International Journal of Computer-Assisted Language Learning and Teaching
Volume 9 • Issue 3 • July-September 2019

Exploring Collocations with The Prime Machine

Stephen Jeaco, Xi'an Jiaotong-Liverpool University, Suzhou, China

ABSTRACT

One of the greatest impacts of corpus linguistics on language teaching has been in the recognition of the importance of collocation. A very influential guide for language teachers with regard to teaching collocation has been the Lexical Approach. Activities pointing students to rich collocational information in monolingual dictionaries, in texts and specifically in collocation dictionaries provided ways for language learners to engage with collocation information: to notice, to remember and to acquire. In recent years, there has been a growing interest in Data Driven Learning and new tools are now available to allow students to access collocation information from corpora for themselves. After introducing some pedagogic considerations, this article presents some of the features of The Prime Machine which were developed to support DDL activities focussed on collocation.

KEYWORDS

Collocation, Concordance Lines, Data Driven Learning, Language Learning

INTRODUCTION

One of the greatest impacts of corpus linguistics on language teaching has been in the recognition of the importance of collocation for effective language learning. As computer searchable databases of authentic language texts became available to linguists and lexicographers, evidence for the patterning and co-selection of word choices required greater attention (Barnbrook, Mason, & Krishnamurthy, 2013; Sinclair, 1991). A very influential guide for language teachers with regard to collocation has been the Lexical Approach (Lewis, 2008) and its activities pointing students to the rich collocational information in monolingual dictionaries, in the examples in their texts and specifically in collocation dictionaries provided ways for language learners to engage with collocation information: to notice, to remember and to acquire (Hill, Lewis, & Lewis, 2000) [1].

While mobile phone technology and the internet have brought increased ease for checking and finding simple meanings of words, some of the pedagogic basis for spending time, thought and energy on retrieving, digesting and recalling this information has perhaps been lost. While web corpora have grown in size and scope, some of the opportunities for exploring collocations in specific contexts, perhaps for specific fields, have also to some extent been overlooked. However, one use of technology in the language learning classroom in particular – Data Driven Learning (DDL) – can offer opportunities for learners to engage directly with language data for a range of language learning pursuits (Flowerdew, 2015; Thomas, 2015; Tsui, 2004).

This paper presents an overview of some of the features of the newest version of *The Prime Machine* which have been developed to support the exploration of collocations as part of Data Driven Learning activities. Inspired by the theory of Lexical Priming (Hoey, 2005), *The Prime Machine* was developed as a learner-friendly corpus tool (Jeaco, 2015; Jeaco, 2017a)[2]. After introducing the

DOI: 10.4018/IJCALLT.2019070103

pedagogical and theoretical background of collocation and DDL, this paper goes on to introduce the way collocation is measured and presented in several corpus tools, before describing the ways users of *The Prime Machine* can view and interact with collocation data.

Defining Collocation

From a language teaching perspective, definitions of collocation can be fairly broad with the aim being to make learners more aware generally of the patterns of words around them. Developers of the Lexical Approach have introduced the concept to learners by describing the variation in the strength of relationships between words as being similar to the variations found in relationships between people (Hill, et al., 2000). Other suggestions include comparing composition to the use of a model airplane kit or drawing on expectations of their mother tongue through reflection or translation of technical phrases (Conzett, 2000; Hill, et al., 2000). With all of these explanations, the aim is to help learners start noticing collocation in the language they encounter and to encourage them to invest time in this enterprise because of the practical advantages collocations can offer. It seems that many learners find it unusual at first to examine language in front of them in units beyond individual words. Indeed, Conzett (2000) claims that explicitly making students aware of the term "collocation" will speed up class activities based around collocation. Woolard argues that definitions based on statistical information do not "guide my students' attention to specific elements of text in a clear and directed way"; saying that for the purposes of teaching a definition focussing on expectation is more useful (Woolard, 2000, p. 29) and he also restricts the patterns covered by the term to specific combinations of word classes.

While the introduction of the term "collocation" in the classroom may be made through metaphor, reflection on translations in the mother tongue, practical applications or repeated methods of annotation, more formal definitions of collocation in linguistic theory and computational linguistics have developed over the years. The differences in the qualification and scope of collocations which mirror the different purposes that different teachers have in mind are part of a general tendency for different researchers to specify the meaning of collocation and similar phenomena in different ways. Linguistic descriptions include a wide variety of ways of limiting what should count and how it should be understood to operate, including strings of characters in raw text (Sinclair, 1991), lexical phrases (Nattinger & DeCarrico, 1992), lexical bundles (Biber, Johansson, Leech, Conrad, & Finegan, 1999), motivated or unmotivated collocations (Hunston, 2002), lexical networks across sections of a book (Phillips, 1985), within Pattern Grammar (Hunston & Francis, 2000), and as a major contribution to the identification of norms (Hanks, 2013). Each of these stipulate such aspects as whether collocation-like phenomena operate on word forms or lemma, only in specific grammatical relations or freely, and for any kind of word or only certain parts of speech. According to Hoey (2005), at times the very definition of collocation has been tied up with the methodological approach for their retrieval. He provides a definition which fuses the psychological importance of collocation with the means of detection and evaluation:

> *So our definition of collocation is that it 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).*

Hoey introduces collocation using the example inevitable + consequence (2005, p. 2). As a piece of software purposefully designed to support the examination of the kinds of relationship between words which are introduced in Hoey's theory of Lexical Priming, collocations are defined for *The Prime Machine* based on his 2005 definition[3]. In this paper, collocation will be used as it is in the software to refer to combinations of two, three, four or five words in a four-word window either side of a node. The term "multi-word unit" will occasionally be used where combinations beyond two words in length are the focus of the discussion.

Collocation and Language Teaching

New measures and new terminology from corpus linguistics may have had indirect impacts on language teaching through discussion of pedagogic applications or through inclusion of corpus-derived patterning in language reference resources and teaching materials. However, collocation stands out as a clear example of a linguistic phenomenon that draws heavily on corpus linguistics and also requires increasing explicit attention from teachers. Teachers are expected to know what collocation means and looking at the presentation of collocation information in dictionaries and textbooks it seems that their students are also being encouraged to appreciate its importance too. This section introduces the role of collocation in language teaching as seen in textbooks, dictionaries and collocation lists.

The power of teaching collocation in language learning can be seen in some of the literature. As the pillar of the Lexical Approach, collocation is claimed to be a way to break through the "intermediate plateau" (Morgan Lewis, 2000, p. 14). Hill puts the lack of knowledge of collocations down as the root of many errors in learner language which are caused "because they create longer utterances because they do not know the collocations which express precisely what they want to say." (Hill, 2000, p. 49). In a book giving guidance to teachers on English for Academic Purposes (EAP), Alexander, Argent and Spencer make the following claim about the importance of collocation for a non-native speaking student's acceptance into the academic community:

In Academic English particularly, where writing is expected to conform to predictable patterns, mis-collocation can be one of the most distracting advertisements that the writer is not a competent writer in English, and can lead to a different meaning from that intended (Alexander, Argent, & Spencer, 2008, p. 163).

Ackermann and Chen (2013) summarize some of the problems related to student learning and use of collocations that have been described in the literature, stating, "if learners aim for advanced proficiency, achieving a high level of collocational competence is essential" (p. 236).

While not all language teachers will be following methods connected with the Lexical Approach or teach EAP, the ubiquity of the term across different aspects of mainstream language teaching is obvious. For twenty years or more, collocation activities have formed part of general English course books. *Cutting Edge* is a widely used course and it includes exercises matching verb-noun combinations and reflections on how best to note "words that go together" (Cunningham & Moor, 1999, p. 53). The *Touchstone* series of books is heavily promoted as being corpus-informed and understandably puts access to collocation information as one of the key aspects of the selection process for items and examples (McCarthy, 2004). From Level 3 in *Touchstone,* a main section is devoted to teaching students how to "Learn new words in combination with other words that often go with them" (McCarthy, McCarten, & Sandiford, 2006a, p. vii). From Level 4, the term "collocation" is introduced and subsequently used again in the skills summary of the scope and sequence pages (McCarthy, McCarten, & Sandiford, 2006b, p. ix). Moving to Business English, in the very popular course, *Market Leader*, students have exercises where they mark "word partnerships" (Cotton, Falvey, & Kent, 2006, p. 9), while in the teacher's book they are called "collocations" and it is explained that one collocation for each word in the exercise can be found in the text (Mascull & Heitler, 2006, p. 12). In EAP, the use of the actual term "collocation" and teaching about its meaning seems to be more common. In an exam preparation book for the *International English Language Testing System* (IELTS) written by one of the key figures behind the exam itself, the need to "Choose words that go well together" is listed in the assessment criteria for both writing and speaking and reflective tasks based on this are provided (Jakeman & McDowell, 2008, pp. 92, 138). The public band descriptors for this test (Speaking, Writing Task 1 and Writing Task 2) all include "collocation" as a requirement for Band 7 "Good User" (available from www.ielts.org)[4]. Other leading courses for Academic English incorporate information about collocations as well as activities encouraging students to notice and record them. The *Academic Encounters* series has reading activities where students are given an explanation of collocation, told how knowing collocations makes reading easier and instructed to scan the text to match nouns to verbs (Brown & Hood, 2002, p. 89). *EAP Now!* includes a definition of collocation explaining that they "just fit together" and emphasising there is no "special reason" for these combinations, encouraging students to ask their teacher if words form a "good collocation" and whether they "sound natural" (Cox & Hill, 2011, p. 31). The *Oxford EAP* series has instructions for teachers for the Pre-Intermediate level stating "Students should recognize words that go with – collocate with – the word *cognitive"* (de Chazal & Rogers, 2014, p. 17). At the Upper-Intermediate and Advanced levels, there are activities in the Student's Books about creating and using collocations (de Chazal & McCarter, 2012), and identifying discipline-specific collocations (de Chazal & Moore, 2013). In textbooks such as these, the teaching of collocation involves awareness raising; attention to the combinations of nouns and verbs or adjectives and nouns in a text; and tips for noting collocations for vocabulary building or for reading.

As well as being fairly well represented in teaching materials, learner dictionaries also seem to draw an increasing amount of attention to collocation information. As the pioneer of corpus-driven lexicography, the *COBUILD* dictionary has always emphasised collocation by design. The *Collins COBUILD Advanced Dictionary of English* (2009) includes collocations in its full sentence definitions and also has prominent "Word Partnership" boxes "giving the complete collocation with the headword in place to clearly demonstrate use" (p. viii). The *Macmillan English Dictionary for Advanced Learners* (2007) has "Word Partnership" boxes showing full collocations with word class information. Words which have "many collocations" have an additional "Collocation Box" grouping collocations by sense and word class combination (p. x). In the Second Edition, there were more than 500 entries with these boxes (Rundell). A similar two tier system is used in the *Longman Dictionary of Contemporary English* (2009), with collocations shown in bold type in the main block and an additional collocation box for those words which have "a lot of collocations" (p. xiii). Obviously, there is not sufficient space to include collocation information for most headwords and publishers also have dedicated collocation dictionaries. In the learner dictionaries listed above, although longer explanations of the meaning and importance of collocation are available, the short usage guides typically explain collocation in simple terms as being: "words that are often used with a particular word" (*Longman Dictionary of Contemporary English*, 2009, p. xiii); "how words combine" (*Macmillan English Dictionary for Advanced Learners*, 2007, p. x) or "… high-frequency word patterns" (*Collins COBUILD Advanced Dictionary of English*, 2009, p. viii).

In addition to textbook and dictionary coverage, lists of collocations have been developed and proposed to help guide language teachers and materials developers in their selection and presentation of texts (Ackermann & Chen, 2013; Durrant, 2009; Simpson-Vlach & Ellis, 2010). These lists are separate resources, removed from the corpora from which they were developed, meaning that to give examples or explanation teachers would need to notice or highlight them in other texts, in much the same ways that the Academic Word List (Coxhead, 2000) has been used by teachers to pick out academic words in a text or to evaluate a text in terms of its coverage of common academic vocabulary.

Data Driven Learning

It is important to consider the role collocations can play as a resource in the classroom and for self-study activities for students of a foreign language. For all aspects of *Lexical Priming*, Hoey (2005) emphasises that associations will be specific to the domain and genre. Although it has been illustrated above that dictionaries and text books highlight collocation fairly prominently, when considering learners' needs for collocations across different disciplines, resources are still very limited. Alexander, Argent and Spencer (2008) give a case study in which students ask whether "arrive at" and "come to" are suitable to be used with "conclusion" in academic writing and the teacher responds in the affirmative immediately. They discuss the fact that when giving students information about collocation it is best to check resources first and that a dictionary would probably not be much help. They recommend looking at concordance lines and presenting these as evidence to learners in the next class. Since collocations are often discipline specific, although textbooks and dictionaries do include some useful information, concordance lines and concordancing software are important resources for EAP. O'Keeffe, McCarthy and Carter (2007) also argue that corpus data can provide teachers with strong support for explaining the collocations for more "banal" or "everyday" words which are less easily retrieved through intuition, adding that providing learners with evidence from multiple texts in a corpus gives a teacher much more confidence than just looking at one example in a class text.

Jeaco (2017a) introduces some of the links between classroom concordancing and language learning and teaching. The use of hands-on classroom concordancing is best known as Data Driven Learning (DDL), first associated with Johns (Johns, 1986, 1988, 1991, 2002). DDL typically involves activities where students type words into a concordancer and then explore the patterning of language through looking at the Key Word in Context display of concordance lines. In this way, rather than being instructed on typical collocations or grammatical patterns, the learners draw their own conclusions from the evidence presented in the corpus data. In his work, Johns emphasized the way in which these activities can help learners and teachers with problem areas such as prepositions and comparing synonyms, and as a reference tool for feedback in tutorials. Bernardini (2004) recognised a wide range of language teaching goals that can be achieved through classroom concordancing. Tasks designed by teachers that guide language learners in the discovery of patterns of language data in corpora may take some time, but actually incorporate study skills, learning about language and acquisition (Thomas, 2015). Links between theories of second language acquisition and the kinds of language processing activities required of language learners when engaged in DDL have also been proposed (Flowerdew, 2015; Jeaco 2017a). For a review of the use of corpora with learners see Yoon (2008) and Kennedy and Miceli (2010). A recent meta-analysis of research into DDL showed that there are positive effects for this approach (Boulton & Cobb, 2017). In a more qualitative review, Chen and Flowerdew (2018) also note that the number of published studies for DDL in EAP is increasing, and one of the main areas of interest is the use of DDL for self-correction. They note that many of these are studies located in Asia and/or involving students from an Asian L1 background. Indeed, corpus-based approaches are also included in Zhang and Cheung's (2018) review of innovations in writing teaching in China, and there seems to be an increasing number of publications exploring the application of DDL in various English-related teaching contexts in China in high school and university contexts (He, 2015; Mao, Liu, & Zhang, 2018). Jin and Lu (2018) have also called for basic corpus literacy skills to become a part of preservice training, and also suggested that "intuitive and teacher-friendly" corpus interfaces are needed (p. 462).

Collocation in Well-Known Corpus Tools

Some proponents of hands-on learning activities with concordancers claim that producing lists of collocations is straightforward. When version 3 of *WordSmith Tools* was current and before *Sketch Engine* or *AntConc* had been developed, Woolard asserted:

concordancers like WordSmith … are not complex and it only takes one short induction lesson to train students to use them for collocation exploration (Woolard, 2000, p. 42).

However, when introducing *AntConc*, Anthony (2004) suggested that some of the more basic information shown in concordancers including collocation tables can be confusing for learners. Even if the learning session is set up so students are expected to act "like a researcher" (Johns, 1988, p. 14), the user of concordancing software needs to be aware of several principles and have made several research design decisions before clicking the button and getting the results. First, they need to decide on which corpus or which sub-corpora to use. Corpus selection is the first screen presented to the user in *Sketch Engine*, and each of the tools in *WordSmith Tools* requires the user to first select the corpus text files to be used for the analysis. In *WordSmith Tools* and *Sketch Engine*, collocations can be calculated after a set of concordance lines has been retrieved. The pathway learners need to follow in order to view collocations is from (1) the selection of texts to (2) the formulation of a search query to (3) the display of concordance lines, and then from the concordance list to (4) the display of a list of collocations. *AntConc* provides a slightly different route, with all the different types of concordancing analysis visible in tabs across the top of the program's window. Learners first have to (1) select the texts, then (2) generate word lists and then (3) create collocation lists. The default collocation settings in *AntConc* are for a minimum frequency of 1, just one word window either side of the node and for matches to be case sensitive. No doubt students from other disciplines who do not have the prerequisite linguistic and software knowledge needed in order to use concordancers to produce collocation lists will find these steps to be quite an obstacle. The list of collocations for these packages, whether the bare-bones list of words and statistical measures, or the grammar-function detail of a word sketch in *Sketch Engine,* all lead towards further questions which the researcher-user of the software needs to consider. It is easy to see why electronic or online dictionaries seem to offer learners a rather more straight-forward query process for the casual user of a system who is looking for answers on how to use a word or phrase. All three software packages cater well for advanced users, with many options available for statistical measures and window size.

In terms of the software design of a concordancer, the choice of statistical measures to make available is an important consideration. One of the earliest and still highly cited measures for collocation is the "association ratio" of Church and Hanks (1990) based on mutual information scores. Interestingly, they note that this was implemented so as to be asymmetrical, so different results would be produced according to the order of the two words being studied. It was proposed as a method to help dictionary writers as "an index to the concordances" (Church & Hanks, 1990, p. 29). However, by the time Oakes (1998) provided a summary of collocation measures, of the dozen measures listed only two were asymmetrical. Gries (2013) presents an overview of the parameters and main statistics for collocation, noting the importance of direction and word order. Other than the "relative frequency" measure as provided in *Sketch Engine*, and DeltaP in *LancsBox,* the collocation measures which are available today in the main corpus packages give symmetrical results and are not sensitive to position.

WordSmith Tools, *Sketch Engine*, *AntConc* and *LancsBox* provide different coverage of statistics based on mutual information, but the summary in Table 1 clearly shows MI, T Score and Log-Likelihood to be the most common.

In terms of the complexity of the presentation of the results, the packages also have some notable differences. With *WordSmith Tools* and *AntConc*, the name of the statistic is hidden on the results page, while with *Sketch Engine,* clickable columns are provided for each statistic chosen. One element contributing to the complexity of *WordSmith Tools* is that it shows counts for different positions (L4, L3, L2, L1, etc.). This information is, of course, very useful to someone familiar with the language and able to do the mental gymnastics of creating phrases in their mind to fit each pattern. Word Sketches in *Sketch Engine* make the grammatical relationship between collocates and the node word accessible, but take demands on mental processing to a higher level. With practice, no doubt advanced *users* can conjure up the appropriate phraseology for "object_of", "subject_of", "modifier" and "modifies", but it is hard to imagine how high intermediate or advanced language learners could. These Sketches are very powerful, but in order to see how these are used in order and position, the user would need to click on the frequencies of each to reveal concordance lines. *LancsBox* provides a different interface and interactivity for exploring collocations. A table of results is visible alongside collocation network graphs. Clicking on different elements in the table or in the collocation network graph will highlight the corresponding item in the other display. As additional searches are made, the collocation network graph adds the network for the new node word to the existing screen (Brezina, et al., 2015). Another notable corpus tool which uses DeltaP is *The Collocation Explorer* (Liang, 2014).

Other kinds of collocation tools that draw on corpora include those that show simplified concordance line data and those that show common patterns (multi-word units, n-grams or collocation lists) rather than concordance lines. An example of a simplified concordance line tool is *SKELL*

Table 1. Collocation measures available in concordancing software

Measure	WordSmith Tools (Scott, 2010)	Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004)	AntConc (Anthony, 2004)	LancsBox[5] (Brezina, McEnery, & Wattam, 2015)
MI	✓	✓	✓	✓
MI3	✓	✓		✓
T Score	✓	✓	✓	✓
Z score	✓			✓
Dice coefficient	✓			✓
Log DICE		✓		✓
Log-likelihood	✓	✓	✓	
Relative freq.		✓		
MI-log-prod		✓		
Minimum sensitivity		✓		

(Kilgarriff, Marcowitz, Smith, & Thomas, 2015). Writing aids based showing multi-word patterns of words based on corpus data include *The Corpus Based Collocation Tutoring System* (Shei & Pain, 2000), *Linggle* (Boisson, Kao, Wu, Yen, & Chang, 2013), *FlaxCLS* (Shaoqun, Liang, Witten, & Yu, 2016) and *AWSuM* (H. H.-J. Chen & Tsai, 2018). With many of these writing aids, raw frequency rather than a statistical measure is used for ranking the results. It seems that they tend to either be based on a single highly specialized corpus or draw from n-grams in broad datasets and therefore results are not divided across different disciplines or contexts.

It can be seen that concordancing packages offer different statistics but they are almost all symmetrical and do not take into account ordering or positioning. Through working repeatedly through the collocation settings and generating multiple pages of results, it is possible through these packages to obtain some word order information. Unlike the learner dictionaries which promote their collocation panels as clearly showing the word partnerships in the order in which they appear, results from the packages introduced above show lists of collocates isolated from the node. On the other hand, writing aids built up from n-grams perhaps focus too heavily on uninterrupted strings of words (n-grams) and do not provide ample information about the source texts or contexts in which these are used. Taken with the points raised regarding collocation in textbooks and dictionaries, five key issues can be summarized as follows:

1. More attention to collocations for specific text types is needed; resources in dictionaries and textbooks for specific academic disciplines and specific domains are quite limited;
2. Ordering and positioning of words in a collocation is important; dictionaries and textbooks tend to present full collocations, but concordancers typically do not;
3. Examples with their contexts of use are needed; dictionaries can only provide a limited number of examples; lists of collocations do not give access to the texts or corpora from which they were derived; writing aids typically provide strings of words and raw frequencies, but do not give attention to genre or domain;
4. Support mechanisms for formulating searches are needed; pathways to find collocations in a suitable corpus can be quite complicated;
5. Sometimes additional elements need to be considered; there is a need to address the tension between a computer-science oriented approach where multi-word units are felt to perform better

with stop lists against the fact that function words are often part of recognizable chunks and that language learners need to see how function words operate as important elements in longer structures;

COLLOCATION IN *THE PRIME MACHINE*

The previous sections have introduced definitions of collocation and the ways collocations are presented in textbooks, dictionaries and well-known corpus tools. The remainder of this paper will describe the ways in which collocations are calculated, stored and presented in *The Prime Machine.* After introducing the motivations for the design of this aspect of the software, the method of calculation and storage in the database will be described. After that, attention turns to the user interface, describing the presentation of collocation data on the different tabs inside the software: the Search Tab, the Collocations Tab, the Lines Tab and the Frequencies Tab.

A number of aspirations related to collocation and multi-word units shaped the design of *The Prime Machine,* and these can be summarized with reference to the five key issues (that were introduced in the previous section) as follows:

- to provide language learners with a means of exploring collocations and multi-word units in a range of different corpora by themselves (1 & 3);
- to help the learner not only see relationships clearly, but also to find and select useful starting points and to avoid unfruitful starting points (4);
- rather than giving learners lists or clouds merely containing isolated collocates, to present full collocations in typical word-order with the aim of improving understanding and retention (2);
- to make the concordance lines central to the user's experience with the software, being a way to help explain the collocation list, and also for strong collocations to offer one way of sorting concordance lines (3);
- since learners are unlikely to read exhaustive lists, to implement a collocation measure which could offer multi-word units of more than 2 words where this could be helpful, but also to show common words around collocations and how these form part of larger units (5);
- to build into *The Prime Machine* the facility to demonstrate to learners the differences between the primings of words and the nesting of those words (3, 4 & 5).

Collocations in the Database

Collocation measures are concerned with the co-occurrence of a word and a candidate collocate and occurrences elsewhere in the corpus. Different measures and methods not only use different statistical tests, but also make different use of stop lists, grammatical information (if available) and different parameters. As others have noted, considerations about window size, the inclusion of exclusion of punctuation and whether to include co-occurrence across sentence breaks are needed (Garretson, 2010; Sinclair, 1991). Collocation measures usually work with the frequencies rather than the number of slots available in the collocation windows. If windows are restricted by sentence boundaries, however, the span will be uneven for many instances where the node is located less than 4 words from the beginning or end of the sentence. That is to say, by limiting the measurement to words within the same sentence, it is obvious that some of the windows will not be a full 8 words in length. In his thesis, Collier (1999) mentions this problem and explains it as one of the reasons for choosing +/- 4 for his concordance line selection. He says that since some of the lines will be -4 but only +3, it will lead to statistical problems. Keeping the window smaller than 5 avoids an ever-increasing number of these problems, but also lessens the measurement of the position of words in the sentence. With a relational database, counting the total number of words actually taking positions in available slots is trivial, although it is time-consuming when repeated for each *type* as a node in a large corpus. The method used in *The Prime Machine* takes the actual sum of the available window slots rather than simply multiplying the node frequency by 8. What this means is that the contingency table can be seen as measuring the relative difference in frequency between a potential collocate in a specific position (or range of positions), balanced against all the other combinations of words in those windows, and its frequency outside the windows. In this way, the frequency of each potential collocate for two word collocations is divided between:

L4/L3/L2 vs. L1 v.s R1 vs. R2/R3/R4

This provides the learner with up to 4 significant collocations for any node and collocate pair:

X .. Y vs X Y vs Y X vs Y .. X

The measurement is effectively asking "Is the proportion of windows where collocate Y occurs in this position significant statistically compared to its occurrence outside the windows". The contingency table for "on the left with a gap" is shown in Table 2.

Collocations are stored in the database's summary table if they meet the minimum threshold of a Bayesian Information Criterion (BIC) score of 0, and the interpretation of these BIC scores is also stored in the table. The use of BIC follows the proposal by Wilson (2013) for how it can be used in conjunction with log-likelihood key word analysis, but here it is applied to collocation. Typically, only those with marked as having "Very Strong Evidence" or "Strong Evidence" appear as collocates in the results screens, but those with a lower strength are also stored so if users look these specific collocations up, they get the other summary information about typical contexts of use. As Jeaco (2017a) explains, information about statistically significant environments for collocations as well as individual items is stored in the database, so the software needs to determine whether or not these data will be available.

For two-word collocations, this approach may seem interesting or perhaps eccentrically novel, but it does have further power when applied to multi-word units. With the system for calculating multi-word units which has been developed, a particular ordering of 3, 4 or 5 words has to compete with all the other possible orderings of those words as well as the occurrence of the words away from the main node. This is rather different from the approach adopted by Danielsson (2007) where initial selection is more open and ordering is only considered after the extraction has been completed. It also contrasts with approaches based on raw frequency only (Wermter & Hahn, 2006). Using the approach proposed here, the result is that in order for a multi-word unit to make it through the "barrier" of significance, it needs to account for a greater proportion of the combinations in a window. Since BIC values are obtained for all of the multi-word units stored in the summary tables, these can be directly compared. In earlier versions of the software, sequences for multi-word collocations of more than 2 words were limited to consecutive words, so for 3 item multi-word units, for example, the slots were L2+L1+node, L1+node+R1 and node + R1 + R2. However, since Version 3, patterns of 3 and 4 word collocations can contain one gap of one or two consecutive slots. Thus, patterns such as L3 + L2 + gap + node and L4 + gap + L2 + L1 + node are also stored in the database.

Table 2. Contingency table for Log-likelihood Collocations for a specific set of slots

	Corpus One	**Corpus Two**
Freq. of word	A = *In slot L4, L3 or L2*	B = *Outside the +/- 4 word window*
TOTAL	C = *Count of all slots in +/- 4 word windows*	D = Whole corpus – C

Collocations on the Search Tab

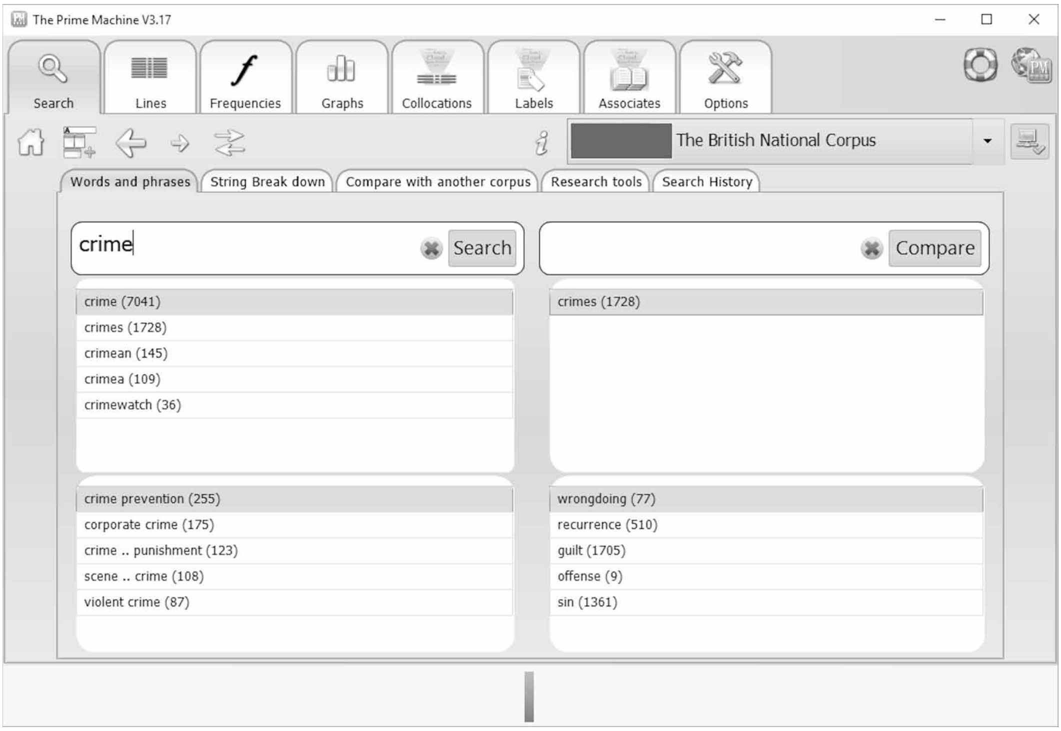
In *The Prime Machine*, since they are stored and indexed in database tables alongside each corpus on the server in advance, short lists of collocations can be retrieved very quickly, allowing text prediction beyond single words to be implemented. Just as Auto-Complete on a word level provides a way of preventing spelling errors, Auto-Complete on the phrase level helps prevent users from making further typos or spelling mistakes and can also provide almost instantaneous feedback on the collocation strength of two or more items. If more than one word has been entered in the search box, when the "Search" button is clicked, the system performs an additional check to ensure that (1) all the words in the box occur together in a 5 word span (i.e. node +/- 4 words) at least once in the corpus, and (2) to check whether the multi-word unit has been stored as being statistically significant using the log-likelihood measure. Within *The Prime Machine*, the following algorithm is used for search strings containing at least one word break:

1. The string is checked to ensure only one occurrence of double period is included. This has a special meaning in *The Prime Machine,* which follows the display of multi-word units in that ".." indicates a required gap between two items in a two word collocation pair of at least one word.
2. The words and symbols contained in the search string are passed to the server, where a check is made for all multi-word units of that length contained in the corpus in any order, with or without gaps. A search is also performed to see whether there is at least one occurrence of the words co-occurring in a 5 word window in any order, whether they occur in order with or without gaps, and whether they occur in order with no gaps.
3. The results are then checked to see whether the original string is included in the list. If so, the search is permitted and concordance lines, collocations (with extensions) and all the other data will be retrieved for the multi-word unit. If the original string is not included in the list, the user will be presented with a list of phrases containing the same words but in different orders, and information about whether they occur at least once in each of the three raw window searches.

This lookup procedure provides a way to give very quick feedback on whether or not words collocate and whether there are any instances of the phrase in the corpus at all. The results of a study by Römer (2009) into other software for language learning and teaching include a suggestion from a teacher that it would be helpful to have very quick feedback on whether words collocate or not. Lewis (2000) argues that the development of collocation knowledge includes greater awareness of words which should not be used together as well as those which should. While lack of a collocation relationship is not something visible on any of the results tabs in *The Prime Machine*, the immediate feedback goes some way to meet these needs and should help learners see if a phrase they are considering using may not be appropriate. The look-up phase also has a gate-keeping role, preventing the fruitless waiting period which would occur if users requested a phrase which simply did not occur. The drop-down boxes for collocations appear as soon as the word-level Auto-Complete routine encounters a string of letters which is listed in the lexicon as a complete word. The top collocations are then retrieved for this word as the node and ordered by log-likelihood. In this way, three, four or five word MWUs will be included in the drop-down list if their statistical significance rather than their raw frequency places them higher in the ranking. As with the lists of individual words, Auto-Complete suggestions include the frequency of the collocations in brackets, providing instant information about the number of instances available. Figure 1 shows an example of these suggestion boxes.

Language learners using concordancing software may not be aware of the best corpus to use for their search; since collocations are often different in different disciplines and registers, it could be the collocation a student wants to examine is available in a different corpus. *The Prime Machine* also allows users to check whether a multiword search term is stored as a collocation in any of the other

Figure 1. The Search Tab suggestion boxes



crime

crime (7041)
crimes (1728)
crimean (145)
crimea (109)
crimewatch (36)

crimes (1728)

crime prevention (255)
corporate crime (175)
crime .. punishment (123)
scene .. crime (108)
violent crime (87)

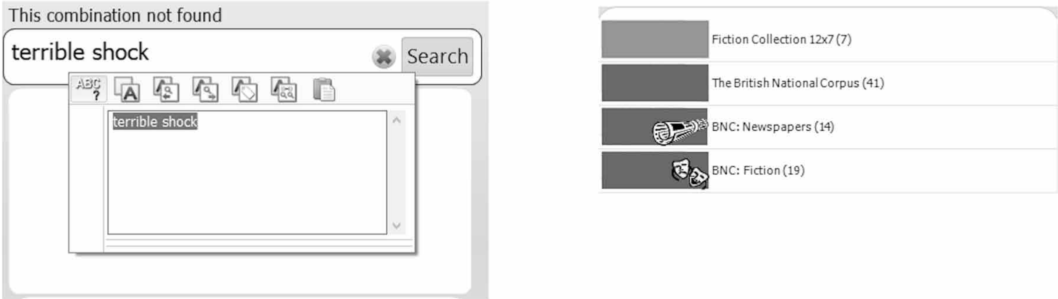
wrongdoing (77)
recurrence (510)
guilt (1705)
offense (9)
sin (1361)

corpora available on the system. This is done by right-clicking on the search box and choosing the "ABC" icon which can be seen in Figure 2.

The Collocation Tab

When corpus data for a word or collocation is retrieved, several different kinds of collocation data can be viewed on the Collocation Tab in two different ways. The default view is to show collocation clouds based on the log-likelihood measure, but tables of results can also be shown. The intention for the clouds is not to provide an exhaustive list, but to help draw attention to some of the strongest collocates. For all the collocation clouds in *The Prime Machine*, the size of the font used for each item is based on the statistic rather than the raw frequency. In order to reduce the magnitude of differences between the statistics when they are displayed in the cloud, cube-roots are used[5] and a multiplier for

Figure 2. Checking other corpora after "terrible shock" has not been found in the BNC: Academic sub-corpus



This combination not found

terrible shock

terrible shock

Fiction Collection 12x7 (7)
The British National Corpus (41)
BNC: Newspapers (14)
BNC: Fiction (19)

each cloud is calculated by placing the highest ranked item first, and then determining the scaling needed to transform the cube-root of its measure of collocation strength into the desired font size. On the Collocations Tab in *The Prime Machine,* collocation clouds and tables are available for several other statistics (T-Score, Dice and MI3) just as they are for Log-likelihood and DeltaP collocations. However, there is an important difference. Since the Log-likelihood and DeltaP collocations are based on specific ordering and proximity of the collocates, it is possible to present each as a complete collocation rather than isolated words. In this way, the items in the cloud should provide a stronger impression and provide learners with the opportunity to experience the phenomena introduced in one of Firth's memorable assertions:

A word in a usual collocation stares you in the face just as it is. (Firth, [1951]1957, p. 182).

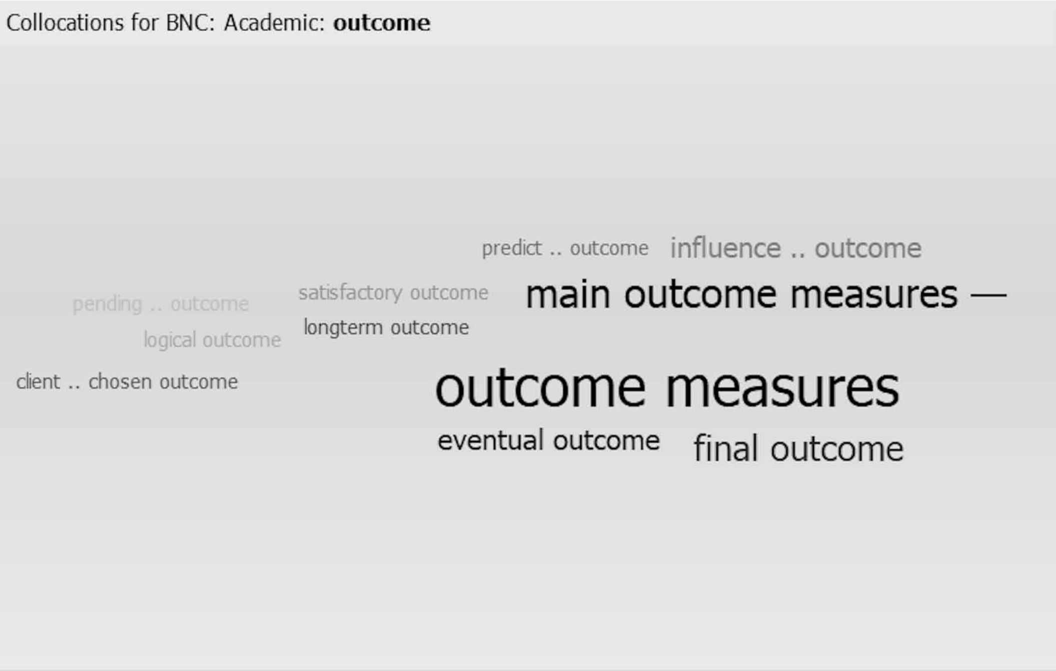
The point is that learners may need to see the words together for these visual representations of the collocations to have an impact. The design is based on the proposition that if only the usual collocation word clouds were presented, the same information may be retrievable, but the user would need to be thinking about the node word and formulating a plausible ordering or grammatical relationship for each link. However, if the node is plainly visible in each element in the cloud, it makes the cloud rather "thicker" but ensures the whole relationship can be seen. Figure 3 and Figure 4 show the cloud and table for the node *outcome* in the BNC: Academic sub-corpus, and Figure 5, Figure 6 and Figure 7 show clouds for this node in several different corpora.

The log-likelihood collocations are also used for a number of other purposes in the software. One of the benefits of looking at collocation lists in a concordancer rather than as a separate resource is that the user can explore the actual concordance lines which were the basis of evidence for the relationship. Other concordancing software packages also try to make links between collocations and concordance lines clear, and in *WordSmithTools* and *Sketch Engine*, as explained earlier, the user first requests concordance lines and then moves on to generate lists of collocates. *Sketch Engine* provides links marked with "+" and "-" so concordance lines can be displayed showing positive or negative evidence for the relationship. In *AntConc*, the list of collocates appear like hyperlinks and clicking on them takes the user to a list of concordance lines containing each one. Generating concordance lines for collocates in *The Prime Machine* would entail right-clicking on the desired collocation and then selecting the menu item to use this as a search term. The context menu is provided for any text on the screen whether it is an item in a cloud, a cell in a table or a line on a card. Since not all the concordance lines for a query are usually downloaded and stored on the user's computer, getting concordance lines for collocations would require a further look-up process. Therefore, immediately jumping to the relevant concordance lines is not possible, but the context menu is consistent across the application and presents simple buttons to copy the text to the operating system clipboard, use the text as a main query, use the text as a query for comparison or to use the text in other kinds of search. The right-most button copies the text to the compare corpus screen on the "Search Tab".

Collocations on the Lines Tab

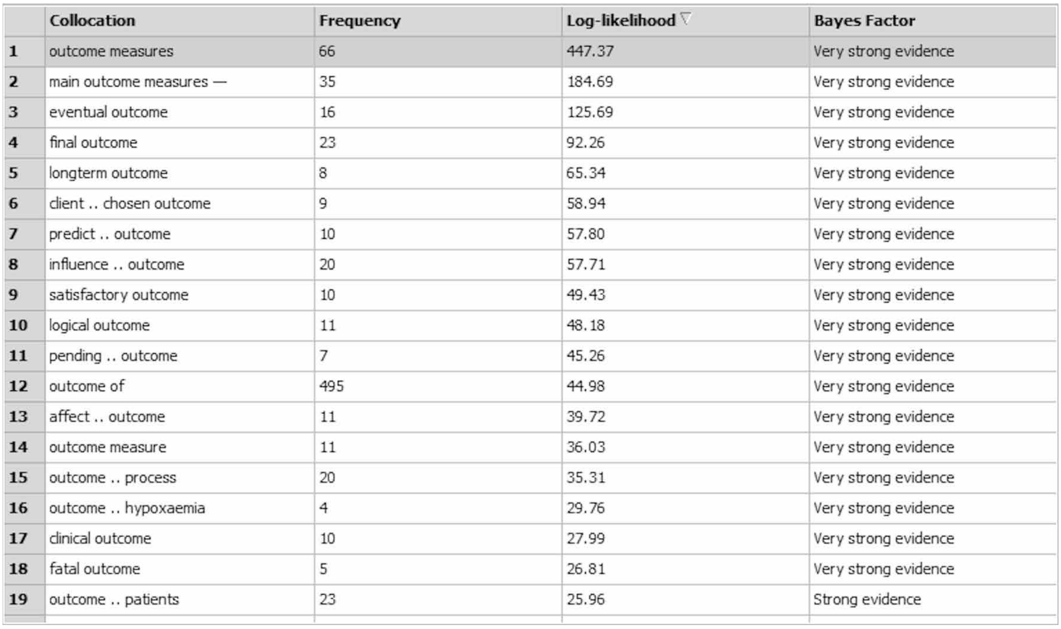
Two other important features of *The Prime Machine* in relation to collocation include the Cards display and the default way in which concordance lines are sorted. As explained in Jeaco (2017b) concordance lines can be viewed as Cards and on these cards collocations in a 4 word window to the left and right of the node are displayed prominently in a caption at the top. The Card view provides a wider context (one sentence before and after the sentence containing the node) and also provides information about the source of each concordance line. When the sample of concordance lines for a search has been retrieved from the server, the default way in which the concordance lines are sorted is also based on collocation information. This provides a means of helping users spot patterns in the concordance lines without needing to sort and resort alphabetically by different columns[7].

Figure 3. Log-likelihood Collocation Cloud in the BNC: Academic sub-corpus for the node outcome



Collocations for BNC: Academic: **outcome**

Figure 4. Log-likelihood Collocation Table in the BNC: Academic sub-corpus for the node outcome



	Collocation	Frequency	Log-likelihood ▽	Bayes Factor
1	outcome measures	66	447.37	Very strong evidence
2	main outcome measures —	35	184.69	Very strong evidence
3	eventual outcome	16	125.69	Very strong evidence
4	final outcome	23	92.26	Very strong evidence
5	longterm outcome	8	65.34	Very strong evidence
6	client .. chosen outcome	9	58.94	Very strong evidence
7	predict .. outcome	10	57.80	Very strong evidence
8	influence .. outcome	20	57.71	Very strong evidence
9	satisfactory outcome	10	49.43	Very strong evidence
10	logical outcome	11	48.18	Very strong evidence
11	pending .. outcome	7	45.26	Very strong evidence
12	outcome of	495	44.98	Very strong evidence
13	affect .. outcome	11	39.72	Very strong evidence
14	outcome measure	11	36.03	Very strong evidence
15	outcome .. process	20	35.31	Very strong evidence
16	outcome .. hypoxaemia	4	29.76	Very strong evidence
17	clinical outcome	10	27.99	Very strong evidence
18	fatal outcome	5	26.81	Very strong evidence
19	outcome .. patients	23	25.96	Strong evidence

Figure 5. Log-likelihood Collocation Clouds and Tables in the Hindawi Biological Sciences corpus for the node outcome

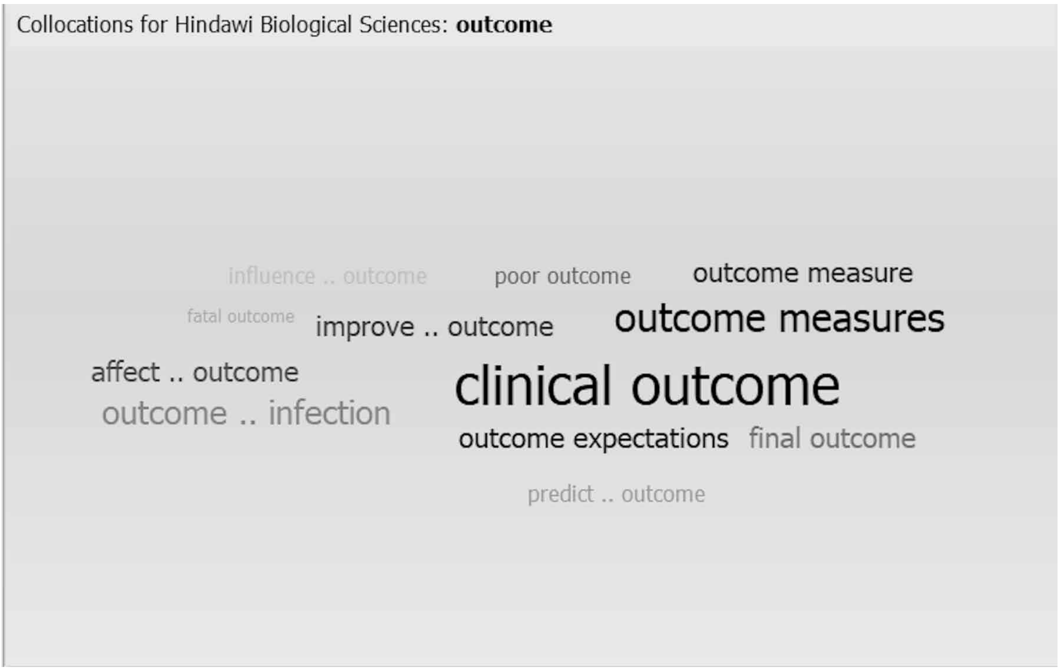


Figure 6. Log-likelihood Collocation Clouds and Tables in the Hindawi Mathematics corpus for the node outcome

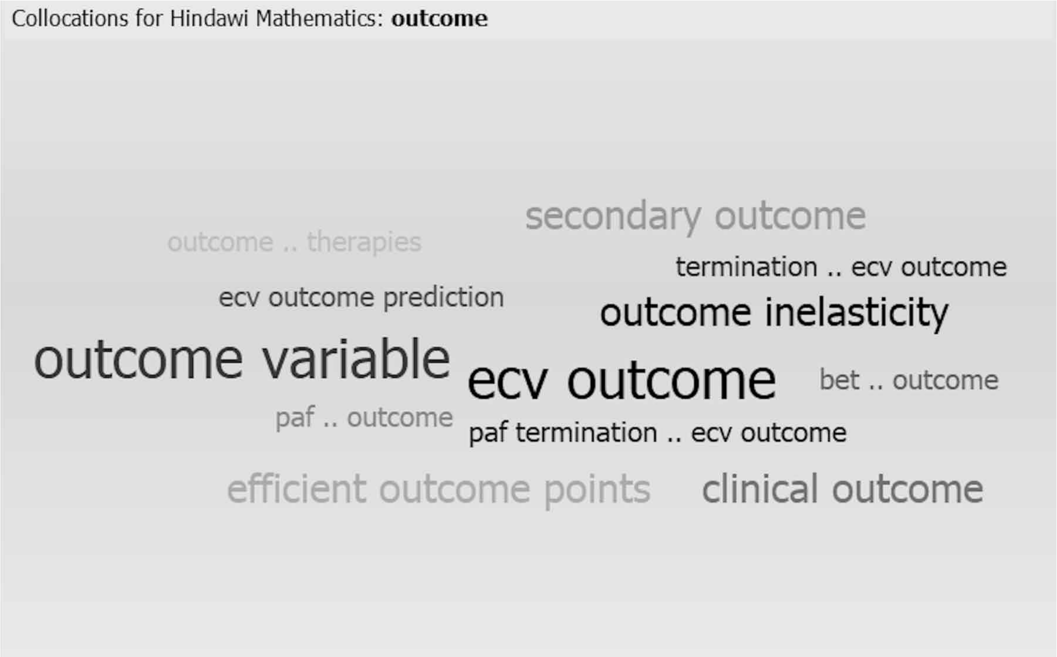
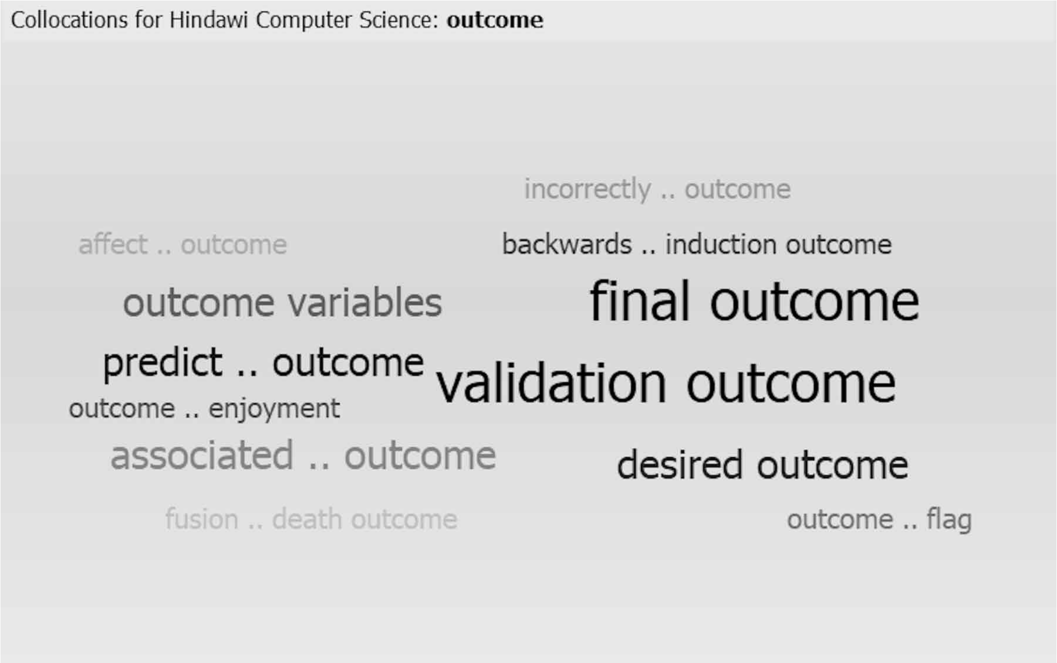


Figure 7. Log-likelihood Collocation Clouds and Tables in the Hindawi Computer Science corpus for the node outcome



Collocations for Hindawi Computer Science: **outcome**

Collocations on the Frequencies Tab

Frequency information for collocations can also be seen on the Frequencies Tab. As well as seeing a graph or table for the raw and normalized frequency across the whole corpus, many of the corpora in *The Prime Machine* have a number of fixed major categories, based on metadata in each corpus, and allow the user to see frequencies across different sections of the corpus. Collocations can also be split into the component parts on this tab to show the distribution of each word in the collocation as well as the distribution of the complete collocation. Table 3 shows a breakdown of a collocation into its components across different parts of the BNC: Academic sub-corpus, along with the overall proportion of words from different categories across the whole corpus.

CONCLUSION

This paper has introduced some of the ways collocation is described and presented in language teaching resources and in several corpus tools and writing aids. It has explained the rationale behind the treatment of collocations in *The Prime Machine*. It has explained how this tool was developed to address some important pedagogical principles and to provide a way for English language learners and teachers to explore collocations and a number of other patterns in language use, taking into account differences across genres, registers and disciplines. Ultimately, it is hoped that others will find that *The Prime Machine* offers a way for language learners to work with a high degree of independence to look up and explore language as a combined reference tool and language learning platform. Exploring concordance lines and corpus data takes some time and effort, but it is hoped that this tool goes some way towards making these kinds of data more accessible and useful for students. While other writing aids and mobile phone dictionaries will provide quick patterns of broad language use, a key advantage of using a concordancer is that examples (and extended concordance lines) can be viewed to see how collocations are used and in what contexts.

Table 3. Results from the Frequencies Tab, showing proportions of hits from different categories in the BNC: Academic sub-corpus for the collocation pilot study

Category	pilot study	pilot	study	Whole corpus
Humanities and Arts	0%	3%	8%	21%
Medicine	31%	23%	34%	9%
Natural Science	0%	2%	4%	7%
Politics, Law and Education	2%	15%	10%	29%
Social Science	66%	54%	43%	30%
Technology and Engineering	0%	3%	2%	3%

The wider community is invited to evaluate this tool through research and teaching practice to consider to what extent concordancing with this tool compares with the use of simpler reference information provided by the writing aids described earlier. Within the author's own teaching context, the author has found that collocation activities are a very good way in for English majors to appreciate the power and usefulness of hands-on concordancing. Some classic (and effective) collocation activities which were initially designed for use with a collocation dictionary can also be used with this new concordancing tool. Examples include activities for brainstorming, comparing synonyms and correcting errors. The collocation features of *The Prime Machine* have been favourably viewed by language learners and teachers in earlier evaluations (Jeaco 2017a, 2017b, 2017c), and in feedback from recent workshops with language teachers and MA TESOL students. However, a number of difficulties still exist. One issue is that some students tend to look words up in the word's base form, as they would in a dictionary. While different word-forms can be combined in searches in *The Prime Machine* by using the Multiple Searches tool, its collocations are based on types (exact strings of letters) rather than lemma. Another issue is that some students try to search for collocations which do not exist in the corpus, and find it hard to gain confidence in the system when their own suggestions deviate so markedly from those in the target variety. A third issue is that when starting with the software, students tend to underestimate the amount of time and effort needed for successful query formulation and for them to gain an understanding the concordance results so they can use them productively. Further research into these issues and useful ways of avoiding them is needed.

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*(4), 235–247. doi:10.1016/j.jeap.2013.08.002

Alexander, O., Argent, S., & Spencer, J. (2008). *EAP Essentials: A Teacher's Guide to Principles and Practice*. Reading: Garnet.

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.

Barnbrook, G., Mason, O., & Krishnamurthy, R. (2013). Collocation [electronic book]; applications and implications: Basingstoke, UK: Palgrave Macmillan.

Bernardini, S. (2004). Corpora in the classroom: An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 15–36). Amsterdam: John Benjamins. doi:10.1075/scl.12.05ber

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Boisson, J., Kao, T.-H., Wu, J.-C., Yen, T.-H., & Chang, J. S. (2013, August 4-9). Linggle: a Web-scale Linguistic Search Engine for Words in Context. *Paper presented at the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria.

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., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. doi:10.1075/ijcl.20.2.01bre

Brown, K., & Hood, S. (2002). *Academic Encounters: Reading, Study Skills, and Writing*. Cambridge: Cambridge University Press.

Chen, H. H.-J., & Tsai, N. Y. (2018). Using a Large Social Science Corpus to Build an Automatic Writing Suggestion System. *Paper presented at the Fourth Asia Pacific Corpus Linguistics Conference*, Takamatsu, Japan.

Chen, M., & Flowerdew, J. (2018). A critical review of research and practice in data-driven learning (DDL) in the academic writing classroom. *International Journal of Corpus Linguistics*, *23*(3), 335–369. doi:10.1075/ijcl.16130.che

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Collier, A. (1999). *The Automatic Selection of Concordance Lines*. Unpublished Ph.D. dissertation, University of Liverpool.

Collins COBUILD Advanced Dictionary of English. (2009). Glasgow: HarperCollins.

Conzett, J. (2000). Integrating collocation into a reading and writing course. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 70–87). Hove: Language Teaching Publications.

Cotton, D., Falvey, D., & Kent, S. (2006). Market Leader Upper Intermediate Course Book (New ed.). Harlow: Longman.

Cox, K., & Hill, D. (2011). *EAP Now!: English for Academic Purposes* (2nd ed.). London: Pearson Longman.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. doi:10.2307/3587951

Cunningham, S., & Moor, P. (1999). *Cutting Edge*. Harlow: Longman.

Danielsson, P. (2007). What constitutes a unit of analysis in language? *Linguistik Online, 31*(2), 18.

de Chazal, E., & McCarter, S. (2012). *Oxford EAP: Upper-Intermediate / B2: A Course in English for Academic Purposes*. Oxford: Oxford University Press.

de Chazal, E., & Moore, J. (2013). *Oxford EAP: A Course in English for Academic Purposes: Advanced C1*. Oxford: Oxford University Press.

de Chazal, E., & Rogers, L. (2014). *Oxford EAP Pre-Intermediate/B1 Teacher's Book*. Oxford: Oxford University Press.

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. doi:10.1016/j.esp.2009.02.002

Firth, J. R. (1957). A synopsis of lingistic theory, 1930-1955. In F. R. Palmer (Ed.), *Selected Papers of J R Firth 1952-59* (pp. 168-205). London: Longman. (originally printed in 1951)

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. doi:10.1075/scl.69.02flo

Garretson, G. (2010). *Corpus-Derived Profiles: A Framework for Studying Word Meaning in Text*. Unpublished Ph.D. dissertation, Boston University.

Gries, S. T. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–165. doi:10.1075/ijcl.18.1.09gri

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press. doi:10.7551/mitpress/9780262018579.001.0001

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. doi:10.1057/9781137440037_6

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 47–69). Hove: Language Teaching Publications.

Hill, J., Lewis, M., & Lewis, M. (2000). Classroom strategies, activities and exercises. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 88–117). Hove: Language Teaching Publications.

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

Hoey, M. (2014). Words and their neighbours. In J. R. Taylor (Ed.), *Oxford Handbook of the Word*. Oxford: Oxford University Press. doi:10.1093/oxfordhb/9780199641604.013.39

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524773

Hunston, S., & Francis, G. (2000). *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins. doi:10.1075/scl.4

IELTS. (n.d.). Researchers - Band descriptors, reporting and interpretation. Retrieved from http://www.ielts.org/researchers/score_processing_and_reporting.aspx

Jakeman, V., & McDowell, C. (2008). *New Insights into IELTS*. Cambridge: Cambridge University Press.

Jeaco, S. (2015). *The Prime Machine: a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language*. Unpublished Ph.D. dissertation. University of Liverpool. Retrieved from https://livrepository.liverpool.ac.uk/2014579/

Jeaco, S. (2017a). Concordancing Lexical Primings. In M. Pace-Sigge & K. J. Patterson (Eds.), *Lexical Priming: Applications and Advances* (pp. 273–296). Amsterdam: John Benjamins. doi:10.1075/scl.79.11jea

Jeaco, S. (2017b). Helping Language Learners Put Concordance Data in Context: Concordance Cards in The Prime Machine. *International Journal of Computer-Assisted Language Learning and Teaching*, *7*(2), 22–39. doi:10.4018/IJCALLT.2017040102

Jeaco, S. (2017c). Helping Language Learners Get Started with Concordancing. *TESOL International Journal*, *12*(1), 91–110.

Jin, T., & Lu, X. (2018). A Data-Driven Approach to Text Adaptation in Teaching Material Preparation: Design, Implementation, and Teacher Professional Development. *TESOL Quarterly*, *52*(2), 457–467. doi:10.1002/tesq.434

Johns, T. (1986). Micro-concord: A language learner's research tool. *System*, *14*(2), 151–162. doi:10.1016/0346-251X(86)90004-7

Johns, T. (1988). Whence and whither classroom concordancing? In T. Bongaerts (Ed.), *Computer Applications in Language Learning* (pp. 9–27). Dordrecht: Foris.

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. doi:10.1163/9789004334236_010

Kang, B.-M. (2018). Collocation and Word Association: Comparing Collocation Measuring Methods. *International Journal of Corpus Linguistics*, *23*(1), 85–113. doi:10.1075/ijcl.15116.kan

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28–44.

Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and Language Learning with the Sketch Engine and SKELL. *Revue Française de Linguistique Appliquée*, (1): 61–80.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. *Paper presented at the 2003 International Conference on Natural Language Processing and Knowledge Engineering*, Beijing.

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 10–27). Hove: Language Teaching Publications.

Lewis, M. (2008). *The Lexical Approach: The State of ELT and a Way Forward.* Andover: Heinle, Cengage Learning.

Liang, M. (2014). Exploring collocations with the Collocation Explorer. *Paper presented at the Second Asia Pacific Corpus Linguistics Conference*, The Hong Kong Polytechnic University, Hong Kong.

Longman Dictionary of Contemporary English (5th ed.). (2009). Harlow: Pearson.

Macmillan English Dictionary for Advanced Learners (New ed.). (2007). Oxford: Macmillan.

Mao, L., Liu, Y., & Zhang, M. (2018). Research on the Effectiveness of College Student English Writing Teaching Based on Data-Driven Learning. *Educational Sciences: Theory and Practice*, *18*(5), 1160–1169. doi:10.12738/estp.2018.5.017

Mascull, B., & Heitler, D. (2006). Market Leader Upper Intermediate Teacher's Book (New ed.). Harlow: Longman.

McCarthy, M. (2004). From Corpus to Course Book. Retrieved from http://www.cambridge.org/us/esl/touchstone/images/pdf/CorpusBooklet.pdf

McCarthy, M., McCarten, J., & Sandiford, H. (2006a). *Touchstone 3 Student's Book*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2006b). *Touchstone 4 Student's Book*. Cambridge: Cambridge University Press.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical Phrases and Language Teaching*. Oxford: Oxford University Press.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511497650

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

Phillips, M. A. (1985). *Aspects of Text Structure: An Investigation of the Lexical Organisation of Text*. Amsterdam: North-Holland.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge University Press. doi:10.1017/CBO9780511667305

Römer, U. (2009). Corpus reseach and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 83–98). Amsterdam: John Benjamins. doi:10.1075/scl.33.09rom

Rundell, M. (n.d.). How it was Created. *Macmillan English Dictionary*. Retrieved from http://www.macmillandictionaries.com/features/how-dictionaries-are-written/med/

Scott, M. (2010). *WordSmith Tools (Version 5.0)*. Oxford: Oxford University Press.

Shaoqun, W., Liang, L., Witten, I. H., & Yu, A. (2016). Constructing a Collocation Learning System from the Wikipedia Corpus. *International Journal of Computer-Assisted Language Learning and Teaching*, *6*(3), 18–35. doi:10.4018/IJCALLT.2016070102

Shei, C. C., & Pain, H. (2000). An ESL Writer's Collocational Aid. *Computer Assisted Language Learning*, *13*(2), 167–182. doi:10.1076/0958-8221(200004)13:2;1-D;FT167

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

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

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

Tsui, A. B. M. (2004). What teachers have always wanted to know - and how corpora can help. In J. M. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 39–61). Amsterdam: John Benjamins. doi:10.1075/scl.12.06tsu

Wattenberg, M., & Viégas, F. B. (2008). The Word Tree, an interactive visual concordance. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6), 1221–1228. doi:10.1109/TVCG.2008.172 PMID:18988967

Wermter, J., & Hahn, U. (2006). You can't beat frequency (unless you use linguistic knowledge): A qualitative evaluation of association measures for collocation and term extraction. *Paper presented at the Annual Meeting of the Association for Computational Linguistics*, Sydney. doi:10.3115/1220175.1220274

Wilson, A. (2013). Embracing Bayes Factors for key item analysis in corpus linguistics. In M. Bieswanger & A. Koll-Stobbe (Eds.), *New Approaches to the Study of Linguistic Variability* (pp. 3–12). Frankfurt: Peter Lang.

Woolard, G. (2000). Collocation - encouraging learner independence. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 28–46). Hove: Language Teaching Publications.

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology*, *12*(2), 31–48.

Zhang, W., & Cheung, Y. L. (2018). Researching Innovations in English Language Writing Instruction: A State-of-the-art Review. *Journal of Language Teaching & Research*, *9*(1), 80. doi:10.17507/jltr.0901.10

ENDNOTES

[1] Richards and Rodgers (2001) define a lexical approach as: "one derived from the belief that the building blocks of language learning and communication are not grammar, functions, notions, or some other unit of planning and teaching but lexis, that is, words and word combinations" (p. 132).

[2] For details of how to access *The Prime Machine* see www.theprimemachine.com.

[3] C.f. Hoey (2014) on collocation operating over greater spans, and C. f. Kang (2018) on paragraph collocations.

[4] While these public descriptors are located in the "Research" part of the website, direct hyperlinks are placed in the "Information for Candidates" pages, effectively embedding them in the student-oriented section too.

[5] *LancsBox* has several other collocation measures including DeltaP. Notes here are version 4.0, accessed 21/01/2019 from http://corpora.lancs.ac.uk/lancsbox/

[6] Unlike the log-likelihood measure, the cube-root is linear and so performing this operation is simply a pragmatic approach to mapping a wide range of values to the range of sizes of text which are legible to the human eye. However, other approaches are also possible and, for example, a metric based on the square-root of the frequency was used for text size in *the Word Tree* (Wattenberg & Viégas, 2008).

[7] Sorting alphabetically in this way is also possible and other sorting methods are also provided in the software.

Stephen Jeaco is an Associate Professor at Xi'an Jiaotong Liverpool University. He has worked in China since 1999 in the fields of EAP, linguistics and TESOL. His PhD was supervised by Professor Michael Hoey and focused on developing a user-friendly corpus tool based on the theory of Lexical Priming.