

Concordancing Lexical Primings

The rationale and design of a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language

Stephen Jeaco

Xi'an Jiaotong-Liverpool University, China; University of Liverpool

Abstract

This author accepted manuscript has been made available for researchers on S. Jeaco's [personal website](#) and should not be redistributed.

The published Version of Record is:

Jeaco, Stephen. 2017. Concordancing lexical primings: The rationale and design of a user-friendly corpus tool for English language teaching and self-tutoring based on the Lexical Priming theory of language. In: *Lexical Priming: Applications and advances*, Pace-Sigge, Michael & Katie J. Patterson (eds.) [Studies in Corpus Linguistics, 79] pp. 273–296.

This material is copyright. © John Benjamins 2017.

<https://benjamins.com/catalog/scl.79.11jea>

Lexical Priming (Hoey, 2005) brings together a range of linguistic patterns that should be an important focus of language learning and teaching. But it also adds an additional load to learners and teachers, demanding attention to different primings of words and nested combinations of words. With many tendencies difficult to observe in dictionary entries or other concordancing software, learners and teachers will face difficulties finding and presenting information about these primings. This chapter introduces the design of a concordancer created for a doctoral degree project and developed to be firmly based on the theory of Lexical Priming. It introduces the pedagogical rationale for the development of some key features, including the search screen interface and the display of concordance lines.

Key words

Lexical Priming, language learning, Data Driven Learning, key word analysis

1. Pedagogical assumptions

A prevalent view of how language operates has been that grammar and vocabulary are separate systems and sentences can be constructed merely by choosing any syntactic structure and slotting in vocabulary. This view is still prevalent in many areas of language teaching. It is evident in China (where the author lives and works) in materials designed to familiarize students with grammatical constructions following sets of rules, and in the wordlists of vocabulary which are frequently used to introduce isolated meanings of individual words, usually with just one or two word by word translations provided. Over the last few decades, corpus linguistics has presented challenges to this view of language, and by drawing on evidence which can be found in the patterning of language choices in texts, it provides both a means of narrowing down the range of items to be taught through an emphasis on the most frequent usage, and also a raising of the bar in the sense of demanding attention be paid to relationships between items in terms of collocation and colligation. From Firth (1957) through to Sinclair (1991), and in a wide variety of corpus linguistic research as well as Systemic Functional Grammar, the necessity for language educators to move away from a belief in a grammar separated from the lexicon is plainly evident. The theory of Lexical Priming (Hoey, 2005) makes a valuable contribution to linguistic theory by building on a range of insights gained from corpus linguistics and establishing a framework and evidence for the existence of other relationships which account for a sense of the naturalness or creativity of produced language. Hoey introduces the theory by providing a cognitive explanation for why collocation is so pervasive and it is clear that the other claims which Lexical Priming makes also challenge prevailing notions of how words and collocations can be used. The concept of textual colligation challenges the idea that words can be used freely

in different positions in the sentence, paragraph or text for example. The concepts of semantic association and pragmatic association challenge the idea that words can be freely slotted into sentence structures purely based on some sort of inherent or isolated meaning. While other theories of language arising from corpus linguistics are clearly aiming to enhance a description of language and to thereby drive developments for language teaching, they are for the most part not particularly concerned with describing a model for the acquisition of language or the processes that underpin language learning.¹ Lexical Priming, however, fills this gap by using insights from corpus linguistics and corpus data as evidence to explain how individuals are primed through exposure and use of language, and explaining how this priming process is the basis for first and second language acquisition. From a pedagogical perspective, the theory could also be used as a powerful metaphor for explaining to adult learners why their understanding of language may need to be adjusted and how they might go about exploring wider relationships between words, context and meaning. Given some of the entrenched views about language which are often held by students regarding vocabulary learning strategies, it would be unrealistic to expect a piece of software to be able to completely shake and remodel their view of language and language learning priorities. Nevertheless, if a simple image of the human brain encountering words and phrases through hearing, reading and production and thereby building up patterns and expectations for how these could and should be used is presented, it could provide an impetus for encouraging language learners to look more deeply at the contexts of the language they encounter and the language that they produce. The idea that traces in the human mind of language which has previously been encountered are similar to concordance lines is a potent analogy, and also promotes a balanced understanding of how corpus resources can be used to find evidence but cannot ever represent the true priming of any one individual. In this sense, students can be

assured of the relevance of corpus data as a way of gaining insights into real language use, while they are also encouraged to be critical and mindful of any resource's limitations.

Although corpora have had an indirect influence on language teaching through the creation of dictionaries and materials which draw on corpus data, the main pedagogic implementation of corpus linguistics is Data Driven Learning (DDL). Johns (2002) listed several advantages of DDL over other types of learning materials, including new ways of approaching problem areas such as prepositions with a main focus on meaning and also helping teachers and learners prioritize what should be learned. Bernardini (2004) argues that concordancing tasks can be used as a means of meeting a variety of language teaching goals. There are several reasons highlighted in the literature that explain why direct use of concordancing software can be especially useful for learners. First, as Sinclair (1991) pointed out, if learners want to learn about common patterns of syntax associated with a particular word, dictionaries do not usually provide this. Secondly, as well as providing more information in an accessible way, it has been argued that concordancers give the learner an "ideal" space to test hypotheses (Kettemann, 1995; cited in Meyer, 2002). Studies have shown that teaching learners to use concordancers and then explore aspects of syntax by themselves can reduce their anxiety, and it has been suggested that this is because they can be freed from a sense of being subject to human judgement (Hunston, 2002). As well as providing the opportunity for learning about language use at the time concordancers are consulted, another advantage of teaching learners to use corpora is that it is a skill which can form part of their life-long learning (Mills, 1994). The procedures learners follow when they systematically perform searches and analyse corpus output help develop disciplines for self-access (Kennedy, 1998). As Thomas explains, "teachers need to be aware of how much studying, learning and acquiring are taking place simultaneously when learners are engaged in corpus-based guided discovery tasks" (Thomas, 2015, p. 17).

Two of the primary aims of using concordancers with language learners are likely to be based on Second Language Acquisition principles: that learners should be exposed to target language in use (see, amongst others, Krashen, 1989); and that “intake is what learners consciously notice” (Schmidt, 1990, p. 149). Tomlinson argues that an important objective in language learning should be for learners to discover for themselves language features which can be found in the authentic texts they encounter, so as to strengthen the positive effects of noticing and recognising a gap in their own language use (Bolitho et al., 2003; Tomlinson, 1994, 2008). When used in language learning contexts, concordancing software leads language learners to read multiple examples from authentic texts, and the potential for concordancers to promote active discovery of patterns is clear.

However, despite some success, only a limited number of teachers and learners of second language seem to make regular use of these tools. Factors which may be holding teachers back from learning to use and teach corpus tools include issues with the context, the level of detail, the means of interpretation, and the time required to get results as well as the design of the software itself. Traditional Key Word in Context (KWIC) concordance output is almost completely cut away from its context (Hunston, 2002). Also, the amount of detail that concordances can provide to a learner can be confusing (Kennedy, 1998). However, Varley (2009) reports some success for students if they can cope with the “overwhelming” amount of corpus data. Another point is that beyond dealing with the amount of raw data, the skills required to actually interpret them in order to understand grammatical patterns are far from simple (Gaskell & Cobb, 2004). Effort is still needed to strive to make concordancers more user-friendly and more suitable for language learners (Horst, Cobb, & Nicolae, 2005; Krishnamurthy & Kosem, 2007).

The motivation for the development of a concordancer for Lexical Priming was twofold. As well as being deeply rooted in an appreciation of some of the struggles and difficulties faced

by English teachers and language teacher managers in terms of helping students in China (and, by extension, any other cohort of L2 language learners) appreciate their language needs and develop their language skills accordingly, the project was also designed to enable teachers and students to explore various features of the theory of Lexical Priming without needing to teach the theory explicitly. It would not be desirable to replace the wordlists and sets of grammar rules that students and teachers may currently use with a complicated exposition of Lexical Priming with all the technical and linguistic background knowledge which that would require. The software is designed, however, to encourage exploration of some of its features and to make it possible to see tendencies of words and phrases which are not usually apparent in either dictionary examples or the output from other concordancing software. The software aims to make insights about the English language based on Lexical Priming accessible and rewarding, by providing a multitude of examples from corpus texts and additional information about the contextual environments in which words and combinations of words tend to occur. While inspiration and methodological approaches have been drawn from other concordancing software, the design of each aspect of the new concordancer, called *The Prime Machine*, has focused first and foremost on how the most basic building blocks of the data structures and the user interface can support pedagogical priorities. The project is also in line with suggestions from two reviewers of Hoey's book on Lexical Priming: Garretson (2007) suggested that technology could provide ways to make analysis of Lexical Priming less time-consuming; Kaszubski highlighted the scope for the theory in the design of "learner concordancing practices" (2007, p. 292).

The rest of this chapter will introduce some of the ways in which design features of *The Prime Machine* were inspired and driven by concepts from the theory of Lexical Priming. A fuller description of the technical procedures; a fuller discussion of some of the background issues; and further examples and evidence are presented in the doctoral thesis (Jeaco, 2015).

What follows here is a list of 3 claims about the software design, with a brief description of some of its features related to: the search query screen; the features of the concordance line displays; and the ways the user is encouraged to interact with the data.

2. Claim 1: The design should help language learners explore differences between words and phrases

When designing the screen that language learners will use to formulate queries in a concordancer, it is important to consider what the main reasons might be for them to perform searches. Looking through the literature on DDL and studies which have evaluated corpus tools with language learners, there seems to be a consensus that comparisons of synonyms, as well as prompts to explore other word forms, would be particularly helpful. In one of the earliest papers on DDL, Johns (1991) explained that students often come to concordancers wanting to compare pairs of words. Corpora are thought to help demonstrate differences between synonyms clearly (Kaltenböck & Mehlmauer-Larcher, 2005). All of the suggested activities given by Coniam (1997) for how corpora could be used in teaching require learners to compare. Three out of the six uses of corpora in the classroom given by Tsui (2004) involve different aspects of synonymy: near synonyms, words which are very close in meaning, and words which have the same translation in the learner's own language. However, student feedback from some studies has also shown that while it can be rewarding, learners find the discovery of differences between synonymous words both difficult and time-consuming (Yeh, Liou, & Li, 2007). There are several other obstacles which learners need to overcome. In order to see a pattern, learners may need to perform two or more searches (Gaskell & Cobb, 2004). Learners are not always ready to call to mind suitable words for

comparisons. They may not be able to come up with further ideas on what to search for (Gabel, 2001). Sun (2003) notes that ineffective search skills also lead to frustration.

Given the importance placed by teachers and researchers on the power of comparisons in DDL, it seems strange that little support is provided in most concordancing software to facilitate this. Both *WordSmith Tools* (Scott, 2010) and *AntConc* (Anthony, 2004) require use of multiple windows or saved results in order to view two sets of concordance results or collocations simultaneously. While *The Sketch Engine* (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) includes the *Sketch-Diff* function, only the summary Word Sketches are available in this view, and comparing actual concordance lines would require moving backwards and forwards between pages or having multiple tabs open in the browser. Each of these tools can provide a rich variety of ways for a researcher to make comparisons between items but the pathway for making these comparisons can be complicated.

A key design feature of *The Prime Machine* was to make comparisons between search terms as easy as possible through the facility to enter two queries at one time, leading to the retrieval of two sets of results; and the provision of pop-up lists giving the user suggestions in terms of alternative word forms, related words and collocations. In the list of claims summarizing the theory of Lexical Priming, Hoey (2005, p. 13) draws attention to several contrasts: differences between synonyms; differences between senses of polysemous words; differences between nested combinations of words; and differences across domain and genre. The user interface was designed to facilitate the selection of items across such contrasts and to enable the user to view results for concordance lines, collocations and other data in a side-by-side view.

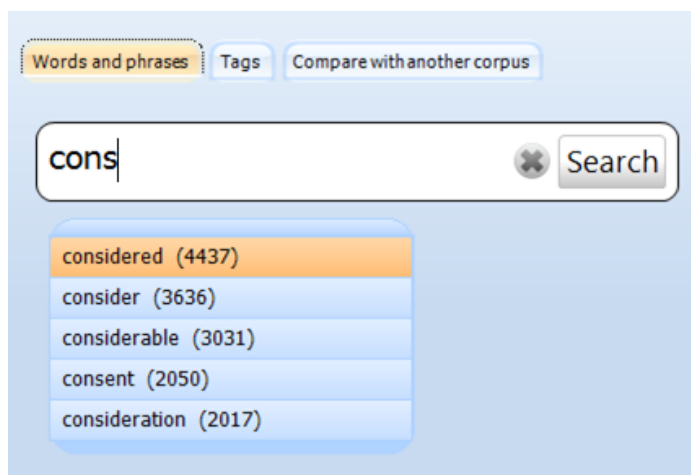


Figure 1: Screenshot showing auto-complete support for a query

The text input box was designed to aid the user with spelling and in choosing between similar strings of words. Figure 1 shows how as the learner starts to type a search term into the box, the words in the currently selected corpus with the same first few letters appear, displayed in descending order of frequency. Once a complete word has been entered or selected, the application provides other pop-up lists, giving suggestions for comparisons which could be made. A stemming process makes links between different word forms held in the database, so these other word forms are presented to the user underneath the right-hand text box. The software also provides a pop-up list of related words. Links between words and related words are based on the words being alternative English translations of Chinese words or on *WordNet* (Miller, 1995). To create the Chinese translation-based links, the *CC-CEDICT* database file (MDBG, 2013), a freely available Chinese-English dictionary file, was downloaded and imported into *Microsoft Excel*. The columns of English words for each Chinese headword were then imported into a database and strings (types) so as to establish links between English words if they occurred in the same row in the original table. Through accessing the database a list of words which are alternative translations for Chinese headwords can be retrieved for any of the words listed as English translations in the dictionary. The concordancer was developed specifically with Chinese learners of English in

mind, but future versions could incorporate lists derived from dictionary mappings for a range of different languages or simple thesaurus data. Links based on *WordNet* are based on semantically related words and include additional links across word forms. A DICE mutual information score is used to provide a ranking for the similar word pairs, so that they appear in the drop-down list with more mutually exclusive items towards the top.

As well as auto-complete for single words, computer users are also familiar with multi-word units appearing as they enter queries into various search boxes across different applications and websites. In *The Prime Machine*, since they are extracted, stored and indexed in advance, short lists of collocations can be retrieved very quickly, allowing suggestions beyond single words to be provided with almost instantaneous feedback on the collocational strength of two or more items.² The pop-up lists for collocations, alternative word forms and words with similar meaning are shown in Figure 2.

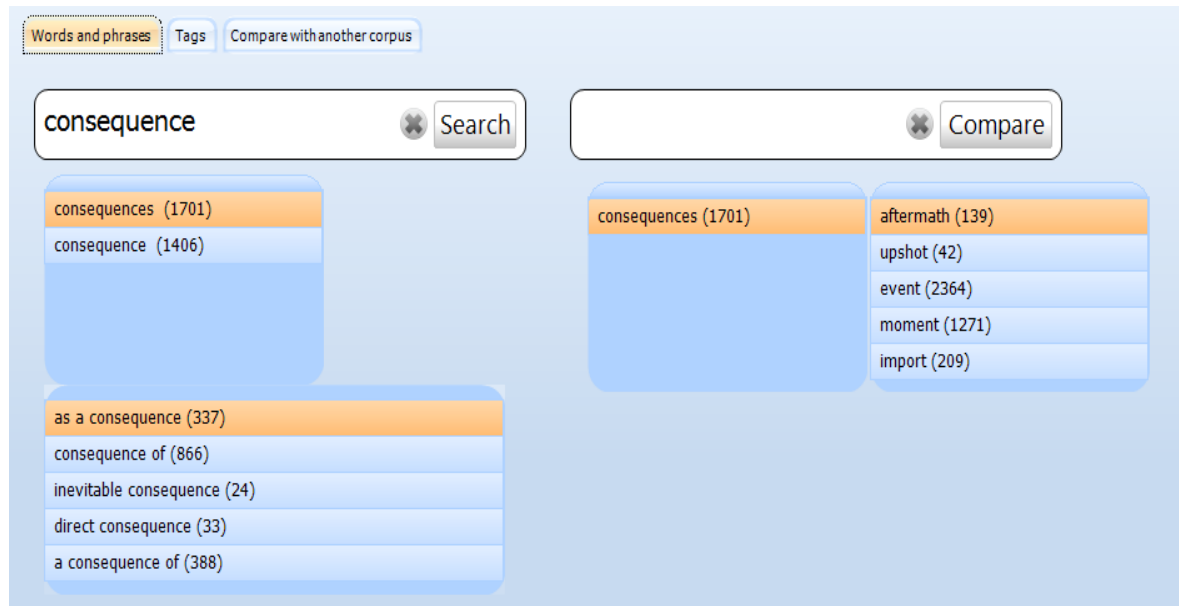


Figure 2: Screenshot showing prompts which appear for *consequence* in the *BNC: Academic* sub-corpus.

Collocations that contain words with the same stem and/or the same words in a different order appear on the right-hand side to encourage users to compare these with their main query. Figure 3 and Figure 4 show how these suggestions appear on screen.

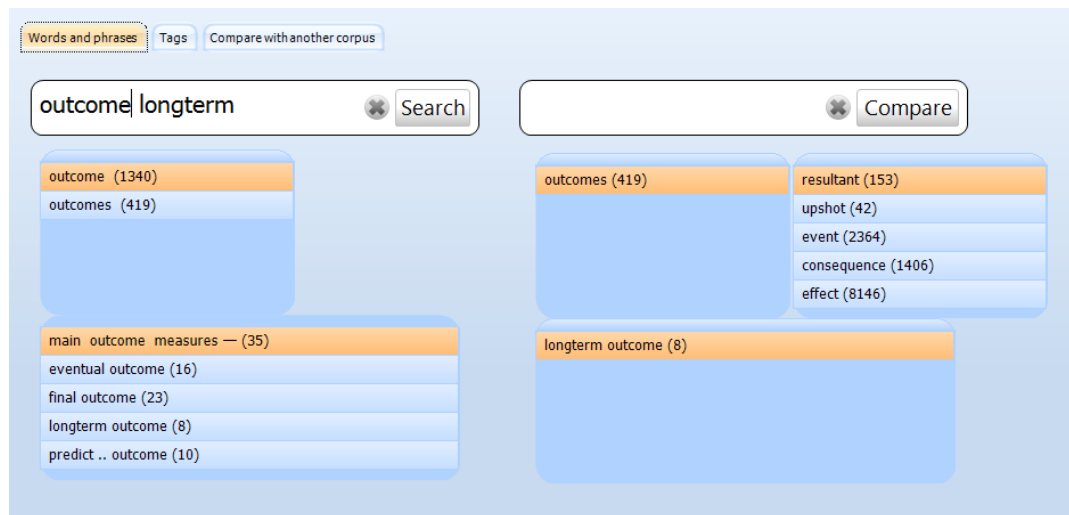


Figure 3: Auto-Complete suggestions showing collocations for data from the *BNC: Academic* sub-corpus for the query *outcome longterm*.

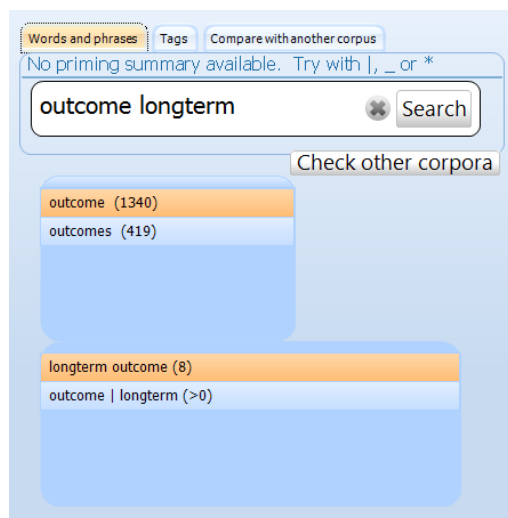


Figure 4: Auto-Complete suggestions showing raw window search queries for data from the *BNC: Academic* sub-corpus for the query *outcome longterm*.

As well as looking at pairs of words, comparisons of a single item across different corpora can be a good way to show how use varies across different registers. Comparing the results of the analyses of two or more language samples is an important part of register analysis,

since it is through comparison with other registers that the characteristics of one register become clear (Biber & Conrad, 2009). Just as most concordancing software does not provide an easy way to view and compare the results of two different items on the same screen, being able to view and compare results from two different corpora is also far from straightforward. *WMatrix* (Rayson, 2008) makes comparisons of two texts or two collections of texts very clear and is an excellent tool for researchers wanting to use differences in frequency between two corpora as a starting point for exploration of differences between the two collections of texts. If, however, a language learner wants to see how a word is used differently in two different corpora, corpus software packages do not provide much support.

In *The Prime Machine*, a comparison between two corpora can be made easily using the “Compare with another corpus” sub-tab. The search box on this screen looks and behaves as before, with auto-complete support at the word and collocation level. To the right of this box, a drop-down menu is provided which contains a list of all the other corpora. When the user clicks on the “Compare” button, the application checks that the word or combination of words is present in both corpora at least once before the query is allowed to proceed. If the words do not appear in either of the two corpora, feedback is provided. **Figure 5** shows the search screen for comparing corpora. In order to allow access to the complete corpus as well as comparisons across its sub-corpora, texts from the *BNC* are stored in the database twice: once as part of the complete corpus and once in a sub-corpus determined according to the text, following the major groups provided by Lee (2001).

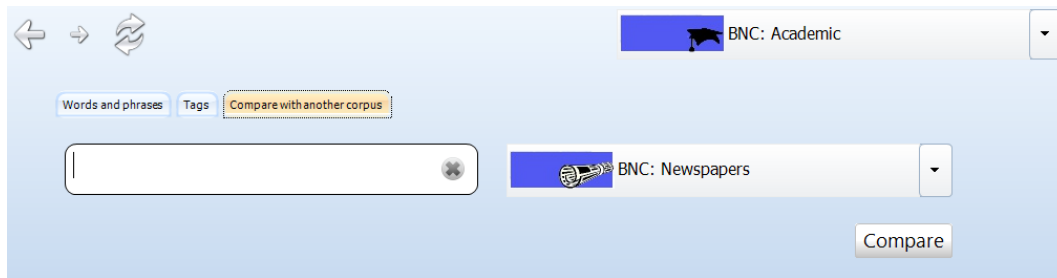


Figure 5: The “Compare with another corpus” sub-tab on the main Search Tab.

3. Claim 2: The design of the display for concordance lines should help language learners notice textual colligation, co-text and contexts

Once the concordance lines for a query or a pair of queries have been retrieved, the results must be presented to the user. While it has been recognised that in order to access some information it may be necessary to have longer contexts than the standard KWIC concordance line (Hunston, 2002; Sinclair, 1991), as many researchers have asserted, there are some advantages of viewing vertical lists of truncated sentences four words either side of the search term. Being able to see a large number of results provides a degree of “safety” for conclusions which the user draws (Mair, 2002). They can provide a “snapshot” of how lexis is usually used (Johns, 2002), can be seen as focusing on the “central” and “typical” (Hunston, 2002), and can be organised in such a way as to highlight patterns (Gaskell & Cobb, 2004). Sinclair (1991) suggested that KWIC provides access to patterns which are not meaning-bearing, allowing the distinction between the physical objects of text in the corpora and their meanings to be clear. However, for a corpus engine built on the theory of Lexical Priming, it would seem that access to wider contexts is important. For all the advantages of KWIC, by showing the node word in the centre of the screen, not only are paragraph breaks

usually masked, but the position of the node in the sentence is not very prominently displayed either. Even if the KWIC window is limited to words occurring in the same sentence, white space to the left of a sentence initial instance gives some indication that the word occurs towards the beginning of a sentence, but then masks whether or not this is a paragraph break. Concordance lines in which the node word is more than 4 or 5 words away from the start of a sentence appear much the same whether or not they are towards the beginning of a long sentence, part of a singleton paragraph, or towards the middle of an average length sentence. One challenge for this project was to find a way to present a much wider context than usual in a way which also facilitates visual scanning of patterns, while at the same time enjoying many of the benefits of KWIC. The Lines Tab in the application provides a KWIC view, and although this is much more similar to other concordancers, the design also incorporated some consideration of the position in paragraph and sentence. However, one of the main differences in the presentation of concordance lines in *The Prime Machine* is the Cards Tab and the single card shown on the Lines Tab for the currently selected line. A screenshot of the Cards Tab showing the paragraph layout, different heights of cards, the collocation captions and the citation information can be seen in **Figure 6**. The card template that is used to organise and present the words in sentences before and after each node will accommodate a fairly wide range of configurations including cards where all three sentences appear as one field, and others where paragraph breaks and headings can be seen before or after the node sentence. The beginning or end of a text is indicated by a blank line at the top or bottom of the card. The card view is intended to be a compromise between the desire to provide additional information about headings and paragraphing and trying to reduce the complexity of both displaying text as it would be shown in the original sources. It is a simplification bringing some order and uniformity to aspects like font size, colour and highlighting, while providing some visual information about the position of words in sentences and sentences in

paragraphs.

One issue regarding cards is that it is rather more difficult to scan across several concordance lines and to see patterns in the co-text. As well as gentle highlighting of the row in the card that contains the node word, the list of collocations for the current node word is also used to provide a visual cue at the top of each card in the form of a caption. This was designed to highlight the relationship between the concordance line and collocations. The caption provides an important way of helping learners see nearby words which have a strong relationship with the node, without disrupting the flow of text. Including collocates in a caption goes some way towards overcoming Kenning's (2000) concern that language learners may need help in seeing how a search term is actually part of a longer unit. It should also support teachers wanting to follow some of the other recommendations in the literature; recommendations such as teaching learners how to note collocations by drawing attention to extra words around a collocation (Lewis, 2000, p. 134) and directing learners away from separate word analysis (Siyanova & Schmitt, 2008).

As well as providing additional data and information in the extended context and the captions, the Cards view in *The Prime Machine* also prominently shows the source of each concordance line. Language learners using a concordancer are much less likely to be aware of the composition of the corpus and also tend to be less sensitive to notions of how language use changes across different genres and registers. However, as mentioned in earlier, an important point Hoey makes regarding all of the claims forming his theory of Lexical Priming is that they are "constrained by domain and/or genre" (2005, p. 13). In the design of the Cards and Lines views for *The Prime Machine*, the question of how best to facilitate clearer information about the source of each concordance line was considered carefully.

Firstly, in order to provide a quick sense of the kind of text from which the concordance line is taken, each individual text in the corpus is assigned to one main text category which is set at the time it is imported and this is used in a heading at the top of each card. Below this heading, other tags or metadata are displayed in the style of an academic reference or other referencing convention.

The Prime Machine - Guest

Search

Cards

Lines

Graphs

Collocations

Tags

Associates

Corpus Info.

project .. a pilot study carried

Social Science
Rapid -- ESRC grant abstracts. u.p.

... It was evaluated by the Edinburgh University Gordon Thompson Unit, with total success, and published as a student course book and tutor's guide, with two videotape material cassettes. The project arose from a pilot study carried out in 1982-3 with the support of the Nuffield Foundation and was one of two parallel projects sponsored by the Scottish Universities French Language Research Association, (the other, Lyon a la une being unfunded).

follows .. a pilot study funded

Social Science
Rapid -- ESRC grant abstracts. u.p.

... This more theoretical aspect should provide a better basis for urban designers and transport planners considering personal security issues. The study, which follows on from a pilot study funded by ESRC, involves the Departments of Civil Engineering and Law.

survey follows a pilot study

Social Science
Rapid -- ESRC grant abstracts. u.p.

... This is the earliest period for which good documentation exists which can be interpreted together with archaeological and linguistic evidence to reconstruct the landscape in detail. This survey follows a pilot study of the West Midlands in the Anglo-Saxon period in which particular use was made of Old English charter material. Here landscape regions were identified which permitted an investigation to be made of the social and economic effects of land use patterns, and which allowed the territorial organisation of the period to be set within its geographical framework.

follows a pilot study .. conducted

Social Science
Rapid -- ESRC grant abstracts. u.p.

... It is hoped that the study's findings will lead to recommendations on how British companies can improve their marketing effectiveness in a manner that will contribute to a greater level of industrial competitiveness. This research follows a pilot study, conducted in 1984, supported by the ESRC. Information will be obtained from personal interviews with senior managers in matched trades of US, Japanese and British companies operating in industries of national significance in terms of size and growth.

a pilot study carried

Social Science
Rapid -- ESRC grant abstracts. u.p.

... This interaction is often culturally determined and therefore whenever it is possible to extend the previously narrow focus on westernised middle-income parents developing spoken English in their infants, then very valuable information can be discerned in the language development. This study extends the work of a pilot study funded by ESRC (C700/23/2220) which examined the interaction between clear parents and their infants in the first 12 months of life. This new study focuses on the transition period from 'pre-language' communication to fully-fledged language competence. ...

study follows .. a pilot study

Social Science
Rapid -- ESRC grant abstracts. u.p.

... This will make it possible for more informed advice to be given to parents of mentally handicapped children about how they can best assist their children's language development. The study follows on a pilot study in 1983-84 supported by the University of York. The present research involves making extensive video recordings in the homes of eight carefully selected families and submitting them to various forms of analysis.

a pilot study carried

Social Science
Rapid -- ESRC grant abstracts. u.p.

Design and society: proceedings of ... an international conference on design policy. Langdon, Richard Cross, Nigel The Design Council UK 1984 51-105

This bears out the results found by Rob Sadler and Liz Spencer in their survey, that people would like more information on, and are enthusiastic about, the advantages of renewable energy. The findings of a pilot study carried out by NATTA on attitudes to the Severn Barrage develop this last point further. A questionnaire was administered after a public meeting on the Barrage in Newport, Wales in April 1982. ...

Figure 6: The Cards Tab with captions at the top and highlighting of the line containing the node; with incidental data from the *BNC: Academic* sub-corpus for a search on the node *pilot*. The currently selected card is shown with a yellow caption.

4. Claim 3: The design should help language learners notice features in the patterning of words and phrases

The tendency for words or phrases to occur in different positions in a text is an interesting and under-researched area, and one which is somewhat difficult to explore using standard concordancing software. Nevertheless, some work has been done looking at some of the possible different text units and the tendency for words and longer phrases to occupy positions at the beginning of these. *WordSkew* is a software tool which allows counts to be performed for items within sentences, paragraphs, sections or texts in terms of absolute slots or by dividing the discourse unit into portions of equal length (Barlow, 2016). The *Concord* tool in *Wordsmith Tools* provides columns of data showing the position of each concordance line as a percentage relative to several text units. *Wordsmith Tools* was the software used by Hoey (2005) and some of the ways it can be used to investigate textual colligation are demonstrated by Scott and Tribble (2006). Garretson's *CenDiPede* software (2010) includes three features under the heading "Pseudo-Colligation", two of which are relevant to textual position. The first uses the results from clausal analysis to report the raw frequency of occurrences of the node occurring before the verb within its clause. This is designed to be a rough mapping to Theme-Rheme. The second is described as a "nod to Hoey's notion of textual colligation" (Garretson, 2010, p. 149), and is the percentage of instances of the node where it is sentence initial. At the text and paragraph level, Hoey and O'Donnell (2008) and O'Donnell et al. (2012) compared the first sentences of texts and paragraphs against the sentences from the remainder of these texts in order to establish which words had a tendency to be used in text initial and paragraph initial position. Their procedure was complicated,

especially for the generation of concgrams, and involved splitting the corpus into sub-corpora according to each of the required set of positions, using concgrams in *Wordsmith Tools* and then a *Python* script before running the wordlist function in *Wordsmith Tools* again.

The use of the key word method to identify words which occur with statistical significance in text initial or paragraph initial position seems very promising. However, concordancing software provides little integration of functions to explore such features and few language learners would be skilled or motivated enough to go through the process of splitting a corpus themselves and then performing key word analysis and interpreting the results. The results from the study by O'Donnell et al. (2012) which found that one in forty individual words showed a tendency to be used in specific positions provides good evidence that this is something worth researching further, but it does also suggest that if the starting point is a word or phrase and the aim to is to discover whether or not this word or phrase has such a tendency, the overwhelming majority of cases are likely to be disappointingly negative.






Writing about concordancing software in more general terms, Cobb (1999) argues that language learners need software which does not assume detailed linguistic knowledge and which also does not assume that the users will be curious enough to explore. It would seem obvious that for phenomena like textual colligation which are less well-understood by both teachers and students, these two aspects of software design are even more important.

Therefore, procedures were developed for *The Prime Machine* to calculate, store and display tendencies of words and nested combinations to occur in various environments. As well as measures related to textual colligation, several other measures were developed to target some of the other features of Lexical Priming. It is hoped that the aim of drawing learners' attention to this selection of features will resonate with language teachers and that will help learners engage with the data in the concordance lines more easily. Although the range of features is limited, some of the well-known trouble-spots for English for Academic Purposes

have been targeted, with the use of articles and propositions, passive voice, and modal verbs included. Rather than looking for specific features and then looking at the words which display a specific tendency, the aim of processing and storing these data is to highlight to the user any tendencies which exist for the specific words or collocations that they have used in their search query. The results of key word analyses for the features are made available in the database so that it is possible to retrieve the tendencies which are key for the search query.

Table 1 shows the list of features and how they are organized into 5 groups.

Table 1: Features of Lexical Priming measured and stored in *The Prime Machine*

Group	Feature	Values	Level
Headings 	Title	Title; Not a title	Sentence
	Heading	Heading; Not a heading	Sentence
Position in text* 	Sentence position in text	Text Initial; Text Ending; Not text initial or text ending	Sentence
	Paragraph position in text	First Paragraph; Last Paragraph; Not first or last paragraph	Sentence
	Sentence position in paragraph	First Sentence; Last Sentence; Not first or last sentence	Sentence
	Word position in sentence	First Fifth; First Third; Last Third; Last Fifth; Not first or last third	Word
	Word position in sentence	Theme; Rheme; (unknown)	Word
Complexity, Modality, Voice & Polarity 	Complexity	Simple Sentence Complex Sentence	Sentence
	Modality	Volition/prediction; Permission/possibility/ability; Obligation/necessity; No modals	Sentence & Word
	Voice	Active Voice/Other; Passive Voice	Sentence & Word
	Polarity	Positive; Negative	Sentence & Word
Determiners & Prepositions 	Determiners	Definite articles / Possessives; Indefinite articles; No articles	Word
	Prepositions	Near Prepositions; Not Near Prepositions	Word
Repetition 	Repetition	Same form Same stem Not repeated	Summary information only

* Not all the values for features in this group are mutually exclusive. For example, words that are in the first fifth of a sentence will also be in the first third.

In order to measure tendencies, features are flagged in the database either at the word or sentence level through a series of processes. Lists of words and collocations are generated according to the proportion of instances in the corpus for each feature of the contextual environment, and a statistical test is applied so that those meeting a threshold will also be stored in list of significant items for each feature. The contingency table used for sentence level measures is shown in Table 2, and that for word level measures is shown in Table 3. For collocations, the contingency tables are based on the number of occurrences of the node of the multi-word unit in each environment. Summary data is stored for all log-likelihood values reaching a BIC value of 2. Following Wilson (2013), Bayes Factors are used as a way of standardizing the cut-off point for the key word method, and the level of significance is stored using the BIC interpretation given there.

Table 2: Contingency table for sentence level features

	Corpus One	Corpus Two
Freq. of word	A = inside sentences with the specific feature	B = <i>Outside the sentences with the specific feature</i>
TOTAL	C = <i>Count of all words inside sentences with the specific feature</i>	D = Whole corpus – C

Table 3: Contingency table for word level features

	Corpus One	Corpus Two
Freq. of word	A = where the specific feature has been marked	B = where the specific feature is absent
TOTAL	C = <i>Count of all words with the specific feature</i>	D = Whole corpus – C

A few examples are presented in Table 4 as they appear in the software's help screens, stripped of all the technical evidence, where they are provided in order to help explain to an advanced learner what each feature was designed to measure. The reader is not being asked to dwell too heavily on whether there is anything remarkable or surprising about the tendencies of the example words to be used in these specific contexts, but rather to consider whether given a learner's interest in the use of one or more of these words it would not be to his or her advantage to have attention drawn to the existence of such tendencies.

Table 4: Selected examples from the help screen

<p><u>Headings: Heading</u></p> <p><u>Examples from the <i>BNC: Academic</i> sub-corpus</u></p> <p>Only 0.6% of words in this corpus are part of a heading.</p> <p>Yet 13% of the occurrences of the word <i>conclusion</i> are paragraph headings and none of the occurrences of the word <i>ending</i> are paragraph headings. Obviously, the heading used for the last section of an academic article is usually <i>Conclusion</i>, but it also occurs very frequently within sentences.</p>
<p><u>Position in text: Paragraph position in text</u></p> <p><u>Examples from the <i>Hindawi Computer Science</i> corpus</u></p> <p>Only around 3 in 100 words are part of the first paragraph of texts.</p> <p>Yet around a quarter of the occurrences of the words <i>advances</i> and <i>increasingly</i> are in the first paragraph of texts. Other words often used in the first paragraph are <i>emerging</i>, <i>novel</i>, and <i>growing</i>. These give a sense of how changes have occurred and progress has been made.</p> <p>Only around 1 in 200 words are part of the last paragraph of texts.</p> <p>Yet words like <i>hope</i> and <i>future</i> occur in the last paragraph much more often than that. Words which frequently occur in the last paragraph of a text often give a sense of looking forward to the future.</p>
<p><u>CMVYN group: Modality</u></p> <p><u>Examples from the <i>BNC: Academic</i> sub-corpus</u></p> <p>Less than 5% of words in the corpus are near modal verbs.</p> <p>Yet words like <i>legitimately</i>, <i>usefully</i>, <i>conceivably</i> and <i>easily</i> are often used with the words <i>can</i>, <i>could</i>, <i>may</i> or <i>might</i>.</p> <p>Words like <i>remembered</i>, <i>noted</i>, <i>emphasised</i> and <i>stressed</i> are often used with the words <i>must</i>, <i>should</i>, <i>need to</i> or <i>ought to</i>. Other words often used with these modals are <i>carefully</i> and <i>surely</i>.</p> <p>Words like <i>suffice</i>, <i>cease</i>, <i>depend</i> and <i>disappear</i> are often used with the words <i>will</i>, <i>would</i> or <i>shall</i>. Other words often used with these modals are <i>examine</i>, <i>argue</i> and <i>discuss</i>.</p>
<p><u>Det. & Prep. group: Prepositions</u></p> <p><u>Examples from the <i>BNC: Academic</i> and <i>BNC: Newspapers</i> sub-corpora</u></p> <p>A little more than half of all words in these corpora are near prepositions.</p> <p>Yet 99% of the occurrences of the word <i>spite</i> are near prepositions while none of the occurrences of the word <i>despite</i> are near prepositions.</p> <p>Sometimes similar words can be quite tricky to use correctly when writing in a foreign language, but a quick search for <i>despite</i> vs. <i>spite</i> in either of these corpora can show preposition patterns very clearly. We would expect the concordance lines to show us <i>despite</i> near verbs and in the phrase “despite the fact”. We would also expect to see <i>spite</i> used in sentences in the phrase “in spite of”.</p>

When the concordance lines are retrieved, the concordancer is able to present information about the proportion of instances for the currently downloaded sample and the proportion of instances in the corpus as a whole, as well as a list of features for which the search term has been pre-calculated as having a statistically significant relationship. This information is displayed in the form of graphs, designed to help the learners appreciate that these primings are almost always representative of relative frequencies rather than absolute restrictions on use. Krishnamurthy and Kosem (2007) make many suggestions about the visual design of a corpus tool and the incorporation of icons and graphs into *The Prime Machine* was in part a response to these. An example of a graph is shown in Figure 7.

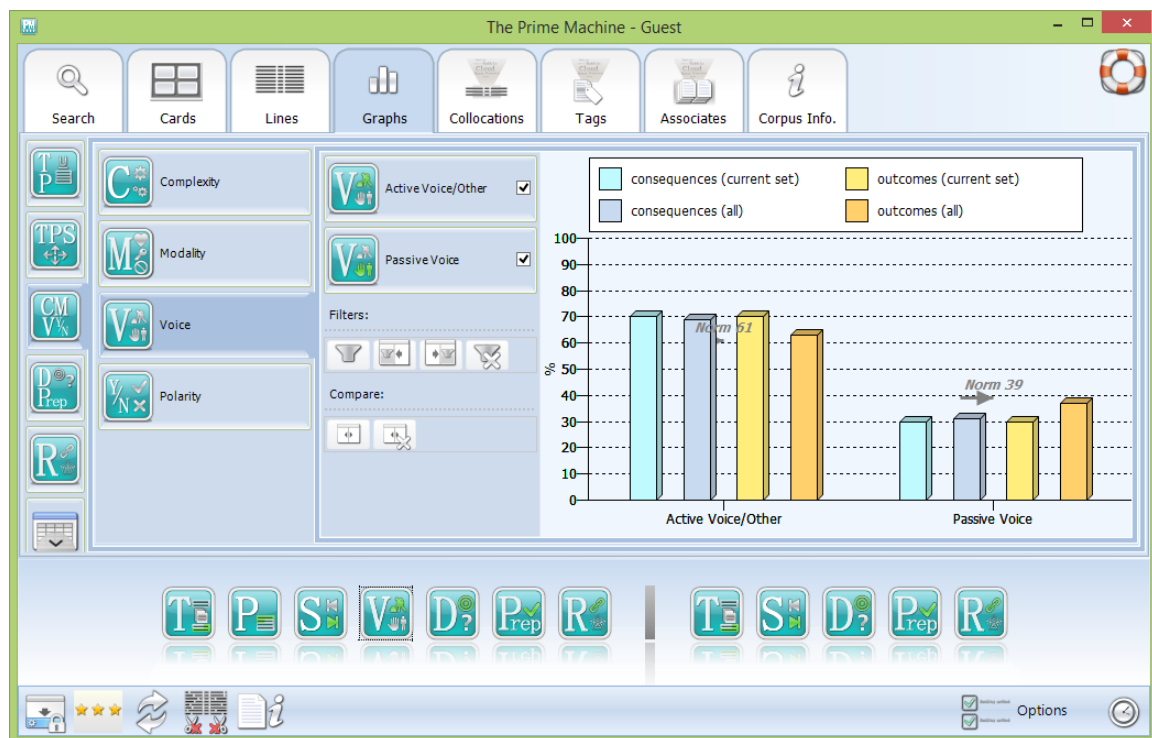


Figure 7: Graph display for compare mode for the Voice submenu on the Graphs Tab with results for *consequences* compared against *outcomes* from the BNC: Academic sub-corpus.

One of the striking things from Hoey's (2005) presentation of the evidence for the priming of words is the need to consider what the expected values or what typical environments for each kind of feature would be. Clearly, the number of text initial sentences will always be very small compared to the whole corpus, yet because of the differences in the length of the texts in different corpora, these proportions can vary. Similarly, some features such as passive voice tend to be much less common in some text types than in others and so it is useful to be able to highlight cases where the proportion is much higher or lower than would be expected based on a collection of texts as a whole. For the graphs, values for expected values are calculated using the total number of words in each priming environment in the whole corpus, and these are displayed using arrows marked "norm".

One of the important goals of the project was to find a way to make tendencies of words and collocations more prominent and *guide* the learner to find interesting and useful patterns. A researcher who is highly motivated to explore exhaustively the evidence for primings of a particular word or phrase based on tendencies revealed through corpus analysis may well be motivated enough to spend time trying different features, not losing too much interest if no relationship is found. However, if a vast array of options is made available to learners without any guidance, they could either waste time filtering the data or become frustrated. Therefore, a means was needed of helping direct the user's attention to priming information which might be explored more fruitfully, and this is the purpose of the "hot" icons which appear on a dock at the bottom of the results screen. When each list of concordance lines and other summary data are retrieved, the application goes through the table of statistically significant priming environments and changes the icons to match the features. Icons representing priming environments which do not reach the "Positive evidence" BIC Factor Score for the current search term are set to be invisible. Clicking on the icon takes the user directly to the sub-section on the Graphs Tab menu corresponding to this feature. **Figure 8**

shows concordance lines with the dock at the bottom showing statistically significant tendencies for position, complexity, indefinite articles and repetition. **Figure 9** shows how the icon grows in size when the mouse is hovered over it.

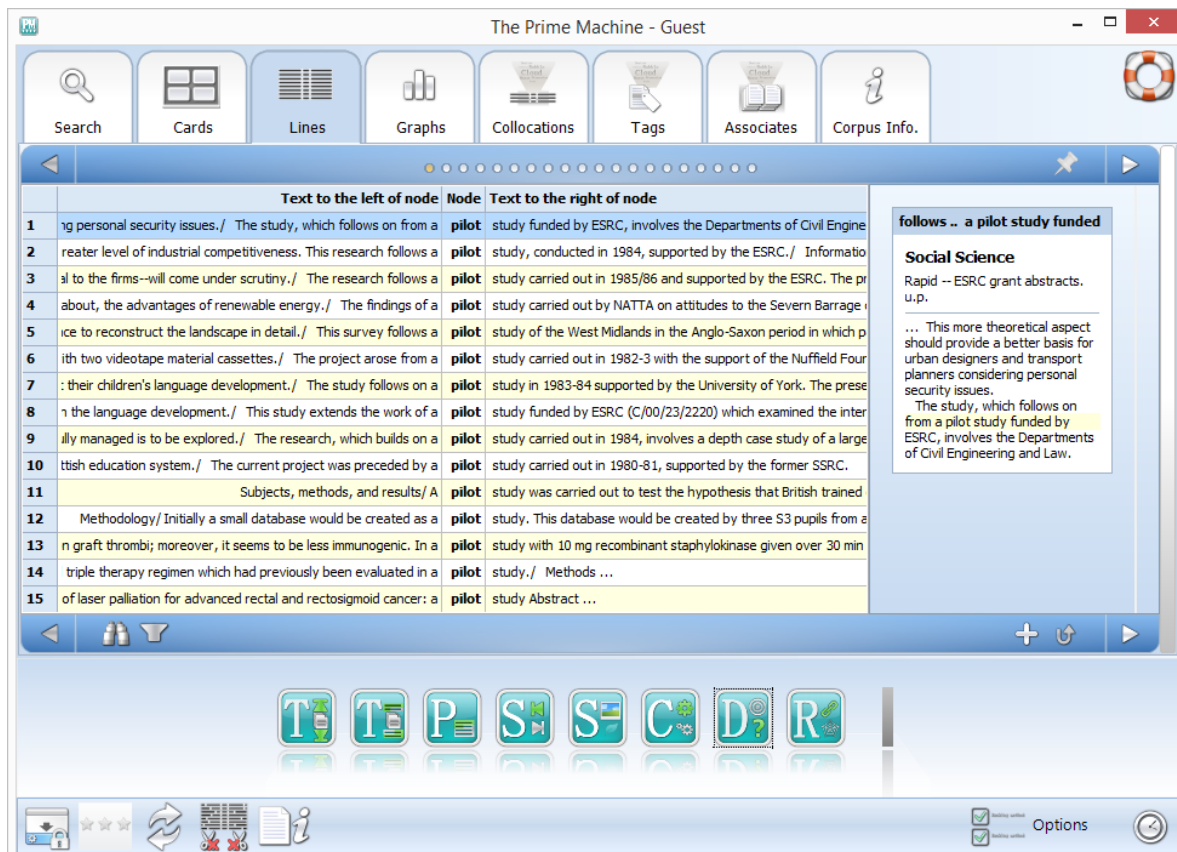


Figure 8: Lines Tab showing the card for the currently selected concordance line and the dock of icons for the node pilot in the *BNC: Academic* sub-corpus.

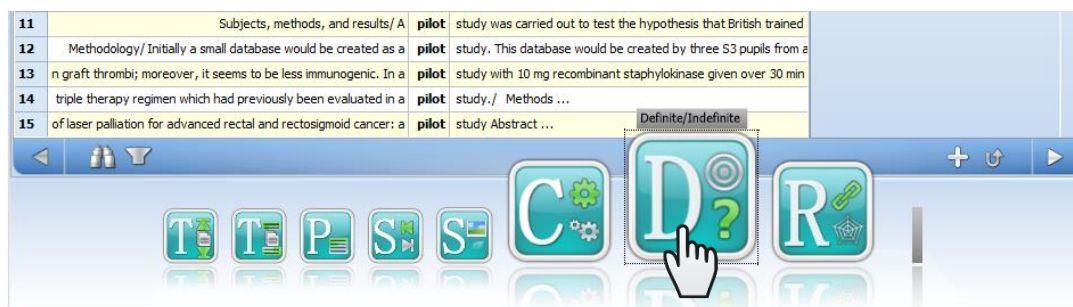


Figure 9: Enlarged icon showing positive evidence for a tendency to occur after indefinite articles. The hand icon represents the mouse cursor position.

An important point is that providing a summary of typical environments for a word or collocation should not be an end in itself; rather the software should encourage learners to consider and explore for themselves whether the words they encounter or want to use in their own writing might be primed to occur with other features. To this end, a system was devised to allow users to move from the list of features on the Graphs Tab to a filtered list of concordance lines matching those features. **Figure 10** shows the checkboxes and filter buttons available for one of the priming menus.

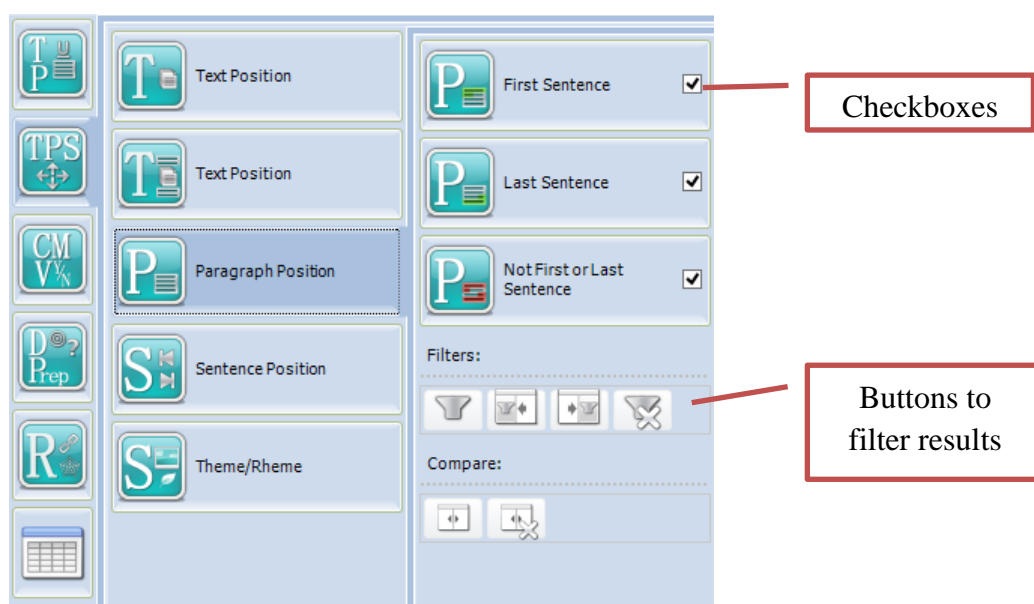


Figure 10: Checkboxes and filter buttons for one of the submenus on the Graphs Tab.

By removing the ticks from some of these boxes, the user can filter down the results.

Looking at filtered results may help to show learners how a word or collocation is used in particular priming environments. The option to compare concordance lines for the same item to see whether patterns can be seen or conclusions can be drawn according to different contexts and to allow learners to see variation as well as common patterns. The complex

categories used for some of the priming features can also be made easier to understand by showing users lines matching the features on the left and lines not matching those features on the right. Figure 11 shows the Lines Tab when in compare mode.

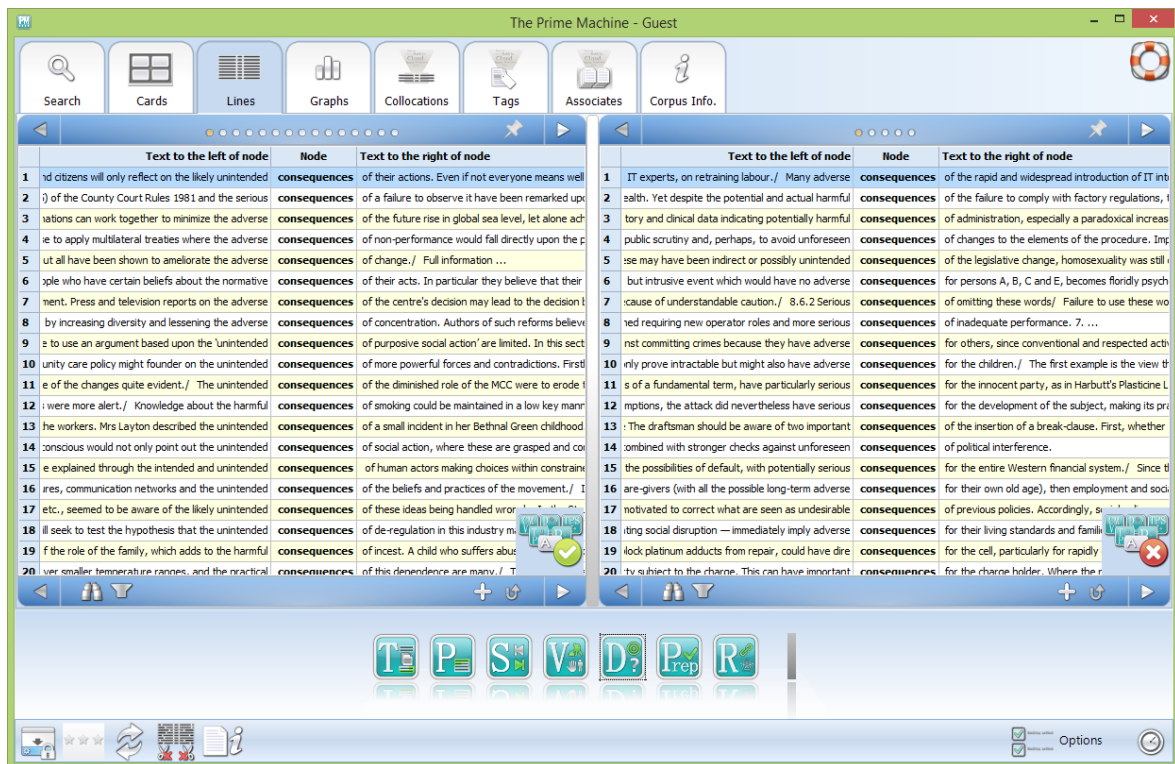


Figure 11: Compare mode for the node *consequences* in the *BNC: Academic* sub-corpus, filtered by definite articles or possessives.

5. Further work and concluding comments

The features presented here were designed to provide language learners with a means of finding and selecting useful starting points for the exploration of words and collocations in a range of different corpora, as well as helping them to avoid unfruitful starting points. In the initial evaluation (Jeaco, 2015), although the actual use of the software was fairly limited and the number of participants was small (25 for the first questionnaire and software session; 23 for the follow-up), respondents to the follow-up questionnaire considered the ability to compare words or phrases side by side to be particularly positive, and a fair proportion (44%)

of the searches performed were using the compare mode. While for the Cards Tab and Lines Tab 74% of respondents to the questionnaire rated them “Useful” or “Very Useful”, it was also evident that different students used and rated the Cards Tab and Lines Tab differently, indicating that each of these different ways of presenting concordance line results may support different kinds of learners. When judging the overall usefulness of the software in the last question of the follow-up questionnaire, the software was received very positively. Twenty-two out of twenty-three students responded positively, and the one student who selected “no” was still positive about the usefulness of the software in the comment, stating that his/her reservation was due to his/her belief that other software packages may be able to provide similar information in a more convenient way. The positive result was especially striking considering that from the results of the first questionnaire it was very clear that very few students had used concordancers before. Work on this project is continuing, both in terms of on-going development of the software features, but also in terms of its evaluation. As of 2016, the system is available to students and staff at the author’s institution and some further evaluation has taken place.³ It is hoped that the software can be made more widely available in the near future.

One of the attractions of the theory of Lexical Priming (Hoey, 2005) is that it brings together a range of features including lexical patterning, grammatical patterning, textual patterning and semantic patterning, highlighting how these interact and suggesting how and why these patterns form in the mind of the language user. From a language teaching perspective, the theory provides insights into features that can easily be recognized as being important: differences across register and genre; differences between synonyms; differences between senses of polysemous words; and differences between nested combinations of words. It is hoped that *The Prime Machine* will prove to be a valuable tool for both students and teachers

as they gain access to corpus information in new and interesting ways, and as they take new opportunities to explore evidence for the wide variety of ways in which words and combinations of words are primed for experienced speakers of the language. The query screen makes it easier for learners to compare items. The Cards design provides a new way for users to view concordance lines with a design offering more context than typically visible in KWIC displays and incorporating features of paragraphing, headings and collocation. The software extends the use of key word analysis for indicating strong tendencies of words and phrases to occur in specific environments on a much wider range of features associated with Lexical Priming. It is also hoped that this chapter makes an interesting contribution to language learning and teaching through the application of Lexical Priming theory to second language learning situations.

List of Corpora

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/>.

Hindawi. (2013). Hindawi's open access full-text corpus for text mining research. Retrieved 6 November, 2013, from <http://www.hindawi.com/corpus/>.

References

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.

Barlow, M. (2016). WordSkew. [Article]. *International Journal of Corpus Linguistics*, 21(1), 105-115. doi: 10.1075/ijcl.21.1.05bar

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.

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

Bolitho, R., Carter, R., Hughes, R., Ivanič, R., Masuhara, H., & Tomlinson, B. (2003). Ten questions about Language Awareness. *ELT Journal*, 57(3), 251-259.

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

Coniam, D. (1997). A practical introduction to corpora in a teacher training language awareness programme. *Language Awareness*, 6(4), 199-207.

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage - based language: Investigating the latent structures that underpin acquisition. *Language Learning*, 63(Suppl 1), 25-51.

Firth, J. R. (1957). Linguistic analysis as a study of meaning. In F. R. Palmer (Ed.), *Selected Papers of J R Firth 1952 - 59* (pp. 12-26). London: Longman.

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.

- Garretson, G. (2007). What your words know: The theory of lexical priming. *International Journal of Corpus Linguistics*, 12(3), 445-452.
- Garretson, G. (2010). Corpus-Derived Profiles: A Framework for Studying Word Meaning in Text. Unpublished Ph.D. dissertation, Boston University.
- Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, 32(3), 301-319.
- Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.
- Hoey, M., & O'Donnell, M. B. (2008). Lexicography, grammar, and textual position. *International Journal of Lexicography*, 21(3), 293-293.
- Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning & Technology*, 9(2), 90-110.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. 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 <http://repository.liv.ac.uk/id/eprint/2014579>
- 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.

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.

Kaszubski, P. (2007). Michael Hoey. Lexical priming: A new theory of words and language.

Functions of Language, 14(2), 283-294.

Kennedy, G. D. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Kenning, M.-M. (2000). Concordancing and comprehension: preliminary observations on using concordance output to predict pitfalls. *ReCALL*, 12(02), 157-169.

Kettemann, B. (1995). On the use of concordancing in ELT. *TELL&CALL*, 4, 4-15.

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.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: additional evidence for the Input Hypothesis. *The Modern Language Journal*, 73(iv), 440-464.

Krishnamurthy, R., & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes*, 6(4), 356-373.

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.

Lewis, M. (2000). Language in the lexical approach. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 126-154). Hove: Language Teaching Publications.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30(2), 333-370 338p.

MacWhinney, B. (2014). *Childes Project. [electronic book] : Tools for Analyzing Talk, Volume I: Transcription format and Programs*: London : Taylor and Francis, 2014. 3rd ed.

Mair, C. (2002). Empowering non-native speakers: the hidden surplus value of corpora in Continental English departments. In B. Kettemann, G. Marko & T. McEnery (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 119-130). Amsterdam: Rodopi.

MDBG. (2013). CC-CEDICT Download page. Retrieved 28 September, 2012, from <http://www.mdbg.net/chindict/chindict.php?page=cedict>

Meyer, C. F. (2002). *English Corpus Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Miller, G. A. (1995). Word Net: A lexical database for English. *Communications of the ACM*, 38(11), 39-41.

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

- O'Donnell, M. B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory*, 8(1), 73-101.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519-549.
- Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11(2), 129-158.
- Scott, M. (2010). *WordSmith Tools* (Version 5.0). Oxford: Oxford University Press.
- Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, 64(3), 429-458.
- 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*.
- Tomasello, M. (2003). *Constructing a language : a usage-based theory of language acquisition*: Cambridge, Mass. ; Harvard University Press, 2003.
- Tomlinson, B. (1994). Pragmatic awareness activities. *Language Awareness*, 3(3-4), 119-129.

Tomlinson, B. (2008). Language acquisition and language learning materials. In B. Tomlinson (Ed.), *English Language Learning Materials: A Critical Review* (pp. 3-13). London: Bloomsbury Publishing.

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.

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.

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.

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.

¹ An exception would be developments in usage-based linguistics. For example, the CHILDES project was originally conceived in the early 1980s as “an archive for typed, handwritten, and computerized transcripts” for researchers in child language (MacWhinney, 2014, p. 24), and its collection of corpus texts has been used for analysis in usage-based linguistics. Corpus methods have been applied to explore first language acquisition of individuals (Lieven, Behrens, Speares, & Tomasello, 2003). Developing areas of usage-based approaches look to greater use of corpora for the exploration of spontaneous speech in child language acquisition (Tomasello, 2003) and there is a recognised need for larger corpora for second language acquisition research (Ellis, O'Donnell, & Römer, 2013).

² As a piece of software purposefully designed to support the examination of the kinds of relationship between words that are introduced in Hoey's theory of Lexical Priming, collocations are defined in this project based on his 2005 definition. In the software, collocations refer to combinations of two, three, four or five words in a four-word window either side of a node. Full details of the way in which these are calculated are provided in Jeaco (2015).

³ The software is supported at the author's institution through the XJTLU Teaching Development Fund (14/15-R9-074).