The definitive version of this article was published as Watson Todd, R. (2017) An opaque engineering word list: Which words should a teacher focus on? English for Specific Purposes vol. 45 pp. 31-39. doi: 10.1016/j.esp.2016.08.003

An opaque engineering word list: Which words should a teacher focus on?

Abstract

Word lists have become influential in the last twenty years, but do not help teachers identify which words to explicitly focus on in the classroom. In this paper, I argue that words chosen for an explicit classroom focus should be words that students are likely to have problems dealing with autonomously, and that these are polysemous words where the meaning required is not the usual meaning; in other words, opaque words. The paper shows how to create a list of opaque words for teaching engineering English at a Thai university by comparing the meanings of words in the context against the main meanings given in the online dictionaries that students often rely on. The resulting list shows that most opaque words are high-frequency words with unusual meanings.

Keywords

word list; polysemous words; opaque words; engineering English

An opaque engineering word list: Which words should a teacher focus on?

The last twenty years has seen the publication of several influential word lists derived from corpora. Based on the assumption that the frequency with which a word is used is an indicator of its utility for learners, these lists primarily aim to help teachers "to set goals for their students' vocabulary learning" (Coxhead, 2011: 357). While the lists are clearly useful in setting targets for vocabulary assessment, in analysing text difficulty, and in modifying reading materials (Gardner & Davies, 2013), since most lists consist of 2,000 or so items, it is difficult to see how they can be used by teachers to select vocabulary for explicit teaching on a course. This paper shows how an initial word list of 500 words generated from a corpus of engineering textbooks can be filtered to produce a list of the most useful words to teach on the basis that words with opaque meanings (i.e. meanings that students are likely to have difficulty accessing by themselves) are those words where an explicit classroom focus is likely to be of most benefit.

1. Introduction

1.1 A brief history of word lists

The earliest major word list which is still influential is the General Service List or GSL (West, 1953; see Gilner, 2011 for a history of its development). This list aimed to identify the 2,000 most frequent words that are useful for students. Since the list reflects its age (for example, by not including any of the technological innovations of the last 60 years), a New General Service List has been developed by Brezina & Gablasova (2015).

While the General Service Lists aim to describe general English use, more recently a plethora of more specific word lists have been developed. The best known of these is Coxhead's (2000) Academic Word List (AWL) which aims to identify the most frequent and widely used words in academic English which are not on the GSL (perceived weaknesses in the AWL have led to two further lists of academic words by Gardner & Davies, 2013 and Paquot, 2007). In the same way that the AWL is more specific than the GSL, there are also several word lists more specific than the AWL, including word lists derived from corpora of engineering textbooks (e.g. Mudraya, 2006; Ward, 2009) and discipline-specific word lists such as the one for environmental science of Liu & Han (2015).

Nearly all of these word lists are generated based on the criteria of frequency, range and dispersion of words through the corpus. Some lists also use stopwords, or words that are necessarily excluded from the list (e.g. the AWL specifically excludes all words on the GSL). While there is little criticism of the idea of word lists in general, there have been numerous criticisms of the decisions made in constructing word lists. These include:

- Criticism of the composition of the corpus used to construct the word list (e.g. Hyland & Tse (2007) argue that the corpus underpinning the AWL does not give full coverage of the range of academic disciplines).
- Criticism of the criteria for identifying words to include on the list (e.g. Gardner & Davies (2013) in designing the New Academic Vocabulary List or New AVL question the decision to use the GSL as stopwords in constructing the AWL).

- Criticism of the nature of items to use as words. For example, the AWL is a list of word families, the new GSL uses lemmas, and Ward's (2009) engineering word list gives word types, with articles about these lists each criticising the choices of the others.
- Criticism of the use of surface forms as the basis for distinguishing between words. This includes issues of homonymy (Ming-Tzu & Nation (2004) found that homonyms are a minor problem in the AWL, but Hyland & Tse (2007) argue that words in the AWL take on different meanings in different disciplines) and how words are used (Hancioğlu, Neufeld & Eldridge (2008) argue that lexico-grammatical patterns of use should be considered in word lists).
- Criticism of the number of words included in a word list (e.g. Ward (2009) argued that lists of 2,000 words are too long for practical use).

This final point is central to the arguments I will make in generating an opaque word list, but all of these criticisms are considered in generating the various word lists in this study.

1.2 The usefulness of word lists for teachers

As mentioned above, word lists are based on the assumption that frequency of use is taken as an indicator of a word's utility for learners. While this may be true, from a teacher's perspective, what is needed is a list of words for which spending valuable classroom time teaching will yield the greatest benefit. These two principles will not necessarily lead to the same list of words.

The language to teach "has to be specified along two parameters: in terms of the objectives to be eventually achieved, and in terms of the process that has to be activated to get there" (Widdowson, 2003: 115). Word lists provide targets for eventual achievement, but say nothing about how those targets are to be reached. When word lists are at least several hundred words long and when classroom time is a scarce resource, teachers need to decide which words to focus on in the limited time available and how much time to spend on each word. These decisions are likely to aim to "provide the best *investment* for learning" (Widdowson, 2013: 11, italics in original). The existing word lists, then, can provide a very long-term goal for vocabulary learning, but in any particular course the teacher needs to choose items from a relevant word list to prioritise for the greatest learning benefit (even for a purposefully short list such as Ward's (2009) 299-word engineering list).

1.3 Prioritising items from a word list

In choosing items from a word list to focus on in teaching, there are two main considerations: how many words to choose, and on what basis they should be chosen. The number of words clearly depends on the amount of time available. For effective learning, each word needs to be focused on explicitly and needs to be retrieved several times (Folse, 2011) with three focused exposures likely to ensure learning (Edwards & Collins, 2013). Given that most courses have numerous objectives in addition to vocabulary learning, we might expect to ensure a word is learnt every hour or two on average (with retrievals and focuses spread over a greater length of time, but with several words ongoing at the same time).

There are several possible bases for choosing words to teach. The most obvious basis for choosing words is the same basis that forms the rationale of word lists, namely, frequency. The AWL, for example, is divided into ten sub-lists in decreasing order of frequency, so perhaps the teacher should focus on those words in the highest frequency sub-list first. However, the actual sequence of words by frequency is an artifact of the exact corpus used to generate the word list and a different corpus will give a different sequence even if the criteria for building the corpus are the same. It is also unclear whether the highest frequency words in a list are those which provide the greatest investment for learning.

An alternative basis is to focus on words with high learnability (i.e. words that are easier to learn) and teachability (i.e. words that can be easily taught) (Thornbury, 2002). Words with similar forms and meanings in English and in the learners' L1 are generally easier to learn, and concrete nouns are generally more easily taught than other vocabulary items. The rationale here is efficiency, with classroom time spent on vocabulary leading to the greatest vocabulary gains.

A third basis, opacity of words, is to some extent the opposite of learnability since for opaque words accessing the meaning is problematic making the words harder to learn. Given that classroom time is scarce, it should be devoted to those words which students will have most problems dealing with by themselves. While students may be able to guess or use dictionaries to find the meanings of words with high learnability or

teachability, for opaque words students may need to rely on the teacher's help to find the meaning and thus classroom time is devoted to issues that cannot be addressed in any other way.

Identifying opaque words is somewhat problematic, but Hsu (2014), in constructing a list of academic formulaic sequences, focused on opaque formulaic sequences and provides some criteria for identifying opacity. Polysemous words (in this article, I am using *polysemous* to mean having many meanings, irrespective of whether the meanings are related or not), words which do not have their usual meaning, and words with a derivational affix that alters the meaning of the base form (e.g. *verse - versed*) are most likely to be opaque. Of these three criteria, the second is a specialised case of polysemy and, although the most difficult of the criteria to evaluate, is the criterion most likely to cause problems for students working autonomously.

Words not having their usual meaning are also likely to be particularly relevant in English for Specific Purposes (ESP), since word meanings are often context-dependent and the specific contexts in ESP result in words taking on 'unusual' meanings. Such 'unusual' meanings may be a low-frequency sense of the word. For example, in engineering English *value* is more likely to mean a particular number rather than its usual meaning of worth, whereas in a random corpus of 100 lines from the British National Corpus (BNC) *value* meaning a particular number only occurs seven times. An 'unusual' meaning may also involve a low-frequency part of speech for a word. For instance, *constant* in engineering English is most commonly a noun meaning a fixed number,

whereas in the BNC only 5.6% of uses of *constant* are as a noun with *constant* as an adjective predominating. A word with an 'unusual' part of speech will also have an 'unusual' meaning, with the part of speech variant adding an extra layer of difficulty for a student dealing with the word.

Identifying words which do not have their usual meaning implies that the usual meaning can be identified. From a linguistic perspective, Cruse (2011) identifies five possible interpretations of the default meaning of words (e.g. the earliest recorded meaning, the most frequent meaning). From a student's perspective, however, for unknown words encountered while reading, the 'usual' meaning is likely to be the first meaning given in a dictionary (often based on one of Cruse's five interpretations of default meaning). In many contexts, students rely on bilingual dictionaries while reading and nowadays these are most commonly online, with research showing that students are especially reliant on Google Translate and bilingual dictionary websites such as wordreference.com (Jin & Deifell, 2013; Nilsen & Mandal, 2015). A brief informal survey in the context examined in this study confirms such student behaviours with Google Translate and the Longdo dictionary website being used most frequently. Relying on bilingual online dictionaries is problematic for polysemous words, especially when the required meaning is not the first meaning given. Nesi and Haill (2002) found that comprehension problems often arise because students focus on an incorrect sub-entry for a headword when using dictionaries, and Boonmoh (2003), investigating the same context as this study, found that students often only read the first sub-entry and ignore the others. If the meaning or part of speech of the word in the reading context is 'unusual', it will not be the first sub-entry in the

dictionary, causing comprehension problems for students relying on the first sub-entry. In such cases, the word being looked up is opaque, and these are the words for which the students will gain the greatest benefit from teacher help.

My key argument, then, is that, although frequency-based word lists provide a useful long-term goal for learners, they do not provide useful information for a teacher deciding which words to focus on in the classroom. For this, a short list of frequent opaque words (i.e. important words which students are likely to have problems dealing with autonomously) is needed. In this paper, I produce such a list for undergraduate engineering students and provide details about how the list was developed.

2. Material and methods

2.1 The context and the corpus

This study uses a corpus constructed to inform a specific teaching context. At a respected Thai university, the Faculty of Engineering runs both Thai-medium and English-medium undergraduate programmes. Most students entering the English-medium programmes do not have the English proficiency levels necessary to perform well in their studies, so a series of preparatory English language courses have been set up to help them. The corpus used in this study was specifically constructed to inform these courses.

The corpus, called the Engineering English Corpus (EEC), consists of the engineering textbooks from all 27 compulsory courses taken by students in the English-medium undergraduate programmes. As such, the corpus consists of the whole population of texts

in this context, so issues of sampling in corpus design are irrelevant. The final corpus is approximately 1.15 million words and details of its composition together with corpus-based materials for teachers and students can be found on the corpus website at http://crs2.kmutt.ac.th/ceem (Osment & Graham, 2013).

2.2 Creating a word list from the corpus

The word lists discussed in Section 1.1 above (with one exception) use raw frequency of vocabulary items as the prime criterion to generate the lists. However, as Gries (2015: 55) argues, "if one wants to teach the English of engineering, it would be useful to have a list of words that are more frequent in an engineering context than they are in general English". This implies that relative frequency may be a better criterion than absolute frequency, and that conducting a keyword analysis (as Paquot (2007) did) is likely to produce a more useful word list. Keyword analysis is a statistical technique, typically using the log-likelihood statistic (Rayson & Garside, 2000), which identifies words occurring significantly more frequently in a target corpus as compared to a benchmark corpus (Scott, 1997). To create a word list from the engineering textbook corpus, a keyword analysis was conducted with the BNC as the benchmark using the KeyBNC program (Graham, 2014). Selecting a cutoff point for words to be considered keywords is problematic since log likelihood values (and thus probability values) are highly sensitive to corpus size (Pojanapunya and Watson Todd, forthcoming). It was therefore decided to take the top 500 words in the keyword list as an initial word list, since, after filtering, this number would be likely to result in a short list of frequent opaque words which is the goal of this study.

This initial list was then filtered to remove inappropriate 'words', to account for word dispersion through the corpus, and to deal with issues of lemmatisation and homonymy. The five filtering stages were:

1. Non-words and abbreviations were removed (e.g. *x*, *p*, *SQL*).
2. Function words were removed (e.g. *of*, *an*).
3. Words appearing in fewer than 15 of the 27 textbooks were removed (e.g. *vertices*, *angular*).
4. The concordance lines of word types falling into a single lemma were checked to see if they had similar meanings and patterns of use. If so, they were combined into a single entry in the word list (e.g. *element*, *elements*).
5. For words with multiple parts of speech, the log likelihood values of the various parts of speech of the word were recalculated to see if they would still fall within the top 500 keywords. For example, *flow* as a noun was retained but as a verb was discounted, whereas *note* as both a noun and a verb was retained (and these were counted as separate items in the word list).

After stage 5, the keyword list consisted of 174 words. For each of these (following a procedure similar to that of Martinez and Schmitt, 2012), a random concordance of 50 lines was generated to identify the main sense of each word in the EEC. Where at least 30 of these lines indicated the same sense for the word, this was taken as the usual meaning of that word in the EEC. Where two different senses both appear common (at least 15 times each), further random concordances were generated to check if this pattern held and

then the two senses were separated as two items. An example 10-line concordance for *scale* is given in Table 1 to illustrate this. In Table 1, for lines 1, 2, 3, 5 and 10, *scale* means size; for lines 4, 6, 7 and 8 it means gradient; and for line 9 it means weighing machine. This pattern of frequency holds roughly across the corpus, suggesting that two meanings of *scale* should be identified and included as separate items in the word list.

[TABLE 1 AROUND HERE]

After these filtering stages had been completed, the word list consisted of 186 items. To give an idea of the composition of this list, the top 10 items are: *determine*, *flow*, *figure*, *temperature*, *energy*, *force(s)*, *pressure*, *function(s)*, *equation(s)* and *shown* (the full list is in the Appendix). The percentages of items occurring in some of the word lists discussed above are given in Table 2.

[TABLE 2 AROUND HERE]

The percentages in Table 2 suggest that the 186-item engineering word list includes a reasonable number of examples of all three of the main types of words traditionally considered in word lists: general (i.e. high frequency) words, academic words, and engineering discipline-specific words. For the purposes of this study, however, these categories are not especially relevant since it is unclear whether the use of words in engineering textbooks from any particular category is more likely to be opaque. Furthermore, 186 items, even though fewer than any of the other word lists, is still too

many items for teachers to deal with. Therefore, a further stage of filtering is needed with the goal of identifying the most opaque words in the list.

2.3 Identifying opaque words

In this study, opaque words are defined as words which do not have their usual meaning. To identify such words, we need to operationalise 'usual meaning' and this is done in two different ways. First, each word's meaning and part of speech in the EEC is compared against the word's usage in general English. To see if the keywords in the EEC have 'unusual' parts of speech, the main part of speech for each word in the 186-item list was checked to see if it was the most frequent part of speech for that word in the BNC. To see if the meaning of the word in the EEC is unusual, in the list of sub-entries for the word in the COBUILD dictionary the location of the main EEC sense was identified, with a focus of whether the sense was near the top (i.e. the more frequently used senses) or near the bottom (i.e. the less frequently used senses). The COBUILD dictionary was chosen since the sequence of sense sub-entries for a headword depends primarily on their frequency of occurrence in a corpus (see Krishnamurthy, 1987).

The second way of operationalising 'usual meaning' concerns how students typically search for meanings of unknown words autonomously. As we saw above, students typically use Google Translate and the Longdo dictionary website and tend to focus on the first sub-entry. Therefore, the parts of speech and meanings of each word in the 186-item list were checked against the parts of speech and meanings of the first sub-entry for the word in the two online resources.

For each word, then, the opacity of the word was evaluated in six different ways with each way generating a score of 0 or 1. The six criteria for opacity are:

1. The main part of speech of the word in the EEC is not the most frequent part of speech for that word in the BNC.
2. The main meaning of the word in the EEC is not in the top half of the sub-entries for that word in the COBUILD dictionary.
3. The main part of speech of the word in the EEC is not the same as the part of speech of the first sub-entry for the word in Google Translate.
4. The main part of speech of the word in the EEC is not the same as the part of speech of the first sub-entry for the word in the Longdo dictionary.
5. The main meaning of the word in the EEC is not the same as the meaning of the first sub-entry for the word in Google Translate.
6. The main meaning of the word in the EEC is not the same as the meaning of the first sub-entry for the word in the Longdo dictionary.

To generate an overall opacity rating from these six criteria, it is unclear whether they should be treated equally or whether any criterion should be given extra weight. Without information about whether usage in general English or position of a sense in a dictionary is more important, I will somewhat arbitrarily treat the six criteria equally, meaning that each word is given an opacity rating from 0 (transparent) to 6 (opaque). To check the reliability of rating words for opacity, a random selection of ten words was given to two raters, and an agreement level of 95% was found, suggesting that opacity ratings are

reliable. Opacity ratings for ten words from the 186-item list are given in Table 3 to illustrate how the opacity rating system works.

[TABLE 3 AROUND HERE]

3. Creating the opaque word list

After applying the six criteria to all of the 186 items, we can sequence the list based on the opacity of the items, with the opaque words like *constant* and *note (V)* at the top and transparent words like *determined* and *note (N)* at the bottom. On the principle that classroom teaching time is best devoted to focusing on opaque words, teachers should cover words near the top of the list in class.

In the teaching context this study is concerned with, 40 hours of classroom time were assigned to teaching reading and vocabulary. On the basis that a word can be learnt every hour or two, the list for teachers in this context should be 20-40 words long. In the opacity-sequenced list, there are 45 items with an opacity rating of two or more, giving a word list of a practical length which would also allow teachers some choice.

Several of the words on this list are homonyms, some by part of speech (indicated by a part of speech tag after the word) and some by sense (indicated by a different superscript number for each sense). Although difficult to assess rigorously, four of these appear to have restricted dispersion through the corpus, with two items (*net (N)* meaning a computer program and *block*^{1} meaning a section of computer code) restricted primarily to

textbooks on computing, and two items (*element*^{2} meaning chemical and *solution*^{2} meaning solvent) restricted primarily to textbooks on chemistry. These items were therefore removed from the list giving a final list of 41 items. These are shown in Table 4 with their rankings in the 186-item list, log likelihood values, opacity ratings, meanings, collocations and sample uses.

[TABLE 4 AROUND HERE]

The list of opaque words in Table 4 is provisional and individual teachers may choose to teach or not to teach certain words based on further criteria. For example, *model*, *column*^{2} and *plate* are sometimes used as loanwords in Thai with the EEC meanings and so may be familiar to learners. Nevertheless, the opaque words in Table 4 are generally the words for which learners are most likely to benefit from a teacher's explicit help.

In Table 2, we saw how the full 186-item engineering keyword list compared to other word lists. Table focuses specifically on the 41 opaque items, showing the extent to which these words appear in other word lists. Compared to the full keyword list, the opaque word list has slightly greater proportions of general and engineering discipline-specific words at the expense of academic words.

[TABLE 5 AROUND HERE]

4. Discussion

In this article, I have created a word list from a corpus based on the opacity of keywords on the principle that these are the words for which learners will gain the greatest benefit from a teacher's help. The list is necessarily much shorter than other existing word lists because there are clear limits to the number of words that teachers can explicitly focus on in the classroom. Its usefulness to teachers in other contexts is limited since a different corpus will produce a different initial keyword list and since the opaque words in this study may be transparent in other contexts. Nevertheless, I believe that the methods used to create the word list, although somewhat laborious, do transfer to other contexts.

Although it is unclear whether the nature of the opaque words selected from the keyword list is generalisable to other contexts, some interesting patterns do emerge. First, there appears to be no relationship between the keyness ranking of a word in the keyword list and whether it is opaque or not. This suggests that word keyness (and, by implication, utility) and word opacity (and by implication, need for learners to rely on the teacher's help) are unrelated issues and both should be taken into account, especially in teaching ESP. Relying solely on the traditional frequency-based word lists is insufficient for the best learning results.

Second, comparing Table 2 and Table 5 we saw that opaque words tend to be either general or discipline-specific words, and are less likely to be academic words. This conclusion, however, needs elucidation. Those word lists in Table 5 with relatively high percentages (e.g. new AVL top 1,000, the engineering word lists) all include high-frequency words (i.e. words in the top 1,000 in the GSL). The word lists with low

percentages (e.g. AWL, new GSL current words) tend to exclude such high-frequency words. The opaque words, then, are more likely to be high-frequency words, a conclusion we should not find surprising since high-frequency words are generally more polysemous than low-frequency words (Zipf, 1945), and only polysemous words can be opaque.

The role of polysemy in opacity highlights a weakness of the word lists discussed at the beginning of this article. Although word lists consisting of word types or lemmas distinguish between parts of speech, none of the word lists distinguish between different meanings of words. Only the original GSL of West (1953) pays any attention to multiple meanings of a word by providing percentages of frequency of the different meanings (for most of the opaque words identified in this article which appear in the GSL, the percentage of the meaning most used in the EEC ranges from 3% to 20%). Even for the GSL, however, whether a word should be included on the list is based on the overall frequency of the surface form. Word lists based on surface forms without accounting for different meanings may promote the teaching of new words over the teaching of new meanings for known words. However, Barchers (1988) argues that both types of teaching are crucial for effective vocabulary learning. For a word list to promote both types of teaching, it would need to be based on the frequency of the various senses of words rather than the frequency of word forms. To show how this might work, I have resequenced the GSL to show frequency by senses (see Watson Todd, 2016). In this sense-based list, we can see that *of* with different meanings appears four times in the top 20 most frequent senses. For our purposes, the word *mass* is included in the GSL as one of the 2,000 most frequent words in English and eight different senses are provided by West (1953). In the

list of all 10,000 senses of the 2,000 GSL words, the highest-ranked sense of *mass* appears at rank number 2,307; the sense in the EEC, however, is at rank number 5,914 suggesting that it is opaque. Although this list only applies to the 2,000-odd words in the GSL, for these words it simplifies the identification of opaque words as a lower-ranked meaning of a word (or a meaning which is too infrequent to be included at all on the list) is likely to be opaque.

5. Conclusion

In this paper, I have attempted to create a short list of words and meanings that a teacher should explicitly focus on in the classroom in a specific context. Although I started the analysis with a longer word list based on the traditional criteria used in creating word lists of frequency and dispersion, the main criterion for choosing which words and meanings should be included on the final list is opacity. This criterion should identify those words for which the learners would gain the greatest benefit from a teacher's help, since these are the words learners are most likely to have problems dealing with autonomously. The opaque word list consists of fairly high-frequency polysemous words where the meaning required is not the usual meaning associated with that word. Such an approach should mean that valuable classroom time spent on teaching vocabulary yields the greatest benefit.

References

Barchers, S. I. (1988) *Teaching Reading: From Process to Practice*. Belmont, CA: Wadsworth.

Boonmoh, A. (2003) Problems with using electronic dictionaries to translate Thai written essays into English. Unpublished MA thesis, King Mongkut's University of Technology Thonburi, Bangkok.

Brezina, V., & Gablasova, D. (2015) Is there a core general vocabulary? Introducing the New General Service List. *Applied Linguistics 36*(1), 1-22. doi: 10.1093/applin/amt018
Coxhead, A. (2000) A new Academic Word List. *TESOL Quarterly 34*(2), 213-238. doi: 10.2307/3587951
Coxhead, A. (2011) The Academic Word List 10 years on: Research and teaching implications. *TESOL Quarterly 45*(2), 355-362. doi: 10.5054/tq.2011.254528
Cruse, A. (2011) *Meaning in Language*. Oxford: Oxford University Press.
Edwards, R., & Collins, L. (2013) Modelling L2 vocabulary learning. In S. Jarvis, & M. Daller (Eds.) *Vocabulary Knowledge: Human Ratings and Automated Measures* (pp. 157-183). Amsterdam: John Benjamins. doi: 10.1075/sibil.47.08ch6
Folse, K. (2011) Applying L2 lexical research findings in ESL teaching. *TESOL Quarterly 45*(2), 362-369. doi: 10.5054/tq.2010.254529
Gardner, D., & Davies, M. (2013) A new Academic Vocabulary List. *Applied Linguistics 35*(3): 305–327. doi: 10.1093/applin/amt015
Gilner, L. (2011) A primer on the General Service List. *Reading in a Foreign Language 23*(1), 65-83.
Graham, D. (2014) *KeyBNC* [Computer Software]. http://crs2.kmutt.ac.th/Key-BNC/ Accessed 24.4.16
Gries, S. T. (2015) Quantitative designs and statistical techniques. In D. Biber, & R. Reppen (Eds.) *The Cambridge Handbook of English Corpus Linguistics* (pp. 50-71). Cambridge: Cambridge University Press.
Hancioğlu, N., Neufeld, S., & Eldridge, J. (2008) Through the looking glass and into the land of lexico-grammar. *English for Specific Purposes 27*(4), 459-479. doi: 10.1016/j.esp.2008.08.001
Hsu, W. (2014) The most frequent opaque formulaic sequences in English-medium college textbooks. *System 46*, 146-161. doi: 10.1016/j.system.2014.10.001
Hyland, K., & Tse, P. (2007) Is there an "academic vocabulary"? *TESOL Quarterly 41*(2), 235-253. doi: 10.1002/j.1545-7249.2007.tb00058.x
Jin, L., & Deifell, E. (2013) Foreign language learners' use and perception of online dictionaries: A survey study. *MERLOT Journal of Online Teaching and Learning 9*(4), 515-532.
Krishnamurthy, R. (1987) The process of compilation. In J. M. Sinclair (Ed.) *Looking Up: An Account of the COBUILD Project in Lexical Computing* (pp. 62-85). London: Collins ELT.
Liu, J., & Han, L. (2015) A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes 39*(1), 1-11. doi: 10.1016/j.esp.2015.03.001
Martinez, R., & Schmitt, N. (2012) A phrasal expressions list. *Applied Linguistics 33*(3), 299-320. doi: 10.1093/applin/ams010
Ming-Tzu, K. W., & Nation, P. (2004) Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics 25*(3), 291-314. doi: 10.1093/applin/25.3.291
Mudraya, O. (2006) Engineering English: A lexical frequency instructional model. *English for Specific Purposes 25*(2), 235-256. doi: 10.1016/j.esp.2005.05.002

Nesi, H., & Haill, R. (2002) A study of dictionary use by international students at a British University. *International Journal of Lexicography 15*(4): 277-305. doi: 10.1093/ijl/15.4.277

Nilsen, T. Ø., & Mandal, R. (2015) "Paper dictionaries are sooo outdated!!!": A study on Norwegian pupils' look-up strategies in the English subject. Fakultet for Humaniora, Norges Arktiske Universitet.

Osment, C., & Graham, D. (2013) *CEEM: Corpus-driven Engineering English Materials*. http://crs2.kmutt.ac.th/ceem Accessed 24.4.16

Paquot, M. (2007) Towards a productively-oriented academic word list. In J. Walinski,, K. Kredens, & S. Gozdz-Roszkowski (Eds.) *Practical Applications in Language and Computers 2005* (pp. 127-140). Frankfurt: Peter Lang.

Pojanapunya, P., & Watson Todd, R. (forthcoming) Log-likelihood and odds ratio: Keyness statistics for different purposes of keyword analysis. *Corpus Linguistics and Linguistic Theory*.

Rayson, P., & Garside, R. (2000) Comparing corpora using frequency profiling. *Proceedings of the Workshop on Comparing Corpora, held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000),* pp. 1–6. Hong Kong. Available at: http://www.comp.lancs.ac.uk/~paul/publications/rg_acl2000.pdf

Scott, M. (1997) PC analysis of key words - and key key words. *System 25*(2), 233-245. doi: 10.1016/s0346-251x(97)00011-0

Thornbury, S. (2002) *How to Teach Vocabulary*. London: Pearson.

Ward, J. (2009) A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes 28*(3), 170-182. doi: 10.1016/j.esp.2009.04.001

Watson Todd, R. (2016) A sense-based General Service List. http://crs2.kmutt.ac.th/sense-based_GSL/ Accessed 24.8.16

West, M. (1953) *A General Service List of English Words*. London: Longman.

Widdowson, H. G. (2003) *Defining Issues in English Language Teaching*. Oxford: Oxford University Press.

Widdowson, H. G. (2013) Closing the gap, changing the subject. In J. Hüttner, B. Mehlmauer-Larcher, S. Reichl, & B. Schiftner (Eds.) *Theory and Practice in EFL Teacher Education: Bridging the Gap* (pp. 3-15). Bristol: Multilingual Matters.

Zipf, G. K. (1945) The meaning-frequency relationship of words. *Journal of General Psychology 33*(2), 251-256. doi: 10.1080/00221309.1945.10544509

Table 1 10 concordance lines for *scale* from the EEC

1	not only on the general	scale	model of the prototype airplane
2	and processes operating at this	scale	focusing on landscape structure function
3	another or on a local	scale,	for example, failure to pass
4	a temperature on the Fahrenheit	scale	that is being discussed. We
5	be drawn to a reduced	scale	to fit on a sheet
6	k in the thermodynamic temperature	scale	in the English system is
7	determine priority if a rating	scale	is used operators should be
8	express temperatures on the Celsius	scale.	Today we no longer use these
9	this force using a spring	scale	such as might be found
10	observed on a small enough	scale	the discrete units or packets

Table 2 Percentages of items in the engineering keyword list appearing in other word lists

Comparative word list (with no. of items in list)	*Percentage*
GSL Top 1,000 word families	41.94%
GSL 2nd 1,000 word families	18.82%
AWL (570 word families)	27.96%
New GSL Top 1,000 lemmas	44.62%
New GSL 2nd 1,000 lemmas	30.11%
New GSL current words (378 lemmas)	4.30%
New AVL Top 1,000 lemmas	48.92%
New AVL Lemmas 1,001-3,000	4.30%
Ward's engineering word list (299 word types)	47.31%
Mudraya's engineering word list (top 100 word families)	39.78%

Table 3 Opacity ratings for ten words in the EEC keyword list

Word	*POS in EEC*	*Sense in EEC*	*Different POS in BNC*	*Infrequent sense in COBUILD*	*Different POS in GT*	*Different POS in Longdo*	*Different sense in GT*	*Different sense in Longdo*	*Opacity rating*
constant	N	fixed number	1	1	1	1	1	1	6
note	V	notice	1	1	1	1	1	1	6
class	N	set	0	1	0	1	1	1	4
plane	N	flat surface	0	1	0	0	1	1	3
section[1]	N	part of a chapter	0	1	0	0	1	1	3
relative	Adj	comparative	0	0	1	0	1	0	2
loop	N	circle	0	0	0	0	1	0	1
section[2]	N	segment	0	0	0	0	1	0	1
determine	V	calculate	0	0	0	0	0	0	0
note	N	brief record	0	0	0	0	0	0	0

Note: POS: part of speech
GT: Google Translate
V: verb
N: noun
Adj: adjective

Table 4 The list of opaque words

Word	*Keyword list ranking*	*Log likelihood*	*Opacity rating*	*Meaning in EEC*	*Collocation patterns*	*Sample use*
constant	22	3006.99	6	fixed number	is a -; the - k	a spring has a force constant k
note (V)	132	652.22	6	point out	- that	in this case note that when the element is supplying energy
acting	129	668.48	5	influencing	loads -; forces -; - on	the external forces acting on the system have no horizontal components
given	130	657.08	5	nominated	a -; - time; - value	shows that for a given value of g mass
used	25	2636.68	4	implemented	can be -; widely -; - as; - for; - to	a calibrated spring scale can be used for this purpose a carbon filter is used to remove volatile organic compounds
class	76	1192.46	4	set	base -; derived -; - definition	place the entire class definition in the interface
required	96	904.74	4	needed	time -; - to; - for	the time constant t determines the time required for the circuit to reach steady state
mass	15	3425.29	3	weight	of -; total -; - balance	equal to the total mass or energy leaving it
value(s)	26	2624.83	3	quantity	a - of; maximum -; - returned	what is the maximum value of tensile strength
point(s)	28	2575.28	3	location	floating -; melting -; - p	let the position of point p be x
plane	30	2535.02	3	flat surface	picture -; horizontal -; - stress	to rotate in a horizontal plane about a vertical axis
section	32	2278.94	3	part of chapter	see -; previous -; in -	as used in the previous section to obtain qualitative descriptions
load(s)	36	2181.14	3	weight	distributed -; - factor	the total reduced by the distributed load in doing this
reaction(s)	77	1189.73	3	interaction	chemical -; - rate; - between	corrosion is a result of the reaction between the pipe material and its environment
consider	80	1187.14	3	ponder	for example -; let us - ; we will -; - what	let us consider what the algorithm has done

string	102	847.82	3	sequence of characters	- containing	formulate a string containing the values of r
net (Adj)	104	838.12	3	remaining total	no -; - energy; - force	to estimate the net energy change per mole
tree	118	739.17	3	hierarchy	spanning -; search -; binary -	can be represented by the binary tree in figure 2
reference	123	690.36	3	benchmark	inertial -; secondary - ; - frame; - state	the procedure based on this reference state is sometimes called
resistance	128	674.28	3	process of opposing	air -; corrosion -; electric -; rolling -	the rolling resistance of a bicycle is negligible
body	144	584.42	3	object	free -; rigid -; - diagram	the two figures are not free-body diagrams
standard (Adj)	147	582.80	3	normal	- atmospheric pressure; - conditions	assume the air is at standard atmospheric pressure
path	149	575.53	3	course of travel	shortest -; Euler -; - compression	it is possible to find a shortest path from s to any other vertex
moves	183	439.67	3	travels	particle -; - along; - through	a particle moves along an x-axis in such a way that
frame(s)[1]	184	435.93	3	theoretical system	reference -; inertial -	in one reference frame can they be found at rest
flow	2	8113.49	2	stream	maximal -; laminar -; mass -; steady -; - augmentation	the value of a maximal flow is limited by the capacities
figure	3	7338.09	2	diagram	accompanying -; see - ; - shows; - illustrates	the accompanying figure shows the velocity versus time
force(s)	6	6029.94	2	strength	driving-; frictional -; gravitational -; - exerted	the conservative gravitational force exerted by the surface
obtain	88	1052.53	2	find	we -; to -; - information	from this method we obtain information about the flow
relative	90	1032.66	2	comparative	- humidity; - to; - velocity	the bar slides relative to the sleeve
model	101	865.15	2	system of	data -; your -; - tests	you must infer the data model indirectly

				principle		
differential	120	718.26	2	derivative (maths)	partial -; - equations; - element	reduces the partial differential equations to ordinary differential equations
static	133	651.72	2	at rest	- determinacy; - friction; - pressure	the coefficient of static friction between the sheets of wood
column[1]	135	635.14	2	pillar	water -; - chromatography	this reduces light in the water column
column[2]	136	635.14	2	vertical row	each -; - name	assign an appropriate column name to the computed results
block	143	592.93	2	solid mass	inner -; - spring	by considering the block-spring combination shown
standard (N)	146	582.80	2	level of quality	national -	for American national standard abbreviations, see
find	155	548.77	2	answer	to - out	to find out which one is lighter
plate	158	529.01	2	sheet	flat -; steel -	the wing can be regarded as a flat plate
scale	174	452.46	2	gradient	Celsius -; Kelvin -; temperature -	Kelvin is the base unit of temperature on the Kelvin scale
frames[2]	185	435.93	2	scaffold	building -	when designing continuous beams or building frames subjected to uniformly distributed live loads

Table 5 Percentages of items in the opaque word list appearing in other word lists

Comparative word list (with no. of items in list)	*Percentage*
GSL Top 1,000 word families	46.34%
GSL 2nd 1,000 word families	29.27%
AWL (570 word families)	12.20%
New GSL Top 1,000 lemmas	48.78%
New GSL 2nd 1,000 lemmas	39.02%
New GSL current words (378 lemmas)	0.00%
New AVL Top 1,000 lemmas	41.46%
New AVL Lemmas 1,001-3,000	4.88%
Ward's engineering word list (299 word types)	53.66%
Mudraya's engineering word list (top 100 word families)	46.34%

Appendix 186-item keyword list for the EEC

1 determine
2 flow
3 figure
4 temperature
5 energy
6 force(s)
7 pressure
8 function(s)
9 equation(s)
10 shown
11 chapter
12 water
13 surface(s)
14 system
15 mass
16 data
17 equilibrium
18 heat
19 can
20 axis
21 object(s)
22 constant
23 graph
24 beam
25 used
26 value(s)
27 gas(es)
28 point(s)
29 using
30 plane
31 example(s)
32 section[1]
33 section[2]
34 diameter
35 diagram
36 load(s)
37 motion
38 process(es)
39 horizontal
40 zero
41 liquid
42 method
43 program
44 pipe
45 line
46 properties
47 design
48 solution[1]

49 solution^{2}
50 air
51 particle(s)
52 component(s)
53 direction
54 volume
55 use
56 distance
57 assume
58 compute
59 maximum
60 density
61 length
62 phase
63 calculate
64 coefficient
65 angle
66 material
67 vertical
68 number
69 rate
70 variable(s)
71 coordinate(s)
72 problem
73 solid
74 input
75 engineering
76 class
77 reaction(s)
78 element(s)^{1}
79 element(s)^{2}
80 consider
81 equal
82 analysis
83 type(s)
84 define
85 called
86 ratio
87 structure
88 obtain
89 table
90 relative
91 explain
92 describe
93 write
94 defined
95 cycle
96 required
97 code
98 weight

99 depth
100 member
101 model
102 string
103 net[1]
104 net[2]
105 total
106 measured
107 positive
108 earth
109 applications
110 size
111 location
112 solve
113 operation(s)
114 specified
115 following
116 interval
117 shows
118 tree
119 corresponding
120 differential
121 cell[1]
122 cell[2]
123 reference
124 approximate
125 applied
126 loop
127 occurs
128 resistance
129 acting
130 given
131 note[1]
132 note[2]
133 static
134 definition
135 column[1]
136 column[2]
137 change
138 frequency
139 height
140 same
141 wave
142 block[1]
143 block[2]
144 body
145 analyze
146 standard[1]
147 standard[2]
148 unit

149 path
150 decreases
151 measurements
152 tool
153 tank
154 average
155 find
156 increases
157 network
158 plate
159 procedure
160 file
161 efficiency
162 source
163 minimum
164 shape
165 form
166 balance
167 integral
168 desired
169 output
170 strength
171 quantity
172 layer
173 typically
174 scale[1]
175 scale[2]
176 supports[1]
177 supports[2]
178 consists
179 negative
180 specific
181 plot
182 conditions
183 moves
184 frame(s)[1]
185 frame(s)[2]
186 equivalent