DIY needs analysis and specific text types: using *The Prime Machine* to explore vocabulary in readymade and homemade English corpora

Dr. Stephen Jeaco

Xi'an Jiaotong-Liverpool University, China

Abstract

Corpus tools offer various methods that can be harnessed in vocabulary needs analysis. This chapter presents an introduction to several methods, providing suggestions on how they can be used specifically for this purpose and identifying steps involved in various well-known corpus tools. It then considers the potential of hands on corpus work with students (known as Data Driven Learning), along with some common challenges. Finally, it introduces a free and user-friendly English corpus tool, *The Prime Machine*, and takes examples from two undergraduate corpus assignments to show how language learners can successfully start to explore their own vocabulary needs in readymade corpora and in collections of texts of specific varieties they have gathered themselves. It is demonstrated that these corpus-driven techniques can help language learners engage in some needs analyses of their own.

Keywords: Specialized Corpora; Data Driven Learning; Concordancing

Introduction

For many decades, analysis of lexical features of collections of texts have formed a basis for helping syllabus designers, textbook writers and language teachers make decisions about the relative importance and usefulness of teaching specific vocabulary. In English language teaching, since the development of the General Service List (West, 1953), through corpus-based developments in learner dictionaries in the 1980s and 1990s (Moon, 2007; Rundell, 1999), various applications of the Academic Word List (Coxhead, 2000), and more recent papers looking at words and collocations in specific academic fields (Ackermann & Chen, 2013; Durrant, 2009), the frequency, distribution and patterning of words have formed a foundation for determining vocabulary levels, text difficulty, and ultimately what to include in language courses small and large.

While exploration of balanced corpora containing multiple text types across genres and domains can help in the decision making processes for general curricula and courses for mixed disciplines, corpus tools can also reveal useful patterns of language use in highly specific fields and collections of homogenous texts. In most situations, teachers and course designers need to take on the vital task of needs analysis, and corpus data (and corpus-derived examples) play an important role. However, one of the goals of lifelong learning (particularly for English majors, translation students and trainee language teachers) must be to equip students with the skills to perform needs analysis in future (unknown) situations. User-friendly corpus tools can provide a test-bed for the development of these kinds of skills.

This chapter introduces some of the ways a new English corpus tool (*The Prime Machine*) can be used with linguistically-oriented language students to explore differences between

their own use of vocabulary and uses in existing corpora, and to uncover lexical features in their own Do-it-yourself (DIY) corpora. It will introduce some of the main ways readymade and homemade collections of texts can be compared and how further exploration of concordance lines can help language learners gain insights into specific lexical patterning. Following Dodigovic's distinction between "development oriented" and "effect oriented" research on Computer Assisted Language Learning (2005a, p. 48), this chapter is primarily concerned with providing background to concepts, techniques and issues that have informed the development of *The Prime Machine* corpus tool. A limited evaluation of its effects will be offered by drawing on examples from Chinese English majors in a Sino-British University in China.

Background

Using loose definition of a corpus as a collection of electronic texts, and defining corpus tools as software (applications, APPs or websites) that provide means of calculating and presenting data derived from these, this section will explore some of the ways in which corpus tools can be used to explore vocabulary. While corpus linguistics often focuses on interactions between vocabulary and grammar – the lexicogrammar (Hoey, 2005; Hunston & Francis, 2000) – the purpose here is to present methods which lend themselves particularly to gathering insights about specialist vocabulary and specialized uses of what appear to be more familiar vocabulary items. Each method will be introduced in turn, starting with definitions and purposes, considering potentials for vocabulary analysis, and then explaining procedures in well-known corpus tools. These tools include *WordSmith Tools* (Scott, 2016), *AntConc* (Anthony, 2004), *LancsBox* (Brezina, McEnery, & Wattam, 2015) and *Lextutor* (Cobb, 2000).

The use of hands-on corpus activities with language learners will then be reviewed, and some potential difficulties will be summarized.

Corpus Wordlist

One basic function of a corpus tool is to provide a list of different words (types) and their frequencies (token counts). In this chapter, this will be referred to as a *corpus wordlist* to distinguish it from some of the other wordlists described later. Corpus wordlists may be sorted in different ways, but here they are defined as lists of types in a corpus, sorted by descending frequency. In corpus linguistics, each *type* is a unique combination of characters making up a word and typically for English these would be extracted from text using spacing and punctuation as boundaries. Here, issues related to inclusion or exclusion of numbers and symbols, treatment of hyphenated words and possible groupings of word-forms together will be overlooked; for further discussion about such issues see Scott and Tribble (2006) and Jones and Durrant (2010). Scott and Tribble (2006) introduce the Wordlist function in *WordSmith Tools*, explaining that when sorted by descending frequency, almost any corpus wordlist will first contain a relatively small number of very high frequency words that form the top hundred or so, mostly consisting of grammatical units that hold a text together; then there will be a set of medium frequency words which typically contain what might be considered fairly common nouns, verbs and adjectives; and then at the end of the list there will be "… an enormous tail of hapax legomena (words that occur once only in a corpus)" (p. 11). Indeed, an important concept to understand with regard to the frequencies of different words in text (long or short; single or collective) is Zipf's Law (Scott & Tribble, 2006; Zipf, 1935). Put simply, Zipf demonstrated that when ranked by descending frequency, there is an extremely sharp decline in the frequencies of each word, and moving down a corpus wordlist

shows great differences between adjacently ranked items, especially at the top end. Through examples, Scott and Tribble (2006) demonstrate that corpus wordlists are useful as starting points, and may be useful for exploring authorship or indicating that texts were produced in different contexts. Corpus wordlists are often used as a starting point for vocabulary needs analysis, either through the construction of a corpus of target texts, or as a way of evaluating the potential usefulness of a text for teaching. Jones and Durrant (2010) point out that frequency can provide a useful means of determining what could be considered important (by virtue of vocabulary items being widely and frequently used in texts), but note other considerations may mean rankings should not be used to exclude items which could best be taught together (e.g. days of the week) or items that build useful communicative phrases. Generating corpus wordlists is a relatively simple in all well-known corpus tools, involving pointing the application at a set of files on the user's computer and selecting the Wordlist tool in *WordSmith Tools, Antconc* and *LancsBox*; or selecting to view the corpus wordlist after texts have been uploaded to the web server in the case of *Lextutor*. However, interpretation of the data is not so straightforward. Since corpus wordlists have potential for authorship identification and provide traces of contexts and production circumstances, word frequencies must be influenced greatly by the style or language habits of the speakers and writers who contributed to the texts they contain and the circumstances of their production. While corpora have much to offer, results and their rankings always need to be treated with caution as corpora are rarely as truly representative as would be desired.

The General Service List, the Academic Word List and Vocabulary Profiles

When wanting to analyse vocabulary in a text or a collection of texts, as well as using the corpus wordlist (derived from the texts being studied), it is also often helpful to match the

words against other wordlists. A number of wordlists are publicly available and the best known of these is probably the General Service List (GSL) (West, 1953) which was created to give an overview of the core vocabulary for English. Other lists of core vocabulary have been generated more recently (Brezina & Gablasova, 2015; Nation, 2000). Corpus wordlists derived from large collections of national corpora such as the British National Corpus (BNC, 2007) and the Corpus of Contemporary American English (COCA) (Davies, 2008-) can also be used for comparison. In order to generate a list of words for general academic language teaching (across disciplines), Coxhead (Coxhead, 2000) created a corpus of academic texts and then created a new wordlist of academic words (AWL) by excluding items already on GSL, and including words with a high frequency across academic disciplines. When provided as computer-readable lists of words, corpus tools can make use of such lists for a number of purposes, including to estimate text difficulty by calculating proportions of items which are frequent in the language generally, and so are likely to be well known by students. They can also be used to consider the generalizability of vocabulary in a text by revealing whether items from target wordlists are well represented. Dodigovic (2005b) and others in this volume introduce the power of such profiling for the evaluation of course books and individual texts. Lists for pedagogical application are not all formed with the same methodology: they may or may not exclude general vocabulary and/or general academic vocabulary (Gardner & Davies, 2014); and they may or may not attempt to focus on specialized terminology as opposed to specialized uses of familiar-looking items (Todd, 2017). Nevertheless, despite methodological differences, with the realization of the importance of specialist vocabulary, in recent years there have been further developments in academic and specialist wordlists for specific academic disciplines, including engineering (Khamis & Ho-Abdullah, 2017; Todd, 2017), medicine (Lei & Liu, 2016; Quero, 2017),

environmental sciences (Liu & Han, 2015) and core subjects in secondary school (Green & Lambert, 2018).

Tools generating or using such lists have been available for some time, with *Lextutor* being an excellent example of an interface that is not only easy to use, but also provides clearly presented results. One reason for the popularity of the *LexTutor* tool is probably the way the results are presented on a long single webpage in several useful ways: summary results, types grouped by wordlist and colour-coded running text. As well as results based on the GSL and AWL, *LexTutor* can also provide results for the top 10,000 items in the BNC and/or COCA, and top items from a graded reader collection. Use of such wordlists in other corpus tools is not usually as straight-forward; comparing a corpus wordlist with another wordlist is possible in *WordSmith Tools*, for example, but perhaps it is not frequently used for this kind of work.

Key Words and related methods

Results from corpus wordlists and vocabulary profiles are usually based on raw frequencies and will include a high proportion of high frequency grammatical words. In order to try to approximate the sense a human reader might have of what is prominent in a text (or a collection of texts), there is another corpus method – known as the Key Words (KW) method – which compares the frequencies of words in the corpus of interest with the frequencies of the same items in a reference corpus. The computation behind this is based on a simple cross-tabulation of the frequencies and total sizes of the item and the two corpora and Scott and Tribble (2006) provide examples from different text types for different purposes. In recent years, there has been some debate about how results from KW should be ranked and presented to users (Brezina, 2018; Gabrielatos, 2018; Jeaco, Accepted). Jones and Durrant

(2010) present KWs sorted by descending frequency. In terms of vocabulary selection, however, this tends to lead to a similar situation as that of the corpus wordlists when grammatical and common words tend to dominate the top. In whatever way they are ranked, KW lists are simply the results of an automated procedure, and interpretation of the importance of items in the list (and why the computer process may have promoted some items) is the responsibility of the user (Scott & Tribble, 2006).

KWs provide data-driven ways into the analysis of prominent topics, themes and heavy use of lexical items for studies of specific genres, registers and styles. Results based on collections of texts from a specific genre and/or from a specific discourse community will usually give lists containing genre markers (words associated with some essential moves for the genre); and topics and themes that give clues as to what the texts are about. The status of being a KW means there is data-driven evidence that this word is likely to be important and provides justification for selecting specific words for closer analysis. KWs often also indicate important aspects of register as features such as personal pronouns may indicate interesting aspects of the situational contexts in which the texts were produced. Similarly, words associated with stance and appraisal in a KW list may indicate interesting points about the way ideas and opinions are typically communicated within a specific domain.

To generate a list of KWs, the corpus tool needs a corpus wordlist for the text (or texts) being studied and a second corpus wordlist as a reference corpus. *WordSmith Tools* requires the user to use the WordList Tool to create a special Wordlist file; to either create a second Wordlist file for the reference corpus or to obtain one previously prepared; and then to use a separate tool within the application to load these two files and present the results. In *AntConc*, the selection of the user's own corpus texts is more straightforward (as loading the texts on

the left-hand panel makes them available across all the other tools), but the reference corpus must be specifically loaded from a specially formatted text file or a complete reference text. *LancsBox* provides a slightly different route, with buttons to download a small selection of reference corpus wordlists, so these can be used to generate KWs for the user's locally stored text files. These steps mean that choosing a reference corpus can be based on practical questions of what is ready to use, what is available on the user's own computer, and for larger reference corpora how well the machine can process large amounts of text. When trying to explore specific text types within a larger text variety; for the identification of academic vocabulary associated with a specific academic field for example, being able to select a very general corpus (such as the BNC) or a more specialized corpus (such as the BNC: Academic sub-corpus) can be very useful.

If the corpus contains many texts and the intention is to get a sense of what many of the texts are about, another related method can be employed. The calculation of Key Key Words (KKW) involves first calculating KWs on a text by text basis, and then ordering the results of these batches of KWs according to the number of texts in which each candidate KW is key (Scott & Tribble, 2006). When a DIY corpus is viewed as a dataset of target text types for needs analysis, it is clear that the KKW list can be particularly useful. For example, to determine words associated with themes of environmental news articles over an entire year, a KKW list will contain words for the major themes or agencies, with scores based on prominence within individual texts so as to screen out KWs which are heavily concentrated in some parts of the corpus but not others, and to screen out KWs which may have a relatively high frequency overall, but actually not be particularly prominent in any of the individual texts. KKWs is not widely available in popular corpus tools, the exception being *WordSmith Tools* (where the technique was first developed). The first step is to create a batch of

wordlists with the Wordlist tool. These can then be loaded together in the second step within the KeyWords tool, to create a KW database. Finally, these are ranked according to the number of texts in which they are Key.

Concordance lines

The methods described above begin with a whole corpus and then provide an overview and possible insights into some prominent or marked uses of vocabulary. The other functions to be described here relate to the way corpus tools can retrieve and calculate patterns based on queries for specific words or phrases. The primary output of corpus tools is typically concordance lines, presented as a list of horizontal lines of text containing the search term with a few words of context to the left and right of the target word. By presenting multiple examples on a single screen, important patterns in the usage of words can be revealed. An important difference between tools is the amount of co-text and contextual information that can be viewed. A vital step when exploring concordance lines is to control the way in which they are ordered as different ways of sorting the results will help make different kinds of patterning more noticeable. Repeated patterns of lexical words in the nearby co-text can help clarify common collocations and/or the use of word in semi-fixed phrases and help users identify semantic associations of a word. Semantic association is defined by Hoey as "when a word or word sequence is associated in the mind of a language user with a semantic set or class, some members of which are also collocations for that user" (2005, p. 24). From the perspective of vocabulary needs, this term can include related but distinct concepts of semantic prosody (Louw, 1993) and semantic preference (Sinclair, 2004). Linguistic research may distinguish between these kinds of feature, but here the main point is that words may have certain connotations or hidden meaning that is evident in multiple examples of

actual use, but may not be evident in dictionary definitions (Shinwoong, 2011). Repeated patterns of grammatical items in the nearby co-text can also show how vocabulary items may frequently be used in certain grammatical structures, and can help students identify patterns of prepositions. For a good overview of how concordance lines can be analysed see Hunston (2002). In terms of how corpus tools provide access to concordance lines, the process typically involves selecting a corpus, typing in a word and tapping a button to retrieve the results.

Collocations

Another way of exploring patterning of specific items is through calculating collocations. As described in Author (2019) while collocation is now known to be an essential aspect of language and is well-established as a component of vocabulary teaching, there are many different definitions and means of calculating collocations. For the purpose of this chapter, collocation will be defined using Hoey's definition:

> … collocation is … a psychological association between words (rather than lemmas) up to four words apart and is evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution.
>
> (Hoey, 2005, p. 5)

The definition here provides two important considerations for measurement: first being based on words rather than lemmas means that different word forms will have different collocation lists. Lemmas are normally understood to be the different word forms of a word within a word class. In vocabulary teaching terminology, following this definition, results are based on specific word forms, not grouped by word families. The second point is that some sort of statistical test is used to determine the likelihood of repeated co-occurrence of candidate

collocations being due to non-random influences. This definition provides information as to how collocations are retrieved as a means of approximating strengths of relationships between words that must exist in the minds of language users. Different statistical tests tend to prioritize (or exclude) words from different points on the Zipf curve; T-Test and to some extent MI and related measures may include more grammatical items, while DICE and related measures may include lower frequency lexical items. Some tests do not consider the order of the words, others take positioning into account. Some discussion of these differences can be found in Oakes (1998), Gries (2013) and Author (2019). Collocation lists may be generated from concordance lines (*WordSmith Tools*) or from a separate menu (*Antconc* and *Lancsbox*).

Data-driven learning in the classroom

Having presented some corpus methods that can be used to explore vocabulary in texts, the use of corpus tools in the classroom will now be introduced. Learning language through classroom activities related to corpus work is known as Data Driven Learning (DDL), and the processing of texts by language learners themselves for exploration of language features in which they are interested is not a new activity. The pioneer of DDL was undoubtedly Tim Johns and in his work with postgraduate students the corpora used were created by the students themselves (Johns, 1986). High level students who are highly motivated have found work with self-compiled corpora to be particularly rewarding (Charles, 2012b; Yoon, 2011). More broadly, studies on language learning through DDL using readymade or homemade corpora have shown that it is effective (Boulton & Cobb, 2017) and it is considered a fruitful means of providing opportunities for engagement language-learning processes (Flowerdew, 2015; Thomas, 2015). Some ways these activities can help in terms of vocabulary can be

showing a "snapshot" of vocabulary use (Johns, 2002), exploring differences between synonyms (Johns, 1991; Kaltenböck & Mehlmauer-Larcher, 2005) and deepening their word knowledge (Cobb, 1999). Examples of effective DDL work at postgraduate level include Johns (1991) and Charles (2012a) and undergraduate level examples include Fligelstone (1993) and Cheng, Warren and Xu (2003). There have been some recent explorations of its potential in China (Guan, 2013; He, 2015). DDL activities not only offer effective ways to engage language learners critically in language exploration in class, they also afford longer-term advantages as they are skills for self-study and life-long learning (Kaltenböck & Mehlmauer-Larcher, 2005; Mills, 1994).

Despite these benefits, using corpus tools in language learning contexts is not always easy or straightforward and a number of issues have been identified in the literature. Problems include getting hold of corpus texts (Ädel, 2010), being able to think of fruitful starting points, formulating queries and obtaining results (Ädel, 2010; Gabel, 2001; Sun, 2003), needing to spend time analysing and evaluating results (Ädel, 2010; Thurstun, 1996; Yeh, Liou, & Li, 2007), dealing with too much data (Ädel, 2010; Varley, 2009) and keeping a balance between focus on form and focus on meaning (Ädel, 2010).

The Prime Machine (tPM) was initially developed to provide user-friendly corpus access to online corpora for language learners and language teachers. Its interface was designed to address some of these difficulties, and to help language learners get started with concordancing and some of the special features of *tPM* for English language learning in terms of search support and highlighting patterning have been presented by Author (2017). Through working with several cohorts of English majors, the developer added new functions for the investigation of patterns in students' own corpora. Being based the lexical priming

theory of language (Hoey, 2005), *tPM* encourages the exploration of specific vocabulary items to compare these with words with similar meanings, to consider different uses and usage of different word forms and to explore differences between use is different kinds of text. The purpose of this section is to introduce some ways in which the readymade online corpora and DIY corpora constructed by students can be used for vocabulary needs analysis. To illustrate these techniques, the tasks used in an undergraduate module for English majors at a Sino-British university will be introduced, with particular attention paid to the ways in which these tasks foster self-awareness of vocabulary needs. The students taking this module were sophomores, and most had little or no prior experience of corpus work. Students at the university typically come from fairly traditional schooling, where grammar-translation methods are most frequently used. After the assessment period, students were invited to give permission for their assignments to be analysed and twenty students from the cohort of sixty-seven students agreed. The performance of these twenty students covered a wide range of marks so in that respect they can be considered representative of the cohort as a whole.

Task 1: Using readymade corpora for a reflective writing task

The first task covered the first six weeks, with weekly two hour lectures on the background of corpus linguistics and weekly one hour computer workshops to introduce and practice using the corpus tool. Students had to produce three reflective summaries based on language points in their own writing or speech, using concordance lines and other corpus data to justify possible choices. Suggestions were given for how to select language points and students used a variety of methods. Self-transcribed speech can aid noticing (Lynch, 2001) and 10 students used this for selection of items in 18 reflections. A student from a previous cohort had suggested using machine translation to translate an essay into Chinese and back into English

as a way of identifying some possible re-wordings, and 2 students used this for a total of 4 reflections. Feedback from a teacher on a former assignment (2 students, 4 reflections), feedback from friends or peers (4 students, 8 reflections), and general rules of thumb (4 students, 5 reflections) were also starting points for some students. In the other 22 cases, a specific reason for selecting items was not stated, but included having seen something recently, some thoughts on creative writing and re-reading of an assignment written several months previously. One student presented only 2 of the 3 required analyses and 2 students presented additional analyses, so a total of 61 reflections were analysed. Table 1 shows the kinds of linguistic data they drew on as they created their own notes on language use in a readymade corpus.

Table 1: Kinds of linguistic analyses

Student #	Syno-nyms	Collo c-ations	Sign-posts	Other phrases	Collig -ation	Seman-tic assoc-iation	Nega-tive meanin g	Contexts of use	Fill ers
1				1	1	1		1	
2	2	1				2			
3	2	2	1			2		1	
4	1		1			2	2	1	
5	2	2			1	1			
6		1	1	1		2			
7	2	2			1				1
8	1	2	1		2	1			
9	3				1	2	1		

10	2	3		2	1	1			
11	2	2			1			1	
12	3	3	2			1			
13	3	1				2			
14	1	1		2	2	3		1	
15	2	2	1	1	1	1		1	
16	2	3		1		3	1	3	
17	1	3			1	1			
18	2	2			1	2		2	
19		2	1		2				
20	2	3		2		2			
Total	**33**	**35**	**8**	**10**	**15**	**29**	**4**	**11**	**1**
Students	17 (85%)	17 (85%)	7 (35%)	7 (35%)	12 (60%)	17 (85%)	3 (15%)	8 (40%)	1 (5%)

When counting different kinds of linguistic analysis, a single reflective summary may have covered multiple points, with collocations often being used to distinguish between synonyms, and commentary on collocation patterns often including analysis of semantic sets that were evident in the examples. For this reason, from the 61 reflective summaries, 146 linguistic analyses were counted. From Table 1 it can be see that synonyms and collocations were analysed by many students, but there was also some engagement with semantic associations and contexts of use, as evidenced in exploration in different corpora or through noting different kinds of sources.

Task 2: Creating homemade corpora to explore specialized language use

The second task ran over the remaining six weeks, and involved the creation of one or two homemade corpora. Students had to complete a number of smaller tasks to explain the design of their study, to consider the situational contexts (following Biber & Conrad, 2009, p. 40), to produce various kinds of corpus data, and then to summarize their findings. One of the main distinguishing features of *tPM*, compared with other corpus tools is that users can access the data from the ready-made corpora in their analysis of their own DIY corpora; the currently selected online corpus can not only be used as a reference corpus for KW and KKW analyses, searches of specific words can be displayed with results from the DIY corpus displayed side by side with the online corpus. Figure 1 shows the DIY corpus wordlist screen, with the one step buttons that are used to obtain results using the currently selected readymade corpus as reference.

Figure 1: DIY Wordlist Tools tab in *tPM*

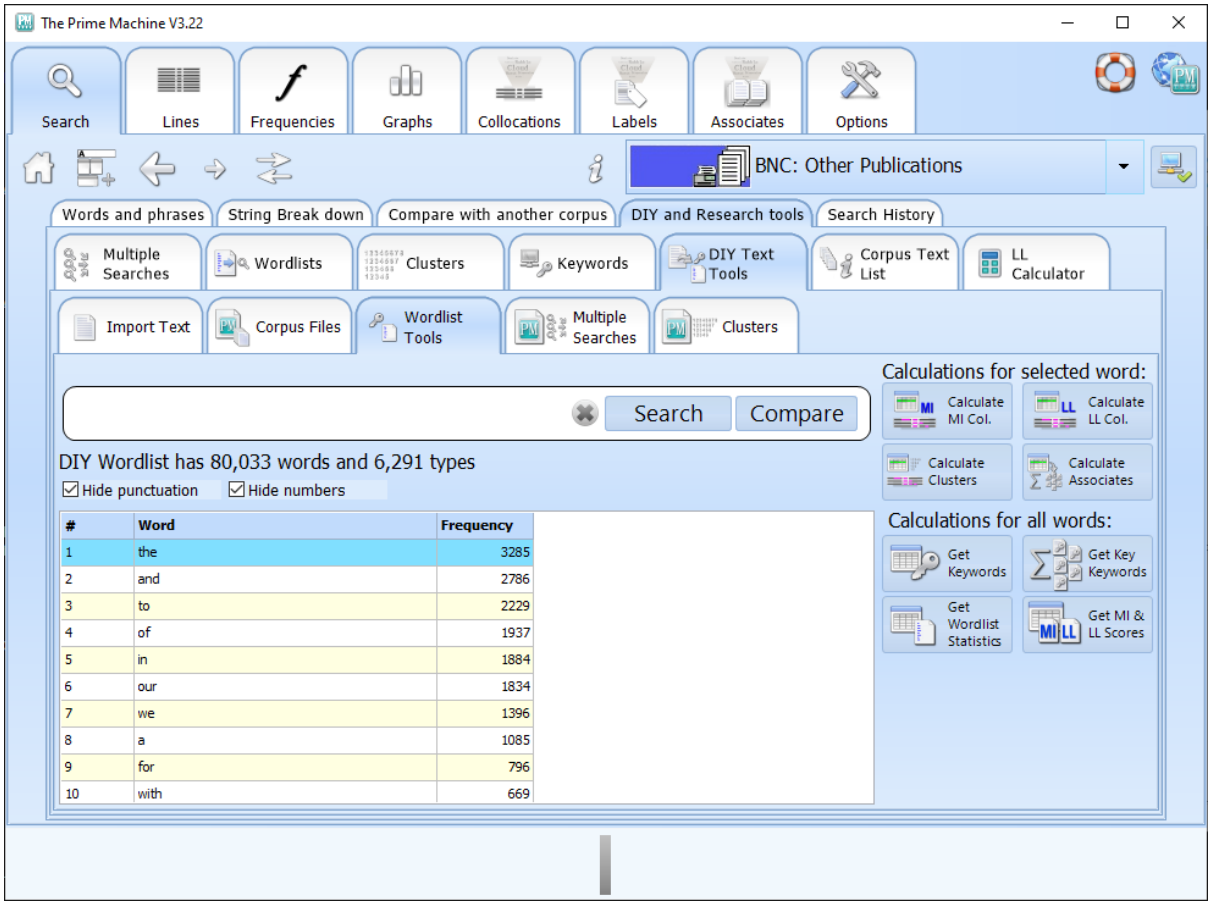

Figure 2: Concordance lines for *innovation* in a DIY corpus of Chairman's Statements and BNC: Other Publications

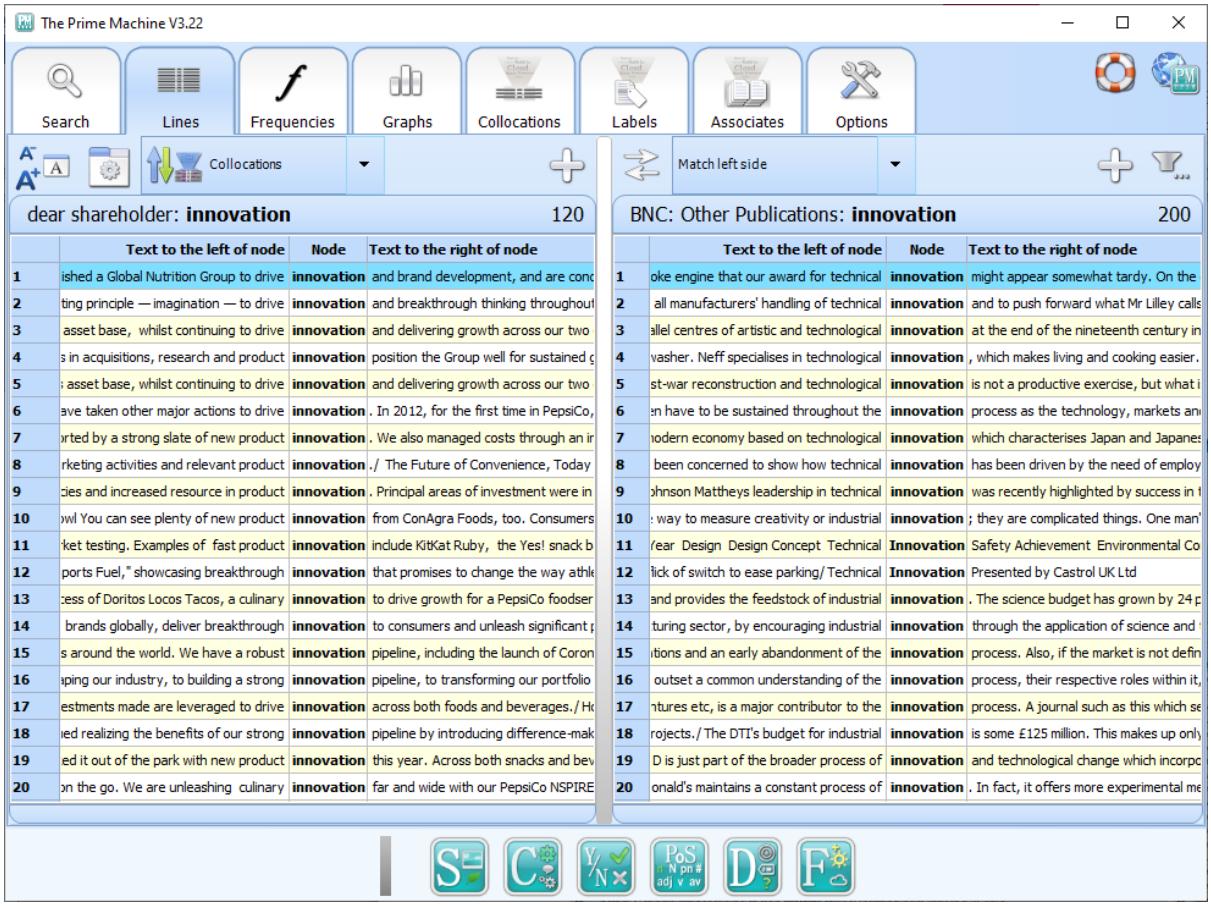

Figure 2 shows concordance lines for *innovation* in a student's DIY corpus on the left with results from a subsection of the BNC on the right. *tPM* allows users to access the whole BNC or to select subsections following Lee's (Lee, 2001) classifications.

Vocabulary profiles

The wordlist statistics function was developed to draw students' attention to differences in the proportion of commonly used words, academic words and other sets of words within their DIY corpora. The ready-made wordlists include GSL, AWL, positive and negative words, modal verbs, personal pronouns, and first and second person pronouns. One button retrieves

summary statistics of matches to these lists. The lists also include comparisons with the reference corpus, with a log-likelihood calculation like that used for KWs, and arrow indicators to show the directions and degrees of difference. Table 2 shows results obtained by the student who compared her Chairman's Statements corpus with the BNC.

Table 2: Wordlist statistics for a DIY corpus compared with the BNC

	Wordlist	**Study Freq.**	**Study Per Thousand**	**Reference Freq.**	**Ref. Per Thousand**	**Arrows**	**LL**	**Bayes**
1	Academic Word List	6998	87.44	4913845	42.63	≧ 2x ↑	2730.92	Very strong evidence
2	1st & 2nd Pers. Pronouns	3785	47.29	2509362	21.77	≧ 2x ↑	1737.53	Very strong evidence
3	Positive words	2537	31.7	1426341	12.37	≧ 2x ↑	1650.73	Very strong evidence
4	General Service List 2	4092	51.13	5414865	46.97	↑	27.3	Strong Evidence
5	Modals Subgroup 1	260	3.25	647649	5.62	↓		
6	Modals	438	5.47	1469299	12.75	≧ 2x ↓		
7	Modals Subgroup 3	54	0.67	240052	2.08	≧ 3x ↓		
8	Archaic	0	0	3940	0.03	↓		

	Pronouns							
9	Personal Pronouns	4441	55.49	6514354	56.51	↓		
10	Function Words	30030	375.22	49484588	429.25	↓		
11	Punctuation	4293	53.64	6659675	57.77	↓		
12	General Service List 1	48045	600.31	73117987	634.26	↓		
13	Modals Subgroup 2	124	1.55	579255	5.02	≧ 3x ↓		
14	Negative words	327	4.09	1649527	14.31	≧ 3x ↓		

The results for GSL and AWL can be used to give an indication of the proportion of words beyond these lists, giving a clue as to how familiar the students are likely to find these. Through learning about features of academic English in EAP classes, students are often ready to note differences in text varieties when it comes to the use of personal pronouns and modal verbs. The results for first and second person pronouns and positive words in Table 2, give some useful insights into features of Chairman's Statements, where *I, we* and *our* and a range of positive words often work together to project an image of strong company performance.

Key Words and Key Key Words

Two of the most useful functions for exploring vocabulary in a DIY corpus are KWs and the related function KKWs. Because the DIY text tools are integrated into a client-server corpus tool, generating KWs and KKWs is extremely easy in *tPM*. After importing texts, any of the readymade corpora can be selected from the prominent drop-down menu to be used as a reference corpus, and then KWs or KKWs can be generated through the click of a single button. *tPM* makes the process simple, but the interpretation of the results is still for students to work on by themselves. Most students export the KW and KKW lists to spreadsheets and they were encouraged to use colour to categorise words. Prompting students to come up with their own classifications of KWs and to shade cells in their spreadsheets has been a good way to ensure they understand the need for interpretation of such lists. Table 3 shows the students own classification of other KWs.

Table 3: Top 15 Key Words for a DIY Corpus with manual categories created by a student.

	Word	Study Freq	Study Per Thousand	Ref Freq	Ref. Per Thousand	Arrows	LL
1	our	1834	22.92	93240	0.81	≧ 10x ↑	8693.94
2	PepsiCo	273	3.41	5	0	≧100x ↑	3921.18
3	we	1396	17.44	350582	3.04	≧ 5x ↑	2567.79
4	growth	408	5.1	12895	0.11	≧ 10x ↑	2306.28
5	brands	191	2.39	773	0.01	≧100x ↑	1819.74
6	clover	132	1.65	221	0	≧100x ↑	1453.8
7	beverage	103	1.29	112	0	≧100x ↑	1200.8
8	business	332	4.15	35430	0.31	≧ 10x ↑	1110.6

9	products	227	2.84	10676	0.09	≧ 10x ↑	1109.87
10	portfolio	142	1.77	1583	0.01	≧100x ↑	1086.67
11	consumers	151	1.89	2294	0.02	≧ 10x ↑	1066.34
12	foods	139	1.74	2085	0.02	≧ 10x ↑	984.99
13	innovation	120	1.5	1694	0.01	≧100x ↑	864.29
14	year	411	5.14	88309	0.77	≧ 5x ↑	863.24
15	ConAgra	59	0.74	0	0	☀ ↑	858.26

Personal Pronoun
company/product name

As well as producing some results based on the whole corpus, students were requested to perform some analysis of concordance lines and to draw on collocation or other data to demonstrate special features of the vocabulary use in specific contexts. Through using different sorting methods, students were able to identify some specialized uses of vocabulary with which they were already familiar, as well as uses of new vocabulary in a specific domain. Table 4 summarizes the corpus methods students completed in this second task.

Table 4: Corpus methods used by students in the second task

	Wordlists	KWs	KKWs	Conc. lines	Collocations
Results presented	15 (75%)	20 (100%)	6 (30%)	19 (95%)	4 (20%)
Analysed	12 (60%)	17 (85%)	5 (25%)	19 (95%)	4 (20%)

As can be seen there was only one student who did not present concordance line data. Of the other students, a majority not only included Wordlist and KWs data, but also presented analysis of these (some through colour classifications, others by highlighting data in figures, and some by describing features in prose). Other corpus methods used in the assignment were almost always presented with some analysis. 12 students presented analysis of three or more kinds of data, and 6 students analysed four or more.

Student response

Both assignments also included a short reflective piece about the overall tasks. Responses included many positive comments about the usefulness of the overall learning process and the insights they gained. Several comments related to increased language awareness and insights into contexts of use.

Task 1:

> *Learning a new word entails more than knowing its meaning but its surroundings as well.* (Student #5)

> *… corpus linguistic analysis can provide learners with practical opportunities to focus on word choice and collocations in authentic examples, thus, develop language expertise.* (Student #13)

Task 2:

> *I found tPM is a useful tool in both analysing semantic features of words and finding the similarities or differences between two text varieties or a variety with a more general one.* (Student #9)

> *Thus, tPM is an extremely convenient and valuable tool for analysing linguistic features, which from my perspective would be a good choice of the topic of my final year project.* (Student #11).

As can be seen, students commented favourably on the task and the software and also demonstrated their increased awareness of language use. It is important to consider that comments were part of assessment, and this could have led to the expression of overly positive views. However, anonymized feedback from the cohort obtained through an institution level module evaluation platform revealed higher than average responses for "a valuable learning experience" (43 responses, mean 4.79 out of 5, 0.47 standard deviation). In response to an open ended question about what they enjoyed in the module and why, there were 4 related to a sense of engagement and/or achievement in the second task, 10 related to the software (all positive), 6 related to insights they gained into language use and 7 related to intentions for future use of corpus tools. Only one response was negative (simply "nothing"), while the other comments were positive reflections on the teacher (18), feedback (4) or assessment design (4).

Conclusion

This chapter has introduced some corpus techniques that have been incorporated and developed in *The Prime Machine* and that can be used to help linguistically oriented English language learners explore their own vocabulary needs. It has provided an overview of ways in which different kinds of corpus data can be used to inform this process, including wordlists, KWs, concordance lines and collocations. Work done by Chinese students majoring in English has illustrated ways the tool can be used and noted a positive response. Future research will need to focus on evaluation of the depths of insight gained by such learners and the extent to which it actually contributes to on-going learning. Nevertheless, with greater availability of free tools such as *The Prime Machine,* it is hoped that more language learners will have the opportunity to steer their own vocabulary needs analyses in future.

The Prime Machine is available for Windows and MacOS from www.theprimemachine.net.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12,* 235-247.

Ädel, A. (2010). Using corpora to teach English for Specific Purposes. In M. C. Campoy & B. M. L. Belles-Fortuno (Eds.), *Corpus-based approaches to English language teaching* (pp. 39-55). London: Continuum.

Anthony, L. (2004). *AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit.* Paper presented at the Interactive Workshop on Language e-Learning, Waseda University, Tokyo.

Biber, D., & Conrad, S. M. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

BNC. (2007). The British National Corpus (Version 3 BNC XML ed.): Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning, 67*(2), 348-393. doi: 10.1111/lang.12224

Brezina, V. (2018). *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

Brezina, V., & Gablasova, D. (2015). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics, 36*(1), 1.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*(2), 139-173. doi: 10.1075/ijcl.20.2.01bre

Charles, M. (2012a). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes, 31*, 93-102. doi: 10.1016/j.esp.2011.12.003

Charles, M. (2012b). *Student corpus use: Giving up or keeping on?* Paper presented at the TaLC10 Conference, Warsaw.

Cheng, W., Warren, M., & Xu, X.-f. (2003). The language learner as language researcher: putting corpus linguistics on the timetable. *System, 31*(2), 173-186. doi: 10.1016/s0346-251x(03)00019-8

Cobb, T. (1999). *Giving learners something to do with concordance output*. Paper presented at the ITMELT '99 Conference, Hong Kong.

Cobb, T. (2000). The Compleat Lexical Tutor, from http://www.lextutor.ca

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Davies, M. (2008-). The Corpus of Contemporary American English (COCA): 520 million words, 1990-present. Retrieved 3 April, 2017, from http://corpus.byu.edu/coca/

Dodigovic, M. (2005a). *Artificial intelligence in second language learning : raising error awareness*. Buffalo, NY: Multilingual Matters Ltd.

Dodigovic, M. (2005b). Vocabulary Profiling with Electronic Corpora: A Case Study in Computer Assisted Needs Analysis. *Computer Assisted Language Learning, 18*(5), 443-455.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for Academic Purposes. *English for Specific Purposes, 28*(3), 157-169.

Fligelstone, S. (1993). Some reflections on the question of teaching, from a corpus linguistics perspective. *ICAME, 17*, 97-109.

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. Amsterdam: John Benjamins.

Gabel, S. (2001). Over-indulgence and under-representation in interlanguage: Reflections on the utilization of concordancers in self-directed foreign language learning. *Computer Assisted Language Learning, 14*(3-4), 269-288.

Gabrielatos, C. (2018). Keyness Analysis: nature, metrics and techniques. In C. Taylor & A. Marchi (Eds.), *Corpus approaches to discourse : a critical review* (pp. 225-258): Routledge, Taylor & Francis Group.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List (Vol. 35, pp. 305-327): Oxford University Press.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for

eight secondary subjects. *Journal of English for Academic Purposes, 35*, 105-115. doi: 10.1016/j.jeap.2018.07.004

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics, 18*(1), 137-165.

Guan, X. (2013). A Study on the Application of Data-driven Learning in Vocabulary Teaching and Learning in China's EFL Class. *Journal of Language Teaching & Research, 4*(1), 105-112.

He, A. (2015). Corpus Pedagogic Processing of Phraseology for EFL Teaching: A Case of Implementation. In B. Zou, M. Hoey & S. Smith (Eds.), *Corpus linguistics in Chinese contexts* (pp. 98-113). Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jeaco, S. (2017). Concordancing Lexical Primings. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 273-296). Amsterdam: John Benjamins.

Jeaco, S. (2019). Exploring Collocations with The Prime Machine. *International Journal of Computer-Assisted Language Learning & Teaching, 9*(3), 29-49.

Jeaco, S. (Accepted). Key words when text forms the unit of study: Sizing up the effects of different measures. *International Journal of Corpus Linguistics*.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System, 14*(2), 151-162.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (Vol. 4, pp. 1-13). Birmingham: Centre for English Language Studies, University of Birmingham.

Johns, T. (2002). Data-driven Learning: The perpetual change. In B. Kettemann, G. Marko & T. McEnery (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 107-117). Amsterdam: Rodopi.

Jones, M., & Durrant, P. (2010). What can a corpus tell us about vocabulary teaching materials? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 387-400). Abingdon: Routledge.

Kaltenböck, G., & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL, 17*(01), 65-84.

Khamis, N., & Ho-Abdullah, I. (2017). Lexical Features of Engineering English vs. General English. *GEMA Online Journal of Language Studies*, *17*(3), 106-119. doi: 10.17576/gema-2017-1703-07

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, *5*(3), 37-72.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42-53. doi: 10.1016/j.jeap.2016.01.008

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes, 39*, 1-11. doi: 10.1016/j.esp.2015.03.001

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, E. Tognini-Bonelli & J. Sinclair (Eds.), *Text and technology* (pp. 157-176). Amsterdam: John Benjamins.

Lynch, T. (2001). Seeing what they meant: transcribing as a route to noticing. *ELT Journal: English Language Teachers Journal, 55*(2), 124.

Mills, J. (1994). *Learner autonomy through the use of a concordancer*. Paper presented at the Meeting of EUROCALL, Karlsruhe, Germany.

Moon, R. (2007). Sinclair, Lexicography, and the Cobuild Project: The Application of Theory. *International Journal of Corpus Linguistics, 12*(2), 159-181.

Nation, I. S. P. (2000). Vocabulary (pp. 665-667): Routledge.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Quero, B. (2017). A Corpus Comparison Approach for Estimating the Vocabulary Load of Medical Textbooks Using The GSL, AWL, and EAP Science Lists. *TESOL International Journal, 12*(1), 177-190.

Rundell, M. (1999). Dictionary use in production. *International Journal of Lexicography, 12*(1), 35-54.

Scott, M. (2016). WordSmith Tools (Version 7). Oxford: Oxford University Press.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.

Shinwoong, L. (2011). Semantic Prosody in Bilingual Dictionaries and EFL Learners' Sentence Writings. *English Teaching, 66*(2), 253-272.

Sinclair, J. M. (2004). *Trust the Text: Language, Corpus and Discourse*: London : Routledge, 2004.

Sun, Y.-C. (2003). Learning process, strategies and web-based concordancers: a case study. *British Journal of Educational Technology, 34*, 601-613.

Thomas, J. (2015). *Discovering English with Sketch Engine*: Versatile.

Thurstun, J. (1996). *Teaching the vocabulary of academic English via concordances*. Paper presented at the Annual Meeting of the Teachers of English to Speakers of Other Languages, Chicago.

Todd, R. W. (2017). An opaque engineering word list: Which words should a teacher focus on? *English for Specific Purposes, 45*, 31-39. doi: 10.1016/j.esp.2016.08.003

Varley, S. (2009). I'll just look that up in the concordancer: integrating corpus consultation into the language learning environment. *Computer Assisted Language Learning, 22*(2), 133-152.

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

Yeh, Y., Liou, H.-C., & Li, Y.-H. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning, 20*(2), 131-152.

Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes, 10*(3), 130-139.

Zipf, G. K. (1935). *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Boston, MA: Houghton Mifflin.