Interactive web-based learning of corpus-generated phrases

Dougal Graham
Christopher Osment
King Mongkut's University of Technology Thonburi, Thailand
Corresponding author: dougal.gra@kmutt.ac.th

Submitted 30 January, 2014; accepted in final form 15 May, 2014

Abstract

This article will review the development of a partially-automated online teaching system developed to help intermediate-level engineering students learn appropriate English phrases in an English as a Foreign Language (EFL) environment. The corpus contains approximately 1.2 million words. Rather than focusing solely on keywords, these materials highlight useful collocations, and genre-specific uses of common words and phrases. The online interactive materials are able to generate a relatively large number of unique exercises for students. These exercise allow students to not only practice content that they have previously learned, but in fact to learn new material that they have not previously studied. In this way, the materials can be used either by students alone, or in conjunction with in-class use of the paper-based materials.

***Keywords**: Corpus, English for Specific Purposes, Online, Data-Driven Learning, Formulaic Language*

Introduction

A large body of research has demonstrated the importance of formulaic language (collocations, formulaic phrases, multi-word units, lexical bundles, etc.) in the production and comprehension of fluent natural language. Schmitt (2010, pp. 117–120, 142) and others (Nattinger & DeCarrico, 2009) explain that formulaic language is important for conveying routine meanings, lexicalising various functions, social functions, discourse organization, and precise information transfer as well as promoting fluency and Biber (2007; 2004)'s lexical bundles focus on the use of formulaic language for discourse and pragmatic functions. Formulaic language also brings cognitive processing benefits as it is easier and faster to process and generate once learned than non-formulaic expressions and increases retention of content (Tremblay, Derwing, Libben, & Westbury, 2011) compared to incorrect use of formulaic language which increases processing demands (Millar, 2011).

Formulaic language, then, has clear benefits for English as a Second or Foreign Language learners. If a learner can master even some formulaic language, it will greatly increase their ability to both produce and comprehend language quickly and easily, resulting in greater fluency, comprehension and comprehensibility. This language is also useful for learners in an English for Specific Purposes (ESP) context. Formulaic language differs between registers (Biber & Barbieri, 2007) and contexts (Conrad & Biber, 2004) and knowing the correct formulaic language to use within a given community can not only help a speaker communicate clearly and effectively within the group, but also help the learner mark themselves as a member of the language community or show social solidarity (Norbert Schmitt, 2010, p. 10) increasing trust and social cohesion.

When learning formulaic language, frequency of exposure, and frequency of use are the two most important factors in determining whether a learner will acquire a given piece of formulaic language or not (Ellis et al., 2008; Huang, Wible, & Ko, 2012). The result of the need for both repeated exposure and repeated use is then a focus on repetition and rote-learning style activities for the study of formulaic language. For example, Yu (2009) recommends explicit memorization as a method by which to improve student performance for formulaic language. Similarly Lewis (1997, p. 51) explains that many researchers recommend "meeting" a word at least seven distinct times in order to learn it properly, and one can reasonably imagine extending this concept to formulaic language. While such an approach may be useful for learners, it does have certain drawbacks.

Some of the drawbacks of a memorization, or rote-learning approach are time-consumption (Swan, 2006), student demotivation and disengagement (Bowen & Marks, 1994, p. 101), and a focus on surface-level memorization, rather than understanding (Biggs & Moore, 1993, p. 215). Bowen (1994) states that students often feel that memorization and rote learning tasks make the learning seem like a "difficult, even painful process". Such demotivation could lead to students who do not try, or feel that the effort to learn the language is not worthwhile. Furthermore, focus only on the surface-level form of the structures, while ignoring their context and usage makes the target language more difficult to learn by ignoring valuable information about when the language is appropriate to use.

An alternative to rote-learning is to take a data-driven learning (DDL) approach. DDL is an approach to language learning that focuses on using genuine language examples (data) as the core from which a student is then guided to "learn how to learn" the language (Johns, 1991, p. 31) cited in (Boulton, 2011). Because of this focus on the student's ability to learn for themselves, the direct access to the data, and the focus on induction (Johns, 1991, p. 29, as cited in Boulton, 2011), DDL can be considered to fall within the category of a student-centered approach to learning. A data-driven learning approach could be used to provide a student with the target language they need and are immediately interested in learning, couched within a variety of real contexts. This would then allow the student to learn and discover the formulaic language inductively and within its appropriate context, rather than in a decontextualized rote manner.

This paper will present an alternative automated approach to the teaching of formulaic language in an ESP setting which avoids the problems associated with rote learning. This project makes up the Corpus-based Engineering English Materials (CEEM) website which is available at (http://crs2.kmutt.ac.th/ceem). We will first outline our general theoretical principles for the creation of the CEEM materials following which we will give a brief overview of some currently available materials, before finally giving a detailed discussion of our approach to the generation of materials.

Principles

In an attempt to achieve the necessary repetition for the learning of formulaic language without resorting to rote learning, and keeping in line with the principles of DDL, we have determined five key principles to use in the development of our materials whose rationale we will discuss in more depth below. The first two principles are a direct attempt to avoid the issues with rote learning while the final 3 principles are more closely associated with DDL and a student-centered approach. The first principle is to avoid pure memorization, and the second to employ variation in the materials presented to students. This will help to prevent rote memorization as the exact context of presentation will be constantly varied and changing. The final three principles which fall under the rubric of a student-centered DDL approach are an attempt to foster student independence, the use of authentic language, and the fostering of self-motivation in the students. The goal of the third and fourth principles are to foster independence by creating materials usable either in or outside of the class with or without teacher support, and by presenting authentic language, to force the student away from a rote focus on the formulaic language unit as an entity separate from context and to instead see it within the context of the various texts within which it appears. The goal of the final principle is find ways to encourage students to take charge of their own learning and not be entirely dependent on the teacher or classroom as motivation for learning.

Independence

In order to allow students greater freedom and self-direction, it is necessary that students be able to access the learning materials either in or out of class. That is to say they should encourage student independence and work equally well with or without a teacher's support. This implies the need for materials to be freely available online, and formatted such that they are equally useful and interesting on a personal computer, a notebook, a tablet, or a small smart phone. In this way, the materials can be accessed either outside of class, or within a classroom setting with additional teacher support.

Secondly, in order to foster true independent student-centred learning there must be choices available for students. We have taken this to imply the need for both a choice of multiple types of activities to suit a variety of learning goals and styles, as well as choice within each activity for the content that can be accessed.

Finally, we understand that not all students will be ready or able to immediately use these exercises. For that reason, we have developed complementary introductory exercises that teachers can employ in class. These exercises serve as both awareness-raising opportunities and a chance to practice useful language. Once students have taken part in these activities with a teacher’s guidance they will be more ready to reinforce their new knowledge by independently taking part in the online activities which we have created.

Avoid Pure Memorization

While frequent exposure is clearly necessary for the development of a powerful and useful vocabulary of formulaic language as discussed above, it does not necessarily entail that pure memorization is either sufficient for or necessary to the acquisition of formulaic language. An alternative approach is to ensure that the learner is actively engaged in the language, not only in reading it and looking for their own patterns and rules, but taking part in interactive activities which help to highlight important information and require active input and consideration from the user. These exercises can be used in conjunction with scaffolding to promote inductive learning of formulaic language, lexico-grammatical profiles or other concepts, and to ensure that students will be able to use these skills independently in the future.

Employ Variation

While variation helps to avoid pure memorization by exposing students to a variety of different contexts and situations for use, it also addresses the problem of over-structured exercises that occurs with rote learning. We propose that a useful source of exercises should produce a large number of different exercises for students with a low probability of a specific set of exercises being repeated. With most existing exercises, both online and paper-based materials, it is extremely difficult to avoid the explicit repetition of exercises and examples. Some online activities such as FLAX (Witten et al., 2013) which promotes language exploration, do employ significant amounts of variation by dynamically tapping into corpora. However, these activities do not actively help to foster a student's ability to recognize linguistic patterns, or to put into use the linguistic patterns that they may be able to recognize. Moreover, the abundance of information that many corpora possess can be daunting to learners who are unfamiliar with using corpora for language learning.

Motivation

Finally, none of the above principles will be useful if students are uninterested in the activities presented. To address this issue, we believe that the activities and exercises should fall on a continuum from fun to challenging, and should incorporate aspects of gamification or educational games. Gamification

(Deterding et al., 2011) is the introduction of game-like aspects to what is traditionally not considered a game situation. This can be the introduction of scoring, comparative rankings, or other systems commonly used within the realm of games. These systems can, if used correctly, encourage repeated use, and provide a more relaxed approach to the use of the materials. The simplest and most useful type of gamification is the use of instantaneous feedback to let the user know how they are progressing, and/or to score them. This fits very well with the need to provide useful scaffolding to students as they progress through an activity. Currently, the project only supports very loose gamification in the sense that it contains both educational games and a partially gamified set of exercises without any of the long-term scoring accumulation, heuristics, or directly competitive aspects of fully gamified applications.

Authentic Language

The use of authentic useful language is a key principle of DDL so it is necessary here to discuss our target audience, corpus, and general learning goals for the CEEM project. This project is intended to serve the interests of first or second-year undergraduate students at engineering universities, especially in developing countries. The work is based upon a corpus of first-year English textbooks from a Thai Engineering University's “International Programme”. Work by several scholars (Evans & Green, 2007; Nurweni & Read, 1999) has found that students at such institutions often lack critical linguistic ability necessary for their academic studies, which are frequently entirely in English. Students have great difficulty in comprehending both their textbooks and their exam questions, and given the importance of formulaic language, the teaching of formulaic language seems an appropriate starting point to address the issue of poor comprehension of basic instructional materials.

In order to create exercises using meaningful in-context examples of engineering English that students could expect to encounter in their academic studies, the Engineering English Corpus (EEC) was created using, on average, 45,000 word representative samples from each of the textbooks that the students would be using during their course of study in the international programme. These samples included sections from each chapter and from a variety of styles of writing, from explanations to practice exercises and questions. The books were digitized, and Adobe Acrobat's optical character recognition software was used to translate the files to text format. The final corpus comprises 29 textbooks and approximately 1.15 million words of text (See appendices for full details of corpus composition).

There are many theoretical approaches to formulaic language. We have not discussed them in much detail as they are not the primary focus of this paper, but they do bear mentioning in order to make clear what our learning goals were in the development of the CEEM Project. Approaches to formulaic language range from solely frequency based lists, such as Biber's (2007; 2004) lexical bundles, to general formulaic language described in the Phrasal Expressions List (Martinez & Schmitt, 2012), comparative EAP studies such as the Academic Formulas List (Simpson-Vlach & Ellis, 2010), to gapped phrase approaches such as phrase-frames (Fletcher, 2011), or concgrams (Cheng, Greaves, & Warren, 2006). For the purpose of this work, we have focused on high-frequency four-word sequences, which either do not frequently occur in other contexts or, in the context of Engineering English, have different lexico-grammatical profiles. We have taken a partially intuitive approach, to this task. This approach will be outlined in detail below.

Currently available materials

Generally, there are two types of DDL materials that are currently available online: “discovery materials” or “testing materials”. Discovery materials are almost completely unstructured and are designed only to facilitate a learner's ability to locate useful corpus data with which to inform their language use. For example, FLAX and the Compleat Lexical Tutor (Cobb, 2013) focus primarily on allowing students to access concordance lines that might be useful to them, either by writing and searching for a word about which they are unsure, or by choosing words from a word list. Such materials are useful to advanced learners who know what they want to study and learn, but they may not provide enough guidance for lower level learners.

Testing materials, on the other hand, test a student's knowledge and provide instant feedback, at least partially meeting our criteria for motivation. Ideally, they also incorporate some sort of scaffolding to enable learning throughout the testing process. The Lexical Tutor site contains a set of exercises that test collocational knowledge comprising approximately 30 items. The collocation tests are interesting in that students are provided with scaffolding in the form of concordance lines from which to induce a correct answer. Such materials then, meet our requirements for authentic language, self-motivation, and independence, but they are unchanging and do not contain sufficient variation for a learner to be able to practice difficult language on more than one occasion. The powerful FLAX tool allows teachers to quickly and easily generate activities for students from given readings and articles, however, as a tool which works on a single text, it is unclear whether these activities will lead to useful generalizable language rules for students.

This pattern of many small sets of materials which are useful, but not fully meeting our requirements demonstrates the need to look for easier ways of creating reusable and automatically variable materials, rather than attempting to create a large number of static non-reusable exercises.

Creation of Materials

Using the EEC described above, the Engineering Word List (EWL) was created. The EWL, which can be accessed online at http://crs2.kmutt.ac.th/ceem/ewl was created using a combination of an empirical and intuitive approach. First, a raw word list was created using the AntConc concordancing application (Anthony, 2013). The list was restricted to words occurring in at least 10 textbooks, or in one of the core textbooks (physics or calculus) used by all departments for their engineering students. This list was sorted from highest to lowest frequency words, and was then triaged by removing non-content words including determiners, prepositions, conjunctions, pronouns, and single-character symbols used in various mathematical formulae. For a sample comparison of the ten most frequent words in the corpus as a whole compared with the ten most frequent words in the EWL, please see Table 1 below.

Table 1: *Ten most common words in the EEC and the EWL*

Rank in EEC	**EEC Raw Frequency**
1	*the*
2	*of*
3	*a*
4	*and*
5	is
6	*to*
7	*in*
8	*that*
9	*for*
10	are

Rank in EEC	**EWL**
5	is
10	are
11	be
34	have
41	used
45	figure
48	system
52	force
54	use
56	energy

Following the triage, the words were organized into 418 word family groups, containing the top 1,416 content words. This list was examined intuitively in order to locate 12 highly frequent words showing significant variation from standard usage to use a starting point for the creation of exercises with the goal of eventually producing exercises for all the words on the list. These words are: applied, based, common, consider, defined, difference, done, equal, following, given, have, and how. For each word, two or three of the most common 4-word MWUs (phrases) containing that word were selected from the corpus as exemplifying engineering-specific language that was likely to cause difficulties for the students.

Phrases were selected not on the basis of non-standard lexis, but on the basis of non-standard lexico-grammatical profiles. A phrase such as “a bipartite graph” is relatively common in the data (LL 53.24 vs. BNC), however it exemplifies a technical term which will certainly be covered in the course of their regular lectures. The interesting and useful phrases for the students will be ones that contain everyday common words, but that use them in ways that the students are not familiar with such as “is defined to be”. Students are all generally familiar with the individual words in this phrase, but unfamiliar with the use of them together in this ordering and to achieve this function. These phrases, and their genuine contexts from the corpus were used to generate random DDL exercises for the students.

Explanation of CEEM Materials

This section will contain an in-depth description of the goals for each of the four types of exercises that are currently available from the CEEM project website including materials available for teachers' use. First we will discuss the in-class activities, then we will describe the three current on-line activities: sight words, hangman, and the phrase challenge. It is important to keep in mind that these materials are intended to be holistic and incremental. While the primary activity does in fact take into account all the principles that we have discussed above, some of the introductory exercises do not. These issues and the reasons for our decisions will be discussed below.

In-class exercises

The printable in-class exercises (available at http://crs2.kmutt.ac.th/ceem/teacher/exercises) are based on the 1,000 most frequent vocabulary from the EWL. As such, they adhere to our principle of using genuine language in that they are words which students will definitely encounter frequently throughout their studies, and in addition, the exercises use genuine sentence-length excerpts from their texts. These exercises have three main pedagogic goals which are to teach collocations, to contrast apparent synonyms, and to raise learners' awareness of these concepts and of skills necessary to make these inferences on their own from natural language.

It is necessary to clarify the place of these exercises within the context of CEEM. These exercises, as classroom-based, printable exercises are neither highly varied (and are clearly single-use), nor particularly student-centered. However, the purpose of these exercises is to raise a student’s awareness of corpus-based approaches to learning, and to provide them a chance to learn basic skills for inducing usage from real-world examples. These exercises are particularly intended for lower level learners who may not have as much language learning experience as a more advanced student.

These exercises take collocations into consideration by trying to asking students to utilize noticing so that students will become aware of common collocations in engineering writing. We begin with simple collocations but also note that some of them can build up into a longer multi-word units.

As well, there are exercises that contrast apparent synonyms. Students are often unaware of how apparent synonyms frequently have significantly different usage patterns. These types of exercises are meant to foster an understanding in students that even though two words may have synonymous meanings, the usage of such words can be quite different. Comparisons of apparent synonyms show that their actual usage is quite different, and these exercises help students to see those differences. For example, “make” and “create” appear to many students to be completely synonymous, however, in our data “create” is often used when discussing data modelling or mathematics as in the sentence fragment “…and the capability to create complex geometries…” while “make” is more often used with the creation of a physical entity or to communicate how to do something as in “to make your drawings clear and easy to read”.

Awareness raising is an important element in our exercises. These exercises raise students' awareness of the above concepts and help them to develop the tools to learn more efficiently in future study. They give the students access to a variety of examples from real usage and try to ask questions which will lead the students to draw inferences about the words and how they are used in natural language.

On-line Exercises: Sight Words

The “sight words” activity is an online low-demand game-like reading practice exercise. While the primary focus of the CEEM activities is the teaching of phrases, it is important for students to be aware of the highly frequent words in the language they are reading, and then to make the connection between those words, and their usage in phrases. By given students a chance to focus only on words, before seeing those words in the context of phrases, we hope that we will help to raise a student’s awareness that words are not used in isolation, and to build incrementally from a small simple activity, to a more complex one, as will be described below.

In this activity, students choose between three and six words, which they would like to practice. These words have been chosen from among the most frequent words in the EWL. This activity attempts to develop automaticity (Logan, 1985) in student reading to improve reading speed and proficiency. Proficient readers are able to distinguish words more quickly and efficiently than beginner readers who must explicitly sound out the spelling in their heads as they go (Biggs & Moore, 1993, p. 341). The goal of this activity is three-fold: firstly, it aims to increase student awareness of some of the most frequent words in the corpus, secondly it aims to help increase reading speed by helping students develop automaticity in recognition of the target words, and thirdly it is meant to be a fun and engaging activity to help students become engaged with the material.

Student autonomy is promoted by asking students to choose which words they would like to practice. They are also given the option to randomly select a set of words. As can be seen in Figure 1, the activity presents students with a countdown timer incremented by two and a half seconds for each word. A target word is shown in the top center of the screen above the timer. Upon clicking the “start” button, students are presented with a grid containing six buttons. One button contains the target word, and the 5 other contain similarly spelled words and non-words. The student must select the correct word from the grid as quickly as possible. The timer is not fixed per word, but per game session, so that if one word takes slightly longer, but another less time, the student will still be able to complete the activity in the time allotted. When a potential answer is selected, the countdown timer stops, and the student is presented with feedback displaying the correct answer if their answer was incorrect (see Figure 1, below).

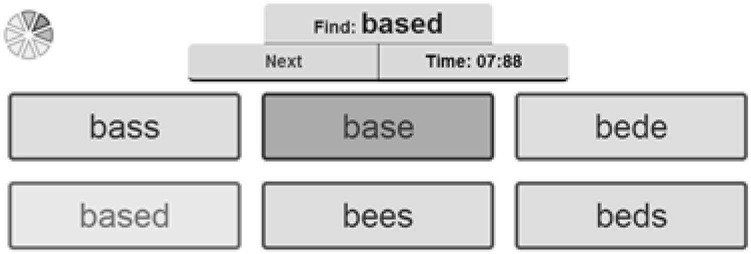


Figure 1: *Sight words activity with incorrect word darkened and correct word lightened*

As students progress through the activity, a coloured wheel in the top-left hand side of the screen displays their score, and at the end of the activity they are given a full feedback report detailing which words took the longest for them to find, and which they answered incorrectly allowing them to note those words need further practice.

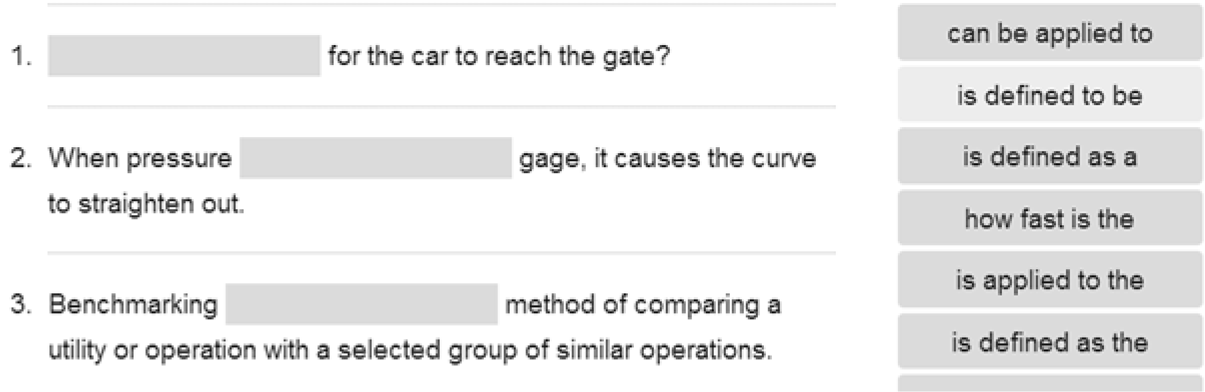
This activity meets our general criteria described in the first section in that it is classroom-agnostic (online), it is not focused on pure memorization, but rather on exposure, it fosters self-motivation through immediate direct feedback and scoring, and it employs variation in both presentation and content. The variation in this activity warrants some brief clarification. Firstly, the six tiles are randomized, to ensure that the presentation is not predictable. Secondly, the list of five similarly spelled words is selected at random from the top twelve similarly spelled words for a given target word, ensuring that the student will be presented with a novel experience each time the game is played.

On-line Exercises: Hangman

Hangman is a classic educational tool used to practice spelling, create an environment for repeated exposure and to foster motivation through game-like aspects. In the CEEM Project website, we use this familiar activity not only for the purposes just mentioned but also to raise students' awareness of formulaic language sequences as single units. Rather than presenting words for students to spell, they are presented with four word phrases. As discussed earlier, these phrases were chosen because they contain the core EWL words that we are taking as our focal points for the activities. These phrases then also allow the students to see that these words (which they have practised in the Sight Words activity) do not occur on their own, but as part of larger units. This again reinforces those words through another exposure vector, this time with greater context. Again, this is a fun activity, which, although it is unscored, certainly contains game-like elements to increase student interest.

On-line Exercises: Phrase Challenge

The final activity developed for students is significantly more challenging than the two previous activities described. As with the sight words activity, students are prompted to choose words that they would like to learn more about; however, in this case, 2-3 phrases are selected that contain each word. The students are then presented with a series of gapped sentences on the left side of their screen and a list of phrases on the right as shown in Figure 2.

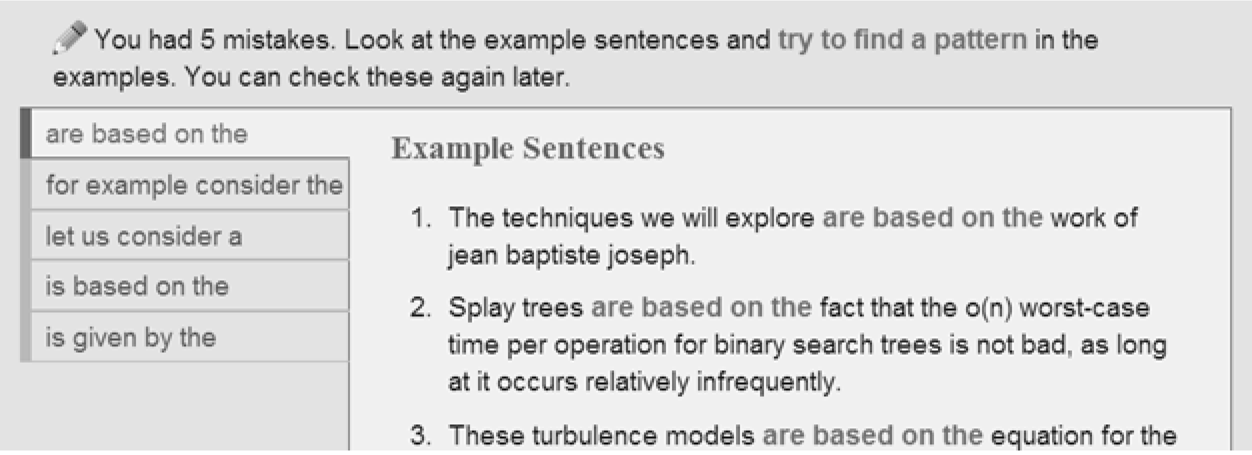


1. ______ for the car to reach the gate?
2. When pressure ______ gage, it causes the curve to straighten out.
3. Benchmarking ______ method of comparing a utility or operation with a selected group of similar operations.

- can be applied to
- is defined to be
- is defined as a
- how fast is the
- is applied to the
- is defined as the

Figure 2: *Phrase Challenge activity with gapped sentences and possible answers*

The student then has a maximum of three attempts to match the phrases with the sentences via drag and drop. After each of the first two attempts, error correction scaffolding is presented to the student alongside their errors. For the first failed attempt, the student will be presented with three sample sentences containing each phrase (Figure 3). The student is encouraged to try to find a pattern in the examples.



You had 5 mistakes. Look at the example sentences and try to find a pattern in the examples. You can check these again later.

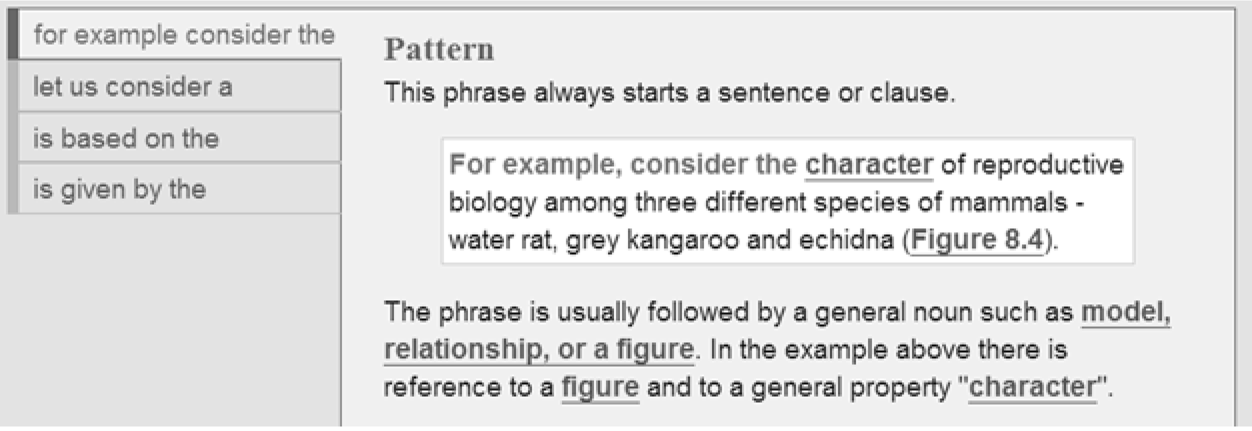
- are based on the
- for example consider the
- let us consider a
- is based on the
- is given by the

Example Sentences

1. The techniques we will explore are based on the work of jean baptiste joseph.
2. Splay trees are based on the fact that the o(n) worst-case time per operation for binary search trees is not bad, as long at it occurs relatively infrequently.
3. These turbulence models are based on the equation for the

Figure 3: *First round of scaffolding in the Phrase Challenge: Example Sentences*

If, after a second attempt, the student is still experiencing difficulty, an example partial lexico-grammatical profile is presented to the student (Figure 4). The student is also presented with 3 more example sentences to examine.



for example consider the

let us consider a

is based on the

is given by the

Pattern

This phrase always starts a sentence or clause.

For example, consider the character of reproductive biology among three different species of mammals - water rat, grey kangaroo and echidna (Figure 8.4).

The phrase is usually followed by a general noun such as model, relationship, or a figure. In the example above there is reference to a figure and to a general property "character".

Figure 4: *Second round of scaffolding with a "pattern" for the phrase*

This scaffolding is intended to serve two purposes. In the first example, it gives context with which learners can begin to attempt to induce a pattern of usage (lexico-grammatical profile) for the phrase in question. However, since some students may not be aware of what a lexico-grammatical profile is, or how to find one, the second type of scaffolding is provided to raise learner awareness of some possible types of patterns, which they can identify for the phrases.

In order to promote variation, the gapped sentences are randomly selected from the database. This combined with user selection of words and the number of sentences stored for each phrase and a randomized sentence ordering provides an astonishing number of possible exercises for students. There are 12 words per section, of which 3 are chosen to generate phrases, resulting in 220 different combinations of phrases that one can study per section. Each word is associated with from 1-4 useful phrases averaging ~2.16 phrases per word. Within each of the 220 combinations of approximately 7 phrases, each phrase is randomly associated with 2 sentences resulting in ~1.9 x 1014 possible exercises. Of course, a student would likely notice some repetition well before attempting use of the exercises that many times, but it should be sufficient for 3-5 attempts distributed over some time, especially if the student varies their target words.

Conclusion

The teaching and studying of formulaic language is an important consideration in language teaching. The findings of corpora from the last twenty years have brought its existence in a variety of genres clearly into view. Educators need to address the teaching of formulaic language. Simply teaching lexical lists based on frequency counts is not enough.

Moreover, with the recognition that learning does not necessarily require a classroom, especially with language, the idea of making materials easily available to learners online is vital. This is particularly true for younger learners who have grown up along with the internet and for mature learners who may not have time to attend traditional classes, but retain a need and desire to learn a language.

Despite the development of improved and versatile features on the internet, many online learning materials still tend to lack interactivity and randomness. If we as educators wish to truly motivate and engage our students, then we needed to design language materials that harness this interactivity and randomness. Furthermore, the usage of randomness better models the actual variety that students encounter when they read.

References

Anthony, L. (2013). AntConc (Version 3.3.5w). Tokyo, Japan: Waseda University. Retrieved from http://www.antlab.sci.waseda.ac.jp/

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.

Biggs, J., B., & Moore, P., J. (1993). *The Process of Learning* (Third Edition.). Prentice Hall of Australia.

Boulton, A. (2011). Data-driven learning: the perpetual enigma. *Explorations across Languages and Corpora*, 563–580.

Bowen, T., & Marks, J. (1994). *Inside Teaching*. Heinemann.

Cheng, W., Greaves, C., & Warren, M. (2006). From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, *11*(4), 411–433.

Cobb, T. (2013). Compleat Lexical Tutor (Version 6.2). Retrieved from http://www.lextutor.ca/

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

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). Retrieved from http://dl.acm.org/citation.cfm?id=2181040

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

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

Fletcher, W., H. (2011). *Phrases in English*. Retrieved from http://phrasesinenglish.org/

Huang, P.-Y., Wible, D., & Ko, H.-W. (2012). Frequency Effects and Transitional Probabilities in L1 and L2 Speakers' Processing of Multiword Expressions. In S. T. Gries & D. Divjak (Eds.), *Frequency Effects in Language Learning and Processing* (pp. 145–176). de Gruyter Mouton.

Johns, T. (1991). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.), *Classroom Concordancing* (pp. 27–45). *English Language Research Journal* (as cited in Boulton, 2011).

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *39*(2), 367.

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

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

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

Norbert Schmitt. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave MacMillan.

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

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

Swan, M. (2006). Chunks in the Classroom: Let's not go Overboard. *The Teacher Trainer Journal*, *20*(3), 5–6.

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.

Witten, I., Brine, J., Finch, L., Franken, M., Johnson, M., Te Taka, Wu, S, et al. (2013). FLAX: Interactive Language Learning (Version 2.6). University of Waikato. Retrieved from http://flax.nzdl.org/

Yu, X. (2009). A formal criterion for identifying lexical phrases: Implication from a classroom experiment. *System, 37*(4), 689–699.

Appendices

Appendix 1: *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 2: *Textbooks of the CEEM corpus by subject and number of words included*

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