000 most frequent types account for 87.09% of all tokens in the BNC.

It is relevant to question if the observed frequency distributions are limited to large corpora and whether texts of smaller size exhibit similar metrics. Furthermore, it is pertinent to ask if ELT materials are equally served by this information. Table 2 answers this question affirmatively by analyzing eight texts of different sizes and types of discourse (spoken and written). These texts have been randomly chosen from the ELT corpus that was compiled for this study and that will be introduced later on. For now, note that these texts have been arranged so that "beginner" materials are displayed in the upper rows and "advanced" materials in the lower rows, that is, the *Interview script* is a transcription of a dialogue of little complexity while the *NYT article* is a piece of news from the New York Times and, therefore, of reasonable difficulty for advanced learners.

The first observation is that the coverage provided by the most frequent types (words) in English is superior for the ELT materials than it is for the BNC (82.65% over 79.25%). The second observation is that while values fluctuate, trends are uniform and correlate. Approximately half (or

more) of all tokens of any text are confined to the 100 most frequent types in the language while approximately three-fourths (or more) of all tokens of any text are confined to the 2,000 most frequent types. The third observation is that, together, the 3,000 most frequent types offer better coverage of the *Interview script* than of the *NYT article.* The fourth observation is that, as we

Source	Tokens	100	500	1,000	2,000	3,000
BNC	100,000,000	45.89%	60.61%	67.57%	74.95%	79.25%
All texts	52,169	49.91%	63.97%	69.63%	78.29%	82.65%
Interview script	560	62.11%	80.33%	84.88%	88.89%	91.62%
Short story	15,935	54.95%	69.65%	74.97%	80.95%	84.76%
Family movie script	17,730	45.81%	58.65%	63.39%	75.93%	80.48%
News article (intermediate)	448	43.95%	61.16%	69.30%	80.00%	85.81%
Novel	10,542	51.77%	66.85%	72.74%	79.70%	84.16%
ESP reading (technology)	1,602	45.56%	58.78%	65.15%	72.14%	77.02%
IHT article	1,555	45.71%	58.80%	67.64%	75.61%	79.00%
NYT article	1,265	42.28%	54.68%	64.45%	72.25%	79.89%

Table 2. Profile of a variety of texts based on BNC raw frequencies.

look from row to row in descending order, we can see that profiling does not provide a definite correlation between frequency and difficulty even though the data shows a tendency in that direction. In other words, better coverage (in this case, fewer infrequent words) does not *necessarily* imply lesser difficulty.

The word lists employed in the two previous analyses were manufactured by identifying the most frequent unlemmatized types (words) in the English language (for example, the first ten types are *the*, *of*, *and*, *a*, *in*, *to*, *it*, *is*, *to*, and *was*). For the purposes of ELT, such a list is not as useful as a list where items are lemmatized and, more importantly, clustered into word families. A word family refers to a grouping containing a headword, its inflections, and its closest derivations (Nation, 2001). It is posited that awareness of word family relationships can "greatly decrease the learning burden of derived words containing known base forms," (Nation, 2001, p. 8) although the extent to which this is true will depend on a given learner's experience and linguistic background among other things (Mochizuki and Aizawa, 2000; Sakata, 2007).

One of the strengths of the BNC-ELT list is that types are clustered into families. The BNC-ELT is comprised of a total of 50,598 types (words), grouped into 14 sublists of 1,000 families each which in turn are ranked by descending frequency. That is, the first sublist contains the 1,000 most frequent word families in the English language, the second sublist contains the following 1,000 most frequent families, and so on. The detail and scope of the

BNC-ELT list makes this contribution an unparalleled resource for the identification of lexical distributions in ELT materials.

Table 3 shows some of the structural characteristics of the BNC-ELT sublists together with the code that will be used to refer to them in the discussion that follows. As mentioned, each sublist contains 1,000 word families. An example of a family from the sublist SL-01 is *ABLE*: *able*, *ability*, *abler*, *ablest*, *ably*, *abilities*, *unable*, and *inability* while an example of a family from the sublist SL-14 is *ALLURE*: *allure*, *allured*, *allures*, *alluring*, and *alluringly*.

Code	Families	Words	Average
SL-01	1,000	6,348	6.35
SL-02	1,000	5,593	5.59
SL-03	1,000	4,517	4.52
SL-04	1,000	4,287	4.29
SL-05	1,000	3,992	3.99
SL-06	1,000	3,494	3.49
SL-07	1,000	3,272	3.27
SL-08	1,000	3,192	3.19
SL-09	1,000	3,050	3.05
SL-10	1,000	2,840	2.84
SL-11	1,000	2,794	2.79
SL-12	1,000	2,568	2.57
SL-13	1,000	2,426	2.43
SL-14	1,000	2,225	2.23
BNC-ELT	14,000	50,598	3.61

Table 3. Characteristics of the BNC-ELT list.

We can see from Table 3 that less frequent families have fewer members (Nation, 2007). The column labeled 'Average' quantifies this trend by presenting the average number of types per family for each of the sublists and for the BNC-ELT list as a whole.

The ELT corpus compiled for this investigation consists of eight collections of texts, each taken from the kinds of sources (newspaper articles, movie scripts, short stories, novels, etc.) generally referred to as authentic material (Gilmore, 2007). Table 4 presents some general statistics which serve to inform on the composition of the ELT corpus compiled for this study. The collections were compiled from the kind of sources that are often used when the desire is to provide students with authentic models of naturally-occurring, fluent language use (Brown and Yule, 1983) and that, in our experience teaching at university level, often find their way into the classroom. These collections can be said to illustrate a natural grading (Gilmore, 2007) in terms of content, presentation, and register (Carter and McCarthy, 1994), ranging from *Interview scripts* and *Short stories* which are likely to be used with less experienced students (i.e. first-year university) to articles from the

International Herald Tribune (IHT) and the *New York Times* (NYT) which might be selected for more experienced and advanced students.

Code	Description	Items	Tokens	Avg. length
CLT-01	Interview scripts	116	49,613	453
CLT-02	Short stories	18	52,413	2,958
CLT-03	Family movie scripts	19	325,581	18,937
CLT-04	News articles (intermediate)	134	72,187	538
CLT-05	Novels	14	398,854	29,614
CLT-06	ESP readings (technology)	31	55,486	1,866
CLT-07	International Herald Tribune	164	154,658	1,010
CLT-08	New York Times	58	48,701	903
ELT Corpus	All collections	554	1,157,493	2,226

Table 4. Characteristics of the ELT corpus.

The design of the ELT corpus was approached from an ELT practitioner's perspective in as much as we wanted to compile an assortment of material that might reflect the choices made by colleagues in the field. The ELT corpus deliberately contains collections that have markedly more tokens than others (i.e. *Family movie scripts vs*. *Interview scripts*) in order to highlight the effect text length has on profiling results. We now describe each collection in turn.

The *Interview scripts* collection is comprised of transcriptions of 116 interviews taking place between speakers of different backgrounds. The materials can be described as modeling naturally-occurring interactions, as when people are getting to know each other, and in which English is used as an international language. The *Interview scripts* have an average length of about 450 tokens (occurrences), making them relatively short. From the ELT practitioner's perspective, the length, breadth, and depth of the interviews make them appropriate for less experienced students. Topics are discussed in general terms and speakers often provide narratives about personal experiences.

The *Short stories* collection contains 18 children's stories (i.e. *The Tale of Peter Rabbit*, *The Emperor's New Clothes*, *Rapunzel*). The stories are written for a young L1 reading audience but the fictional, imaginative aspects of the texts can make them entertaining and engaging for L2 learners of an older age. The stories are lengthy, on average about 3,000 tokens, and could be a resource for extensive reading in lower and intermediate levels.

The *Family movie scripts* collection contains 19 scripts, on average about 19,000 tokens long, from movies that seem to be widely-recognized, such as *E.T.* and *Back to the Future*. The nature of the genre implies that a substantial

amount of the discourse comes in the form of dialogues. To a large extent, the structural and conceptual complexity of the material is bound by the target audience (parents and children). As family movies are less likely to involve in-depth development of ideas or elaborate argumentation, they are deemed most appropriate for intermediate-level students.

The *News articles* (*intermediate*) collection includes 134 newspaper articles from Voice of America and the English version of a well-known Japanese newspaper. Topics vary from economics and politics to health and education. They are relatively short, on average about 500 tokens long, and differ from the articles in the IHT and the NYT collections in terms of lexical and structural complexity as well as depth of exposition; these factors combine to make this collection accessible to intermediate-level students.

The *Novels* collection contains 14 full-length fiction stories written for adult audiences, each averaging 30,000 tokens in length. The nature of these kinds of texts implies a more in-depth development of plot and characters than the other collections and is likely to include examples of spoken discourse in the form of dialogue interwoven in the narrative. Full-length novels are also likely to make use of a larger and more varied vocabulary than shorter texts. Given these characteristics, the texts in this collection represent extensive reading material for students at advanced-levels.

The *ESP readings* (*technology*) collections contain 31 texts which describe aspects and constructs related to computers, the Internet, and electronics. The texts average about 2,000 tokens, are procedural in nature, and make use of domain-specific, specialized vocabulary. This collection is deemed to represent intermediate- and advanced-level material for Science and Engineering majors.

The *International Herald Tribune* (IHT) and *New York Times* (NYT) collections contain 164 and 58 news articles and special reports, respectively. In both cases, the average length is about 1,000 tokens. Topics vary widely and include: politics, economics, culture, society, travel, sports, health, fashion, etc. As the name suggests, the IHT targets an international audience while the NYT, although of international repute, is thought of as a newspaper of record in the U.S. Thus, even though the two newspapers are owned by the same company, the role of each may influence not only the treatment and perspective provided in the texts, but also the style of discourse and expression. Either collection might be used with advanced-level students.

With this outline of the collections and corpus in mind, we move on to the analysis of the ELT corpus by means of the BNC-ELT list. Results for all eight collections are presented throughout. However, we will focus the discussion on CLT-01 and CLT-08 as these collections are similar in size while clearly distinct in difficulty. This narrow focus will allow us to formulate (weak) propositions regarding the insights into “difficulty” that

profiling affords. Once the data has been presented and discussed, we will proceed to take into account the results obtained from the other collections. The reader is encouraged to consider all results as analyses are presented.

The reason why the discussion elaborates on the assessment of difficulty by means of frequency profiling is because we find that it is a relationship that is established intuitively yet not addressed in the literature on vocabulary frequency lists. There seems to be an assumption or tendency to assume, for instance, that frequent words are "easier" than infrequent ones. Intuitively, it makes sense to consider a word that is rarely used as being, one, of very specialized application and/or, two, of such low occurrence that it is difficult for a learner to obtain repeated exposure to it. The interpretation of results presented hereafter hopes to provide some observations regarding the equivocal relationship between frequency profiling and learning burden.

A profile of the ELT corpus using the BNC-ELT list is presented in Table 5. The lexical material in the ELT corpus belonging to the BNC-ELT list is 98.63% (last column, bottom row). By collection, the extremes are 99.64% coverage of CLT-01 ('lower-level' texts) and 97.42% coverage of CLT-8 ('advanced-level' texts).

Code	CLT-01	CLT-02	CLT-03	CLT-04	CLT-05	CLT-06	CLT-07	CLT-08	ELT Corpus
SL01	91.45%	82.23%	79.27%	80.04%	82.25%	76.18%	76.84%	76.94%	80.43%
SL02	4.23%	7.01%	7.07%	9.59%	7.01%	10.28%	9.47%	9.35%	7.65%
SL03	1.62%	3.72%	4.67%	2.93%	3.94%	3.62%	3.18%	3.32%	3.83%
SL04	0.80%	1.78%	2.43%	2.24%	1.63%	2.71%	2.80%	2.95%	2.13%
SL05	0.42%	1.36%	1.39%	1.10%	1.34%	1.36%	1.47%	1.36%	1.32%
SL06	0.34%	0.76%	0.97%	0.94%	0.72%	0.96%	1.03%	0.86%	0.85%
SL07	0.17%	0.51%	0.63%	0.46%	0.53%	0.57%	0.63%	0.64%	0.56%
SL08	0.15%	0.46%	0.39%	0.44%	0.35%	0.82%	0.54%	0.53%	0.42%
SL09	0.13%	0.23%	0.42%	0.21%	0.30%	0.34%	0.41%	0.41%	0.34%
SL10	0.11%	0.33%	0.58%	0.25%	0.26%	0.59%	0.37%	0.27%	0.38%
SL11	0.09%	0.15%	0.28%	0.16%	0.28%	0.22%	0.26%	0.27%	0.25%
SL12	0.09%	0.16%	0.17%	0.15%	0.15%	0.25%	0.21%	0.17%	0.17%
SL13	0.02%	0.15%	0.34%	0.13%	0.17%	0.17%	0.21%	0.22%	0.22%
SL14	0.02%	0.16%	0.08%	0.06%	0.07%	0.23%	0.15%	0.15%	0.09%
BNC-ELT	99.64%	99.03%	98.70%	98.69%	98.99%	98.28%	97.58%	97.42%	98.63%

Table 5. Token Coverage of the ELT corpus by the BNC-ELT list and sublists.

Inspection of the column labeled 'ELT Corpus' shows that sublist SL-01 (the first 1,000 families and, thus, the most frequent in the language) accounts for 80.43% of words in the ELT corpus. The second sublist of families, SL-

02, accounts for 7.65% of the vocabulary and, together with the first sublist, amounts to 88.08% of all words in the ELT corpus. A clear drop in use is evident as families become more infrequent; a trend that is disrupted in only two occasions (SL-10 and SL-13).

The coverage analysis also reveals that collections deemed more adequate for lower-level learners have higher concentrations of vocabulary in the first sublist (SL-01) than those collections containing advanced material. The data for SL-01 indicates that this sublist accounts for 91.45% of CLT-01 yet for a much smaller share of CLT-06, CLT-07, and CLT-08 (76.18%, 76.84%, and 76.94% respectively). From this data, it is possible to propose that, in general, there might be a connection between lexical frequency and level of difficulty. In other words, a characteristic of advanced texts might reside in the use of a comparatively infrequent vocabulary and, conversely, that a characteristic of beginner texts might reside in the limitation of vocabulary to frequently occurring - i.e. more common - words. The proposition might be intuitively correct but it is of importance to note that the BNC-ELT list provides a means for the quantification of this characteristic.

We now examine the amount of families (and types) from each sublist in the BNC-ELT list that is used by each of the collections in the ELT corpus. The data is first presented globally, that is, regarding the ELT corpus as a whole and without detailing use by collection.

Code	Families	Percent	Types	Percent	Tokens
SL-01	1,000	100.00%	4,563	71.88%	930,981
SL-02	994	99.40%	3,822	68.34%	88,559
SL-03	981	98.10%	2,972	65.80%	44,325
SL-04	954	95.40%	2,486	57.99%	24,642
SL-05	902	90.20%	2,018	50.55%	15,271
SL-06	860	86.00%	1,640	46.94%	9,817
SL-07	763	76.30%	1,284	39.24%	6,430
SL-08	698	69.80%	1,149	36.00%	4,837
SL-09	654	65.40%	1,030	33.77%	3,943
SL-10	626	62.60%	925	32.57%	4,386
SL-11	588	58.80%	820	29.35%	2,914
SL-12	481	48.10%	633	24.65%	1,958
SL-13	469	46.90%	627	25.85%	2,493
SL-14	313	31.30%	388	17.44%	1,097
BNC-ELT	10,283	73.45%	24,357	48.14%	1,141,653

Table 6. Use of BNC-ELT lists in ELT corpus (global data).

Table 6 shows that all families in sublist SL-01 (the most frequent in the language) are used in the ELT corpus. As a family is a collection of inflected and derived forms, it is important to note that 71.88% of all types in SL-01 are found in the ELT corpus and that these types account for 930,981 of all of the tokens (80.43% as shown in Table 5). As in all other analyses, sublists of

more infrequent families are used progressively less both at the family and type (word) level, a fact that correlates with the amount of tokens they account for in the ELT corpus. The data, again, supports the validity and adequacy of the BNC-ELT list for the purpose of assessing ELT materials beyond the scope of the GSL.

From this data, it is possible to formulate a second proposition, namely, that there might be a connection between lexical variety and level of difficulty. In other words, one way to characterize advanced texts might be in terms of the use of a comparatively rich vocabulary. Table 7 makes a clear case for this proposition

	CLT-01	**CLT-02**	**CLT-03**	**CLT-04**	**CLT-05**	**CLT-06**	**CLT-07**	**CLT-08**	**ELT Corpus**
SL-01	85.00%	83.60%	99.00%	96.50%	98.60%	92.80%	98.40%	95.90%	100.00%
SL-02	47.80%	57.50%	92.40%	78.80%	91.80%	77.50%	94.70%	79.70%	99.40%
SL-03	23.10%	43.70%	83.70%	49.30%	82.30%	50.80%	80.20%	52.90%	98.10%
SL-04	14.20%	29.20%	69.90%	39.10%	67.20%	39.30%	71.80%	42.60%	95.40%
SL-05	7.90%	21.40%	59.40%	26.70%	54.50%	27.70%	58.30%	29.80%	90.20%
SL-06	6.10%	17.30%	48.80%	19.60%	45.70%	20.40%	45.90%	21.40%	86.00%
SL-07	3.60%	11.10%	37.60%	11.70%	36.10%	14.30%	38.00%	17.00%	76.30%
SL-08	3.00%	8.60%	33.60%	12.10%	32.50%	11.70%	30.90%	12.90%	69.80%
SL-09	2.20%	7.40%	31.60%	8.30%	30.40%	8.40%	26.20%	11.80%	65.40%
SL-10	1.80%	6.30%	28.40%	8.60%	26.50%	7.70%	24.30%	9.70%	62.60%
SL-11	1.10%	4.80%	25.50%	6.60%	24.80%	6.70%	24.10%	7.70%	58.80%
SL-12	1.00%	4.20%	17.60%	4.10%	18.70%	5.70%	15.20%	6.00%	48.10%
SL-13	0.70%	5.10%	18.70%	3.80%	18.90%	5.50%	15.20%	6.30%	46.90%
SL-14	0.50%	2.40%	10.50%	2.60%	10.70%	3.80%	11.30%	4.50%	31.30%
BNC-ELT	14.14%	21.61%	46.91%	26.27%	45.62%	26.59%	45.32%	28.44%	73.45%

Table 7. Use of BNC-ELT words in ELT corpus (family level data).

Collection CLT-01 makes use of 85.00% of the families in sublist SL-01, 47.80% of the families in SL-02, 23.10% of those in SL-03, 14.20% of those in SL-04, and, beyond this point, from 7.90% to 0.50% of the families in the remaining sublists. In contrast, collection CLT-08, makes use of significantly more families in every one of the sublists in the BNC-ELT list, about double for SL-02 and SL-03, triple for SL-04 through SL-06, quadruple for SL-07 through SL-09, and so on.

When considering the use of actual types (words) from each sublist in the BNC-ELT list, the data shown in Table 8 becomes more uniform although it still correlates with trends seen in Table 6 and 7.

From the data shown in Table 3, we know that the amount of types per sublist decreases according to the relative frequency of the sublist. Sublist SL-01 contains 6,384 types and sublist SL-02 contains 5,593 types while sublists

	CLT-01	CLT-02	CLT-03	CLT-04	CLT-05	CLT-06	CLT-07	CLT-08	ELT Corpus
SL-01	26.54%	28.39%	48.72%	41.73%	52.03%	37.15%	53.04%	39.74%	71.88%
SL-02	12.05%	17.63%	40.78%	27.96%	44.06%	25.73%	44.09%	27.71%	68.34%
SL-03	6.24%	14.43%	38.90%	16.29%	40.85%	16.67%	32.79%	16.27%	65.80%
SL-04	3.76%	9.07%	28.57%	13.32%	29.44%	12.74%	28.97%	13.58%	57.99%
SL-05	2.33%	6.99%	23.12%	8.59%	23.72%	8.92%	22.62%	9.44%	50.55%
SL-06	2.12%	6.01%	20.89%	6.78%	20.78%	7.04%	18.20%	7.33%	46.94%
SL-07	1.28%	4.22%	15.77%	4.10%	16.44%	5.17%	14.52%	5.93%	39.24%
SL-08	1.07%	3.07%	14.10%	4.42%	14.25%	4.23%	12.06%	4.73%	36.00%
SL-09	0.95%	2.46%	13.90%	2.92%	13.57%	3.05%	10.43%	4.33%	33.77%
SL-10	0.63%	2.64%	12.68%	3.35%	12.43%	3.20%	9.89%	3.70%	32.57%
SL-11	0.47%	1.90%	10.95%	2.58%	11.42%	2.61%	9.81%	2.86%	29.35%
SL-12	0.39%	1.75%	8.14%	1.64%	9.27%	2.49%	6.50%	2.45%	24.65%
SL-13	0.33%	2.39%	9.27%	1.94%	9.93%	2.47%	7.09%	2.76%	25.85%
SL-14	0.22%	1.17%	5.44%	1.30%	5.48%	1.89%	5.53%	2.25%	17.44%
BNC-ELT	6.18%	9.66%	24.94%	13.34%	26.14%	12.70%	24.29%	13.57%	48.14%

Table 8. Use of BNC-ELT types in ELT corpus (type level data).

SL-13 and SL-14 contain 2,426 and 2,225 types, respectively. Resolving the percentages shown in Table 7, we see that 1,694 types from sublist SL-01 are used in collection CLT-01 while 2,523 types from the same sublist are used in collection CLT-08. Considering all 14 sublists together, collection CLT-01 uses 3,127 types from the BNC-ELT list and this number accounts for 99.64% of the collection (see Table 5). In contrast, collection CLT-08 uses more than double that amount - specifically 6,866 types - from the BNC-ELT list, this amount accounting for 97.42% of the collection. It is easy to see that collection CLT-08 uses a larger vocabulary than collection CLT-01.

As we mentioned, the interpretation of results must be done with caution as the relationship between the frequency of words and their "difficulty" is not necessarily unequivocal. With this in mind, we now take into account the results from all collections. The profiling information obtained does not show a uniform evolution from CLT-01 to CLT-08 (Tables 5, 7, and 8). For example, CLT-03 uses 24.94% of the types in the BNC-ELT list while CLT-8 uses a little more than half that amount (Table 8). When one takes into consideration that CLT-03 (*Family movie scripts*) makes use of a more varied vocabulary than CLT-08 (*New York Times* articles), our second proposition proves invalid. It couldn't be any other way, a New York Times article is almost necessarily more difficult for a learner than a family movie and while, this might not be so in particular instances, it is certainly so in general. Since these two collections contain not one but 19 scripts and 58 articles respectively, we must find reason for the counter-intuitive results.

The explanation and the reason for caution in interpretation reside in, at least, two observations: first, frequency is a measure of probability of occurrence; second, frequency can influence semantic precision and collocational variation. Regarding probability, the word *astronaut* has a frequency of less than one occurrence per million words (Leech et al., 2008), meaning that it has about 1/1,000,000 chance of occurring in a randomly chosen text (in contrast with the word *people* which has a chance of occurrence above 1/1,000). If a text (ELT or otherwise) of, say, 300 tokens in length contains several occurrences of the type *astronaut*, we know that this word is behaving abnormally, i.e. it is defying its probability of occurrence. Profiling allows us to automatically detect these divergences and as Nation (2006) points out, there is cause to consider if the appearance of infrequent types in a text might not imply they are 'topic words' or, simply, words of particular import. A text that includes several occurrences of the word *astronaut* is likely to be related, in some way or other, to space exploration.

The flip side of probability of occurrence is that it increases in tandem with the size of a text, that is, the larger the text the better are the chances that infrequent words appear. Simply put, the collection CLT-03 contains 325,581 tokens while CLT-08 contains 48,701 tokens, that is, CLT-08 amounts to 14.95% the size of CLT-03. And so, the probability of occurrence of infrequent words is much larger in CLT-03 than it is in CLT-08 even though the former collection is deemed to be easier for learners. The data shown in Tables 5, 7, and 8 reflects the effect of size.

Semantic precision and collocational variation generally go hand in hand. A word such as 'point' can be used as a noun and as a (transitive and intransitive) verb, each part of speech dictating different collocational partners. Furthermore, the word *point* has 46 different meanings (Webster's New World, 2006) and its highly polysemous nature implies a wide range of collocational relationships. In contrast, the word *astronaut* has but one single meaning and part of speech (Webster's New World, 2006) implying that its collocational complexity will be limited. Frequency-wise, the word *point* occurs 484 times per million words as a noun and 142 times per million as a verb, that is, the word *point* (SL-01) is over 600 times more frequent than the word *astronaut* (SL-11). Comparatively speaking, therefore, a higher frequency might imply a heavier learning burden. However, there are cases were the converse may also hold. The precision of meaning that infrequent words exhibit can also imply a greater degree of difficulty for a learner because of the fine distinction of meaning they might convey. Examples can be *fallacy* (SL-08), *assiduous* (SL-11), *adulation* (SL14), or *maladroit* (unlisted, 0.08 times per million) all of which have one or more approximate synonyms of relatively general meaning and applicability in more frequent

words, for example, *lie* (SL-01), *persistent* (SL-04), *praise* (SL-03), or *unskillful* (SL-02), respectively.

These observations have greatly simplified the notion of learning burden as it is outside the scope of the paper. The intention has been to show that profiling should not be used to determine the difficulty of a text in general or its vocabulary in particular unless the results are interpreted with caution. Granted, results of this investigation have shown that, given similarly sized texts, it is possible to use profiling to differentiate the extreme cases, i.e. the easiest from the most difficult.

Summing up, the BNC-ELT list provides detailed and exhaustive information about the lexical composition of the ELT corpus. Unlike the GSL, it leaves minimal uninformative gaps - regardless of the difficulty of the text - ranging from 0.46% to 2.58% for each collection and 1.47% for the entire ELT corpus of 1,157,493 tokens (Table 5). Results also make it possible to formulate two (weak) propositions: first, comparatively more difficult texts demonstrate a tendency to use a larger amount of infrequent vocabulary; second, comparatively more difficult texts demonstrate a tendency to use a wider, more varied vocabulary.

The discussion so far has shown that the extension of coverage provided by the BNC-ELT (over the GSL) is informative. We now turn to the first two sublists (SL-01 and SL-02) to see what to expect when migrating from the GSL to the BNC-ELT. As previously mentioned, Nation (2004) conducted a comparison of the 3,000 most frequent families in the BNC against the GSL and AWL. His results showed that much of the content of the lists was shared and that the coverage provided (of the corpora he employed) by each was quite similar. In general, the 2,000 most frequent families from the BNC provided marginally better coverage than the GSL with the exception of fiction texts, in which case, the coverage provided by the GSL was superior (again marginally) than that of the 2,000 families from the BNC.

CODE	CLT-01	CLT-02	CLT-03	CLT-04	CLT-05	CLT-06	CLT-07	CLT-08	ELT Corpus
SL-01	91.45%	82.23%	79.27%	80.04%	82.25%	76.18%	76.84%	76.94%	80.43%
SL-02	4.23%	7.01%	7.07%	9.59%	7.01%	10.28%	9.47%	9.35%	7.65%
GSL-01	87.25%	84.39%	78.04%	79.33%	83.59%	73.74%	75.32%	75.08%	80.02%
GSL-02	5.11%	6.60%	7.87%	5.24%	6.58%	7.92%	5.65%	5.96%	6.71%

Table 9. Coverage of the ELT corpus by the GSL and BNC-ELT first two sublists.

The results shown in Table 9 concur with Nation's. We used the first 2 sublists (SL-01: 1,000 families; 6,348 types. SL-02: 1,000 families; 5,593 types) from the BNC-ELT and the GSL (GSL-01: 998 families; 4,119 types. GSL-02: 988 families; 3,708 types). Overall coverage of the ELT corpus by both lists is strikingly similar despite the fact that the GSL-01 amounts to only 64.88% the size of the SL-01 and GSL-02 amounts to only 66.29% of the size of SL-02. This will not come as a surprise to those familiar with the origin and content of the GSL. It is a remarkably well-manufactured word list. Note that collections CLT-02 (short stories) and CLT-05 (Novels) are marginally better served by the GSL, again in accord with Nation's data.

In regards to range, it is of interest to see how these two lists (SL-01 + SL-02: 2,000 families; 11,941 types. GSL: 1,986 families; 7,827 types) work across collections. Table 10 shows the percentage of each word list that appears in all 8 collections (right-most column), only 7 collections, only 6 collections, and so on, until we are left with the percentage of words that do not appear in any collection (left-most column).

# of collections	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
SL-01 + SL-02	0.30%	0.60%	0.90%	2.05%	3.80%	7.45%	15.25%	24.70%	44.95%
GSL	0.76%	0.50%	1.21%	4.03%	5.74%	10.32%	12.99%	21.90%	42.55%

Table 10. Amount of words from each word list that appears in up to 8 collections (range).

Again, results are strikingly similar for both lists, especially when one takes into consideration that only 3.4% of SL-03 is found in all 8 collections, the trend continuing to decline sharply, 0.9% of SL-04, 0.3% of SL-05, 0.1% of SL-06, and zero beyond this point. It is with this data in mind that we echo Nation's comment regarding the difficulty of finding a replacement for the GSL in regards to a core vocabulary of general service.

Despite the agreement in coverage and range between the GSL and SL-01+SL-02, differences in content exist. Nation (2004) provided information from the GSL perspective, that is, showing how many of its word families could be found in the 3,000 most frequent BNC word families. His analysis also revealed, as we mentioned previously, that 80% of the AWL was also present among these 3,000 word families. Unsurprisingly, our analyses again reveal corresponding results, the only exception regarding the inclusion of academic vocabulary which in our data is lowered to 67.36% (the remainder of the AWL is present in the BNC-ELT but at lower frequency levels). It appears that the reorganization of the 14,000 word families according to the

spoken subcorpus of the BNC has, indeed, produced a word list more appropriate for ELT.

alright	Christ	kid	score
America	client	lad	Scotland
awful	county	London	switch
bet	Europe	minus	television
bloke	feed	non	thou
bother	France	okay	traffic
brilliant	Germany	pence	video
Britain	guy	pension	wee
budget	hell	quid	x
chap	Jesus	reckon	

Table 11. The 39 families from SL-01 not present in the GSL + AWL.

For those familiar with and migrating from the GSL, Table 11 shows the 39 families in SL-01 that are not found in the GSL or AWL. The underlined words are proper nouns that the GSL makers intentionally excluded as were words such as *bloke*, *chap*, *wee*, or *pence*, on account of lacking universality (Faucett et al, 1936). Differences increase when considering SL-02 in which 212 new families are introduced.

In conclusion, comparison of content and coverage between SL-01 + SL-02 and the GSL do not make a clear case as to which word list might be "better" in regards to the identification or isolation of a vocabulary of general service. However, the scope of the BNC-ELT is so much larger that there is no question that it provides a more detailed characterization and classification when used in vocabulary profiling. Moreover, from the perspective of the ELT practitioner, migrating from the GSL to the BNC-ELT does not involve a sacrifice of established expertise or practices as the GSL can be considered a sublist of the BNC-ELT in terms of content and application.

Bibliography

Brown, G. and Yule, G. 1983. Teaching the spoken language. Cambridge: *Cambridge University Press*.

Carter, R. and McCarthy, M. 1988. Vocabulary and Language Teaching. New York: *Longman*.

Coxhead, A. 2000. A new Academic Word List. *TESOL Quarterly*, 34, 2, pp. 213-238.

Ellis, N. 2002. 'Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition.' Studies in Second Language Acquisition 24/ 2: 143-188.

Faucett, L., H. Palmer, E.L.Thorndike, and M. West. 1936. Interim Report on Vocabulary Selection. London: *P.S. King and Son*, Ltd.

Gilmore, A. 2007. Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40, pp. 97-118.

Hyland, K. and Tse, P. 2007. Is there an "Academic Vocabulary"? *TESOL Quarterly*, 41, 2, pp. 235-253.

Leech, G., P. Rayson, and A. Wilson. 2001. Word Frequencies in Written and Spoken English. Harlow: *Pearson Education Limited*.

Lorge, I. 1949. The Semantic Count of the 570 Commonest Words. New York: Teachers College, Columbia University.

Mochizuki, M. and K. Aizawa. 2000. An affix acquisitional order for EFL learners: An exploratory study. *System*, 28 pp. 291-304.

Nation, P. and R. Waring. 1997. Vocabulary size, text coverage, and word lists. In Schmitt and McCarthy (eds.).

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

Nation, I.S.P. 2004. A study of the most frequent word families in the British National Corpus. In Bogaards and Laufer (eds.).

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

Sakata, N. 2007. How do Japanese EFL learners comprehend derivatives?: A qualitative analysis from the perspective of vocabulary expansion. *JACET Journal*, 45, pp. 15-29.

Webster's New World College Dictionary (4th ed.) 2006. Cleveland, OH: *Wiley Publishing, Inc*.

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

Xue, G. and Nation, I.S.P. 1984. A university word list. *Language Learning and Communication*, 3, pp. 215-229.