รหัสโครงการ SUT 2-203-47-12-39

รายงานการวิจัย

A vocabulary syllabus for five engineering disciplines

หัวหน้าโครงการ

Assoc. Prof. Dr. Jeremy Ward

สาขาวิชาภาษาอังกฤษ

สำนักวิชาเทคโนโลยีสังคม

ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี ปีงบประมาณ 2547

ผลงานวิจัยเป็นความรับผิดชอบของหัวหน้าโครงการวิจัยแต่เพียงผู้เดียว

มิถุนายน 2552

ABSTRACT

The purpose of this project is the establishment of a vocabulary syllabus for five engineering subdisciplines. The first stage of the project involved the creation of a 250000 word corpus of textbook material in five engineering disciplines. This material was then analyzed to identify the most frequent items. It was also established, incidentally, that the degree of specialisation in vocabulary was such as to necessitate such disciplinary-specific lists. The second stage involved analyzing the words to establish what aspects of vocabulary knowledge were especially salient in this type of textual material. It was found, upon examination of a sample from the corpus, that collocation was by far the most frequent feature of the many common words. A sample of a collocationary syllabus for chemical engineering was prepared as an example for the other four disciplines.

CONTENTS

SECTION 1. INTRODUCTION 5

SECTION 2. OBJECTIVES 5

SECTION 3. METHODOLOGY 6

3.1 (Objective 2.1) Establish a corpus of 250000 words in five engineering disciplines 6

3.2 (Objective 2.1) Create wordlists for each of the five disciplines. 6

3.3 (Objective 2.2) Establish important aspects of word knowledge for engineering textbooks 6

3.4 (Objective 2.3) To refine the syllabus in the light of objective 2.2 6

SECTION 4. RESULTS AND DISCUSSION 7

4.1 Objective 2.1: Creation of the wordlists 7

4.2 Objective 2.2: Aspects of vocabulary knowledge 12

4.3 (Objective 2.3) Refining the lexical syllabus 21

SECTION 5: A COLLOCATIONAL APPROACH 42

5.1 Collocation and the choice of texts: definitional texts 42

5.2 Item selection 45

5.3 A collocation list for chemical engineering 51

REFERENCES 55

APPENDICES 57

Tables

Table 1: number of content words in the corpus of different ranges 7
Table 2: common words with range of 5, hapax legomena in certain subcorpora 8
Table 3: common words with range of 4, non-occurring and hapax legomena in certain subcorpora 8
Table 4: peakratio for words of different ranges 9
Table 5: coverage by listCHE of 10000-word and longer texts 10
Table 6: type-token ratios and infrequent words in five subdisciplines of corpuseng 12
Table 7: distribution of rate over corpuseng 13
Table 8: distribution of *rate* and subforms over corpusche 13
Table 9: gas (g) collocations in corpustxt 24
Table 10: frequency and proportion of collocations in corpustxt 25
Table 11: *liquid*'s (l) collocations in corpustxt 26
Table 12: *heat* (h) collocations in corpustxt 26
Table 13: collocation data for other common words in corpustxt 27
Table 14: distribution of three words across corpuseng 28
Table 15: distribution of four other words across corpuseng 29
Table 16: collocates' distributions over corpuseng 29
Table 17: comparative definitions given by a layman and an expert 40
Table 18: text 1 collocations' occurrence in other corpora 45
Table 19: *heat* collocates in corpusche and corpustxt 47
Table 20: *gas* collocates in corpusche and corpustxt 47
Table 21: *temperature* collocates in corpusche and corpustxt 48
Table 22: most frequent corpusche collocations' occurrence in corpustxt 49
Table 23: other middle-frequency collocations in corpusche and corpustxt 50

Section 1 Introduction

In their 3rd and 4th years of study it becomes increasingly important for Thai engineering undergraduates to have access to academic material in English, as the amount of Thai language material available decreases. With their somewhat low level of competence in English, it is suggested that vocabulary is the first problem that needs to be dealt with. A previous project (Ward 1999) established a word list for foundation engineering students, based on textbooks used in their second year of study. This project showed that, contrary to the established wisdom (e.g Laufer 1998), a vocabulary of 2000 base words was sufficient for foundation engineering students. However this list has various weaknesses as far as more advanced students are concerned. These are, briefly, that the list fails to account for much of the specialisation present in the 3rd/4th year textbooks, and that in any case SUT students cannot, for the most part, attain knowledge of 2000 word families during the time that they study foundation engineering. Ward (2004) showed that many of them know only half that number and a learning target of 1000 words in a year is unrealistic. Teaching the list would therefore be largely futile: students would have moved into their specialist field long before they mastered it. It is therefore desirable to establish a specialist vocabulary syllabus for each separate engineering discipline.

Section 2 Objectives

2.1 To establish a vocabulary syllabus for five engineering disciplines - chemical, civil, electrical, industrial and mechanical.

2.2 To investigate further what type of knowledge is necessary for the words in the list

2.3 To refine the syllabus in the light of objective 2.2

Section 3 Methodology

3.1 (Objective 2.1) Establish a corpus of 250000 words in five engineering disciplines

This was done by analysing a random sample of about 10000 words from each of 25 engineering textbooks used in 3rd/4th year courses at SUT. Five textbooks were chosen from each of the five disciplines mentioned in section 2.1. These books were chosen in consultation with engineering faculty members. Permission was obtained from publishes to scan a certain number of pages from each book, and the scanned files were then edited and abridged to a length of 10000 words from each book.

The 250000-word advanced engineering corpus will be referred to as corpusENG. The chemical engineering part of it, 50000 words in size, will be referred to as corpusCHE. An additional corpus was established to check the findings from corpusENG and corpusCHE, using three book-length texts in chemical engineering; this will be referred to as corpusTXT, consisting of 387000 words.

3.2 (Objective 2.1) Create wordlists for each of the five disciplines.

CorpusENG was analysed using Wordsmith Version 2.0, producing word frequency lists for the five disciplines. The wordlist from corpusCHE was lemmatized for further analysis.

3.3 (Objective 2.2) Establish important aspects of word knowledge for engineering textbooks

A random sample of 20 common words was taken from corpusENG and concordances were examined to establish what it was about about each word that was important for students to learn. The framework for doing this was Liu & Shaw's (2001) list of qualitative aspects of word knowledge, focusing on polysemy (multiple meaning), collocation, and grammatical behaviour (including derivations and inflections).

3.4 (Objective 2.3) To refine the syllabus in the light of objective 2.2

In the light of the findings in 3.3, corpusTXT was examined to establish the frequency (or otherwise) of a certain type of collocation. The collocation patterns of the most common nouns were established and a collocationary syllabus, based on frequency in the corpus, was drawn up for chemical engineering.

Section 4 Results and discussion

4.1 Objective 2.1: Creation of the wordlists

4.1.1 Specialisation

Specialisation appears from the corpus data to be a key feature of the engineering subdisciplines, as shown by the narrow distribution of many comon words across the five sub-corpora of corpusENG; this justifies the creation of the word lists. A simple measure of distribution (or dispersion) was used called PEAKRATIO, from Yang (1998). This was calculated in the following way. We take the five 50000-word subcorpora which make up corpusENG, then divide the maximum frequency (in any of the five subcorpora) by the mean frequency over all of them. A word, for example, with a distribution of 10-2-1-2-1 over the five subcorpora has a maximum frequency of 10, and a mean frequency of 3.2 (total 16 divided by 5): this gives a PEAKRATIO of 10/3.2, or 3.13. A word with a flattish distribution of 5-2-3-4-2 gives a PEAKRATIO of 5/3.2, or 1.56. The lower the figure, the flatter the distribution. With five subcorpora, the maximum PEAKRATIO value is 5.0, as for example in a 5-0-0-0-0 distribution; and the minimum is 1.0, since the maximum equals the mean (e.g. 1-1-1-1-1).

For an overview of distribution let us take all the content words in corpusENG which occur at a frequency greater than 9. For convenience's sake we will consider word types rather than word families. It is not considered worthwhile to lemmatise the whole of corpusENG as we are not going to use it to create one single wordlist.

There are 2312 word types in the corpus with frequency greater than 9. Table 1 below shows the range of these word types.

Table 1: Number of content words in the corpus of different ranges

Range	No of types in corpus
1	233
2	256
3	346
4	519
5	958

Less than half of them (i.e. the 958 with a range of 5) occur in all subdisciplines. Of course this proportion would decrease substantially if we also included words at frequency lower than 10, but this may be largely a function of corpus size. More than 1300 do not occur in all subdisciplines, and each of these represents some kind of evidence in favour of a specialised syllabus (being redundant in at least one subdiscipline).

But the range statistics give only a crude impression. We would be misled if we considered as typical words like *divide,* which is distributed 3,1,2,3,1 over the five subcorpora (PEAKRATIO = 3.33) and obviously relates to all different engineering subdisciplines. A closer look shows that there are plenty of words with a range of 5 where the distribution is very skewed. If we look among the most common 100 words in corpusENG we find the following words, all with a range of 5, which occur only once in at least one subcorpus. In all, no less than 350 of words with a range of 5 are hapax legomenon (occurring only once) in one or more subdisciplinary corpora. In the selection shown in Table 2 below, we see that many also tend to congregate in two of the subcorpora.

Table 2: Common words with range of 5, hapax legomena in certain subcorpora

word	range	freq	chem	civil	elect	indu	mech	PEAKRATIO
load	5	537	7	265	138	1	126	2.47
beam	5	404	1	222	1	5	175	2.75
power	5	352	33	1	272	15	31	3.87
shear	5	337	4	195	1	1	136	2.89
energy	5	281	98	12	120	1	50	2.14
velocity	5	239	82	4	11	1	141	2.95
flux	5	222	42	1	171	3	5	3.85
strength	5	216	1	173	3	22	17	4.00
reaction	5	201	189	8	1	1	2	4.71
phase	5	178	58	1	108	2	9	3.03

The phenomenon naturally is even more marked among words with a range of 4: consider ten of the most common words of this range, in Table 3 below. Note especially the PEAKRATIO values in the last column.

Table 3: Common words with range of 4, non-occurring and hapax legomena in certain subcorpora

word	range	freq	chem	civil	elec	indu	mech	PR
stress	4	507	6	253	0	1	247	2.49
levels	4	125	1	8	9	107	0	4.28
service	4	108	7	44	0	56	1	2.59
soil	4	100	1	97	0	1	1	4.85
moments	4	97	0	73	3	1	20	3.76
tension	4	83	2	64	0	1	16	3.86
compression	4	82	2	74	0	1	5	4.51
observations	4	82	2	1	3	76	0	4.46
vector	4	74	1	6	21	0	46	3.11
event	4	73	3	1	1	68	0	4.66

All the above are among the most common 50 words with a range of 4. They are practically absent from at least two of the five subdisciplinary corpora.

Of course the words with a range of 5 tend to have lower PEAKRATIO values, reflecting flatter distributions. Table 4 below shows the number of tokens (running words) accounted for by words in each range, and the average PEAKRATIO values of those words. The right-hand column gives examples, just for illustration, of the distributions expressed by the mean PEAKRATIO values.

Table 4: PEAKRATIO for words of different ranges

Range	no. of wd types	no. of wd tokens	Mean PEAKRATIO	Sample average distributions
1	233	5002	(5.00)	(100,0,0,0,0)
2	256	6608	4.09	9,0,0,2,0
3	346	10366	3.27	8,1,3,0,0: 0,0,12,1,5
4	519	17899	2.68	16,0,2,7,28: 1,2,0,11,11
5	958	74124	2.06	1,11,5,3,9: 27,22,56,11,12

Subdisciplinary specialisation is clearly a substantial factor.

The dictionary tells us that engineering is

> The work done by, or the occupation of, an engineer; the application of science for directly useful purposes, such as construction, propulsion, communication, or manufacture. [1]

If we leave out the examples, the definition above is rather vague, and one also notes that engineering students study core engineering for only two terms, while they study their specialist subdisciplines for about seven. Also, consider the proliferation of engineering schools and subdisciplines (of which SUT provides a good example). It is perhaps inevitable that the more technology advances, the more these schools find to divide themselves from each other. Looking at just the lexical aspect, we find that there are only 300 content words (word types, not lemmas or families) that occur more than four times in all the subcorpora. A stratified sample of these words (freq>4) is shown below. By distributional criteria this is an engineering vocabulary par excellence, but the list is perhaps surprisingly lacking in what the laymen might think of as an engineering "flavour", especially if we exclude the items in the first line.

---

[1] Excerpted from *Oxford Talking Dictionary*. Copyright © 1998 The Learning Company, Inc. All Rights Reserved.

system section rate process data following method surface large applied length average normal considered usually necessary position limit initial taken single well equivalent yields calculated appropriate cases properties characteristic period designed presented variation elements require reduce practical calculations close provides steps exist

All this suggests, albeit tentatively, that lexically speaking what separates engineers may be more salient than what unites them: that a specialised approach appropriately reflects the content of the discipline.

4.1.2 Text coverage

The wordlist derived from corpusCHE, which we will refer to as listCHE, was lemmatised and Paul Nation's RANGE software was used to establish what coverage the list gives of chemical engineering text. Naturally, since the list was created from corpusCHE, it gives good coverage of corpusCHE: hapax legomena and twice-occurring words account for 6.1% of the tokens in corpusCHE, so the words at a frequency >2 actually account for 93.9% of the corpus; but when we include into the families the low-frequency subforms like *adds, adjusting*, the coverage of corpusCHE by the listCHE rises to 95%. This list would thus be quite adequate if the only material that students were required to read was that of corpusCHE; but how does this list fare against other chemical engineering material?

First we test the new list against a different 50000-word corpus (corpusCHE2) derived from five 10000-word extracts from textbooks supposedly on the same subtopics as corpusENG. In other words, the content base of the two corpora is similar; one might therefore expect a list derived from one (corpusCHE) to give good coverage of the other (corpusCHE2). In fact the token coverage is of corpusCHE2 91.2%. This is not so far from the 93.9% coverage that the list gave of corpusCHE.

The evidence below suggests that this 91.2% figure is reasonably reliable. Two of the 10000-word extracts in corpusCHE2 are from books (referred to as Chembk1 and Chembk2) of which nearly the entire text is available for examination. The coverage by the chemeng list of the two 10000-word extracts is very similar to its coverage of the longer texts from which the two 10000-word samples are taken. The results of this comparison are shown in table 5 below.

Table 5: Coverage by listCHE of 10000-word and longer texts

	Chembk1	Chembk2
coverage of 10000-wds (%)	90.6	90.5
coverage of full texts (%)	90.8	90.1

The upper row shows the list's coverage of the two 10000-word extracts; the lower row its coverage of the longer texts. Any discrepancy between the figures in the two rows would suggest that the smaller texts were unreliable as predictors of coverage, but no such discrepancy exists.

We thus conclude that listCHE probably does give about 91.2% coverage of any chemical engineering text *as long as it is on the subtopics represented in* corpusCHE.

ListCHE, being based on a comparatively small corpus, lacks all types of vocabulary. For example in its coverage of corpusCHE2, which supposedly covers the same topics, it misses the following:

bifurcation hexane flowsheet pentane phase-plane heptane algorithm glycol cumene flowrates plug-flow substrate ammonia eigenvalue eigenvalues ethyl molal convective hysteresis propane eigenvector heuristics polynomial acetaldehyde butanol downcomers eigenvectors jacobian linearly low-viscosity newtonian numerator steady-states stoichiometry acetate annulus bubble-point cyclohexane der interfacial ordinate prandtl subspace

all of which occur at frequency >4. These are by any measure technical terms. It is with technical terms that one would expect the list to give especially good coverage of corpusCHE2, since the topic base is supposed to be the same. On the other hand if we consider words at the other end of the range of difficulty, words which are completely non-technical, the following words are examples of those frequent (frequency >4) in corpusCHE2, which are not covered by the list.

notice understand guess cold growth concepts obviously significantly objective watch decide little qualitative strategies wire dominate explicit investment lowest match phenomena plan calculator characteristic extremely indeed past roughly tear affects connected demands goal learn leave likely manufacture map million realize stop together

These words and the technical words suggest that 50000-word corpora like corpusCHE have various shortcomings for creating a comprehensive wordlist for chemical engineers. This is perhaps slightly less so, however, in the case of the other subdisciplines. This is because chemical engineering introduces new vocabulary at the fastest rate of all the five subdisciplines represented in the corpus. Table 6 below shows that chemical engineering introduces proportionally more word types than other engineering subdisciplines. The ratio is called the type-token ratio - the number of word types divided into the number of running words (tokens).

Table 6: Type-token ratios and infrequent words in five subdisciplines of CorpusENG

	tokens	types	type/ token	types fre>2	types fre<3	% of missing tokens
chem eng	50671	4262	8.41	1875	2385	6.1
civi eng	51775	3397	6.56	1722	1675	4.1
elec eng	57882	3741	6.46	1819	1922	4.3
indu eng	56224	4003	7.12	1898	2105	4.8
Mech	54630	3866	7.08	1895	1971	4.6

This phenomenon is extremely marked in the table above, with the full 50000-word corpora. CorpusCHE has fewer, in some cases 15% fewer, tokens than other subdisciplines but more, in some cases as much as 25+% more, word types. So while listCHE, by selecting only words with a frequency greater than 2, misses 6.1% of tokens in corpusENG (the reader may recall this figure from the beginning of this section), corresponding lists from other four subdisciplines would miss only between 4.1 and 4.8%, as shown in the right-hand column. This result of this would be better coverage of the four (non-chemical engineering) corpora by the corresponding wordlists. These (non-chemical engineering) wordlists would also be roughly the same length as listCHE (see the column "types fre>2" above) i.e. short enough to be perhaps manageable by SUT students.

4.2 Objective 2.2: Aspects of vocabulary knowledge

We have established that the 50000-word corpora may be used to create wordlists for the five subdisciplines and that such wordlists would provide reasonable, though not excellent, coverage of engineering textbook material. We now turn to the question of what aspects of vocabulary knowledge are important for engineering undergraduates. We begin with the example of *rate*.

4.2.1 *Rate*

Rate is a high frequency word in corpusENG, occurring 413 times, or roughly once every two or three pages on average. The *rate* family, including *rates, rated, rating* and *ratings*, occurs 605 times. The distribution of the baseword *rate*, shown in Table 7 below, indicates that it is useful for students of all subdisciplines, but in very differing degree.

Table 7: Distribution of *rate* over CorpusENG

	range	total freq	Che eng	Civ eng	Ele eng	Ind eng	Me
rate	5	413	246	8	18	89	

More than half the occurrences are in the corpusCHE while the word occurs only eight times in civil engineering[2]. We might suppose from this that learning *rate* was a matter of great urgency for the chemical engineering student, but less for the civil engineering student. This is information that will not be found from the foundation engineering wordlist. It represents another argument for specialisation.

As far as what *rate* means is concerned, the situation for engineering students is far less daunting than it might appear. In English generally it is capable of carrying any of a number of meanings: Collins COBUILD Essential English Dictionary distinguishes between *rate* as the *speed* and *rate* as the *frequency* with which something happens. They also give *rate* as *level* (tax rate). In addition the CCEED gives the verb *rate*, meaning *evaluate*, as in *she rated herself below average*; and rate meaning *deserve* (*...who rated a long obituary in the Times...*). The dictionary ignores other meanings such as *criticise*. Larger dictionaries of course give a much wider range of senses. Examination of corpusCHE, however, reveals that *rate* occurs exclusively with the meaning of the speed or frequency with which something happens. Both are be translated into Thai by the same word, <*atra*>. All the occurrences in corpusCHE are nominal.

However when we consider whether students need to know any more than the basic information above, complications quickly ensue. The distributional figures for the family are given below.

Table 8: Distribution of *rate* and subforms over corpusCHE

	range	total freq	chem	civ	elec	indu	mech
rate	5	413	246	8	18	89	52
rated	2	107	0	0	106	1	0
rates	5	62	50	3	3	5	1
rating	3	14	4	0	9	1	0
ratings	2	9	0	0	7	2	0

We are confronted with the apparently verbal form *rated*, whose behaviour we could not have guessed from the data on the headword as it occurs in corpusENG. *Rated* occurs very commonly but practically exclusively in electrical engineering - 106 out of 107 total occurrences. *Rated* is clearly an essential word for electrical engineers, but is apparently not used elsewhere.

---

[2] By "civil engineering" I of course mean the civil engineering subcorpus section of corpusENG.

The important question is the semantic one: what does *rated* mean? Our (hypothetical) engineering reader has not yet encountered the base verb *rate*, so he cannot deduce the meaning from that, and even if he did he would be somewhat misled. The meaning of *rated* is connected to the verbal meaning of *rate* (to estimate or evaluate) but not quite directly.

The rated value of x is greater than the rated value of y...

In fact (as enquiries in the electrical engineering department elicited) *rated* has a specialist meaning - something like *assessed as being the most efficient value for operation*. The *rated current* is the current at which the system or machine in question operates most efficiently.

Thus, grammatically speaking, it seems more accurate to describe *rated* as an adjective, as in *rated current, rated operation* (the most frequent collocates). *Rate* has an interesting etymology which shows how the two meanings (corresponding to the verbal/adjectival and nominal forms) came into being from a common source. But for present purposes, they are semantically and syntactically quite different. Collocationally they are also distinct. *Rated* is frequently followed by *current, speed, value, operation* and *value*. However none of these words occur within a 5-word horizon of the headword *rate* (at least with any frequency). *Rate* is frequently preceded by the attributive nouns *flow* and *reaction*, neither of which are found near *rated*. *The rate of reaction* occurs 18 times and is confined exclusively to chemical engineering. *The mass flow rate* occurs 11 times but only three of these are in chemical engineering: the other eight are in mechanical engineering. Next, *the rate of change of...* with 10 occurrences is divided between these two subdisciplines. *The specific reaction rate* is confined to chemical engineering and *the volume flow rate* to mechanical engineering.

Rate is thus semantically, syntactically, morphologically and collocationally quite distinct from *rated*. We note also that much of the difference corresponds to subdisciplinary specialisation.

With the other family members there is much less data to play with but we note that *rating* is always nominal *(the rating...the voltage rating)*. Its use in electrical engineering is in the specialist sense (described above) of *rated*, but in chemical engineering it is used in the general sense of *evaluating*. *Ratings* is the same. But there is not enough evidence here for us to speak with any great certainty.

The main problem is perhaps caused by the fact that *rate* is a conversion - a word that has come to be used as two separate parts of speech. This has exacerbated especially the morphological and syntactic difficulties. But over 10% of the words in corpusENG are like this, including many of the most common ones. How can we deal with the mass of idiosyncratic information thrown up by a fairly cursory examination of a word like *rate*?

In so far as one word entitles us to reach any kind of conclusion, what *rate* suggests is that:

1. Specialisation in teaching is strongly indicated, more so than is suggested by the basic headword distributions. It apparently influences all the different categories of word knowledge.

2. Clearly there is a great deal of hidden length in lists of word families; the headword reveals only a fraction of the information necessary to understand the word. Indeed *rate* suggests that the appropriate unit for the lexical syllabus is not the family, nor even the lemma, but the word type.

3. Given that the amount to know about *rate* increases exponentially on even a cursory examination such as the above, it appears to be a gross oversimplification to even think of vocabulary learning in terms of "learning new words" – at least if learning new words means learning single L1 equivalents.

4.2.2 Are many words like *rate*?

In order to gain further information about the words in corpusCHE, and how specialisation affects their behaviour, a random selection of 20 from the most common 1000 word types was made.

One reason for sampling from the most common 1000 is obvious; the behaviour of these words is more germane to the teaching programme than that of the less frequent words. In addition, it is statistically inevitable that specialisation tends to increase as frequency declines (the more frequent a word, the more likely it is to appear in different places); treating less frequent words would be tendentious and probably misleading, especially with this small corpus.

This section, then, summarises the semantic, morphological, grammatical and collocational behaviour of 20 randomly selected common words in order to establish, first, how much further knowledge is needed (depth of knowledge) of these words and, second, whether their behaviour justifies teaching and learning vocabulary in subdisciplinary groups.

The 20 words are types, not families. Since we will be examining morphological behaviour anyway, it does not matter whether we start with types or families.

1. ***TORSION*** occurs 42 times.

Distributionally: This is a specialised word (distribution 0,26,0,0,16 over five subcorpora).
Semantically: Only one of the two technical meanings given in the OTD,

The action of twisting an object by the operation of two opposing turning forces acting at right angles to its axis; a twisted condition (as) produced by this action.

occurs in corpusENG. Polysemy appears not to be an issue.

Morphologically: Torsional occurs (33 times) with an almost identical distribution: specialisation is not a factor in morphology *per se*. There is thus no evidence here against the utility of word families.

Collocationally: *Torsion* has only *simple torsion,* which occurs seven times, exclusively in mechanical engineering. There are no longer clusters. *Torsional* enters into four collocations with frequency greater than two[3], with *system* (5 occurrences exclusively in Mechanical engineering), *strength* (4 exclusively in civil engineering), *buckling* (3 exclusively in civil engineering) and *stiffness* (3, in two subdisciplines). Collocations are thus specialised..

Summary: Apart from the simple morphological change, there seems not much to know about *torsion* beyond its basic meaning. The collocations are almost all specialised, but it is uncertain whether they involve learning more than just the two constituent words, as it is difficult to estimate their opacity or transparency. However the word already represents a strong case for specialisation by virtue of its distribution.

2. ***SUBSTITUTION*** occurs 33 times.

Distributionally: The distribution is 10,0,8,2,13. The absence of *substitution* from civil; engineering looks like justification for specialisation, but this is doubtful (see below).

Semantically: One of the several dictionary meanings, *The replacement of one algebraic quantity by another of equal value but differently expressed,* accounts for all occurrences.

If this substitution is made in Equation 10...

Again, polysemy is not an issue.

Morphologically: Substitute and *substituted* display no distinct semantic or distributional properties. The only mystery is why *substituting* should occur 11 times in civil engineering while *substitution* fails to appear in that subdiscipline. The meaning is the same in each case, except that *substituting* gives us the additional phrase *...by substituting x into y, we obtain...*, which corresponds to

[3] Remember that from now on I will mention only those collocations occurring at least three times.

substitution of x into y gives/yields.... As seven of the occurrences of *substituting* are from one textbook one is tempted to put this down to authorial style or idiolect.

Collocationally: There are no content-word collocations. The cluster *substitution of (x) into (y)*... occurs 20 times.

Summary: *Substitution* presents no compelling evidence for specialisation and apart from the normal inflected forms there is nothing to learn except for the basic meaning. This, surely, is because it is a mathematical term.

3. ***PRESSURES*** occurs 38 times.

Distributionally: The distribution of 22,10,0,0,6 is proportionally similar to that of the much more common *pressure*.

Semantically: The dictionary definition *Force exerted on an object by something in contact with it, regarded as a measurable quantity and usu. expressed as weight per unit area* (OTD) accounts for all occurrences. Polysemy does not arise.

Morphologically: The only other subform is *pressurized*, occurring 4 times exclusively in mechanical engineering.

Collocationally: There are no collocations or clusters at all with the plural form of the word but many with the singular. The most common is *water pressure* which occurs 22 times, all in *pore water pressure*, all in civil engineering and all in the same book. *High pressure* and *vapor pressure* occur overwhelmingly in chemical engineering (9/11 and 11/11 occurrences respectively); *pressure distribution* and *pressure drop* in both chemical and mechanical (with distributions of 4/0/0/0/7 and 13/0/0/0/4 respectively).

Summary: There seems little extra to learn about *pressure/s*, provided the collocations are as transparent as they look. These, however, and the distribution of the word, make a strong case for specialisation

4. ***LAYER*** occurs 50 times and an extra 18 times in *boundary-layer*.

Distributionally: The spread 9,16,21,0,4 (excluding *boundary-layer*) gives some evidence for specialisation.

Semantically: The dictionary definition (*A region of matter that is thin in relation to its lateral extent.*) is vague and so covers a wide semantic range. The different "meanings", however, are largely distinguished by collocation (see below).

Morphologically: *Layers* is uncomplicated and fairly rare.

Collocationally: All occurrences in chemical engineering are in the collocation *boundary layer* which seems somewhat opaque (*a thin layer of fluid that occurs at the point of interface between two phases (e.g. gas and liquid). Boundary-layer,* which occurs almost exclusively in a single book in corpusCHE, is used in the attributive position like an adjective *(boundary-layer thickness, boundary-layer theory)*. In mechanical engineering *boundary layer* occurs three times and *boundary-layer* nine, all in the single same book with similar collocations to chemical engineering. In civil engineering *layer* occurs 13 times in one book, mainly in collocation with *clay, soil, artesian*, and *sand*. In electrical engineering layer occurs as *inversion layer* and *depletion layer*, which look technical and account for 12 of 21 occurrences. These meanings of *layer* all fall within the dictionary definition and would be translated by the same Thai word.

Summary: *Layer* being a somewhat vague word, there is more to learn about how its meaning is modified by collocation then about the word itself. These collocations are specialised. *Layer*'s absence from industrial enginering is additional support for specialisation.

5. ***EDGE*** occurs 38 times.

Distributionally: Edge is specialised, at 4,18,0,0,16.

Semantically: Difficulty is hard to assess without a more thorough investigation than this section merits. *Edge* seems to retain the general meaning *The narrow surface or side of a thin object* in most cases, except for *edge beam* (see below) .

Morphologically: We see only *edges*, with (proportionally) similar distribution.

Collocationally: In corpusCHE three of the four occurrences are in *the leading edge* and this phrase also occurs 4 times in one book in mechanical engineering, which also contains the only four occurrences of *trailing edge*. In civil engineering *edge* contributes to the phrase *edge beam/s* seven times, all in one book, where the meaning appears to be *The boundary line of a surface or region.*

Summary: *Edge* seems to have semantic and collocational complications. Although it is not a very common word, there seems to be enough to learn about it that is specialised to justify specialised teaching.

6. ***LIFE*** occurs 34 times.

Distributionally: 5,3,1,14 and 12 occurrences.

Semantically: Virtually all occurrences are related to

The time for which an inanimate thing exists or continues to function or be saleable or valid

This covers *life* as in *human life*, but Thai has a different word for the latter.

Morphologically: There are no subforms occurring in corpusCHE.

Collocationally: *Fatigue life* is slightly opaque, being *...the period during which a beam can withstand certain loads before failing* (OTD), and occurs in only mechanical engineering. *Half-life* is opaque and technical and is in only one book in chemical engineering (6 times). *Battery life* (7 occurrences all in one book in industrial engineering) and *tool life* (3 in mechanical), are specialised though fairly transparent.

Summary: After learning that *life* is used in English with a meaning it does not have in Thai, the Thai learner seems to have done the hard work. But this word would inevitably be taught with examples, which involve collocation, which are specialised.

Specialisation is not demanded but is convenient.

7. ***PROVIDED*** has 79 occurrences.

Distributionally: distributed 13, 43, 14, 4 and 3.

Semantically: There is the obvious polysemy (meaning of *...on condition that....* or *...Equip(ped) or fit out with what is necessary for a certain purpose; furnish or supply with something...* The former conditional meaning (which does not apply to *providing*) occurs 13 times. 10 of these are in three individual books where the authors do not use the other meaning. Similarly, the authors who use the latter sense seem reluctant to use the former. This sirenical pathway into idiolectal variation must remain unexplored here.

Morphologically: All of the non-conditional occurrences of *provided* are passive. The other members of the family (*provide, provides, providing*) are fairly flatly distributed. *Provisions*, however, is almost exclusively in industrial engineering, occurring 9 times with the legal meaning.

Collocationally: There are no collocations in the present use of the term.

Summary: Knowledge of the common meaning of *provide* will take the learner a long way (especially if he knows passive inflections). He may be mystified by some occurrences of *provided*, but specialisation does not seem to be a factor here.

8. ***VELOCITY*** occurs 239 times.

Distributionally: 82,4,11,1,141. This word is essential in some subdisciplines but marginal in others. This justifies specialisation.

Semantically: Alone, the word is straightforward. But see collocation below.

Morphologically: *Velocities* mirrors the singular form fairly exactly.

Collocationally: There are 22 collocations most of which look daunting to the layman. Even simple-looking ones like *average velocity* and *constant velocity* occur mostly in single books (4/5 occurrences with the former in chemical engineering: 7/10 with latter in mechanical). The (more common) opaque or technical ones (*velocity profile, velocity distribution*) appear to behave in the same way. The term *angular velocity* occurs 37 times, 32 in mechanical, meaning the *rate of change of angular position* (OTD). And while *relative velocity* (11 occurrences with 9 in one book in mechanical engineering) looks straightforward, it turns out to be a method of calculating (again, this phrase is confined to mechanical engineering).

Summary: The learner armed only with the knowledge that *velocity* means *speed* will soon face large obstacles. These arise from the many collocations containing the word, which tend to be opaque and are on the whole specialised.

9. ***CUSTOMERS*** occurs 61 times

Distributionally: 0,0,1,61,0, with the word practically restricted to industrial engineering.

Semantically: Invariant.

Morphologically: The singular form behaves like the plural.

Collocationally: There are no collocations to speak of.

Summary: There is nothing more to know about *customers*, unless you count *customer*. *Customers* is however a prime example of specialisation.

10. ***LOW*** occurs 116 times.

Distributionally: 42,13,35,16,10.

Semantically: *small or reduced in quantity, degree, or intensity* covers all occurrences.

Morphologically: *Lower* and *lowest* are also flatly distributed. The verb *lower* occurs only three times.

Collocationally: In transparent phrases like *low temperature, low pressure*. These are somewhat specialised but this is not because of *low*.

Summary: There is not much to know about *low* except its basic meaning.

The reader may by this time see advantages in accelerating the process of this examination. Let us summarise the data from the other ten words more briefly.

In terms of extra information or specialisation, there is nothing to say about (11) *several*. (12) *Static* enters into some specialised collocations in mechanical engineering, while *statics* is absent from three subdisciplines. The corpusENG frequencies (29,7,5,0,11) say all there is to say about (13) *solid*. (14) *Present* has three different meanings in corpusCHE, with three different word classes, which causes morphological and semantic difficulties, which are however largely unconnected with specialisation. *Shown*'s (15) different meanings in English are not distinguished in Thai either; it is very common indeed (93,211,245,92,212), more so where the subdiscipline relies heavily on charts and diagrams. *Specification* (16), distributed 3,23,0,10,4, has one or two specialised collocations and the full range of seven family members. *Apply*'s (17) transitive and intransitive senses are not distinguished in Thai. Six family members are present. *Applied* has one or two specialised collocations. *Respect* (18)occurs widely (74 times) but practically exclusively in the phrase *with respect to*. *Excitation* (19) occurs only in two subdisciplines and has a number of technical collocations. *Said* (20) is difficult only because of *say* meaning *for example*.

Only tentative conclusions are possible from these twenty words, but they do point in a certain direction.

Firstly, these corpusENG words are perhaps surprisingly well-behaved. In general, a knowledge of one meaning of the word and the basic inflections will carry a reader quite a long way. We might say that just as engineering uses a limited set of words it uses them in limited ways. Polysemy is not prevalent. The example of *rate* led us to somewhat overestimate the irregularities stemming from morphological change. Some of the words are somewhat technical but this aspect may be less of a problem for the engineering student than for the English teacher.

Second, the justification for specialisation is confirmed in terms of collocation, especially for nouns. Indeed the next section will show how this feature, supposedly of depth of vocabulary knowledge, may provide a solution to the problem of teaching subdisciplinary vocabulary to students at the low level of SUT engineering undergraduates.

4.3 (Objective 2.3) Refining the lexical syllabus

According to the evidence of the previous section, by dealing with collocation we are dealing with the single most important aspect of depth of vocabulary knowledge in English for engineering. But this evidence is not yet sufficient. The words examined were in some cases of low frequency and were of various grammatical classes. Both of these factors might affect the amount of collocation involved. In addition, we only looked at collocations occurring three or more times; in a small corpus like corpusENG this might miss a lot of relevant data. And in any case, some of the words examined in section 4.2 did not show any great tendency to enter into collocations. The claim about collocation needs further support.

In fact when we examine the corpus data in more detail it becomes clear that collocation is an essential element in the sub-disciplinary word frequency figures. This means first, that collocations account for a lot of the word frequency data (section 4.3.1 below); and second, that while the single word frequency figures show a large degree of specialisation, the collocational data do so to an even greater extent (section 4.3.2). These two factors strongly support an approach to teaching engineering lexis in sub-disciplinary groups and with an emphasis on collocations.

These are the distributional (frequency) facts about collocation; but there is equally strong support on the qualitative side. Section 4.3.3 shows that dealing with collocations enables us to deal with technicality at an appropriate level of difficulty - appropriate to students entering the sub-discipline and appropriate to language teachers with limited knowledge of the specialist subject matter. This is because of the part collocations, in the sense of noun groups, plays in two essential aspects of technicality, namely compression (section 4.3.3.1), and precision (section 4.3.3.2).

In defining collocations we exclude all non-phrases, clauses and all combinations containing function words. We are left with (recurrent) combinations of noun-noun, and adjective-noun. Without going into an unprofitable argument about what exactly constitutes an adjective as opposed to a participle, we will use a loose definition of the word adjective so that we may include phrases like *recycled gas* and *entering gas*.

4.3.1 The ubiquity of collocations

For reasons connected to its size, corpusCHEa is not the best guide to collocations. For examining the real frequency of collocations, it is more reliable to look at the larger book-length corpusTXT, where we can examine collocations in their natural habitat, so to speak.

The investigation begins with one of the more common content words in chemical engineering, *gas*. Following the criteria established in the previous section, the following list (Table 9) of gas collocations was derived from corpusTXT.

The data is shown in order of length (first the four-, then the three-word collocations, etc) and frequency, starting with the left-hand column. Note that we have included only collocations with a frequency greater than 2. This may seem a slightly arbitrary application of the recurrence criterion but it at least serves to eliminate typographical and punctuation errors.

The total number of occurrences of *gas* in CorpusTXT is 1089. There are in total 78 collocation types in the table below. If we added those occurring at a frequency of 2, this would give 26 more (thus 52 tokens). Counting only those in the table, the total number of occurrences of these items is 727. These collocations therefore account for about 66% (727/1089) of the occurrences of *gas*. (With those occurring once or twice the figure probably rises to about 75%.). This suggests that the reader needs to know not *gas*, but *gas+*.

This gives a very different perspective on the word *gas* from that employed so far in this thesis. From being one item in a list (like listCHE) of 1000 or so, *gas* is now 78 items in a list of we know not what length. Before tackling this issue, we need to know whether *gas* is typical or not.

Table 9: Gas (G) collocations in CorpusTXT

phrase	tks	phrase	tks	phrase	tks
elementary G phase reaction	4	*ideal G*	23	*G temperature*	5
irreversible G phase reaction	4	*stack G*	22	*original G*	5
bulk G phase concentration	3	*G stream*	19	*pure G*	5
outlet flue G temperature	3	*G analysis*	16	*saturated G*	5
ideal G law	35	*G absorption*	15	*synthesis G*	5
G phase reaction	27	*G-liquid*	14	*recycled G*	5
G phase reactions	18	*real G*	14	*perfect G*	4
ideal G constant	8	*noncondensable G*	13	*G composition*	4
dry flue G	8	*exit G*	11	*feed G*	4
G flow rate	6	*G pressure*	11	*G analysis*	4
ideal G laws	6	*G mixtures*	10	*propane G*	4
flue G analysis	6	*G oil*	9	*G chromatograph*	4
G phase concentration	5	*fuel G*	9	*G chromatographs*	4
stack G analysis	5	*reactant G*	8	*G bubbles*	3
superficial G velocity	4	*G density*	8	*G velocity*	3
G phase decomposition	4	*entering G*	8	*reacting G*	3
G absorption resistance	3	*G law*	8	*G concentration*	3
G phase constant	3	*G oil*	7	*hydrogen G*	3
homogeneous G phase	3	*G flow*	7	*G behaviour*	3
ideal G equation	3	*G constant*	7	*G cylinder*	3
partially saturated G	3	*G properties*	7	*G vapor*	3
G absorption column	3	*exhaust G*	7	*inlet G*	3
G storage tank	3	*dry G*	6	*perfect G law*	3
G phase	97	*bulk G*	5	*product G*	3
natural G	41	*stagnant G*	5	*waste G*	3
flue G	25	*chlorine G*	5	*water G*	3
G mixture	24				

Two other words chosen randomly from the top of the frequency list show rather similar behaviour. *Liquid* occurs 835 times in the corpus, and enters into the collocations (frequency >2) shown in Table 11.

These 46 collocation types account for 476 tokens in CorpusTXT. This is 57% of the occurrences of *liquid* in CorpusTXT.

Heat (see Table 6.3) is similar: 540 of *heat*'s 846 occurrences are in these collocations. This is 64% of the total occurrences of *heat*.

Table 10 below summarises the data from the above three tables.

Table 10: Frequency and proportion of collocations in CorpusTXT

	A	B	C	D	E (%)
gas	1089	78	4/19/55	727	66
heat	846	35	1/12/22	540	64
liquid	835	45	1/8/36	476	57

The columns show the following:

A: Number of tokens (single word)
B: Number of collocations (types)
C: Number of 2/3/4-word collocations (types)
D: Number of collocations (tokens)
E: Percentage of tokens (collocations) as against tokens (single word).

The table above reveals a high degree of consistency in the proportion o collocations to occurrences (column E). Not surprisingly, 2-word collocations are much more common than longer ones, although the relative proportions are no very consistent. There seems little consistency in the number of collocation type and there seems no sense in trying to identify a central tendency in the number o tokens per collocation.

Table 11: ***Liquid*****'s (L) collocations in CorpusTXT**

phrase	tks	phrase	tks	phrase	tks
Ele'tary L-phase react'n	6	*L mixture*	14	*pure L*	4
L-phase reaction	19	*bulk L*	12	*L streams*	4
L-phase reactions	13	*L-L*	11	*L benzene*	3
L level control	8	*L holdup*	10	*L composition*	3
L flow rate	7	*L feed*	7	*L concentrations*	3
L static pressure	5	*L holdups*	7	*L film*	3
vapor-L equilibrium	5	*L phases*	7	*L heat*	3
gas-L interface	3	*L-solid, solid-L*	7	*homogeneous L*	3
vapor-L equilibria	3	*L density*	6	*L properties*	3
L level	102	*L flows*	6	*reactant L*	3
L phase	56	*L systems*	6	*subcooled L*	3
L water	32	*boiling L*	4	*L system*	3
L vapor	17	*L compositions*	4	*L volume*	3
L stream	19	*L diffusivity*	4		
saturated L	16	*L mixtures*	4		
gas-L	11	*L product*	4		

Table 12: ***Heat*** **(H) collocations in CorpusTXT**

phrase	tks	phrase	tks
overall H-transfer-coeff	33	*H exchangers*	17
H transfer coefficient	15	*H exchange*	16
H capacity equations	14	*H flow*	11
mean H capacity	5	*latent H*	9
mean H capa'ies	5	*specific H*	9
H transfer fluid	5	*sensible H*	9
H exchange area	5	*H duty*	9
H-tr correlation	4	*H input*	9
H-tr coefficientss	3	*H balance*	8
H-tr surface	3	*integral H*	8
sensible H terms	3	*H losses*	8
standard integral H	3	waste H	5
tubular H exchange	3	H loss	5
H capacity	100	H flux	5
H transfer	96	constant H	4
H exchanger	58	humid H	4
H capacities	22	total H	3
standard H	22		

If we take another sample from the most common content words (nouns) in CorpusTXT, we find a more variable but, generally, similarly high proportion of collocation tokens to total word occurrences. The following words were chosen at random from the more common content words in CorpusTXT (frequency>100). The recurrence rate was set at 3.

Table 13: Collocation data for other common words in CorpusTXT

word	Collocations (types)	Proportion of total (tokens)	Proportion as % (tokens)
reaction	86	1538/2311	67
rate	48	1782/2125	84
mixture	15	153/335	40
time	66	918/1400	66
boundary	5	97/121	80
equilibrium	16	150/256	59
surface	23	252/358	70
system	64	838/1626	51
vapor	23	363/478	76
pressure	48	767/1158	66
energy	29	635/792	80
capacity	5	160/177	90
absorption	3	27/34	79
performance	3	40/67	60

This data from tables 6.1 to 6.5 show that collocations account for between 50% and 80% of the occurrences of many common content words. This suggests that a lexical syllabus cannot ignore the phenomenon of collocation. On the other hand we saw with *gas* etc. that there are large numbers of collocation types, and that many of them are fairly infrequent. This poses the problem of choosing which ones to deal with. The obvious criterion for selection is frequency, but this presupposes that the corpus we use to identify frequency is representative of the sub-discipline as a whole. In fact this can only partly be the case, as is shown below.

What this section has shown is that collocation, in the particular sense of two-, three- and four- content-word noun-phrases, is ubiquitous in chemical engineering textbooks and thus accounts for a large part of the word frequency data. This is one good reason for writing the lexical syllabus on the basis of this particular aspect of depth of vocabulary knowledge.

4.3.2 The specialisation of collocations

The appearance of noun-phrase collocations in particular specialist sub-disciplines - in other words, their narrow distributions - was one of the tentative findings of section 4.2, and this was seen as a possible justification for teaching lexis to those sub-disciplines individually. However a very limited amount of data was used. This section will show that the tentative conclusion was in fact correct.

For this part of the investigation it is obviously necessary to switch from the monodisciplinary corpus CorpusTXT to the sub-disciplinary corpora, which are only 50000 words in size. As we saw with *gas, heat* and *liquid*, there are a lot of collocations with low frequencies. A lot of these must be missed by the sampling procedure involved in creating the sub-disciplinary corpora, which come from five 10000-word textbook samples, which in turn constitute only 10% or less of the total text from which they were derived. For this reason a fairly large number of words are examined below, to give better support to the general idea of specialisation in collocations.

In the previous section we looked in some detail at the collocations of *gas, heat* and *liquid*. Table 14 below shows how these words are distributed over CorpusENG:

Table 14: distribution of three words across CorpusENG

	chemical	civil	electrical	industrial	mechical	Total
gas	162	0	2	0	11	175
heat	139	4	7	6	10	166
liquid	153	5	6	0	9	173

Since these words are overwhelmingly concentrated in chemical engineering, it is obvious that hardly any of their collocations appear in sub-disciplines other than CorpusCHE. In fact no more than three of all the collocation types with *gas, heat* and *liquid* appear in other sub-disciplines, and none with any frequency[4]. In these cases it may be said that it is the headwords of the collocations that reflect specialisation, not the collocations. While this is clearly true in some sense, it remains the case that the reason for the overwhelming relative frequency figures is in large part precisely because of the number of collocations entered into by these words, as we saw in the last section. However a better way to demonstrate the

[4] Bear in mind that the 50000-word corpora are far from infallible in providing collocational data so such statements must be somewhat tentative if applied to the sub-discipline as a whole.

centrality of collocation to specialisation is to show data from words that are common in more than one sub-discipline. Let us look at some common CorpusCHE content words with flatter distributions across CorpusENG, like those in the left-hand column of Table 15 below.

Table 15: distribution of four other words across CorpusENG

	Che eng	Civ eng	Ele eng	Ind eng	Mec en	Total	PR
system	159	31	390	167	177	924	2.11
time	104	34	119	189	54	500	1.89
value	88	128	160	113	79	568	1.41
factor	34	66	70	110	42	322	1.71

As these words' realtively low PEAKRATIOS, shown in the right-hand column, indicate, they are well distributed over the five sub-disciplinary corpora. Remember that with 5 subcorpora, the maximum value of PR is 5.0 and this indicates the narrowest distribution; a PR of 1 shows a perfectly balanced distribution.

The table below shows the collocation/PR figure for the fourt unspecialised nouns (frequency>4) in corpusENG. Note that in complete contrast to the PR figures in table 16, the collocation PR figures are very high.

Table 16: Collocates' distributions over corpusENG

Collocates	**Mean PR over corpusENG**
System (left-hand only)	3.46
System (right-hand only	3.48
Time (left-hand)	4.27
Time (right-hand	3.77
Factor (left-hand only)	4.23
Factor (right-hand only)	4.90
Value (left-hand only)	2.74
Value (right-hand only)	(none)
Process (left/right)	4.64
Form (left/right)	3.32
Component (left/right)	5.00
Flow (left/right)	4.65
Condition (left/right)	3.96

These data suggest that although not all words exhibit the same tendencies, many collocations are strongly associated with specialisation at the level of sub-discipline. The case for taking a sub-disciplinary approach with collocations is thus made

4.3.3 Technicality

It is all too easy for the EAP practitioner to throw up his hands when confronted with something like

Therefore, the component balance yields a stripping operating line (SOL)..(equation omitted)....the slope of the SOL is the L/V ratio in the stripping section. The SOL intersects the 45 ° line at xb, as shown in ....

Language teachers sometimes have little knowledge of scientific and engineering fields and so cannot be expected to teach field-related vocabulary, even if the subject specialist teachers wanted them to. As Nation (2001: 204) says,

> "...learning technical words is closely connected with learning the subject.....English teachers are not usually well equipped to work with technical texts and the technical vocabulary they contain."

Similarly Jordan (1997:152)

> "...the (technical) words are closely connected with learning the subject and may present conceptual difficulties"

One way to tackle this problem has been (essentially) to avoid it, by claiming that it is not the language teacher's business. In this view, the student is assumed either to know the technical terms because he knows the subject (see e.g. Cobb & Horst 2001:315) or to be able to learn them directly from the specialist teacher (Cohen et al. 1988:Cowan 1974:391). In either case the EAP practitioner's job is to identify technical vocabulary so that it can be excluded from the language syllabus. The view is reinforced by studies claiming that technical words constitute only 5% of the tokens in technical text, as opposed to the 10% comprised by "academic" vocabulary (see e.g. Nation 2001). In this way quantitative linguistic data, as well as the practicalities of teaching, support the decision to leave technical vocabulary alone. Moody (1975) goes so far as to say that any teacher ignoring this maxim "...will arouse amusement, if not contempt, in his students."

If, however, we consider why technicality is avoided in this way, we see that it is because it is difficult. The word itself implies difficulty. The following statement is made up but I believe plausible:

a) *this article is very technical*

It can be read as purely descriptive. It could also have the implication that "I cannot read it"; but it is difficult to take it (as it stands) as a recommendation to read the paper in question. "You should read this paper – it's very technical" seems, pragmatically, slightly odd.

Similarly the following statement,

b) this paper is too technical for me

which again is made up but I think unexceptionable, makes no sense unless *technical* carries the connotation of difficult to understand.

Distributional views of technicality seem to be based on the idea of placing technicality in quarantine, as far as the language teacher is concerned. A more considered approach is to examine what it is about technicality that makes it difficult. Among those who have attempted to do this, and whose work we will draw upon here, are Halliday (1985), and Pueyo & Val (1996), who in turn draw upon Martin (1990, 1993).

4.3.3.1 Technicality as compression

Technicality arises as a result of the creation of new knowledge. The birth of technicality, technicalisation, is a process. One important aspect of this process is compression.

New knowledge must initially be expressed in the form of propositions. Relativity is not new knowledge: the new knowledge is that observations depend on the relative motion of the observer and the observed object. E=mc2 is not new knowledge: the new knowledge is that energy equals the product of the mass and the square of the speed of light.

These propositions are then compressed. This is not a feature peculiar to technicalisation, but it is characteristic of it. As one encyclopedia puts it:

> Technical terminology evolves due to the need for experts in a field to communicate with precision and brevity.....the terms of technical terminology are used to express a great deal of information in a compact form. This makes it possible for professionals to speak to each other without having to exhaustively describe each concept; they can simply use the terms whose definitions are already known in the profession.

(http://www.internet-encyclopedia.org/wiki.php?title=Technical_terminology)

The mechanics of the process of compression are described by Quirk et al. (1985):

> "However new and unfamiliar it is, an entity, activity or quality can be stated or described in sentences:
>
> Let us convert our railways from having steam-engines to using engines powered by diesel oil.
>
> The nub of this suggestion might then be expressed by means of a nominalization such as:
>
> The use of diesel-powered engines (is being investigated).
>
> This already presupposes some discussion of and familiarization with the notion. But as the proposal becomes more widely accepted as viable, such a nominalization will seem too clumsy......(and too undercommitted). At this stage we will not be surprised to find the notion institutionalized by means of (the) word
>
> Dieselization (is feasible).
>
> We have now lexicalized a notion that could previously be discussed only in sentences and periphrases which varied from person to person. In lay terms, we now "have a word for it"" (authors' italics).
>
> (Quirk et al. 1985:1526)

The process as described above, then, involves nominalisation of propositions, either by the formation of new nominal phrases, or by the creation of new single words (called *lexicalisation* by Quirk et al. (loc. cit.)., or by the two processes in sequence (as in the example of *dieselization*).

Technicalisation of this type is also explained in terms of grammatical metaphor. This phrase, coined by Halliday (1985), is explained as follows. The natural relationship is between processes or (as above) events on the one hand and verbs on the other; but the process of nominalisation allows the treatment of processes as nouns, which in turn allows both the attribution to them of properties and also their participation in other events.

> "...nominalization refers to the tendency in English to represent events, qualities of objects, and qualities of events not as verbs, adjectives and adverbs, but as nouns. Nominalizations are important in science because they allow complex phenomena to be summarized in a few words. However their main advantage is the ability they have to

> construct meaning and to build new knowledge upon previous knowledge." (Pueyo & Val 1996:257)

As Banks (2005:351) puts it, reflecting "..on the usefulness of nominalization..",

> "From a grammatical point of view, nominalizing a process allows the addition of modifiers and qualifiers. Thus *The government announced new benefits* can be nominalized as *the government's announcement of new benefits*. The former can be considered as the unpacked version of the latter...Adding a new modifier, we might produce something like *the government's long-awaited annnouncement of new benefits*. It will be noted that this new version is much more difficult to unpack, at least without using a fairly lengthy paraphrase. Thus the nominalization permits the concentration of information in this sense."

Banks here introduces[5] the notion of unpacking, the reverse process of compression, whereby nominals (phrases or words) are expanded into the propositions from which they are derived. The attempt to unpack leads to the problem of what Quirk et al. (loc. cit.) call *unrecoverability*. We cannot recover the meaning of *dieselization* from its form: we may guess, but there is no way of being sure that it means what it does rather than (say) *lubricating railway lines with diesel oil*. It is necessary to revert to the phrasal, or further, propositional equivalent of the word in order to disambiguate.

The point here is that, all other things being equal, the creation of new nominal phrases (what we are calling collocations) represents an easier stage of technicalisation than the creation of new words. Consider two examples.

The term *enthalpy*[6] is defined in the dictionary as follows:

> The total heat content of a system, expressed as a thermodynamic quantity obtained by adding its free energy to the product of its pressure and volume (Oxford Talking Dictionary)

This definition, this explanatory nominalisation, represents an unpacking of the term and must, logically, be easier than the word itself. But it contains three further nominal phrases which themselves require explanation. We are all used to dealing with the idea of heat, but what can *total heat content* be? Is it a temperature, a volume, or what? The answer to these questions is presumably contained in the phrase *thermodynamic quantity*, whatever that may be. Another difficulty is with

---

[5] This is not to say he invented the term.

[6] My assumption here is that the reader has no more idea than I do about what this word means.

free energy: a two-minute conversation with a chemical engineer failed to produce a definition which I was able to grasp with any certainty, and the dictionary definition

> "..a thermodynamic property of a system that represents ability to do work." (Oxford Talking Dictionary)

is useless without understanding what *thermodynamics* is (or are). The point is that a word like *enthalpy* represents the culmination of the process of compression: compression of any number of propositions into noun phrases of increasing degrees of difficulty to the non-specialist. We cannot encapsulate any further the proposition - the idea - from which the word is derived (barring, of course, by abbreviation or by symbolic representation as in an equation).

We do not claim that there is a single, two-word phrase for *enthalpy* which is easier than the word itself. Neither do we claim that two- or three-word phrases are always easier than the single words contained in them (we find, for example, *partial molar enthalpy*). But as a general principle, phrases represent an earlier and thus easier stage of technicalisation than single words, partly because the burden of unpacking and unrecoverability is smaller.

The second example illustrates the same point in reverse. The phrase *flue gas* is fairly common in CorpusCHE. We could in theory trace *flue gas* from several propositions:

there is some gas remaining in the flue
there is some gas exiting from the flue
there is some gas entering the flue

etc. etc.

and from several nominal groups:

the gas (which is) remaining in the flue

etc. etc.

Unpacking *flue gas*, however, depends on, or arises from, knowing which proposition or nominalisation lies behind the phrase in question. This is obviously one aspect of technical knowledge. But it would be possible to make *flue gas* more compressed and thus more difficult, just by deciding that two words are too cumbersome and that you will in future refer to it as (say) *flug* or some other neologism. Now the burden is on the reader to "unpack" *flug* as *flue gas*, a burden additional to that of relating *flue gas* to its correct nominal and propositional antecedents, the burden of technical understanding. So *flue gas* is relatively easy, because the difficulty of recoverability is less than with *flug*.

The data presented concerning compression accords with Quirk et al (1985:1531), who speak of:

> "Scientific writing ...having a distinctly higher proportion of noun phrases with complexity (and multiple complexity); .... and the weakest association of simple with subject and complex with non-subject."

(The authors' "simple noun phrases" are single words).

In terms of teaching, then, we can summarise by saying that dealing with collocations helps tackle the difficulty of technicality caused by compression. This is because they are less compressed (in the sense described) than single technical words and because we can unpack them into propositions, which is the original form of the new knowledge that they represent.

4.3.3.2 Technicality as precision

So far we have been talking about how technicality comes about through the naming of phenomena: the creation of nominal groups is seen as a natural product of this naming. The next aspect to be examined is precision.

> "Technical terms are those for which there is a congruity of concept between all scientists, whatever the language used....the properties or characteristics can be enumerated to define the object in an unambiguous manner." (Selinker & Tarone 1981:24).

This relates to items collectively as well as individually:

> "...Each term is dependent for a full appreciation of its meaning on the meaning of (the) other terms.) (Loc. cit. p.28)

A similar point is made by Pueyo & Val (1996:251):

> "..what makes written language so difficult is not the technical terms themselves but the complex relationships they establish with each other, that is, the "wording" as a whole."

Enthalpy in section 4.3.3.1 was a clear example of this. We saw that understanding the definition depended on knowing two other terms (*free energy* and *heat content*, which were themselves interrelated) and that both of these rested on the technical concept of *thermodynamics*.

But the point also applies to seemingly simple words. Few people, one imagines, would confess to being baffled by the word *gas*; it is in a sense an "easy" word. If taxed for a definition, the layman might talk about states of matter, of solids and liquids; we are still in the realm of everyday experience. But it is doubtful whether the layman will produce anything as concise, or precise as the chemical engineer's *a species at above boiling point* as a definition. It is precise because it depends on the two concepts *species* and *boiling point*: no doubt the reader will know exactly what a boiling point is, but the assumption (made by myself) that *species* just means *substance* is mistaken. The difference between the two could not be made clear to me in a short conversation with a chemical engineer: again, technicality is involved.

Gas, then, is basic and everyday, but in chemical engineering it is a) technical because it is precise, and b) precise because it is technical. If we look at the most common nouns in CorpusTXT, listed below, we see that they too are used in everyday language with meanings similar to those found in chemical engineering (with the exception of *reactor*).

reaction rate reactor system control equation temperature process time pressure gas concentration function heat liquid energy volume feed stream vapor

In a sense, then, these words are at the threshold of technicality. This is the point at which the everyday becomes technical, reflected in the precision of these terms' definitions. By unpacking these terms, we define them, and vice versa. The definition may be taxonomisation - placing the term in some kind of hierarchical relationship with other terms (as in straightforward classification) - or simply explanation of how a process occurs, a process involving participants which are in themselves technical. But in either case it involves relating the term to other terms. The definition establishes the precision of the term. Through this process of unpacking we create mini-texts at a low level of technicality - precisely what is needed for students at the beginning second-year stage of study.

But there are strong arguments in favour of creating mini-texts based on the collocations that these words enter into rather than the words themselves.

First, the collocations are more specialised and sub-disciplinary specific than the than the individual words themselves. This was shown in section 6.3. It is considered something of a motivating factor for specialising students.

Second, being more specialised, the collocations involve more specialised vocabulary when they are unpacked. Indeed, the next section will show that the mini-texts created are packed with relevant lexis. This vocabulary will be recycled many times in the definition texts that are created.

Third, by treating lexis at a level above that of the single word, we encourage better reading habits. By teaching students to process words in groups (noun groups of two, three or even four words) we take their reading to a higher level. The arguments behind this are based on the idea of verbal efficiency - the essentially commonsense idea that better readers are capable of processing text in larger chunks.

Incidentally, in treating noun groups we tackle a substantial contrastive problem. That is the tendency of the Thai language to postmodify rather than premodify head nouns in noun phrases, in contrast to English - for example the phrase *gas temperature* is translated as */oomaphoom gad/*, literally *temperature gas*. This also goes for adjectives (*pure gas* = */gad thae/* = *gas pure*)

4.3.3.2.1 Unpacking collocations

The following exercise was designed to illustrate the precision discussed above, how it is manifested in the process of unpacking, and how the process produces the type of text that can be used in a language classroom.

I first wrote definitions of some gas collocations, based purely on what the everyday meanings of the components would suggest to the layman (taking, of course, myself as a typical layman). This list of definitions was then passed to a chemical engineering lecturer here to be (as it were) marked. He was asked to write his own definitions for comparison. See below the lay and specialist definitions juxtaposed. The specialist's definitions are in bold face type, my own in ordinary.

The collocations were selected at random from a full list of those in CorpusTXT. Not all appear in CorpusCHE.

natural gas: the gas we cook with in UK, which comes via pipes into our kitchens from the gas mains.

natural gas: **A gas for cooking, heating, and also used in some vehicles as a fuel (often stored in the liquid state). It is a fossil fuel which is in the gas phase. The main component is methane, but it also contains other light hydrocarbons as ethane, propane, and butane, and some contaminants, such as carbon dioxide, and hydrogen sulfide.**

Comment: The differences in the definitions well illustrate the difference between the everyday and the technical. The layman's definition correctly identifies the referent of the phrase, but is made to look mundane by the greater accuracy and amount of detail - precision - in the expert definition. Pedagogically, the point is that the expert definition gives us a text which is comprehensible to the layman teacher while being appropriate for the beginning chemical engineer - subject specific, perhaps slightly challenging for many, and not patronising.

Of the 25 content words in the expert definition of natural gas above, 18 occur more than 50 times in the total chemical engineering corpus. This corpus is an amalgam of CorpusCHE, corpusCHE2, corpusCHE3[7] with CorpusTXT, a total of just over 500000 words. These 18 words (all except *stored hydrocarbons contaminants cooking sulfide vehicles* and *fossil*) thus would each occur on average a minimum of 10 times in a fairly short book of 100000 words, or once every 10000 words. They account for 1.4% of the entire corpus. We can thus claim that this text, the expert definition which unpacks the collocation, is packed with very frequent chemical engineering words.

Perhaps it hardly needs pointing out that the text is also lexically highly specialised. Only five of the 25 content words in the expert definition (*used phase state component main*) occur at this frequency (once every 10000 words) in the other engineering sub-disciplines of the cae corpus. With the exception of *phase* these five words are broadly distributed and may thus be called "general" engineering words. Of the other twenty, nine do not occur at all elsewhere in cae. We are justified in calling this list of 25 words, overall, a specialised vocabulary. The text is a technical text.

entering gas: a gas entering a piece of equipment where a particular process takes place.

entering gas: **a gas which enters the equipment in which a particular process or phase of a process is carried out.**

Comment: The lay definition is virtually correct, as the term is quite transparent. There is thus no conceptual difficulty in the phrase; the two words are used in their everyday senses. But the combination is unusual to the layman. The only refinement added by the expert is "or phase of a process..". Incidentally this meaning of phase is not the most common one in the sub-discipline.

Let us again look at the frequency elsewhere of the content words in the expert definition. *Enters* occurs 140 times in the total corpus. This represents 30 times in a short book. All the other words are more frequent than this - *gas* (1482 occurrences), *phase* (759), *carried* (294, of which 244 are in the phrase *carried out*), *particular* (195) and *equipment* (176). The coverage of the total corpus by all the six words is 0.9%. Again, the text is packed with relevant vocabulary.

feed gas: a type of entering gas.

feed gas: **A gaseous feed (entering) stream into chemical process equipment.**

---

[7] This is a different 50000-word chemical engineering corpus on topics separate from those in corpusCHE1 and corpusCHE2.

Comment: the layman's guess is substantially correct. The term is more or less transparent and this use of *feed* is everyday. The expert's 9-word definition incorporates two function words and seven content words, the latter of which account for no less than 0.8% of the whole chemical engineering corpus. All these words occur at a frequency of more than 1:10000.

ideal gas: a gas with no extraneous contents, i.e. apart from what is expressed in its formula. CO2 is a perfect gas when it contains nothing but carbon and oxygen.

ideal gas: **A gas in which the molecules are of negligible size, and the density of the gas is sufficiently low that the molecules are very far apart, and thus have no molecular interaction with other molecules in the same gas.**

Comment: While the layman's guess is wrong, the correct definition is not difficult to understand and not highly opaque. The 11 content words in the expert definition cover no less than 3% of the 518000-word corpus.

A kind of pattern seems to have emerged. The expert definitions tend to be longer and more precise than the lay ones. They are not however like *enthalpy* - technical in the sense of being quite inaccessible to the layman. In other words, they are just difficult enough.

Another twelve phrases were "tested" and the conclusion was generally supported strongly. Four collocations from each of *gas*, *heat* and *liquid* were taken and the results may be seen in Table 17 below.

Table 17: Comparative definitions given by a layman and an expert

term	layman's definition	expert's definition
product gas	the gas which a process is supposed to produce, e.g. for sale.	**A gaseous product (leaving) stream from a piece of chemical process equipment.**
bulk gas	gas in large quantities which is refined or used by a process to produce a purer form of this gas or something else	**Gas, in a gas phase, and which is sufficiently far from the interface with another phase to have properties that are constant with respect to spatial position.**[8]
inlet gas	a gas through which entering gases must pass as soon as they enter the process equipment	**The same as a feed gas.**
carrier gas	a gas in which vapors or particles or other gases are transported within some particular process equipment	**A chemical component or species in a gas phase which plays no part in the reaction or separation operation, other than to contain another component that does participate in the reaction or separation**
heat capacity	The degree of heat that can be contained within a given container	**The ability of a substance to absorb heat.**
heat-transfer coefficient	A number indicating the degree of heat moving from one place to another	**A constant which specifies the amount of heat transfer relative to a temperature difference**[9]
heat duty	The amount of heat required to perform a particular task or process	The amount of heat energy required to carry out a process
waste heat	Heat that is generated but not used in the process for which it was generated	Excess or unused heat in a process
boiling liquid	(self-explanatory)	Liquid undergoing a phase transformation to the vapor or gas phase
liquid holdup	delay in the feed of a liquid	**The mass or volume of liquid**

[8] Admittedly this is barely comprehensible to the layman. This definition would require further unravelling; does "constant with respect to spatial position" just mean "which do not move?"

[9] Again, further explanation required.

	into a process	**present on a component of the equipment at a particular time.**
liquid broth	a thickened version of a liquid	**A liquid mixture of chemicals designed to be suitable for growth of cells**
vapor-liquid equilibrium	A condition where vapor and liquid are stable in relation to each other	**A state where vapor and liquid are in contact with each other and the properties of both phases are constant.**

Bulk gas and *heat transfer coefficient* present some difficulty of comprehension but the claim here is that the other definitions do not, and that this feature of "easy technicality" makes the expert definitions highly appropriate textual material for engineering undergraduates beginning to specialise.

The 60 content words in these definitions account for no less than 5% of the 500000+-word chemical engineering corpus: the reader will on average see a word from this list ten times on every page!

Two tentative comparisons may be made here. While it is hardly worth making a detailed analysis, we may recall that the University Word List, with 850 words (and 2800 lemmas) gives about 11% coverage. The lesser efficiency is very marked. Second, the layman's definitions contain a total of 40 content words, which cover 2.8% of the total corpus (cf. 5% with the expert definitions).

4.3.4 Collocations: conclusion

The picture of collocation that emerges from the data above is as follows. As far as we can judge from chemical engineering, a substantial majority of the most frequent words in engineering sub-disciplines are significantly more frequent in that sub-discipline than in others. More than half, generally speaking, of the occurrences of these words are in complex noun phrases which we have called collocations. These collocations are thus key factors in these words' frequency and (since the collocations relatively rarely appear in other sub-disciplines) in their statistically significant specialisation. It seems therefore that they deserve attention in a lexical instruction programme. Furthermore, on examination, it emerges that as a class they represent certain features which make them pedagogically especially useful. This is partly to do with their relationship with the idea of technicality, which is in turn to do with their level of difficulty

The most frequent words are not at first sight a difficult set, in the sense of being highly technical. Some are both non-technical and transparent and thus seem easy, while others appear at least semi-transparent and very few look highly technical to the extent of being intimidating to the lay language teacher. But this is a slightly misleading impression. When one examines the collocations it becomes apparent that understanding them depends on a more precise knowledge of the individual words involved than is obvious to the layman. At the same time, because the words involved are not highly technical, the definitions are generally fairly accessible to the layman. We may therefore claim that the collocations, which often represent combinations of these frequent words, represent an intermediate stage of difficulty. They are accessible to teacher and student alike and are neither too easy or too difficult for either.

We therefore see texts defining noun phrases (collocations) as the appropriate vehicle for teaching lexis to second-year undergraduate engineering students at SUT.

There are a number of arguments against this approach. First, the texts as they stand do not present enough textual material. This is easily remedied by expanding the definitions and supplying more background. Second, there are far too many collocations for them all to be dealt with in a teaching programme. This question is discussed in section 5, where we examine how the corpus CorpusCHE is used to design the lexical syllabus. Perhaps the most obvious objection, however, is that the approach ignores so many words.

The first point in answer to this is that about 50% of text coverage comes from function words, especially determiners and prepositions. There is no reason to believe that these will be under-represented in texts like those seen in this section. It is not intended that there be much explicit teaching of these words, but that would be the case whatever the word list that was used. Virtually all the most common words in CorpusCHE will be represented in the texts defining collocations. Second, by ignoring "general" vocabulary we at least reduce one hit-or-miss element, in that different students will all have very different knowledges of any general word list. By treating these types of collocation through defining (unpacking) texts, we constantly recycle the most common chemical engineering words, and show how they combine with other words. In other words, we have moved from breadth to depth of knowledge.

Section 5: A collocational approach

Let us now see how we may describe the findings of this project as the basis for what we may call a collocational approach to teaching vocabulary and reading.

5.1 Collocation and the choice of texts: definitional texts

The data in section 4 showed that in undergraduate engineering textbooks the most frequent words (e.g. *heat*) occur so often because they are frequently modified (*specific heat, latent heat*) or used to modify other words (*heat exchanger, heat capacity*). The modification is a part of technicalisation, part of the creation of new knowledge, about different kinds or uses or measurements or characteristics of heat. Although there is of course a technical definition of *heat*, the difficulty of these phrases does not generally arise just from *heat* but from the combination with other words. The combinations may involve some opacity, or idiomaticity (*specific heat* is not heat which is specific). but even if they do not both of the constituent words are usually precise and thus the technicality is multiplied. Words like *heat* may be modified to the extent that the phrases thus formed are considered cumbersome (*the total heat content of a system*), with the result that they are compressed into single words (*enthalpy*). These in turn may be modified again (*partial enthalpy*), creating a higher level of technicality. But the most frequent words in corpusCHE, as we have seen, are not generally of this level of difficulty. We are thus entitled to at least guess that many of the collocations that the frequent words enter into may be within or near the conceptual range of SUT undergraduates and also perhaps the language teachers too. The intention, then, is that students are presented with texts whose function is to provide extended definitions of basic chemical engineering terms, in the form of collocations. It was shown in section 4 that such definitional texts are full of vocabulary which is frequent all over chemical engineering textbooks. They would also serve as introductions to fundamental chemical engineering concepts. This, then, is the first sense in which we may describe our approach as collocational.

5.1.1 Collocations generated

Let us consider an example of a definitional text, Text 1 below. It comprises an explanation of the term *gas phase*. It was written by an SUT chemical engineering professor. The only instruction given to him was to keep the level of technicality within the reach of an educated layman (nothing was said about collocation). From the evidence in section 4 we might expect that since the text is technical, albeit at a low level, it will generate a great deal of collocation in the form of complex noun phrases. This expectation is strongly borne out.

Text 1

A *phase* is a state of matter. *Gas* is one of these states. Other common states include solids, liquids, vapours, and plasmas. Gases have a very low concentration of molecules (number of molecules per cubic meter) compared to solid and liquid phases, meaning that gases tend to have low density, low viscosity, high diffusion rates, and the ability to expand and change shape to fill the space they are contained in. Individual chemical species, and mixtures of chemical species, may be gases depending on the temperature and pressure of the system that they are in. An example of usage is: "air is a gas mixture at atmospheric temperatures and pressures". *Gas phase* is a phrase that suggests what is being described occurs with all participating mixtures or elements being gases. For instance the description "...the conversion of hydrogen sulfide to hydrogen and sulfur at 1100 Kelvin is a gas phase reaction" indicates that both the reagent species (hydrogen sulfide), and the product species (hydrogen and sulfur) are gases throughout the reaction. The phrase may appear instead as a hyphenated word in some instances, such as *gas-phase equilibria*.

The text is about 200 words long. We find at least ten and perhaps sixteen relevant noun phrases, relevant in the sense that they are complex and technical in nature. It appears that they arise naturally during the defining, unpacking process. The texts are collocation-based as an inevitable result of being technical.

How the text is exploited by the teaching material for vocabulary learning or reading comprehension is a matter for section 3. At this point we will observe merely that we have to deal with its meaning. This will involve focusing on the following:

- the characteristics of gases
- how air illustrates the meaning of *gas*
- how the conversion of hydrogen sulphide to hydrogen and sulphur illustrates the meaning of *gas phase reaction*

which in turn probably involves dealing with the meaning of the phrases *gas phase, low concentration, low density, low viscosity, diffusion rates, chemical species, gas mixture, atmospheric temperature, reagent species, product species*, and *gas phase reaction*; and perhaps too with *liquid phase, solid phase, participating mixtures,* and *gas-phase equilibria*.

This, then, is another sense in which our approach may be said to be collocational. The choice of text dictates the necessity of dealing with collocation in the teaching material.

Note that text 1 is based on a prescribed syllabus item (*gas phase*) but in a sense creates its own new collocational syllabus, one that emerges from the text. These items behave, distributionally, in wildly varying ways. Consider ten clearly technical collocations from text 1 in Table 18 below. The table shows the frequency of these items in two corpora – corpusCHE (50000 words) and the longer, whole-textbook based corpusTXT (387000 words).

Table 18: Text 1 collocations' occurrence in other corpora

phrase	corpusCHE	corpusTXT
gas phase	19	171
low concentration	0	3
liquid phase	26	94
low density	6	0
low viscosity	1	1
diffusion rates	1	2
gas mixtures	8	24
reagent species	0	0
product species	0	1
atmospheric temperature	0	0

These figures hardly merit statistical analysis but obviously there is some correlation between the items' frequency in the two corpora. We thus have some evidence that only three occur with any frequency elsewhere; most, apparently, may be only of very limited use in other texts. How the teaching material deals with this is the subject of section 3.

5.2 Item selection

The whole thesis so far has been about producing the best list of items for a lexical syllabus: which words, or latterly collocations, to get students to learn. However, if we think in terms of collocations we have a huge number of items to deal with. As was pointed out in section 6, *gas* goes from being one item, one word type, in a list (e.g. Lfe) of 1000 or so, to being about 20 items (the number of collocations we find in corpusCHE, depending on where we set the recurrence criterion) in a list of imponderable length. It is obviously impossible to deal with this number of lexical items. It seems only common sense to choose the most frequent ones from the corpus, and that is the approach we will follow here. But this decision raises a number of questions (see section 3.1.2 below).

To digress slightly, it might quite possibly be better to choose items from the larger corpus corpusTXT, but this is not certain: corpusCHE covers a wider range of topics in the sub-discipline, although it is much smaller in total size. But part of the original purpose of this thesis was to investigate the use of such small corpora as corpusCHE, of which we have five (with the other four in the other engineering subdisciplines.) We will thus continue to use corpusTXT only to check the conclusions arrived at from corpusCHE.

5.2.1 Issues concerning collocational frequency as a selection criterion

Frequency seems to represent a simple formula for choosing syllabus items, but one point of this section is to show the impossibility of establishing an all-embracing and valid formula. Some of the reasons for this are to do with the surprising vagueness of the term *frequency* (section 5.2.1.1 below). Others arise from the small size of the corpusENG and its resulting lack of reliability (section 5.2.1.2). Finally, we must consider certain other miscellaneous aspects of collocation (section 5.2.1.3).

5.2.1.1 The meaning of *frequency*

The data on collocations presented in section 4 was based on frequency in the corpus, but it did not literally show the frequency of collocations. What it showed was in fact was *the frequency of collocations/noun phrases of the most frequent nouns*. To try to illustrate, *gas mixture* appeared not because it is a very frequent collocation, but because *gas* is a very frequent noun and *gas mixture* is one of the more frequent collocates of that noun. This is not quite splitting hairs. There are also other ways of interpreting *frequency*. In the name of frequency we might also take those nouns whose occurrences contain the highest proportion of collocations (as with *gas*, for example, which occurred in collocations 66% of the time); or we could take those nouns with the highest raw number of collocations; we could even select those nouns with the highest frequency relative to the other sub-disciplines in corpusENG, on the grounds of specialisation. Describing the arguments for these different formulae is a lengthy and somewhat tedious process, which we do not propose to undertake in any detail. The different lists produced tend to coincide to a fairly large extent, and similar problems attend each of them. Given our aim to present learners with coherent, 300-word (on average) collocation-rich texts, there is a slight advantage to choosing *frequency of collocations of the most frequent nouns* as our criterion; it makes it slightly easier to write texts which have thematic unity (they can be about different collocations related to, for example, *heat* or *distillation*). But the advantage is marginal and it is far more important to point out that the "formula" must be tempered by judgment, as will become clear below. In fact it is the impossibility of producing a perfect formula that makes the discussion of the different definitions of frequency somewhat academic.

5.2.1.2 Reliability

By using a small corpus like corpusCHE we risk misrepresenting the collocational picture. corpusCHE is derived from only 10% of the text in the books it is based on so there is a danger that it will miss too much relevant data. We can check this, albeit not definitively, by comparing occurrences in corpusCHE with those in corpusTXT. Incidentally the results are similar whichever definition of frequency we apply).

Tables 19, 20 and 21 below show frequency of collocations of three common words, *heat*, *gas* and *temperature*, in corpusCHE and corpusTXT.

Table 19: *Heat* collocates in corpusCHE and corpusTXT

heat collocate	in corpusCHE	in corpusTXT
Transfer	42	156
exchanger	9	61
exchangers	7	17
latent	7	9
capacity	5	124
conduction	5	0
flux	4	5
exchange	3	21
specific	3	9
molar	3	2

Again, there seems little sense in subjecting the figures above to any statistical analysis. The best we can say is that if we select for teaching the *most frequent* corpusCHE items, it appears that we will be teaching items for which students will find a use. But as we move down the corpusCHE frequency column, we see at frequency 5 one phrase (*heat capacity*) which appears very important, and another (*heat conduction*) which is apparently of no use[10]. A similar picture emerges with *gas*:

Table 20: *Gas* collocates in corpusCHE and corpusTXT

gas collocate	in corpusCHE	in corpusTXT
phase	19	171
perfect	15	7
natural	11	41
mixture	8	24
entering	7	9
stream	6	19
flow	6	14
velocity	5	7
feed	4	4
ideal	3	75
absorption	3	18
behavior	3	3
film	3	0
product	3	3
law	3	46
rate	3	1

[10] Actually this is not the case, but we cannot see that from the figures

In fact a *perfect gas* (row 2 above) is the same thing as an *ideal gas* (row 10), but while corpusCHE tells us that the former is more common, the opposite is in fact the case - this is shown by the corpusTXT frequency figures and the fact that my chemical engineering informant (a practicing, publishing engineer) had never come across the phrase *perfect gas*. Again the items at the top of the corpusCHE frequency list are mostly confirmed as useful but those at the lower end are unreliable. To take a final example of this type:

Table 21: *temperature* collocates in corpusCHE and corpusTXT

temperature collocate	in corpusCHE	in corpusTXT
distribution	8	1
profiles	8	3
profile	7	6
reactor	6	24
outlet	5	1
low	4	3
rise	4	4
high	4	10
difference	4	16
limit	4	1
constant	3	28
dependence	3	12
differences	3	3
increasing	3	15
lower	3	4
surface	3	5
wall	3	0
composition	3	4
pressure	3	0
stage	3	0

Perhaps because none of the *temperature* collocates are very frequent in corpusCHE, it is quite unpredictable whether their frequency in corpusCHE is replicated in longer texts.

On the other, literal, interpretation of collocational frequency a similar picture emerges.

Table 22: Most frequent corpusCHE collocations' occurrence in corpusTXT

phrase	corpusCHE	corpusTXT
mass transfer	46	151
flow rate	45	489
heat transfer	42	155
boundary conditions	28	36
liquid phase	26	93
gas phase	19	171
energy balance	18	254
material balance	17	143
perfect gas	15	7
plug flow	15	128
pressure drop	13	80
residence time	13	51
surface area	13	45
rate constant	12	55
(x)-order reaction	12	171
transfer rates	12	1
natural gas	11	40
transfer coefficient	11	94
transfer medium	11	1
vapor pressure	11	114
feed comp'tion	10	12
pore diffusion	10	0
reaction rate	10	254
unit time	10	23
unit volume	10	37

We see that many of these phrases, relatively common in corpusCHE, are also common in other corpora; it is clear from a glance that these collocations are of general use. As we move down the corpusCHE frequency column, however, exceptions to this rule start to appear. For example *pore diffusion, transfer rates* and *transfer medium* hardly occur at all outside corpusCHE.

There must be a element of randomness about the corpusCHE figures, because of the small size of the corpus. As has been pointed out already, includes only 10% of the books from which it is derived. This randomness is multiplied if we compare corpusCHE with corpusCHE2, so it is not surprising that (as is clear from a glance at Table xx above) as frequency in corpusCHE declines, frequency in corpusCHE2 becomes less predictable. This is in spite of the fact that the two corpora are based on books on the same subjects. corpusCHE3, which is on different subjects within the subdiscipline, appears almost as easily predictable from corpusCHE.

Incidentally, on the literal interpretation of frequency, we would probably have to go down the frequency list as far as those items occurring only five times; as Table 23 shows, the occurrence of such items elsewhere is wildly unpredictable.

Table 23: Other middle-frequency collocations in corpusCHE and corpusTXT

	Corpus CHE	Corpus CHE2	Corpus CHE3	Corpus TXT
activity coefficient	5	1	3	1
boundary line	5	0	0	0
distillation boundary	5	0	0	0
heat capacity	5	5	10	124
liquid flow	5	11	0	10
pinch point	5	3	0	0
stage requirements	5	0	0	0
surface reaction	5	4	0	55
turbulent flow	5	4	0	5

This would bring us to the situation where, having chosen items on the basis of frequency, we discover that they are not in fact frequent. If we are basing texts around the most frequent nouns and some of those nouns' collocations turn out to be less frequent elsewhere than we hoped, then this is less inaccurate than basing texts around collocations which we expect to be frequent but which in fact are not. This marginal point inclines us to the former, non-literal interpretation of *frequency*.

5.2.1.3 Miscellaneous factors

The necessity of using judgment in addition to a formula is shown by various other aspects of the data. First, some of the most common nouns have very few collocations at all (e.g. *problem*). Others (e.g. *case, conditions*) carry very little content and it is difficult to imagine basing a thematic unit around them. With others, the obvious problem of technicality arises: It is doubtful, for example, whether teachers will be able to make much sense of the following definition of the common collocation *differential equation:*

> A differential equation is an equation that defines a relationship between a function and one or more derivatives of that function. Let y be some function of the independent variable t. Then following are some differential equations relating y to one or more of its derivatives.
> The equation

(Equation omitted)

states that the first derivative of the function y equals the product of and the function y itself. An additional, implicit statement in this differential equation is that the stated relationship holds only for all t for which both the function and its first derivative are defined.

(http://www.physics.ohio-state.edu/~physedu/mapletutorial/tutorials /diff_eqs/intro.html)

The size of the problem that a text like this presents again depends on how great a part teachers are required to play in the programme, how great their tolerance is for technical language, whether or not other help with technical understanding is available, and other factors. A formula cannot take account of these. Subjective judgement must supplement the quantitative data.

5.3 A collocation list for chemical engineering

Our final objective in relation to the lexical syllabus is to present a list of collocations for undergraduate chemical engineering students, comprising 24[11] clusters of collocations based on a selection of the most frequent nouns in corpusCHE. A list of 24 of the more common nouns that tended to collocate in noun phrases was given to a chemical engineering professor here. He was asked to choose from the collocates under each noun a group of four or five that were clearly related, and that could form the basis for one 1000-word text or four 250-word texts, or some combination in between. Obviously this grouping is better undertaken by a specialist than by a layman. Thus, for example, one way in which the phenomenon of heat transfer occurs is through heat conduction, and this process lies behind the idea of the heat exchanger: this group of collocations forms the basis for a 500-word text, which may be seen in appendix 2.

1. Processes
    a. Chemical processes
    b. Process conditions
    c. Process equipment
    d. Process dynamics
2. Reactors
    a. Flow reactors
    b. Tubular reactors
    c. Batch reactors
    d. Catalytic reactors
3. Balances
    a. Balance calculations
    b. Balance equations

[11] This number is taken from the number of weekly units specified in section xx.

c. Mass balances
d. Energy

4. Equilibrium
    a. Equilibrium conditions
    b. Equilibrium models
    c. Equilibrium phases
    d. Equilibrium stages
5. Chemical Reactions
    a. Reaction equations
    b. Reaction rates
    c. Surface reactions
    d. Single phase reactions
    e. Two phase reactions
6. Separations
    a. Separations processes
    b. Separation operations
    c. Ideal separations
    d. Separation factors
7. Stages
    a. Equilibrium stages
    b. Countercurrent stages
    c. Crosscurrent stages
8. Distillation
    a. Distillation columns
    b. Batch distillation
    c. Azeotropic distillation
    d. Extractive distillation
9. Transfer
    a. Transfer rates
    b. Heat transfer
    c. Heat transfer coefficients
    d. Heat transfer correlations
    e. Mass transfer
    f. Mass transfer coefficients
    g. Mass transfer correlations
    h. Mass transfer operations
10. Heat
    a. Heat transfer
    b. Heat flux
    c. Heat conduction
    d. Latent heat
    e. Specific heat
    f. Heat exchangers
11. Diffusion
    a. Molecular diffusion

REFERENCES

Banks, D. (2005). On the historical origins of nominalized process in scientific text. *English for Specific Purposes* 24/3, 347-357.

Cobb, T., & Horst, M. (2001). Reading academic English: carrying learners across the academic threshold. In J. Flowerdew & M. Peacock (eds.), *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Cowan, J.R. (1974). Lexical and syntactic research for the design of EFL reading materials. *TESOL Quarterly* 8/4, 389-398

Halliday, M.A.K. (1985). *An introduction to functional grammar*. London: Edward Arnold

Jordan, R.R. (1997). *English for Academic Purposes*. Cambridge: Cambridge University Press

Laufer, B. & Sima Paribakht, T. (1998). The relationship between active and passive vocabularies: effects of language learning context. *Language Learning* 48(3), 365-391.

Liu, E.T.K., & Shaw, P.M. (2001). Investigating learner vocabulary: a possible approach to looking at EFL/ESL learners' qualitative knowledge of the word. *International Review of Applied Linguistics* 39, 171-194.

Martin, J.R. (1990). *Literacy for a changing world*. Hawthorn, Victoria: Australia ACER.

Martin, J.R. (1993). Genre and literacy: modeling context in education linguistics. *Annual Review of Applied Linguistics*, 13, 141-172. Cambridge: Cambridge University Press.

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

Pueyo, I.G. & Val, S. (1996). The construction of technicality in the field of plastics: a functional approach towards teaching technical terminology. *English for Specific Purposes* 15/4, 251-278.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

Selinker, L., Tarone E., & Hanzeli V. (eds.) (1981). *English for academic and technical purposes: studies in honour of Louis Trimble*. Rowley, Mass.: Newbury House.

Ward, J.W. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12/2 309-324

Ward J.W. (2004): The vocabulary knowledge of SUT undergraduates. *English Language Studies Forum* 1/1.

Yang, Huizhong (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing* 1 (2), 93-103.

APPENDIX 1

Most frequent words in five subdisciplines

Most frequent (frequency>4) civil engineering words

the of in is and to a for be are as at by that on with or this stress load from will it if beam moment concrete shown can section shear steel force strength an used which not member have design may using value beams loads example than each these column area members when all pressure required determine must end reinforcement use soil we then two such water forces strain one table values between method has case its equation joint should determined only thus where equal maximum also compression given depth length moments other point bars truss there columns sections frame same weight assumed been under factor no stresses since support tension would applied failure live loading total deflection effective tensile vertical but line being more span equations obtained direction due footing axial so some test assume diagram equilibrium lateral structure bending both building capacity floor bar into minimum results through analysis flexural three conditions distribution part size time wall dead limit tests after however nominal solution about face number service because provided ratio acting joints made specimen they thickness top unit very less slab therefore curve distance first simple their width any connections yield following most based cracking cross fixed horizontal increase state usually work zero center compressive computed elastic plastic result supported was above bottom factors now pore prestress stirrups surface cases displacement eccentricity large necessary over particular principal problem taken calculated having net pin prestressed tendon consider considered constant structural tendons those ultimate available buckling crack function girder greater placed procedure system actual below body boundary bracing construction diameter either form initial web axis bearing element flow found full influence internal points rectangular resistance safety standard type carry consolidation critical draw exterior factored material requirements addition clay code connected corresponding flange free loaded possible prestressing reinforced see structures although average certain chapter different larger longitudinal place plane plate plates select torsion up zone before bolts change collapse condition does effect ends external general girders saturated settlement single small subjected virtual were within against along appropriate assuming centroid cover deflections described gives high lower modulus percent provide reactions resulting side smaller spacing terms uniform well base figure follows known reduced respectively resultant specification supports usual without becomes compute connection defined diagrams he here interior least measured midspan near normal note range square stage types way applying buildings do footings grade level lines little obtain occur practical reached spans volume while close developed development dimensions generally inclined increased long negative per positive quite represented solve times torsional unknown upper yields according application approximate approximately coefficient combined during further half important mass often out produce produced represents rigid seepage similar stiffness strengths take amount could edge entire exceed normally properties resisting several show undrained walls common composite creep discussed furthermore instantaneous linear location losses numerical permeability practice reduction relationship rotation selected self shows sizes slope theory transverse until action again biaxial curves degree equivalent even four frames hence layer magnitude manner manual means period shrinkage slabs slip specified stirrup subject sum theorem additional anchor apply bridge cantilever caused consolidated designed displacements estimated foundation higher including loadings low make many non percentage plotted problems rather requirement resist shape steels straight strands trial typical welding wind act allowable angles around characteristic+ data deformation designer distributed done excess expression far information lacing limits local occurs opposite round sand superimposed try unbraced another applies balanced basic causes check control desired detailed directly established few find flat formed formula fully give hand height indicated instead just lightest main model much parameters permitted plus proposed referred relative requires rise sides simply situation states testing transfer treated trusses units weld acts areas called closed compared component computation consideration cracks deck depends directions easily economical energy exceeds expressed final finally floors great left major methods needed overload region removed right roller rupture sectional selection sidesway statically stressed superposition technique uniaxial uniformly welded welds what years accordance angle back behavior braced calculation calculations codes combination commonly compatibility components content cylinder difference effects flexure how immediately inelastic inertia ordinary portion principle procedures real relatively repeat respect series shapes short stages static substituting though throughout true various whole wire actually adequate always appendix approach assumption bend cause changes chosen clays compaction consequently considering consist dependent depending drained eliminated engineer examples excessive follow gravity kips limiting materials mechanism metal nearly next order parallel particularly placing presented pressures pretensioned previous process provides reference representing seen shears specifications suggested symmetrical takes them tied ties across adjacent advantage analyzed arm beyond bolted bridges built calculate causing confinement contact continuous controls criterion days deformed details difficult direct drawn elasticity electrodes envelope extend figs friction hardening indeterminate indicate indicates intermediate isotropic later matrix might multiplied noted outside present probably proportion provisions reinforcing related repeated represent require sands scale situ situations six specific splitting starting stiff strap sustained tensioned triangular upward uses variation written accordingly added advantages alternative apparatus applicable block cables carried cell clear combinations core decreases density depths develop down ductility equipment except fall filter flexible giving gradually group hinges identical immediate imposed increment independent laterally levels lever lies lightweight lips longer need noncompact notice overall panels parts perhaps rate reaction reader referring revised rolled satisfied second segment separate set she sign somewhat specimens split still story sufficient surfaces tables temperature tends together unknowns variable whether wide widths year you account accurate anchored arrangement basis behaviour bolt bound cantilevered clockwise compare complex compound computing concentrated conjugate considerable coordinate coordinates correct correction cup deep deformations dense determining diagonal distributions divided downward drop earth elements embedment estimate excavation exist expressions fact fill followed framing homogeneous hooked hydraulic impossible inadequate include included increments inside interest involves joists magnitudes middepth modified nor once particles paths perfectly permits plot post prevent primarily ranges recommendation+ recommended redesign reduce reducing relation significant simplified soils solid space spaced special stable steps strand student study subsequent suitable tie treatment truck tube unless varies working yielding acceptable appears appreciable assigned axes begin blows buckle cable cap centre circular clearly closely completely computer concept consists constructed continued contribution controlling corrected corrosion couple cured curved desirable determinate determination dimension dissipation double drainage dry earthquake empirical enough exact extreme fail fatigue field five flanged forms formulas frequently going gradient gross ground grout heavy heel hold holes hook idealized illustrate illustrated illustrates inch inches increases increasing individual inspection installation integration interaction keep law leg like likely limited locations loss making mode occurred off overburden panel path peak pedestal perimeter permit plain portions position preceding previously prior properly proportional ratios reasonable redistribution relating restraint retaining rotate rule satisfactory say sec sense severe shaded simultaneously slightly solutions solving start statics systems theorems toward transferred transformed transmitted unloaded vector view weights whereas accompanying accurately adhesion age allow allowed almost alone among analyze anchorage attempt author balance box care carrying centers central centroidal channels checked choice choosing circle comment compact comparison complete continuity damage dashed discussion eccentrically edges enables enclosed erection establish exactly exceeded excellent expected extending extensive fit flanges fracture front geometry global gradual greatest had hooks humidity idealization index initially instance intended intersection involve involving iteration laboratory last layers lightly limitations linearly liquid listed matter me measurements metric minor moves neutral original parabolic passes pi piles pinned plan possibility precast product profile proper providing purposes radial reach refer remain roof samples saving sequence serviceability seven shaped shearing snug spread stability stated stiffener stories strip stronger struts successive summarized supporting symbols tabulated taking techniques thicknesses third tight today toe toughness transmit turn twice us varying wires your

Most frequent (frequency>4) electrical engineering words

the of is a in and to for be that as by current at motor system are this with voltage from speed it power on we an shown can will phase if torque when equation circuit may or load control flux which frequency figure value rotor have field where source determine constant has time state line energy two input not one then inverter rated output three its zero function winding used armature operation thus machine dc example given response so also converter required these than into coil must because would ac between no resistance stator induction therefore magnetic step however loop only all controlled each factor obtained per such core equal high transformer same rectifier using pole synchronous transfer currents ratio systems switch losses other but connected large model terminal obtain air drive maximum root conditions magnitude negative assumed converters flow following gain range region been diagram find off characteristic design point motors during equivalent roots small there total case poles series angle operating section shows single some through closed electric form gap matrix turn values called change fundamental order starting use axis consider direction harmonic harmonics impedance steady supplied terms variable was both differential electrical feedback full low number problem slip very waveform less now over primary turns characteristic+ discussed equations note results voltages first positive supply switching up charge device iron mechanical peak about any applied side solution terminals made mode result since increase locus loss means most part second across chapter error expression overshoot turned type wave density due should signals variables windings continuous does junction limit necessary out performance reduced secondary considered curve described driven inductance leakage more sinusoidal under average calculate complex drift important percent phasor plot they applications assume cage component efficiency method potential proportional unity circuits computer force hence reactance table developed do gate much neglected signal without acceleration actual analysis curves defined effect generator ideal linear possible their us base cycle diode expressed machines network parameters process see similar what components forward half interval new square stable thyristor transistor unit while area capacitor devices increased let real rotating set storage condition delay determined friction many plane relationship respectively rotational saturation short test those usually within above assuming being braking controller designed draw estimate even excited higher how often operate path period provide represented select settling stability stored term were becomes block coils desired electronic end follows found illustrated layer occurs open position problems requires vector compared generated here induced larger link mass reverse temperature transient waveforms additional appendix approximately basic before corresponding data digital discussion effective four internal level parallel produce pulse reduce surface third transmission useful utility well class conductors down driving drop electrons fixed functions further gives having increases law lower material place rate size speeds switches after approach available body cause commutation compensator difference directly dominant drain equipment give initial practice represents running several specified thyristors transform addition amplitude calculated capability copper distance general greater known long operated our physical resistor resonant right squirrel suitable taken transition vary width along amplifier balanced battery common compare compute could different direct domain drives dynamic either free graph great minimum nearly occur opposite polarity standstill start them theory written another antenna approximation assumption bridge capacitors changes cross cylindrical fact get lag latchup leg length lines location magnetizing measured normally noted points provided quantities quite referred represent require resulting simple various ampere application approximate assumptions bandwidth below cannot commutated conductor consists controllable coupling develop displacement distribution drawn excitation independent lagging loads main make produced rectifiers reference regulation relatively shaft show sum variation windage you able according although auxiliary capable constants conversion done entire external filter flowing good indicated inertia leading limits mutual needed practical provides relation relative remains ripple rotation seconds special standard structure take times together train until varying versus work write zeros accurate advantage again appears autotransforme+ based breakdown chopper chosen combination criterion damping degrees elements enough few flows frequencies inductor inversion limited magnetization maintain margin need negligible plant polynomial produces quadrant reaches reactive reason reduction requirement rheostat sectional self separately sequence slots space takes tank typical varied velocity view yields account actuator angular around become capacitance circular coefficients collector command complete correct corresponds coupled decreases depends depletion derive desirable division element energized exceed figs forms frame fully gating heating included increasing keeping kinetic lead magnet making mechanism meters methods models parts permits portion previous qua rapidly reached regenerative resistors rotate setting shunt simply sin smaller solar static topology vehicle wire added advantages anode applying arrangement associated back bipolar capacity changed chapters consist controlling cost course deceleration detail determining dissipated essentially examples exciting fall fed final fluxes freewheeling generating goes gravitational held identical including injected inside inverters left levels listed manner ohmic presence radius rapid rating related remaining replaced resistances resistive rise row salient sampling seen sensitivity separate shorted simplicity still symmetry uniform weakening adjusted aircraft allows almost already appropriate cable calculations camera causes channel clearly critical cycles decrease degree deliver depend derived dimensions discharge dissipation distortion earlier eddy electronics except exist exponential factors far finite generators had hand hold horsepower include indicates infinite introduction inverse lateral maintained marked measure minimize move moves next numbers optimum passes perpendicular phases phasors pilot present quantity quasi ramp reel reluctance representation semiconductor solve states study substitution supplies symbol traction transformation twice utilizing varies wide wish wound achieved adding alternator always amount amplitudes analog appear apply areas balance bound branch calculating center close column commutating comparison compensation composed conduct conducting connecting damped definition dependent describing determines development earth easily equals expect forced formed groups heat hole horizontal impedances importance impulse includes industrial information initially injection instant just like linked little magnitudes maintaining minimized moved my nominal obtaining occurring outer paths permeability permeance permissible plugging program pull ratings readily reduces repeat repeated schematic scheme selected shall significant situation snubber solid solving supplying synthetic turning upper uses utilize utilized vectors vertical via ways wheel whose why words works absorbed act alone analyzed automobile avoid axes bars behavior better bias biased bode calculation caused chart circle cloud combinations combined comparing compensated considering contains delivered designer diameter discontinuous disturbances dividing double electricity equilibrium evaluate exactly examine excessive expected expensive explained expressions faraday fifth forces formula fraction generate goal handling holes ignored illustrate illustrates inductances integral involved itself keep kept kit later leads liquid loading longer matrices might modern modulated modulation moving nature nonlinear nonzero origin overmodulation particular periodic placed plotted plus predict presents procedure product properties purpose remain remainder requirements resultant reversal robot run sampled satellite sides simplified slope sometimes sources specific statements steel student summary taking tape techniques though too toward unchanged understand volts whereas whether accelerate actually add against alternate alternative am answer antiparallel apart applies approaches atom automatic basis book call cases chemical clear complicated compressor consequently conservation considerably consisting demonstrates describe diagrammed diodes disadvantage discrete doping drained drops dual duty effectively effects electromagneti+ eliminate eliminated emitter engineering equally essential evaluated evaluation excess expansion express fluid follow foregoing fortunately fourth front generation geometry group guide hard highly hoop human increasingly instantaneous integration intersection intervals lags largest last least letters lifetime linkages major mathematical mean motion mover nameplate narrow natural near neutral normal north notches numerical observed obviously ohms once overcome own oxide pairs partial perform permanent permit pin preceding presented prime principle processes pulses purposes quadratic quality quickly rather recovery reducing refer regions relationships remember removed representing respect rings rotates rotors said saturated say sections sensors session shaded sign significantly similarly simpler sinusoids sizes sketch steering steps surfaces technology tells tend tests thickness throughout transformers transport traveling true typically unstable user valve variety wheels yield

Most frequent (frequency>4) industrial engineering words

the of a is and in to that for are be this as with we if on an by it or model two each at from can one will design random mean not time then which factor probability these example data system number all table have experiment analysis process simulation distribution may three there such effects factors value when variable variance other more used has first block effect only sample squares levels test given some using any interaction since would between error no but was so state shown than variables standard thus equation four rate order treatment where run set use also event sum independent means now results what following level factorial observations problem suppose its were values deviation follows interactions blocks consider normal average however population should freedom customers methods let poisson made same section total been chapter distributed service function having about into both experimental find fixed per square designs percent quality single they cost degrees different found could second treatments within case large estimate expected times computer randomized regression experiments machine main must five result method hence how often type up customer determine over possible see solution above combinations equations exponential obtain show significant temperature figure most way equal events linear taken terms assume called either note out plot tests whether being chosen control many hypothesis residuals size difference interval models statistical do less point present their very general he new replications response because defined line might observation points program samples selected under after arrival confidence required shows significance arrivals denote language obtained does here numbers queue several small confounded contrasts during hand our randomization real replicate those before chart inventory loss operator operators probabilities procedure replicates testing usually amount considered day experimenter further interest parameters yields anova coin column illustrate important simple term through units algorithm approach assumption available desired flow form investment need output product replication systems them associated cell components hours least measures until another appropriate contrast differences hypotheses incomplete measure occur orthogonal particular randomly runs server strength sums take upon again always balanced better cannot chain days graph heads last layout limits next occurs problems resulting said six zero area construct current defining estimates even generate known languages make multiple repeated states step us whose among arrive calculated cases certain component computed information measured person represent right shall study true well although assuming combination common completely continuous discrete distributions errors exactly final give greater hour interested latin machines normally parameter performance ratio represents tested therefore too unit variability variation aliased change coefficients complete conditions corresponding definition easily fact length much nested part performed purpose research say space who appears assumed ball based car degree described equally gives good half increase just making minutes necessary original populations preceding respectively software solder statement third while without work yield additional alias black clear contains drawn exercise high initial likely low null page position presented production she similarly specified totals useful users you approximately comparisons complex conclusions defective implies improvement include individual left nonhomogeneous observed others power range white words against analyze appear become code coefficient consists decision depends determining directly discussed enough exact expressed few fractional go included involving life long parts people programs quadratic reduce require situation statistic statistics subjects ten usual validation behavior binomial changing course demand equals estimated exponentially finite fit formula fraction gene generating had higher his implementation item material measurements period placed positive produce property restrictions return seen simulating takes technique transition verify application approaches assumptions balls box brands class constant define easy extreme helpful illustrated increases infinite inspection instance letters logic management pair reason refer requires review special steel urn variances verification voltage waiting worth your affect along alternative analyst annual applied batch becomes below bonder brand calculate comparison computing confounding cycles describe develop down due elements end entire equivalent every examples express generated intensity involved letting like manner manufacturing mass measurement median minimum name notation noted once pounds previous price produced proportion quarter queueing quite raw remaining representation row rule seven smaller split starting steady still successive transformation types user wait walk whole write years added analyses assumes central classes collected congestion consisting criteria cumulative curve defects determined development differ divided draw effort eight experimentatio+ fish functions furthermore idea importance increasing indicates individuals interpretation know location mixed motion needed negative non obtaining operation outcomes perform predicted previously proper properties quantitative rather reasonable reduction relative resolution running scores sequence simulate specification surface theorem thickness throughout validating why world written actually addition altitude applications approximation attempt basic best center check choose cited clearly columns command communicates compare compared concerned conditional connect consequently cut denoted departures dependent detectives diagram estimating exist fair fourth frequently get giving independently introduced introduction itself lane larger largest limit lines lot managers mathematical mechanics minimal notice otherwise outcome outline place practice processes provide provides residual sets significantly sometimes specifically speed statements tail trials unbalanced uniform valid accessible according adjusted aliases already answer appendix blocking board cells centered characteristic charts checkout comparing computational compute concentration connected containing curves decided developed die discuss efficient engineer environment equality evidence explain far finally fitted follow genes highest identify illustration inferences instead items job lead lies manufacturer middle normality objective obvious operating option outside permutation plus polynomial pretest principal proposition questions re reader recurrent renewal representing respective restriction rules select selecting situations solving stationary structured subscript supermarket symmetric target techniques terminating tool tossed uniformly upper variates varied versus waits whenever able algorithms analytic analyzing arriving assigned axis basis begin boards cars choice choosing circuit close communications company computers conclusion considering controlled correlation density designed detect dice difficult discussion distance drawing employed estimator examining expensive experience finding formulas forth generally going groups handle help her histogram ideas including joints knowledge lifetime log lower nor off optimal perhaps positions practical procedures products profit proofreader proposed pure qualitative region relationship relationships remains repeat role sampling scheme sections served servers setting settings similar simply simulations sizes slope smallest spends standardized strong subjected success surprising tables taking task theoretical theory top understanding verifying view ways week width wishes year actual adequate allows applying approximate arrangement arrives bank battery breakdown burrs business calculating carry chance characteristic+ checking children claim closely collection combined comment conclude confound continues correlated costs covariance cross cutting department derivation desirable details difficulty doing done ease equipment explained extended extension force formal formulations fortran future gave great hardness heat hold humidity idle immediately improve incremental indicate influence initially input integral intermediate introduce investigate investor jobs joint keep kind law layouts list little logical longer look manager maximum memory memoryless modeled objects occasionally occurred orders ordinary pairs paper periods plots pressure profile programming pulp quantities recent reduced regenerative reject rejected related relatively represented ridge seem selection separate series serviced signs simulated solve somewhat structure studied subintervals subject subscripts suggest summarized text transactions transient try underline vary whereas wish word ability accept accepting accommodate adjustment advantage aid allowed alternatives amplitude analyzed appearance apply areas assembly assess automatically batches beam beyond bound breaking calculations call carefully changed changes chapters chi child coins come communicate communication concentrations concept correct corrected create criterion cube cycle depend depot derived describes designing developing deviations dimension dimensional direction disciplines discount dispersion effective efficiency electrical eliminate enter entities entries entry equispaced evens examine exceeds except execution exercises expanded expect expectation explanation exponentials fails favor fields fixture flips forward full generalized greedy greek group head highly holding identical identically illustrative improved impurities includes integer intended intervals involves issue labeled lack leads leave letter lie limiting losses major makes marble marked mask members metal mixture modeling namely narrow nature needs negligible nine nonterminating numerical odds old ones overall partition pattern percentage physical plotted policy poor portion ports possibly post potential predict presents probably proceed project put quantity quenching question rates readings recall referred refers reflect regardless remain remark removed requiring restricted reward rows scale shared sheet skewed skewness skills slightly specifications specify specimens stated statistically steps stock store strengths strips studies subsequent subtracting suggested suggests temperatures thereby throughs transistor trends trial underlying understand unequal unless uses varies varying wave whatever works

Most frequent (frequency>4) mechanical engineering words

the of and a is in to for are that at by be as with we this on from stress an flow equation shown or can it if section system beam two which determine point force velocity shear cross load maximum where these example when pressure may each equations not has given have then link one must mass obtain all conditions motion its material will between constant into solution strain since through first also displacement only same thus axis design value frequency plane center such forces position used results surface values end plate see body case find problem terms use consider function natural principal direction member using bending matrix applied bar linkage normal small so eq mechanism number figure method any equilibrium spring cam shaft speed zero area equal other time acceleration but coordinates elastic form free three points radius steel assume however stresses along cylinder deflection rate angular energy made subjected about because fixed length loads than was yield angle both rectangular show there circular second vector obtained ratio relative air couple diameter instant linear required would analysis beams boundary magnitude moment more sections under what control

factor following stiffness total volume axial change fluid shows work compute
curve damping gear let sec some state therefore pipe rotation been calculate
complex computer directions found profile frequencies uniform follower hence
known note now becomes called determined due general line slider substituting
components different four inertia parts power simple static amplitude behavior
chapter density diagram effects flows mm order right should stagnation vibration
again data distance modes relation result square synthesis temperature expressed
measured perpendicular real theory unknown wall crank curved here internal often
path respect response straight structure torque deformation input mode model part
particle presented properties they thickness tube were considered critical does
dynamic element fracture initial loading per positive possible problems rod type
criterion effect fatigue gives long plot their us condition defined frame linearly
location materials mechanical no nozzle previous similar specific table written
arbitrary axes cause coordinate distribution exit independent located most necessary
our out over radial respectively shape side step strains supported those acting
assumed corresponding differential fan fig friction links many plastic positions
potential size steady water aluminum cm column equivalent expression ideal
kinematic large need requires unit upon weight well another assumptions basic
before compare contact dimensional displacements evaluate follows less members
methods simply single standard useful above duct estimate law mean moments
moving pump related shock solve solved additional changes coefficients component
curvature done during field increase inlet lower minimum occurs portion procedure
solving systems take term torsion acts approximation attached coupling do ends
freedom functions fundamental gears greater half incompressible left performance
relations resulting several sin subject third window after centers constants
cylindrical depth described difference edge engineering excitation generalized
important itself modulus negative numbers period stage strength suppose symmetry
test thin variation virtual addition advantage applicable associated available base
cases check choose circle convenient coupler curves degrees derivative efficiency
either express extension external generally ground higher interval joint limit lines
locate neutral nonlinear parameters plates prescribed rigid safety sheave springs
tension theorem very viscous wheel width write yields absolute assuming back
based become bottom cannot coefficient composed composite concepts could cycles
developed figs flat fully information inner means momentum notice operating origin
present remain represents rise roller scale sectional six sum symmetric together top
types variable without according action always application card closed concept
crack drag edges elements entire extensional further graphical height imaginary jet
loaded outer parallel replaced represented streamline stroke usually varies within
your accurate although appropriate certain chain channel characteristic chosen
circumferentia+ comparison compressor cycle designed develop dimensions
discussion entropy failure generation horizontal layer manner mechanisms next
numerical occur particular passes plotted process reduced reference remains repeat
require rotate set shapes significant spectral structural substitution tangent torsional
twisting unknowns variables walls way apply average combined computed deformed
determinant double engineer examples face fillet gage geometric give head how life
longitudinal loop mach magnitudes might models move noted output physical pivot
probability property prototype random range reals relationship say similarity
solutions steps tensile times truss uncertainty undamped various versus while
yielding alloy approximate areas building calculated calculations clockwise
compared concentration concrete consequently designer downstream driving excited
fact formula gas gravity he heat indicates lateral letting low machine make

mathematical mechanics negligible opposite orientation outlet peak percent precision product pulse requirements return series smooth solid specified supports tank true typical vertical wide you across actual against algebraic am angles approaches approximations arc around axially being belt bernoulli best block cantilever choice completely core cylinders degree desired direct discussed elasticity error even finally gate geometry instead integral just kinetic lies limiting limits new offset operator pair program quantities quantity rather repeated requirement rest row study substituted sufficient treated upper vectors act adding approximately assumption blade box characteristic+ chart connected contains correct correspond definition derived describe directly drive driver every experience few five forms high lengths liquid much near nearly neglected nominal octahedral oscillation own pendulum phase pitch place polar preceding principle profiles proportional recall red refer resultant rotates rotating rotations selection shaped sketch slightly stationary still twist unity up valid variations whose adjacent analytical applications arm causes centroid clearly common compressive concentrated configuration construction continuity dashed determination distributed down driven errors essential evident exerted extensive fall flange flexure foregoing frictionless generated graphically identical illustrate impeller increased infinite influence initially inside integrating introduced inversion isothermal kg kinematics latter level locations loss main maintain me measure middle my net nonhomogeneous original pathline pi practical practice reduces referred relatively represent ring satisfied satisfy segments sign simultaneously states station stream structures suitable tangential technique teeth tests tip ts vane vanish wish accuracy added appendix approach away ball below clamped complete computation constraints converging course definitions depend depicted designs diagonal diagrams differ dimension disk dividing drawn dwell eigenvalues eigenvector eigenvectors eight employed environmental escape establish etc examine except exciter exists expand expected figures finite flexural follow formed get hand having hinge hollow hose improved increases indeterminate indicated individual integrate integration intersection introduction intuition laboratory later least local movable needed off orthogonal parameter perhaps pin piston placed predicted produce produces ratios reactions read reader reasonable reduce regime regions remaining resonance resonant said sash select similarly simplest simplified source streakline sufficiently taken them throughout too transmitted transverse tunnel until velocities view whereas wind wing wire words accelerations add aids allowed analogy anisotropic appeared applies band basis blocks bodies book brittle buckled carried chains clamp close cofactors concentrations considering consists contour cost counterclockwi+ couples cut define defines denoted depends derivatives device diameters difficult digital downward draw eliminated engine enough enters equally experimental extreme fails filter fit flexibility forced formulated foundation hangs harmonic hold homogeneous illustrated include including increasing integrated interested involute isentropic joined lie like linkages logarithmic lowered masses matrices mercury minimize odds orthotropic oscillating outside particles pass polygon pressures previously provides pure purpose reaches relate relationships representation rotated rubber rules seen segment self sense setting shafts shaker slope smaller space statement statical statically statics story substitute supersonic support taking testing text throat tooth trailing transfer translation uniformly usual viscosity vortex web why writing acceptable accomplished accordance achieve adiabatic allow alone applying approximated arithmetic buckingham buckling bump capacity casement columns combination compression considerations coupled decomposed delivered description devices diffuser dimensionless distributions ductile dyed elbow elevation ensure equating essentially establishes exceeding exceeds exerts exist experiment expressions factors final fix flux formulation frictional functional furthermore geometrically gradient greatly groups growth guided guiding hardening help holds hook hydraulic ideally ii implies importance indeed indicate inertial infinitesimal inspection instantaneous integrals isotropic junction km know knowledge laminar larger last limited localized logic major measurements metal midspan motor neglect neighborhood nondimensional notation obvious operation options oscillations oscillatory paper passed pile pinned planes port portions produced propeller proposed questions reason reducing regardless replace residual rigidity rotor rule satisfies scalar separate separated shall shortest significantly signs slice slide sliding sometimes special spherical stated streamlines subsonic suddenly summing superposition surfaces symmetrical tab task thermodynamics thick train transmission transpose treat treatment uncoupled units unloading vary wedge wheels yoke

APPENDIX 2: A definitional text

Heat is a flow of thermal energy, either within entities or between entities, because of a difference in temperature between two entities, or gradient in temperature in a single entity. *Heat transfer* refers to the process of the transfer of heat from one spatial area to another by the mechanisms of conduction (heat transfer purely by molecular transfer through a temperature gradient), convection (heat transfer via movement or circulation of a fluid), or radiation (heat transfer by electromagnetic waves).

Conduction is the most basic form of heat transfer, and can be modeled with a law that predicts the rate of heat transfer being proportional to the temperature gradient within the object. The temperature gradient is simply the rate of change of temperature with respect to the distance in the direction of heat transfer. This constant of proportionality is referred to as the thermal conductivity. The law is known as Fick's law.

Convective heat transfer involves a more complex mechanism, and is modeled via Newton's law of cooling. An example is heat transfer from a hot wall, at temperature T_W, into a colder, flowing fluid adjacent to the wall, which has a bulk temperature of T_∞. This system can be modeled by assuming that the rate of heat transfer is proportional to the magnitude of the temperature difference between the wall and the fluid, i.e. T_W, - T_∞, multiplied by the area of the surface which comes into contact with the fluid, which is the area through which the heat transfer may occur. The constant of proportionality in this case is a heat transfer coefficient, usually given the variable name h, which has units of W/ $m^2 \cdot {}^\circ C$. The convective heat transfer coefficient depends not only on the properties of the fluid, but also on the properties of the flow, and on the properties of the wall. In a limited number of situations it can be predicted theoretically, but in most realistic situations it must be determined from experimental measurements.

A common application of heat transfer in industrial situations is the use of heat exchangers to alter the temperature of a fluid by contacting it indirectly with another fluid of a different temperature in a heat exchanger. The fluids in a heat exchanger are separated by a wall (typically metal), and heat is transferred from one fluid to the wall, and then from the wall into the second fluid. In this instance the heat transfer is achieved through a combination of conductive and convective mechanisms, and the transfer is modeled using a relationship which predicts that the total rate of heat transfer is proportional to the temperature difference between the two streams multiplied by the area of surface through which the heat transfer occurs. The constant of proportionality in this case is referred to as *U*, the overall heat transfer coefficient (W/ $m^2 \cdot {}^\circ C$). The overall heat transfer coefficient may be determined from the two convective heat transfer coefficients (for each side of the exchanger) and the thermal conductivity of the metal wall, assuming the geometric configuration of the heat exchanger (such as the shape and thickness of the metal wall) is well defined.