The Student Engineering Corpus: Analysing word frequency

Olga Moudraia
Lancaster University, UK

Abstract

This paper presents the findings of a computer-aided research that aimed to establish a frequency-based corpus of student engineering lexis. The Student Engineering Corpus reported here represents the engineering lexis encountered in the English-language textbooks in basic engineering disciplines that are compulsory for all engineering students regardless of their fields of specialization.

The Student Engineering Corpus contains nearly 2,000,000 tokens and over 18,000 types. On its basis, a word list of the most frequent engineering lexis was developed which is organised by word families consisting of over 1,200 word families that comprise about 9,000 words. The word-family-based approach to word list organization may help broaden the EFL / ESL learners' lexical base and, more importantly, raise their awareness of the lexical nature of language. The Student Engineering Word List can also serve as a lexical syllabus foundation of English for Engineering.

In addition, frequency and other statistical information may help teachers and learners make judgements about the importance of frequent usage and core meaning of individual vocabulary items. This paper shows the results of the content analysis and word frequency analysis of the Student Engineering data, in comparison with the COBUILD Bank of English Corpus and the written part of the British National Corpus (BNC). The analysis was concerned with the most frequent word forms in all three corpora, including the most frequent closed-class (grammatical) and open-class (content) word forms, as well as keywords and key concepts.

1. Introduction

Corpus-based studies of language data have become a norm in linguistics. Larger and larger language corpora are being developed, and a corpus containing less than a million words will be considered small now. Various professional corpora are coming into existence. One such project, the Corpus of Professional English (in collaboration between Professional English Research Consortium, Japan, and Lancaster University, UK) is underway, for example, and, when finished, it will consist of a 100-million-word database of English used by professionals in science, engineering, technology and other fields (http://www.perc21.org/cpe_project/index. html). Another one, a monitor engineering corpus of several million words representing the English used by engineers in over 355 professional engineering organizations, has been growing at the University of Aizu in Japan (Orr and Takahashi 2002).

However, parallel to the rapid development of large corpus studies, an interest in the analysis of small corpora has arisen (Chadessy and Roseberry 2001), especially in the area of language teaching where smaller corpora can be more useful than, for example, large, evergrowing, constantly changing monitor corpora. Particularly, smaller corpora are designed to represent the specific part of the language under investigation and are tailored to address the aspects of the language relevant to the needs of the learner. Importantly, well-designed specialist corpora provide practical examples of particular language uses that are of specific inerest / relevance to the language learner. Furthermore, they are more manageable allowing easier and faster access to language data. The examples of the smaller technical corpora designed for language learners are the Guangzhou Petroleum English Corpus of about 400,000 tokens (Qi-bo 1989) and the Hong Kong University of Science and Technology (HKUST) Computer Science Corpus of 1,000,000 tokens (James et al 1994).

The Student Engineering Corpus reported in this paper contains nearly 2,000,000 tokens and is similar in design to the HKUST Computer Science Corpus. It was built in order to establish a corpus of Student Engineering English that represents engineering lexis encountered in compulsory textbooks for engineering students regardless of their fields of specialization, and is quite unique in this respect. I also aimed to provide teachers and learners with a word list that could serve as the lexical syllabus foundation for Engineering English.

This paper presents the findings of my research establishing a frequency-based corpus of student engineering lexis that led to the development of a word list of the most frequent engineering lexis which is organised by word families consisting of over 1,200 word families that comprise about 9,000 words. The paper will also discuss the results of the content analysis and word frequency analysis of the Student Engineering data, in comparison with the COBUILD Bank of English Corpus and the written part of the British National Corpus (BNC).

2. The Student Engineering Corpus

2.1 Rationale behind the study

This project's goal was to develop a reliable lexical syllabus for the engineering students in order to meet the objectives of English teaching for Engineering at Walailak University in Thailand[1] where I had worked for nearly seven years. One of those objectives was to provide students with a solid basis for further study, or for entering careers which require use of English. During their studies, students are supposed to acquire a working knowledge of English - a practical skill valued in many of today's professions. This is a demanding task, and to cope with it, dependable teaching materials are needed. At Walailak University English teachers create teaching materials to suit the particular needs of their students.

We were also in a situation quite common in Southeast Asia: lectures in most subjects were delivered in a local language (Thai, in this case) while textbooks were in English. That is why, in order to build a representative corpus of Student Engineering English, I selected the English-language textbooks in basic engineering disciplines, such as Engineering Mechanics, Engineering Materials, Mechanics of Materials, Mechanics of Fluids, Thermodynamics, Electrical Engineering, Engineering Drawing, Manufacturing Process and Computer Programming that were compulsory for all engineering students at Walailak University regardless of their fields of specialization[2]. The major criterion for the selection was that those textbooks were recommended for the engineering students who had to read them in English.

2.2 Procedures

The stages in the project included gathering a text corpus, processing it onto computer, conducting the computer analysis of the material and building the word list. The processing of the text corpus onto computer ready for analysis required scanning and verifying the texts with the help of Optical Character Recognition (OCR) software[3]. The most laborious stage, however, was the analysis of the engineering lexis that involved lexical computing and frequency count.

The material was analysed with the help of *WordSmith Tools 2.0* software – an integrated suite of programs for examining words' behaviour in texts. The *WordList* tool was employed to generate lists of all the words or word-clusters in the corpus, set out in alphabetical and frequency order, while the *KeyWords* tool was applied to identify keywords and make a database of keyword lists enabling identification of *key keywords* – words that are most frequent over a number of files in the database. The *Concord* tool that produces concordances and finds collocates of the search word was used to differentiate between parts of speech (i.e. *use* as a noun and a verb), homonyms (i.e. *light* = *heavy* / *light* = *dark*), and different senses of the same word (i.e. *impress* meaning a) *press hard into a soft surface leaving a mark* or b) *have a favourable effect on somebody*). This material gave a corpus of about 2 million tokens and over 18,000 types (see Fig. 1).

****

Corpus	
	tokens - 1,986,595
	types - 18,203
	token / type ratio - 109.14
	type / token ratio - 0.0092
Word Families	
	entries in the word list - 1,260
	comprise types - 8,850
	minimum frequency - 0.005 %

Figure 1 Statistics on the Student Engineering Corpus

****

[1] The research on which this article is based was supported by a small grant # 970112 from the Walailak University Research Council.

[2] I would like to thank the Walailak University lecturers of Engineering and especially Asst. Prof. Dr. Rugchart Owasith for help with selecting textbooks.

[3] This step also had to deal with obtaining permission from the publishers for the electronic use of their texts. My acknowledgements go to *McGraw-Hill Australia* (permission issued on October 12, 1998), *McGraw-Hill Companies, Inc.* (December 1, 1998), *Brooks / Cole Publishing Company* (grant No. G-09857, November 17, 1998) and *Addison Wesley Longman Limited* (ref. AP/2743, November 25, 1998) for their understanding of the nature of the project and kind permission to store their texts in an electronic format in order to create a word list.

2.3 The structure of the corpus

The Student Engineering Corpus is composed of thirteen text files as presented in Figure 2.

N	Text File	Bytes	Tokens	Types	Type/ Token Ratio	Standardised Type/Token Ratio	Ave. Word Length
	Overall	11,694,812	1,986,595	18,203	0.92	9.85	4.64
1	Manufact.txt	1,764,178	290,782	10,082	3.47	13.72	4.86
2	Material.txt	1,444,793	232,743	7,056	3.03	10.49	4.98
3	Fluidmech.txt	1,307,973	220,666	5,333	2.42	9.15	4.67
4	Mechmat.txt	1,177,429	202,513	4,125	2.04	7.79	4.48
5	Elec.txt	983,672	167,394	5,626	3.36	10.27	4.62
6	Intofluidmech.txt	860,281	147,028	4,666	3.17	9.54	4.61
7	Dynamics.txt	795,910	142,446	3,205	2.25	7.07	4.38
8	Statics,meriam.txt	710,854	127,623	4,129	3.24	9.57	4.39
9	Statics,beer.txt	668,896	121,696	2,919	2.40	6.94	4.28
10	Chemi.txt	653,622	110,812	4,299	3.88	9.60	4.62
11	Graph.txt	486,152	80,804	5,034	6.23	11.55	4.80
12	Pascal.txt	466,756	77,242	3,124	4.04	8.54	4.73
13	Draw.txt	374,296	64,846	3,030	4.67	9.05	4.52

Figure 2 The structure of the Student Engineering Corpus

2.4 Annotation and tagging

The Student Engineering Corpus has not been annotated yet. The part-of-speech and semantic tagging of the corpus is being carried out at the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster using CLAWS (the Constituent Likelihood Automatic Word-Tagging System) for POS tagging and the SEMTAG program for semantic tagging.

Semantic tagging is a form of corpus annotation that assigns semantic field tags to the words in a corpus. For this task, we will use the SEMTAG program, part of the UCREL Semantic Analysis System (USAS), which was designed and used across a number of research projects over the last ten years at Lancaster University that led to the initial design and implementation of the tools for semantic analysis and their application in the domain of software engineering documentation with a web front end called Wmatrix (for details, see Rayson, Garside, and Sawyer 1999) that we will use for the content analysis of the Student Engineering Corpus.

However, prior to the actual semantic tagging of the Student Engineering Corpus, the current SEMTAG Lexicon will need to be enhanced with the specific lexis occurring in the Student Engineering Corpus that is not currently recognised by the software because of either missing in the Lexicon or having a special meaning in the Student Engineering Corpus which is different from the one in the Lexicon. These unrecognised words will be assigned proper semantic tags according to the established set of semantic fields and subsequently added to the Lexicon. At the current stage, over 6% of the lexis in the Student Engineering Corpus is not recognised.

2.5 Word list organization

The entries in the resulting Student Engineering Word List were lemmatized according to word families. The lemmatization process reduced the number of entries to about 7,700 that were treated according to the cumulative frequency of occurrence of the members of the word families, and the most frequent word families (with the sum total of 100, or 0.005%) were selected. As a result, over 1,200 most frequent word families comprising nearly 9,000 words were included in the Student Engineering Word List. Incidentally, the most frequent word family in engineering textbooks is *use*[4] (see Fig. 3).

As can be seen from Figure 3, the word family here is interpreted in the the most broad sense - in accordance with the Bauer's and Nation's (1993) level 7 of generalization which includes derived and inflected forms as well as compound words. Thus, the word entry *use* lists not only *use, uses, using, used* but also *useful, usable, user, reuse, unused, misuse, abuse, multiuse* and their derivatives, giving details on the 'sub-families' within a family, i.e. *misuse* within *use*. In most modern dictionaries these are separate entries, and an EFL / ESL learner may not necessarily notice the connection between them.

[4] Interestingly, *used* is the most frequent content word form in Kuo's (1999: 10) Corpus of Scientific Journal Articles of a similar size followed quite closely by *using* - the ninth most frequent content word form in his corpus.

****

#	Head word	Freq.	%	Words Joined

ABC order - 1186
Freq. order - 1
	Use	10,313	0.52	use (2784: *n* – 961, *v* – 1823), uses (262: *n* – 48, *v* - 214), using (2100), used (4538);
				useful (341), usefully (1), usefulness (7); useless (6);
				usable (22), useable (2);
				user (149), users (24), user's (2); usage (39);
				reuse (4: *n* – 3, *v* - 1), re-use (3: *n* – 1, *v* - 2), reused (5),
				reusable (7);
				unused (5 - adj), unusable (5);
				misuse (1 – *n*), misusing (1), misused (1);
				abuse (2: *v* – 1, *attrib* – 1);
				multiuse (1 – *attrib*), multi-user (1 – *attrib*)

Figure 3 *Use* - the most frequent word family in the Student Engineering Corpus

****

Moreover, the Student Engineering Word List provides some additional information that allows the learner to find out which word forms of a particular word are used more frequently than the others, what part of speech represented by a particular word form is encountered more commonly (i.e. *use* as a noun or *use* as a verb), and even compare different spelling of the same word (i.e. *usable* and *useable* or *reuse* and *re-use*). Besides, frequency and other statistical information may help teachers and learners make judgements about the importance of frequent usage and core meaning of individual vocabulary items. Figure 4 presents the one hundred most frequent entries listed by headwords – the base word or the most frequent word in the family.

****

N	Headword	Freq.	%	N	Headword	Freq.	%
1	use	10,313	0.52	51	area	2,827	0.14
2	force	9,247	0.46	52	plane	2,820	0.14
3	form	7,075	0.35	53	direction	2,784	0.14
4	flow	7,045	0.35	54	result	2,763	0.14
5	pressure	7,016	0.35	55	move / remove	2,751	0.14
6	show (*v*)	7,002	0.35	56	all	2,741	0.14
7	determine	6,896	0.34	57	follow	2,731	0.14
8	figure / configure	6,650	0.33	58	constant	2,719	0.14
9	section	6,404	0.32	59	unit	2,661	0.13
10	line	5,812	0.29	60	view	2,647	0.13
11	equation	5,771	0.29	61	fluid	2,639	0.13
12	point	5,236	0.26	62	know	2,609	0.13
13	angle	4,923	0.25	63	draw	2,603	0.13
14	act / react / interact / transact / counteract	4,666	0.23	64	operation	2,601	0.13
15	velocity	4,614	0.23	65	component	2,560	0.13
16	system	4,540	0.23	66	expression	2,528	0.13
17	value	4,484	0.23	67	beam	2,513	0.13
18	apply	4,327	0.22	68	end	2,484	0.12
19	problem	4,278	0.21	69	pipe	2,476	0.12
20	work	4,198	0.21	70	make	2,467	0.12
21	give	4,103	0.21	71	steel	2,429	0.12
22	axis	4,053	0.20	72	assume	2,424	0.12
23	stress	4,033	0.20	73	shear	2,409	0.12
24	material	4,014	0.20	74	case (= state)	2,351	0.12
25	center	3,992	0.20	75	find	2,343	0.12
26	length / long	3,890	0.19	76	diameter	2,341	0.12
27	part	3,867	0.19	77	obtain	2,341	0.12
28	surface	3,821	0.19	78	mass	2,337	0.12
29	solution (of a problem)	3,776	0.19	79	air / aero-	2,315	0.12
30	type	3,606	0.18	80	define	2,276	0.11
31	produce	3,582	0.18	81	also	2,267	0.11
32	metal	3,457	0.17	82	calculate	2,266	0.11
33	example	3,447	0.17	83	water	2,262	0.11
34	load	3,406	0.17	84	cut	2,258	0.11

35	other / another	3,371	0.16	85	element	2,254	0.11
36	time	3,299	0.16	86	rotate	2,250	0.11
37	high	3,252	0.16	87	maximum	2,246	0.11
38	energy	3,245	0.16	88	different	2,235	0.11
39	vary	3,232	0.16	89	change	2,205	0.11
40	number	3,216	0.16	90	equilibrium	2,183	0.11
41	temperature	3,119	0.16	91	structure	2,183	0.11
42	body	3,101	0.16	92	position	2,177	0.11
43	process	3,048	0.15	93	base / basic	2,172	0.11
44	chapter	3,016	0.15	94	write	2,167	0.11
45	moment	2,989	0.15	95	consider	2,154	0.11
46	machine	2,979	0.15	96	design	2,125	0.11
47	dimension	2,938	0.15	97	free	2,087	0.10
48	put	2,889	0.14	98	friction	2,086	0.10
49	placement	2,840	0.14	99	low	2,083	0.10
50	require	2,828	0.14	100	method	2,070	0.10

Figure 4 The one hundred most frequent word families in the Student Engineering Corpus

****

In organising the Student Engineering Word List by word families, I accord with J. Nattinger (Nattinger 1988: 69) who has suggested that words built about a particular root are gathered so that the associations among them can be seen, and even though the meanings of these words may be slightly different, clustering them will aid students in remembering their general meaning. P. Nation (Nation 1990: 17) has also pointed out that grouping words under headwords is an attempt to increase the coverage of high-frequency vocabulary; the implication is that learning a word involves learning its derived and inflected forms as well. Nation's statement echoes Carter and McCarthy's (Carter and McCarthy 1988: 44) who have noted that among other things, 'knowing a word' means knowing its underlying forms and derivations. I also believe that this approach may help broaden the EFL / ESL learners' lexical base and, more importantly, raise their awareness of the lexical nature of language[5].

As it was pointed out elsewhere (Nation 1990; Nation & Waring 1997), word lists can be useful in a number of ways – course designers may refer to them when considering the vocabulary component of a language course; teachers may use them to judge whether a particular word deserves attention or not, and whether a text is suitable for a class; and learners may use them as a checklist or even as a goal. The Student Engineering Word List can help English instructors in selecting the vocabulary component of a course for the engineering students or deciding whether a particular text is useful for a class, serving as a lexical syllabus foundation of English for Engineering.

3. Word frequency analysis

3.1 Comparison with the COBUILD and the BNC Written

The word frequency analysis of the Student Engineering data was carried out in comparison with the COBUILD Bank of English Corpus and the written part of the British National Corpus (BNC). The analysis (Fig. 5-7) was concerned with the most frequent word forms in all three corpora, including the most frequent closed-class (grammatical) and open-class (content) word forms. It has revealed, firstly, that the most frequent word forms in all three corpora – being mainly function words – concur (see Fig. 5). The Spearman's rank order correlation between the fifty most frequent closed-class word forms in the Student Engineering Corpus and the COBUILD Bank of English is .778 while in the Student Engineering Corpus and the BNC Written is .802, with both figures significant at the .01 level.

****

Student Engineering Corpus			COBUILD			BNC Written		
N	Word	%	N	Word	%	N	Word	%
1	the	8.50	1	the	5.58	1	the	6.43
2	of	4.19	2	of	2.60	2	of	3.11
3	a	2.84	3	to	2.51	3	and	2.70
4	and	2.72	4	and	2.37	4	to	2.60
5	is	2.43	5	a	2.21	5	a	2.18
6	in	2.07	6	in	1.83	6	in	1.95
7	to	2.06	7	that	1.04	7	is	0.99
8	for	1.08	8	is	0.93	8	that	0.99
9	are	0.88	9	it	0.92	9	was	0.94

[5] For a detailed overview of the lexical approach to foreign language teaching, see Moudraia 2001.

10	be	0.83
11	that	0.80
12	at	0.76
13	as	0.75
14	by	0.71
15	with	0.57
16	on	0.50
17	from	0.48
18	an	0.47
19	this	0.47
20	or	0.46
21	we	0.42
22	which	0.42
23	it	0.38
24	if	0.32
25	figure	0.31
26	flow	0.31
27	can	0.28
28	determine	0.27
29	force	0.27
30	two	0.26
31	shown	0.25
32	will	0.25
33	used	0.23
34	may	0.22
35	velocity	0.22
36	pressure	0.22
37	its	0.20
38	when	0.20
39	have	0.20
40	has	0.19
41	equation	0.19
42	not	0.19
43	one	0.18
44	each	0.18
45	point	0.18
46	where	0.18
47	system	0.17
48	forces	0.17
49	these	0.16
50	between	0.16

10	for	0.87
11	i	0.78
12	was	0.76
13	on	0.70
14	he	0.65
15	with	0.64
16	as	0.57
17	you	0.54
18	be	0.53
19	at	0.52
20	by	0.50
21	but	0.47
22	have	0.46
23	are	0.44
24	his	0.43
25	from	0.43
26	they	0.43
27	this	0.39
28	not	0.38
29	had	0.35
30	has	0.34
31	an	0.32
32	we	0.32
33	or	0.29
34	said	0.28
35	one	0.28
36	there	0.27
37	will	0.27
38	their	0.27
39	which	0.27
40	she	0.26
41	were	0.26
42	all	0.25
43	been	0.25
44	who	0.25
45	her	0.24
46	would	0.23
47	up	0.22
48	if	0.22
49	more	0.22
50	when	0.22

10	it	0.93
11	for	0.88
12	on	0.72
13	with	0.67
14	he	0.67
15	be	0.67
16	i	0.66
17	by	0.55
18	as	0.55
19	at	0.49
20	you	0.47
21	are	0.47
22	his	0.47
23	had	0.46
24	not	0.46
25	this	0.45
26	have	0.44
27	from	0.44
28	but	0.43
29	which	0.39
30	she	0.38
31	they	0.37
32	or	0.37
33	an	0.36
34	her	0.35
35	were	0.33
36	there	0.28
37	we	0.28
38	their	0.28
39	been	0.28
40	has	0.27
41	will	0.26
42	one	0.26
43	all	0.25
44	would	0.25
45	can	0.22
46	if	0.21
47	who	0.21
48	more	0.21
49	when	0.21
50	said	0.20

Figure 5 The fifty most frequent word forms in the Student Engineering Corpus, the COBUILD Bank of English Corpus and the BNC Written

****

Secondly, the comparison of the fifty most frequent open-class (content) word forms (Fig. 6) has indicated that the content word forms in the Student Engineering Corpus are predominantly from the scientific register while the most frequent content word forms in COBUILD and BNC Written are of a general nature.

****

Student Engineering Corpus			
N	**Rank**	**Word**	**%**
1.	5	Is	2.43
2.	9	are	0.88
3.	10	Be	0.83
4.	25	figure	0.31
5.	26	flow	0.31
6.	27	can	0.28
7.	28	determine	0.27
8.	29	force	0.27
9.	30	two	0.26
10.	31	shown	0.25
11.	32	will	0.25
12.	33	used	0.23
13.	34	may	0.22
14.	35	velocity	0.22
15.	36	pressure	0.22
16.	39	have	0.20

COBUILD			
N	**Rank**	**Word**	**%**
1.	8	is	0.93
2.	12	was	0.76
3.	18	be	0.53
4.	22	have	0.46
5.	23	are	0.44
6.	29	had	0.35
7.	30	has	0.34
8.	34	said	0.28
9.	35	one	0.28
10.	37	will	0.27
11.	41	were	0.26
12.	43	been	0.25
13.	46	would	0.23
14.	55	can	0.20
15.	58	new	0.16
16.	59	do	0.16

BNC Written			
N	**Rank**	**Word**	**%**
1.	7	is	0.99
2.	9	was	0.94
3.	10	that	0.99
4.	15	be	0.67
5.	21	are	0.47
6.	23	had	0.46
7.	26	have	0.44
8.	35	were	0.33
9.	39	been	0.28
10.	40	has	0.27
11.	41	will	0.26
12.	42	one	0.26
13.	44	would	0.25
14.	45	can	0.22
15.	50	said	0.20
16.	51	do	0.20

17.	40	has	0.19	17.	60	two	0.16	17.	61	could	0.16
18.	41	equation	0.19	18.	62	time	0.15	18.	64	time	0.15
19.	43	one	0.18	19.	63	people	0.15	19.	67	two	0.14
20.	45	point	0.18	20.	64	like	0.15	20.	70	may	0.14
21.	47	system	0.17	21.	68	now	0.15	21.	73	new	0.13
22.	48	forces	0.17	22.	71	year	0.14	22.	74	like	0.13
23.	51	surface	0.16	23.	75	first	0.13	23.	78	first	0.12
24.	52	energy	0.16	24.	76	could	0.13	24.	80	did	0.12
25.	53	stress	0.16	25.	81	last	0.12	25.	81	now	0.12
26.	54	section	0.15	26.	83	well	0.12	26.	83	people	0.11
27.	55	example	0.15	27.	85	years	0.11	27.	85	should	0.11
28.	57	line	0.14	28.	86	know	0.11	28.	86	very	0.11
29.	58	chapter	0.14	29.	89	very	0.10	29.	88	see	0.10
30.	60	use	0.14	30.	91	pound	0.10	30.	91	made	0.10
31.	63	temperature	0.13	31.	92	back	0.10	31.	93	back	0.10
32.	64	problem	0.13	32.	94	get	0.10	32.	94	way	0.09
33.	65	must	0.13	33.	95	may	0.10	33.	96	years	0.09
34.	66	given	0.13	34.	97	think	0.09	34.	97	being	0.09
35.	67	time	0.13	35.	98	even	0.09	35.	100	work	0.09
36.	68	body	0.12	36.	100	way	0.09	36.	107	make	0.08
37.	72	area	0.12	37.	101	right	0.09	37.	108	even	0.07
38.	73	constant	0.12	38.	102	three	0.09	38.	111	must	0.07
39.	75	value	0.12	39.	104	don’t	0.09	39.	112	own	0.07
40.	77	number	0.12	40.	106	world	0.09	40.	113	know	0.07
41.	78	solution	0.12	41.	110	being	0.09	41.	115	year	0.07
42.	79	fluid	0.12	42.	111	says	0.09	42.	116	good	0.07
43.	80	shear	0.12	43.	112	government	0.09	43.	119	last	0.07
44.	81	length	0.12	44.	114	dollar	0.08	44.	120	get	0.07
45.	82	moment	0.11	45.	115	should	0.08	45.	121	three	0.07
46.	84	mass	0.11	46.	116	made	0.08	46.	122	well	0.07
47.	85	axis	0.11	47.	117	good	0.08	47.	123	take	0.07
48.	86	maximum	0.11	48.	119	see	0.08	48.	125	go	0.07
49.	87	thus	0.11	49.	120	go	0.08	49.	126	government	0.07
50.	88	work	0.11	50.	121	did	0.08	50.	129	man	0.06

Figure 6 The fifty most frequent open-class (content) word forms in the Student Engineering Corpus, the COBUILD Bank of English Corpus and the BNC Written

****

However, most frequently encountered words in the Student Engineering Corpus appear to be *sub-technical*, i.e. words with non-technical as well as technical senses, common in most kinds of technical writing. Words in general non-technical sense appear to be more frequently used in the Student Engineering Corpus than the specialist terms, as can be seen in Figure 7. It seems to me an interesting finding worthy of notice. It also concurs with Lynn’s (Lynn 1973) observation about the absence of technical terms in his resulting specialized word list for commercial students.

****

Rank		Headword	Freq.	%
ABC order	Freq. order			
1032	29	**solution** (of a problem)	3,776	0.19
1033	242	**solution** (liquid)	1,025	0.05

Figure 7 Technical vs. non-technical senses

****

3.2 Keyword analysis

The keyword analysis of the Student Engineering data was carried out by using the *WordSmith Tools 2.0* software. To locate and identify keywords in a given text, the *KeyWords* tool compares the words in the text with a reference set of words taken from a large corpus of text that acts as a norm. Characteristically, keywords are not the most frequent words but the words which are most unusually frequent in a given body of text against the reference corpus. Another interesting feature provided by the *KeyWords* tool is the *key-keyword analysis* that allows us to see the most frequent keywords over a number of files in the database (and not just, for example, in one or two texts only) ensuring even dispersion. The key-keyword analysis of the Student Engineering data against the written part of the BNC Sampler has shown some interesting information on the key verbs in the Student Engineering Corpus – they appear to be predominantly from the academic register. The key verbs in the Student Engineering Corpus are as follows: *be, show, determine, use, require, obtain, apply, assume, calculate,* *correspond to, define, give, act, illustrate, occur, become, consider, exert, indicate, locate, sketch, solve,* and *substitute* (see Fig. 8).

****

N	Word	of 13 texts	as % of 13 texts
1	is / are	13 / 10	100.00 / 76.92
2	shown	12	92.31
3	determine	12	92.31
4	used / using / use	11 / 9 / 5	84.62 / 69.23 / 38.46
5	required	10	76.92
6	obtained / obtain	9 / 6	69.23 / 46.15
7	applied	7	53.85
8	assume	7	53.85
9	calculate	7	53.85
10	corresponding (to)	7	53.85
11	defined	7	53.85
12	given	7	53.85
13	acting	6	46.15
14	illustrates	6	46.15
15	occurs	6	46.15
16	becomes	5	38.46
17	consider	5	38.46
18	exerted	5	38.46
19	indicated	5	38.46
20	located	5	38.46
21	sketch	5	38.46
22	solve / solving	5 / 5	38.46 / 38.46
23	substituting	5	38.46

Figure 8 Key-key verbs in the Student Engineering Corpus compared against the BNC Sampler Written

****

This finding has an important implication for teaching Engineering English – more attention in the ESP classrooms should be given to non-technical than technical vocabulary. It agrees with Hutchinson and Waters' (Hutchinson & Waters 1981: 66) opinion who have challenged the assumption that technical vocabulary is the most useful for ESP students and arrived at the conclusion that students of ESP require not a corpus of technical language but the ability to mobilize the resources of general English in the solving of technical problems, and that technical English is only a development or extended application of general English while a wide-ranging knowledge of everyday vocabulary and the ability to mobilize this knowledge in the interpretation of technical discourse are important aids to comprehension and memory.

4. The semantic profile of the Student Engineering Corpus

As I have mentioned above, the semantic tagging of the Student Engineering Corpus is being carried out with the help of the SEMTAG program, part of the UCREL Semantic Analysis System (USAS), which was developed at Lancaster University, and the content analysis of the Student Engineering Corpus is being performed using a web front end called Wmatrix.

The semantic tagset employed by SEMTAG was loosely based on the major categories in McArthur's Longman Lexicon of Contemporary English (McArthur 1981) and refined by Andrew Wilson (available online: http://www.comp.lancs.ac.uk/ucrel/usas/semtags.txt). It has a multi-tier structure with 21 major discourse fields, subdivided, and with the possibility of further fine-grained subdivision in certain cases, for example:

T1 Time
T1.1 Time: General
T1.1.1 Time: General: Past
T1.1.2 Time: General: Present; simultaneous
T1.1.3 Time: General: Future
T1.2 Time: Momentary
T1.3 Time: Period
T2 Time: Beginning and ending
T3 Time: Old, new and young; age
T4 Time: Early/late

Wmatrix allows us to see significant concepts in the corpus and the words related to those concepts in frequency order. The Wmatrix semantic frequency list has a very useful option 'compare to normative BNC IT' which compares the concept frequencies against a subcorpus of the BNC that has been semantically tagged. The BNC IT corpus contains 135 files selected from the pure and applied science section of the BNC which are related to Information Technology (IT). Collectively, the files form a corpus of 1.7 million words, of which about 60% are news stories relating to IT.

The results of the key concept comparison against BNC IT are displayed in Wmatrix with the most significant key items towards the top of the list since the results are sorted on the LL (log-likelihood) field which shows how significant the difference is. The items with a '+' code signify overuse in the given text as compared to the standard English corpora. To be statistically significant, the LL value should be over 6.63 which is the cut-off for 99% confidence of significance (p<0.01).

Figure 9 illustrates the most statistically significant data on the overused semantic categories in the Student Engineering Corpus produced in comparison between the Student Engineering semantic frequency list and BNC IT. However, these results are not extremely accurate as 127,870 lexical items (i.e. 6.4% of the lexis) are not currently recognised by SEMTAG. The current SEMTAG Lexicon will need to be enhanced with this specific lexis occurring in the Student Engineering Corpus to be able to assign proper semantic tags to the as yet unmatched lexis according to the established set of semantic fields.

****

Item	Frequency	LL	Semantic category
N1	74870	+12844.38	Numbers
O4.4	21627	+10704.81	Shape
N3.5	7165	+ 7287.83	Measurement: Weight
Z5	679652	+ 7254.53	Grammatical bin
N2	11061	+ 7152.46	Mathematics
O2	53221	+ 6810.75	Objects generally
O1.2	6883	+ 6703.00	Substances and materials: Liquid
O1	8707	+ 6294.31	Substances and materials generally
N3.7	7675	+ 4440.78	Measurement: Length & height
O4.6	3259	+ 3805.63	Temperature
O1.3	3667	+ 3679.58	Substances and materials: Gas
E3-	5407	+ 3102.38	Calm/Violent/Angry
E6-	4666	+ 2762.14	Worry, concern, confident
X6+	6847	+ 2715.36	Deciding
B1	8780	+ 2664.81	Anatomy and physiology
M6	23444	+ 2541.75	Location and direction
O4.6+	2975	+ 2318.45	Temperature
M5	3671	+ 1880.51	Aircraft and flying
N3.2	3734	+ 1758.18	Measurement: Size
O3	7581	+ 1739.22	Electricity and electrical equipment
Z99	*127870*	*+ 1737.24*	*Unmatched*
G3	3934	+ 1600.35	Warfare, defence and the army; weapons
A1.7+	4324	+ 1387.76	Constraint
T3	3171	+ 1371.41	Time: Old, new and young; age
A6.1+++	4155	+ 1352.22	Comparing:- Similar/different
N3.3+	2642	+ 1237.55	Measurement: Distance
T1.2	4270	+ 1206.39	Time: Momentary
M8	1867	+ 1008.59	Remaining/stationary
N5.1-	7570	+ 917.77	Entirety; maximum
N3.8	2039	+ 854.25	Measurement: Speed
A6.3+	4361	+ 835.41	Comparing: Variety
X5.2+	4037	+ 814.79	Interest/boredom/excited/energetic
B5	2923	+ 800.28	Clothes and personal belongings
O4.6-	960	+ 740.58	Temperature
A10+	13370	+ 739.95	Open/closed; Hiding/Hidden; Finding; Showing
F2	1324	+ 678.71	Drinks
N3.1	2767	+ 657.03	Measurement: General
A3+	44725	+ 640.90	Being
O4.5	2919	+ 626.17	Texture
A12-	5182	+ 598.44	Easy/difficult
A1.2+++	760	+ 591.31	Suitability
O4.3	3695	+ 517.15	Colour and colour patterns
N6	1375	+ 492.99	Frequency etc.
N5	22374	+ 489.55	Quantities
A1.6	1060	+ 484.44	Physical/mental
B4	921	+ 476.94	Cleaning and personal care
A4.1	16529	+ 475.69	Generally kinds, groups, examples
A2.2	13157	+ 453.21	Affect: Cause/Connected
S1.2.5+	1307	+ 443.30	Toughness; strong/weak
W3	3929	+ 442.86	Geographical terms

N5+	11644	+ 439.62	Quantities
B2-	1345	+ 428.03	Health and disease
A6.2+	6947	+ 399.49	Comparing: Usual/unusual
N3	338	+ 398.00	Measurement

Figure 9 Comparison between the Student Engineering semantic frequency list and BNC IT

****

5. Further development

This paper has reported on the creation of a Student Engineering Corpus that was originally developed at Walailak University, Thailand. The project had three primary aims: a) to establish a representative corpus of Student Engineering lexis regardless of the fields of specialization; b) to provide teachers and learners with a word list that could serve as the lexical syllabus foundation of English for Engineering; and c) to explore the data for the linguistic analysis of the syntactical, morphological, lexical, and discursive features of Engineering English. The first two aims have been accomplished although the corpus still needs to be annotated, and proper part-of-speech and semantic tagging has to be carried out. The third aim, however, is a long-term. The empirical evidence from the part-of-speech and semantically tagged corpus of Student Engineering English will provide the basis for research into syntax, morphology, vocabulary, and discourse of Engineering English. Tagged data will be beneficial for the studies of semantic fields and grammatical categories. Subsequently, the material is expected to produce valuable information relevant to wide-ranging linguistic analysis.

References

Bauer L, Nation P 1993 Word families. *International Journal of Lexicography* 6(3): 1-27.

Carter R, McCarthy M 1988 *Vocabulary and language teaching*. Harlow, Longman.

Chadessy M, Henry A, Roseberry R L (eds) 2001 *Small corpus studies and ELT: theory and practice.* Amsterdam / Philadelphia, John Benjamins Publishing Co.

Hutchinson T, Waters A 1981 Performance and competence in English for specific purposes. *Applied Linguistics* 2(1): 56-69.

James G, Davidson R, Heung-yeung A C, Deerwester S 1994 *English in Computer Science: a corpus-based lexical analysis*. The Hong Kong University of Science and Technology, Longman Asia Ltd.

Kuo C H 1999 Can numbers talk? Basic data management of a corpus. *RELC Journal* 30(1): 1-17.

Lynn R W 1973 Preparing word lists: a suggested method. *RELC Journal* 4(1): 25-32.

McArthur T 1981 *Longman lexicon of contemporary English.* London, Longman.

Moudraia O June 2001 Lexical approach to second language teaching. *Eric Digest* EDO-FL-01-02. Washington, DC, ERIC Clearinghouse on Languages and Linguistics (http://www.cal.org/ericcll/digest/0102lexical.html).

Nation I S P 1990 *Teaching and learning vocabulary*. Boston, Heinle and Heinle Publishers.

Nation P, Waring J 1997 Vocabulary size, text coverage and word lists. In Schmitt N, McCarthy M (eds), *Vocabulary: description, acquisition and pedagogy*. Cambridge, Cambridge University Press, pp. 6-19.

Nattinger J 1988 Some current trends in vocabulary teaching. In Carter R, McCarthy M, *Vocabulary and language teaching*. Harlow, Longman, pp. 62-82.

Orr T, Takahashi A 2002 Constructing a corpus of fundamental engineering English for nonnative speakers. In Williams J(ed), *Proceedings of the IEEE International Professional Communication Conference*. Oregon, USA, pp. 403-409.

Qi-bo Z. April 1989 A quantitative look at the Guangzhou Petroleum English Corpus. *ICAME Journal* 13: 28-38.

Rayson P, Garside R, Sawyer P. May 1999 *Language engineering for the recovery of requirements from legacy documents*. REVERE project report, Lancaster University.