The Engineering Phrases List: Towards Teachable ESP Phrases

Dougal Graham

Abstract

This article presents the rationale and development of the Engineering Phrases List (EPL). The EPL is a 40-word list of teachable phrases for Engineering students based on a mixed-method empirical and intuitive corpus approach. First, frequent, engineering-specific, widely used phrases were identified in a corpus of engineering English. These phrases were then analyzed for markedness and ranked to determine which would be most useful from a teaching perspective and in terms of productive and receptive usefulness for learners. This research both creates a useful tool for teachers of engineering English as well as presents a methodology which should be useful for those developing similar lists in other ESP/EAP contexts. Furthermore, the implications of identifying and teaching phrases by focusing on markedness are discussed throughout the paper.

Keywords: ESP, EAP, corpus linguistics, markedness, teachability, formulaic language

Introduction

The fields of English for Specific Purposes (ESP), English for Academic Purposes (EAP), and their related subfields (see, eg. Jordan, 1997) have clearly established the need for English to be taught for specific purposes based on learner context and need. For example, work by Evans and Green (2007), and Nurweni and Read (1999) among others has shown that students entering universities in English as a foreign language environments such as Singapore, or Thailand often have difficulty comprehending their textbook materials both for general purpose studies and for specific study areas. This is most likely because learners in academic contexts are lacking in academic literacy and lacking in knowledge of academic discourse appropriate to their field of study (Hyland & Hamp-Lyons, 2002). Similarly, those learning English for work purposes require a specialized set of linguistic knowledge to allow them to properly conduct their business. Because of differences in usage between general purpose English which is usually taught in elementary through secondary schools (English for General Purposes, EGP) and the more specific context that the learners at post-secondary levels are usually targeting (a specific ESP or EAP context) learners

have difficulty applying the general purpose linguistic knowledge that they have studied so far to their new linguistic environment.

Therefore, there is a great need to teach specialized linguistic knowledge to students coming from a general English background and a wide variety of resources and materials have been created. Bowker and Pearson (2002), list vocabulary, collocations, syntax, discursive function, and text and discourse structure as some of the traditional levels of language employed by these materials to teach learners English for specific or academic purposes. The first level at which language can be seen to be highly specialized in both academic and professional fields is in the use of specialized vocabulary. Many materials have been generated in ESP and EAP to attempt to address learners' vocabulary needs, such as the Academic Word List (Coxhead, 1998), and other more specialized word lists (see eg. Mudraya, 2006; Ward, 2009).

Similarly, there can be variation at the syntactic level, the most often cited example being that academic language contains a high degree of passivization and nominalization compared with general English usage (Bowker & Pearson, 2002). Other levels at which language has been analyzed for difference between EGP and a specific context are collocations (eg. Fuentes, 2001), discursive functions (eg. Conrad & Biber, 2004), and the structure of texts and discourse (eg. Swales, 1990).

Recently, another level of language has emerged for study in corpus linguistics. Formulaic language, which takes a wide variety of forms such as multi-word units, phrasal expressions, lexical bundles, phrases, gapped phrases, or concgrams (among others) has provided insights into a new level of variation between genres and registers. These highly frequent formulaic expressions have been shown to be useful to both learners and native speakers in terms of fluency, comprehension, retention, and processing speed (Millar, 2011; Tremblay, Derwing, Libben, & Westbury, 2011), as well as production, and as a means of showing in-group knowledge and membership (Nattinger & DeCarrico, 2009; Schmitt, 2010). Therefore, there has recently been much research into the development of language materials focused on formulaic language such as the Phrasal Expressions List (Martinez & Schmitt, 2012), academic lexical bundles (Conrad & Biber, 2004), and the Academic Formulas List (Simpson-Vlach & Ellis, 2010).

These phrase lists while useful to learners and teachers, contain many phrases such as "the university of Michigan" (Simpson-Vlach & Ellis, 2010, p. 494), or "I don't want to" (Biber &

Barbieri, 2007, p. 271) which are most likely not of interest to the majority of teachers. In the case of these two examples, it is not a problem, as a teacher or learner can easily ignore items which they believe are either not useful in their context such as "the university of Michigan" or are already well understood by the learner such as "I don't want to". However, there may well be other issues. The sheer number of phrases in the lists may be overwhelming for some teachers or learners, or they may be unsure of whether a phrase is indeed useful or not. For that reason, one of the topics of interest in this article will be the question of how to determine not only which phrases represent difference between EGP and ESP, but which of those are interesting from a teacher's perspective. While these lists of phrases are all statistically significant, and the formulaic language is presented, generally, alongside their discourse functions there has been little research done into which phrases will be most useful for teachers in ESP.

This article will describe and explain the theory behind the development of the Engineering Phrases List (EPL), a short (40 phrase) list of teachable, interesting phrases that represent language specific to engineering English that is useful for both English teachers as well as those teaching engineering. While this article will focus on data from an engineering corpus, it is hoped that the concepts and approach taken will be useful and pertinent to those working in ESP and EAP fields in general. The article begins with a discussion of the corpus, statistical methods for determining important phrases, followed by a discussion of the methods used to select useful phrases for teaching. The article will finally propose that a more teaching-oriented approach to research into formulaic language for the purposes of ESP/EAP based on the concept of linguistic markedness, and that formulaic language by its nature carries with it elements of each of the traditional levels of language described above and can provide insight into those aspects of language used within the ESP context to both the teacher and learner.

Corpora Used

The specific language context to be discussed in this article is that of engineering English. In order to study the language that students at an engineering university in Thailand will be using, the Engineering English Corpus (EEC) was developed using the 29 English language engineering and math textbooks the students use in their first year of study. The corpus includes approximately 45,000 word samples from each textbook with an effort made to include all styles of language: explanation, practice exercises, and questions from several chapters in each book resulting in a

corpus that is approximately 1.15 million words in size. These books were scanned and optical character recognition software was used to convert them into plain text documents and make up the target corpus for the purposes of this study. For full details of the corpus' composition, please see appendices B and C.

Given that the aim of this research is to determine difference between English used in a specific context with that used in general English, a comparison corpus is necessary for the purpose of comparing usage. The general English corpus used in this research is a representative sub-set of the British National Corpus (BNC). The BNC is a useful reference point as it contains a wide variety of language from many sources both spoken and written across a range of genres and registers. Slightly more than 38 million words from the BNC were used covering the range of registers, genres, found in the BNC and both written and spoken modes. It may be useful to note here that the textbooks used in the EEC are by and large written in American English, and therefore the phrase lists had to be normalized for some spelling differences (eg: color vs. colour) before comparing phrases between the corpora.

Identifying Common Phrases

In order to determine the interesting formulaic language in the target corpus, first a frequency list of all N-Grams of 3, 4, or 5 words long was created. An N-Gram in the context of formulaic language can be considered to be any immediately consecutive sequence of 2 or more words, however, in this article "phrase" and "N-Gram" will be used interchangeably. The common N-Grams used to create the EPL were first identified on the basis of four main criteria: frequency, occurrence in multiple texts, corpus-specificity, and by co-occurrence. By this it is meant that the N-Grams should occur frequently in multiple textbooks of the target corpus, be used overall more frequently in the target corpus than in general English as represented by the BNC, and that the words that make up the phrases should more often occur together than with other words.

The first criterion for identifying these N-Grams was to set a raw frequency cutoff, in line with previous work (eg. Biber & Barbieri, 2007; Conrad & Biber, 2004; Simpson-Vlach & Ellis, 2010). In formulaic language research it is common to set a hard cutoff point for words with a given frequency of occurrence per million words where phrases that fall below the cutoff will be ignored for further analysis. Biber generally applies a cutoff of 40 occurrences per million words mark for 3 or 4-grams. The goal of this criterion is to ensure that items which only occur rarely

will not be included in the final data. On the other hand, Ellis et al. while compiling the AWL employed a cutoff frequency of 10 occurrences per million words.

When the initial frequency lists were analyzed, it was determined that setting a cutoff of 40 per million would be too restrictive in that only a small number of highly technical phrases would be produced. The technical phrases are generally uninteresting from an English teacher's perspective as they contain straightforward grammar, collocations, and so forth. For example, in the phrase "the magnitude of the" the only novel content for the learner might be the use of the technical term "magnitude" which a learner from an EGP background would not be familiar with. However, these technical terms are taught as part of the standard material for the engineering courses, and are therefore less interesting from an English teacher's perspective as the only problematic point for the learner, the technical term, will already be taught. For this reason lower cutoffs were selected for the purposes of the EPL. Because general frequency per million words declines as N-Grams increase in length, the frequency cutoff was scaled by length of phrase to 15, 10, and 5 occurrences per million words for N-Grams of 3, 4, and 5 words in length.

The second criterion for inclusion in the EPL was that the phrases should occur in multiple disciplines. As there is variation between the language used in different engineering disciplines and the goal is to produce results that apply to the entire field rather than any specific sub-discipline, it is necessary to ensure that the phrases produced for this work are ones which will be useful to a variety of engineers, not only one particular subgroup of engineers who happen to use a phrase very frequently. The criterion was set such that a phrase must occur in at least 10% of all texts in the corpus as well as one of the key sub-corpora (calculus, chemistry, math, and physics). Each of these sub-corpora represent materials which students from all engineering disciplines must study, and were therefore considered to be "key". After applying both of these criteria, an initial list of about 3,000 phrases was created.

Identifying Significant Phrases

A variety of statistical methods are available to determine which phrases represent the statistically significant phrases for a corpus of data. These statistics include Log Likelihood (LL), Mutual Information (MI) (Rayson & Garside, 2000; Schmitt, 2010), and a composite statistic known as the "beta score" (N. C. Ellis, Simpson-Vlach, & Maynard, 2008). Each statistic has a

different function and appropriate applications, and each will be briefly discussed in the context of its use in developing the EPL.

Log Likelihood

The Log Likelihood statistic (Rayson & Garside, 2000) is used to measure how "surprising" an item from one corpus is. The statistic is a hypothesis test to check the difference between expected frequency and actual frequency. The phrases with the highest LL scores, then, are those which show the most significant difference in frequency between the two corpora, and therefore, should be most representative of language that is used in the engineering context. This statistic was used to select phrases occurring with a significance level of $p < 0.0001$ for three and four-word phrases, and $p < 0.001$ for phrases five words in length. Because long phrases occur less frequently in both the target and comparison corpora, the threshold of significance of the results is somewhat reduced.

Mutual Information

The final criterion for the creation of the initial list of N-Grams used to create the EPL was the calculation of mutual information (MI). The MI statistic is used to determine how much a combination of words predicts each other. That is to say that a pair of words with a high MI score occur together very frequently, but rarely occur with other words. In a sense, it can be viewed as how much one word will then predict the words that are to follow. While the LL score is used to determine difference between general English and engineering English, the MI score is used to determine which phrases consist of words that occur together more frequently than would be expected by chance. Using the Collocate tool, the MI scores for each N-Gram were calculated and a cut-off of 3.3 was selected. After applying this statistic as well as LL, there remained 1,289 three-word phrases, 389 four-word phrases and 50 five-word phrases. These phrases were then used as the basis for the formation of the EPL.

The Identification of Teachable Phrases

While there has been much research into statistical significance of formulaic language (Dunning, 1993; N. C. Ellis et al., 2008; Schmitt, 2010), there has been has been less research into how best to locate the teachable language from a set of statistically significant phrases. One approach to the identification of teachable phrases is the beta score (N. C. Ellis et al., 2008). The beta score is calculated as a composite value determined by both the frequency of a phrase and the

phrase's MI score. The Beta Score is based on Ellis et al's investigation of the correlation of both frequency and MI to teacher's ratings of phrases for teachability. It was found that frequency had a greater effect on the ratings, but that MI was also significant and a composite score was developed to attempt to represent these judgments empirically. This beta score was used as the final ranking for the approximately 2,000 phrases that remained after the application of the criteria described above. However, while this score is an interesting and useful first step toward filling the void of metrics of teachability, I believe that it may be useful to look at non-statistical, more intuitive measures by which to identify useful phrases.

Markedness as Criteria for Teachability

While a phrase may be highly significant from a statistical perspective, it may not be interesting from a teaching perspective. For example, the phrase "what is the" (see Table 1) is the highest ranking three-word phrase by beta score from the initial 2,000 phrases. This is a phrase which is exceptionally frequent, it is more frequent in the engineering textbook language than in general English (high LL vs. the BNC), and the words in the phrase strongly predict each other's presence (high MI score). However, this is an uninteresting phrase from a teacher's perspective, as it is still frequent enough in general English that learners from an EGP background should be familiar with it, and there is nothing special about its usage in the engineering English context. Other similar phrases shown in the table below are "how long will it take", or "can be used".

Table 1: Top ten phrases by beta score

Rank	3-Gram	4-Gram	5-Gram
1	what is the	can be used to	at a rate of #
2	the number of	as a function of	you should be able to
3	as shown in	the magnitude of the	beyond the scope of this
4	# and #	as shown in figure	how long will it take
5	can be used	with respect to the	the first law of thermodynamics
6	shown in figure	in this chapter we	in such a way that
7	the value of	the value of the	the rate of change of
8	in terms of	the sum of the	the external forces acting on
9	be used to	newton 's second law	recall from chapter # that
10	# to #	in terms of the	in this section we will

Other phrases shown in Table 1, that may initially appear interesting, are at a second glance, possibly less interesting to teach. For example, the phrase "the first law of thermodynamics" or

"the magnitude of the", are both phrases which could reasonably be expected to be new to first-year engineering students. However, both of these phrases focus on technical terms. In the first case, the entire phrase is a technical term, and in the second case the phrase revolves around the term "magnitude" with standard collocations and syntactic patterns. As stated earlier in this work, technical vocabulary is generally well taught by specialist teachers within the engineering discipline and are already known to be part of the necessary knowledge for students and therefore these phrases are not considered interesting from the perspective of an ESP/EAP English teacher.

In order to determine which phrases are interesting from a teaching perspective, it is proposed to apply markedness criteria in order to differentiate between more teachable and less teachable phrases. Markedness refers to how "standard" a particular linguistic feature is to the grammar of the language. The more different from standard English or language-specific a linguistic feature is, the more highly marked it is said to be (see eg. R. Ellis, 1985). A form which is highly marked should be special in some way it may be a form which is which is less frequent, or more "structurally or conceptually complex" in some way (Saville-Troike, 2005). Saville-Troike explains that marked forms of language are generally acquired later than unmarked forms. As any type of English for a specific purpose can be considered to be a specialized subset of English, it will by necessity contain certain marked forms which occur much more frequently than in general English and help to serve to distinguish it from general English. These marked forms should therefore be indicative of the difference between any specific English, and general English. Furthermore, as these marked forms are more difficult to learn, they are forms that learners are unlikely to be familiar with even though they may already be used in general English.

Here I propose six categories of markedness for the determination of the phrases to include in EPL which will be listed briefly here and described in detail below. For each criteria that is satisfied a phrase can be considered to be more marked. These criteria were then applied to the 300 highest ranked three and four word phrases, and all five-word phrases to create the final 40-word EPL. The criteria are:

(1) **marked part of speech**: any of the words in the phrase do not have their usual part of speech;
(2) **marked word form**: any word in the phrase does not occur in the most common form of that word;

(3) **non-prototypical word meaning**: any word in the phrase does not occur with its most prototypical meaning;
(4) **marked collocations**: the phrase contains any collocations or co-occurrence patterns that differ from general English patterns;
(5) **non-literal phrase meaning**;
(6) **specialized syntax**: the phrase contains or is connected with complex or unclear syntax.

Finally, an additional criterion un-related to markedness was put in place to remove phrases which were only marked due to their inclusion of a technical term. Such phrases were deemed less useful to teachers as technical terms are already taught by specialist teachers. However, it is possible that a phrase containing a technical term might be interesting due to its syntax, or some other aspect. Therefore, phrases containing a word with only a specialized technical use were not considered unless the phrase was marked in at least two of the categories mentioned above. A word such as "axis", for example, is highly technical and unlikely to occur in general English. However, a word such as "function" has both a regular usage, and a special technical usage in mathematics and would be considered acceptable in a marked phrase. All data discussed in this section is from the EEC, unless otherwise specified, and the full list of phrases from the EPL annotated for markedness can be found in Appendix A organized by level of markedness.

Marked part of speech

The first criterion of markedness was if any word in the phrase occurred with a less common part of speech, that it should be considered as marked on that score. Many words may occur with different parts of speech in different contexts and meanings. Let us begin with an example phrase from the EPL:

Ex 1. for a **given**

It is clear that the part of speech of the word "given" in this sentence will be adjective because it follows the determiner "a". In general English "given" can be used as a verb (past participle) (see Ex 2), an adjective (Ex 3), or a noun (Ex 4).

Ex 2. The man was **given** a car.
Ex 3. Using the **given** information, determine the speed of the arrow.

Ex 4. It was a **given**.

In general English, the use of "given" as a verb is by far the most common. In the BNC 91% of uses of the word "given" are coded as "verb", 7% as adjective, and 2% as a noun. We can see that while there is substantial use of "given" as an adjective, it is overwhelmingly used as a verb in general English. However, in the EEC, the usage of "given" is significantly different. In this corpus the usage of given is split almost evenly (approximately 50%/50%) between verb and adjective, a significant difference from usage in general English. This difference in usage is clearly reflected in the top phrases in the EPL. The top ten most common three-word phrases containing the word "given" in the EEC are clearly split between passive and adjectival usage, whereas in the BNC an adjectival usage is not encountered until the 28th phrase, with the previous all being past participles as shown in Table 2.

Table 2: Top six phrases by frequency per million words in the EPL and top 5 and 28th phrases from the BNC

BNC		
Rank	**Freq**	**Phrase**
1	14.07	given to the
2	11.51	to be given
3	9.85	be given to
4	7.55	should be given
5	7.27	given by the
28	2.51	in a given

EEC		
Rank	**Freq**	**Phrase**
1	85.22	is given by
2	63.70	for a given
3	37.01	given by the
4	22.38	in a given
5	20.66	is given in
6	18.94	at a given

A second example is the phrase shown in Ex 5. In this phrase from the EPL, the part of speech of the word "note" is clearly different from the most common usage in general English. Normally, "note" will be used as a noun, rather than as a verb. This principle of markedness not only allows us to locate a phrase which contains usage that is different from that in general English, but also then gives us insight into a general usage pattern that is different within the context of engineering English.

Ex 5. **note** that the

Similarly to the case of "given", above, "note" appears almost exclusively as a verb in the EEC, compared with almost equal noun and verb distributions in the BNC. Again, we see that phrases by their nature make clear not only which words are most frequently used, but in what way they are used.

Marked Word Form

Words can occur in a variety of forms. To continue with the example of "given", this word is the past participle form of the canonical form "give". Similarly, many words used in engineering English use the non-canonical form much more frequently than the canonical form of the word. This can often be seen to link to passive voice usage, which as in general academic English, is very common in engineering English as well.

Ex 6. **acting** on the

Ex 7. **passes** through the

In Ex 6, "acting" is a less common form of the word "act". While "acting" is not a particularly strange word in and of itself, in general English is much more common for it to appear in the bare form "act". In fact, "act" is approximately ten times more common than "acting" in the BNC. The same is true of "pass" and "passes", where "passes" is significantly less common in general English usage than the base form "pass". Again, as was the case when discussing part of speech, the usage of the words in the phrases is representative of the differences between engineering English and general English. In both of these cases, the form used in the phrase is the less common form in general English, but the more common form in engineering English, once again, highlighting a difference in usage between the two types of language.

Non-prototypical Word Meaning

Many words in English have multiple possible meanings. This is most obvious in the case where a word has both a specialized technical meaning, and a non-specialized general meaning that is significantly different from the technical meaning. In Ex 8, the word "function" refers to a mathematical function, rather than "what or how something does what it does". This is a highly specialized technical meaning which is rare to non-existent in general English, but in very common use when discussing math, and physics.

Ex 8. (as) a **function** of

However, there may also be words with several meanings that are equally valid in general English, but one is more common, but that in the context of a specialized field, a different meaning (which is still valid in general English) becomes more common, see Ex 9 and Ex 10.

Ex 9. **under** the action of

Ex 10. **about** an axis

Both Ex's 9 and 10, show a usage of prepositions different from the most common one in general English. In each case, the meaning is not the one typically associated with the word. "under" does not refer to location, but rather refers to an object which is subjected to some action or event ("Merriam-Webster Online," n.d.). This is not a specialized technical meaning as it can be used in common phrases such as "be under pressure", or "be under fire" ("Merriam-Webster Online," n.d.), but this usage is quite rare, and therefore is here considered marked. Similarly, "about" does not have anything to do with something concerning an axis, but rather denotes location rotating around the axis. While this use of "about" is not the most common usage in the EEC (the standard usage to convey information concerning something is most common), it is used with this uncommon meaning much more frequently than in general English usage.

Finally, one difficult case that was selected for the EPL is the phrase in Ex 11, "due to the". This phrase while unmarked, may be considered marked in English of most EFL students arriving at university as they may only be familiar with the use of "due" in the meaning of "time at which work is to be handed in" as this is most likely the most common usage in their learning environment.

Ex 11. due to the

Marked collocation/co-occurrence patterns

The fourth markedness criterion included for the determination of what will occur in the EPL was whether the words occurred in collocational patterns which are different from those in general English, that is, do the words in the phrase normally appear together in general English? While this criterion is less frequently useful than those of word-meaning and word-form, it is still a useful criterion to take into account. If words are used together in different ways, it will certainly be of interest to learners.

Ex 12. **normal to** the

In Ex 12, the word "normal" is used with the preposition "to". In general English, the word "normal" does not have any special prepositional collocates, and the preposition that collocates with it most highly is "under". However, in this more specialized technical usage of the word, it collocates very highly with the preposition "to", on the right-hand side. In general English use,

when “normal” and “to” are used together, “to” will appear on the left-hand side, as in Ex 13 from the BNC:

Ex 13. It's as close **to normal** as it can be.

Again, this phrase shows some insight into the language used in the EEC that would not be readily apparent merely by examining a word list. It becomes clear that the usage of the word “normal” is abnormal relative to general English, and interesting for both its collocational pattern and meaning.

Non-literal Phrase Meaning

Occasionally, a phrase will be used that demonstrates a non-literal meaning. Such phrases are similar to cases in which a word is used with a marked meaning and the two often come together. These phrases are interesting for learners because the actual meaning may not be immediately clear. Often phrases in these categories are employing rhetorical devices for the sake of arguing about theoretical ideal cases.

Ex 14. A function f **is said to be** continuous at x=c provided...

Ex 15. **We see that** for leftist heaps, another strategy is needed.

In Ex 14 and Ex 15, the phrase does not actually mean what it could be literally construed to mean. In Ex 14 no one is in fact saying that a function named ‘f’ is continuous, but the reader is being informed that this is the definition for what constitutes a continuous function. Similarly in Ex 15 there is nothing to be literally seen, but rather information that must be understood.

Specialized syntax

Certain phrases either contain specialized syntax, or exclusively appear in sentences containing specialized syntax. For the general purposes of language learners, the declarative indicative sentence is the standard basic syntax that can be expected to be used, and other more complex structures will be more highly marked. Often, phrases occur in a specific set of grammatical conditions. In Ex 16 and Ex 17 the usage of the subjunctive mood can be observed. The subjunctive mood is often realized when discussing hypothetical situations which students will need to consider for the purpose of understanding theory or solving problems posed in their text. However, it is an aspect of English which many students find difficult to master as it is rarely visible in general English. Ex 16 also contains the use of the imperative verb “let”, another form of marked

syntax wherein the subject is dropped. Ex. 17 may further be confusing for learners of English as it may pose a garden path type problem. Learners may expect a phrase of the form of a noun followed by copula be followed by adjective or noun phrase, but instead be met with an infinitive verb.

Ex 16. **Let x be** the length of a straightaway.

Ex 17. A garden **is to be** laid out.

Another less marked type of syntactic structure common to academic English is the extensive use of the passive voice. This can also be see frequently in the phrases in the EPL. In Ex 18 and Ex 19 it is clear from the use of the past participle form of the verb followed and the following preposition that these phrases most likely occur in the passive voice. And if an investigation of concordance lines is performed, then that hypothesis is born out.

Ex 18. **based** on the

Ex 19. **applied** to the

Finally, a syntactic structure could be marked in terms of its position within a clause or sentence. For example, a number of phrases such as “in this case” (Ex 20) are almost always sentence-initial and are used to introduce a new clause.

Ex 20. **In this case** s increases as t increases.

Results and Implications for teaching

Using markedness criteria to identify teachable phrases is an approach with several benefits. First, because marked phrases are more difficult for learners, we can be sure that these phrases will be at least somewhat useful to teach. Secondly, markedness has potential implications for the teaching of phrases for the purposes of either comprehension or production by learners. Thirdly using the markedness approach allows both teachers and learners to induce patterns of usage that occur in the specific linguistic context being taught.

Generally speaking, learners acquire receptive capabilities earlier than productive ones. That is to say that comprehension of language precedes the ability to reproduce that language effectively. The markedness of phrases has implications then for how a learner will best be able to put to use the phrases in the EPL. It is proposed that a more highly marked phrase can be viewed as something

which a learner might have more difficulty learning to use, but that the learner can learn to understand for the purposes of comprehending their materials. However, it should be easier to acquire productive capabilities for a phrase which is less marked, and therefore these can be taught as phrases which a student can learn to use in their own language early on.

Table 3, below shows a comparison of some highly marked and some less marked phrases. Each phrase is given a markedness score based on how many of the markedness criteria are met. A phrase such as “for a given” or “acting on the” may be more complex or difficult for a student to learn to write correctly as it is more highly marked and may be more suitable for learning for receptive purposes initially. Conversely, a phrase such as “we assume that the” or “we see that” should be fairly straightforward for a learner to learn use it for productive purposes early once their awareness has been raised.

Table 3: Comparison of markedness in phrases (See full list in Appendix A)

Highly Marked Phrases	
Phrase	**Markedness**
can be viewed as	5
for a given	4
acting on the	4
is known as	4

Slightly Marked Phrases	
Phrase	**Markedness**
we assume that the	1
in this case	1
we see that	1
relative to the	1

Markedness and phrases also have implications for the general teaching of language for a specific purpose. The phrases and markedness categories each focus of parts of language that are traditionally taught separately by teachers: vocabulary, parts of speech, collocations, syntax, and discourse function. The phrases bring parts of each of these traditional levels of language along with them, and provide insights into general patterns of use in the specific linguistic context. As described above, the phrases can be taught in terms of vocabulary (word meaning in context, parts of speech, word forms), collocations, or grammatical patterns such as use of passive voice constructions or subjunctive mood. As shown by Biber (2007), phrases can also be used to show functional discursive patterns that occur in a specific type of English and that could be used as a criterion for future work.

This research shows the beginnings of a useful approach to determining useful phrases for teachers of English in a specific context, but it will need further refinement and development to be truly useful. Further research might also be useful to see if the same types of inferences can be

made equally well or better using other types of formulaic language such as gapped phrases, or concgrams. Finally, markedness is not binary, but exists on a scale within types of markedness. For example, one type of abnormal syntax may be considered more marked than another. This would affect judgments of markedness overall and a more detailed metric may need to be developed. Nevertheless, the current research shows that applying markedness criteria can be a useful way to judge teachability of phrases and lends insight into why the phrase is important to teach.

References

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes, 26(3), 263–286. doi:10.1016/j.esp.2006.08.003

Bowker, L., & Pearson, J. (2002). Working with specialized language: A practical guide to using corpora. London: Routledge.

Conrad, S., & Biber, D. (2004). The Frequency and Use of Lexical Bundles in Conversation and Academic Prose. Lexicographica, 20, 56–71.

Coxhead, A. (1998). An academic word list. School of Linguistics and Applied Language Studies, Victoria University of Wellington, 18. Retrieved from http://xa.yimg.com/kq/groups/18074488/2081673666/name/LDOCE_%EE%80%80AWL%EE%80%81.pdf

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. Computational Linguistics, 19(1), 61–74.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic Language in Native and Second-Language Speakers: Psycholinguistics, Corpus Linguistics and TESOL. TESOL Quarterly, 42(3), 375–396. doi:10.1002/j.1545-7249.2008.tb00137.x

Ellis, R. (1985). Understanding second language acquisition (Vol. 1). Oxford University Press Oxford. Retrieved from http://www.getcited.org/pub/102583342

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. English for Academic Purposes, 6(1), 3–17.

Fuentes, A. C. (2001). LEXICAL BEHAVIOUR IN ACADEMIC AND TECHNICAL CORPORA: IMPLICATIONS FOR ESP DEVELOPMENT. Language Learning & Technology, 5(1), 106–121.

Hyland, K., & Hamp-Lyons, L. (2002). EAP: issues and directions. Journal of English for Academic Purposes, 1(1), 1–12.

Jordan, R. R. (1997). English for Academic Purposes: A Guide and Resource Book for Teachers. Cambridge: Cambridge University Press.

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. Applied Linguistics, 33(3), 299–320.

Merriam-Webster Online. (n.d.). Dictionary. Retrieved January 12, 2014, from http://www.merriam-webster.com/dictionary/under

Millar, N. (2011). The Processing of Malformed Formulaic Language. Applied Linguistics, 32(2), 129–148.

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

Nattinger, J., R., & DeCarrico, J., S. (2009). Lexical Phrases and Language Teaching. Oxford University Press.

Nurweni, A., & Read, J. (1999). The English language knowledge of Indonesian university students. English for Specific Purposes, 18(2), 161–175.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora (pp. 1–6). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1604686

Saville-Troike, M. (2005). Introducing Second Language Acquisition. Cambridge: Cambridge University Press.

Schmitt, N. (2010). Researching Vocabulary: A Vocabulary Research Manual. Palgrave MacMillan.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics, 31(4), 487–512. doi:10.1093/applin/amp058

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. Cambridge University Press.

Tremblay, A., Derwing, B., Libben, G., & Westbury, C. (2011). Processing Advantages of Lexical Bundles: Evidence from Self-paced Reading and Sentence Recall Tasks. Language Learning, 61(2), 569–613.

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

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

Appendices

Appendix A: The Engineering Phrase List (EPL) with Markedness Categories (An empty space *is "*un*marked" and an "X" is "marked")*

Phrase	Marked POS	Marked Word Form	Non-Prototypical Meaning	Marked collocations	Non-Literal Meaning	Specialized Syntax	Markedness Score
can be viewed as		X	X	X	X	X	5
for a given	X	X	X		X	X	4
acting on the		X	X	X	X		4
is known as		X	X		X	X	4
which/one/each of the following	X	X	X		X		4
under the action of			X	X	X		3
let us consider (a)			X		X	X	3
can be written				X	X	X	3
about an axis			X	X			2
the degree of			X	X			2
note that the	X		X		X		3
with respect to (the)			X			X	2
based on the		X	X				2
passes through the		X	X				2
the action of			X		X		2
let x be					X	X	2
we can write				X	X		2
for each of			X			X	2
in such a way (that)			X	X			2
suppose that you					X	X	2
such that the			X	X			2
normal to the			X	X			2
beyond the scope of this				X	X		2
(as) a function of			X	X			2
assume that the					X	X	2
applied to the			X			X	2
as shown in (figure)		X				X	2
is assumed to		X				X	2
is said to be (in)					X	X	2
can be expressed as			X			X	2
it can be shown that		X				X	2
is to be				X		X	2
in terms of		X	X				2
relative to the	X		X				1
we say that			X				1

Phrase	Marked POS	Marked Word Form	Non-Prototypical Meaning	Marked collocations	Non-Literal Meaning	Specialized Syntax	Markedness Score
due to the			X				1
we see that					X		1
we will assume that					X		1
we assume that the					X		1
in this case						X	1

Appendix B: Disciplines included in the CEEM Corpus

Disciplines Included in EEC
Civil Engineering
Mechatronics
Mechanical Engineering
Computer Engineering
Chemical Engineering
Environmental Engineering
Electrical Engineering
Materials Engineering
Production Engineering
Tool Engineering
Control Systems and Instrumentation
Electronics and Telecommunication

Appendix C: Textbooks of the CEEM corpus by subject and number of words included

	Textbook Subject	# of Words		Textbook Subject	# of Words
1.	Biology	42,857	15.	Hydraulic fluids	42,174
2.	C++	50,103	16.	Java	28,049
3.	Calculus	59,326	17.	Manufacturing processes	61,837
4.	Chemical engineering	46,509	18.	Material and energy balance	21,950
5.	Chemistry	45,350	19.	Mechanical solids	26,501
6.	Database	52,811	20.	Physics	88,978
7.	Data structure	35,789	21.	Statics and dynamics	50,302
8.	Discrete mathematics	50,991	22.	Statics	36,888
9.	Circuits and circuit analysis	34,585	23.	Structural analysis	36,826
10.	Engineering materials	53,426	24.	Surveying	48,353
11.	Engineering programming	29,165	25.	Technical drawing	69,228
12.	Environmental pollution	34,235	26.	Thermodynamics	54,149
13.	Environmental engineering	40,861	27.	Wastewater management	24,144
14.	Fluid mechanics	39,138		Total	1,204,525