Helping Language Learners Put Concordance Data in Context:
Concordance Cards in The Prime Machine

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

ABSTRACT

While corpus tools provide several different ways to display relationships between words within texts and across texts, the main format for viewing concordance data is Key Word in Context (KWIC). In Computer Aided Language Learning, concordance lines in KWIC format may be accessed inside a concordancer or within other software through links to corpus data. Language learners can and do gain useful insights from exploring concordance data in KWIC format, but some kinds of information may be harder to see, some patterning of use may not be so obvious, and reading of complete examples may not be very easy. The Prime Machine was developed for language learners and aims to make corpus data easier to access and interpret. This paper introduces the design of the Cards Tab, which provides an additional way of viewing concordance data. Results from three evaluations with language learners and teachers show positive attitudes towards this display.

KEYWORDS

INTRODUCTION

The tools of corpus linguistics have made it possible over the last few decades for researchers and lexicographers to access vast quantities of examples for specific search terms, and to discover and analyse patterns of language use. For more than 30 years, this usage-based approach to the analysis of language, drawing on corpora of increasing sizes, has had a huge impact on the way in which dictionaries are constructed (Hanks, 2012; Renouf, 2007). Corpora have also been used to explore the nature of language (Hanks, 2013; Hoey, 2005). Archives of transcripts and specially constructed corpora from child language research have been the basis for corpus work on first language acquisition within Usage Based Linguistics (Lieven, Behrens, Speares, & Tomasello, 2003; MacWhinney, 2000). Lists of words and multiword units have been used for the grading and selection of items in second language learning and teaching (Bauer & Nation, 1993; Coxhead, 2000; Durrant, 2009; Simpson-Vlach & Ellis, 2010). Corpora have been a basis for the analysis of lexical and grammatical differences across genres and registers (Biber & Conrad, 2009; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Thompson, 2004), and in translation (Baker, 1993; Hu, 2016; Teubert, 2004). Corpora can also be used to compare patterns in literature or language as a whole or with those of a specific literary author (Fischer-Starcke, 2010; Mahlberg, 2013; Semino & Short, 2004). In the Chinese context, computerized corpus research has also covered a broad range of linguistic fields over the last few decades (Li & Smith, 2015). Work continues in a host of areas including exploring vocabulary for testing (Jin, Guo, Mak, & Wu, 2017) and discipline-specific teaching (M. Zhang, 2013), exploring

China English (Xia, Xia, Zhang, & Nesi, 2016), exploring methods for extracting n-grams (Wei & Li, 2013), tracking changes in Chinese news media (W. Zhang, 2015), and building a system for opinion classification for Western news (Xiong, Xu, & Liang, 2014).

One of the central tools in corpus linguistics has been the concordance line, typically presented as Key Word in Context (KWIC), with each corpus example presented horizontally across the screen with a number of characters (letters) or words visible to the left and right. However, language teachers and language learners alike can find it hard, particularly at first, to understand how KWIC data can be used and interpreted. This paper presents the rationale and implementation of a complementary concordance line display format (Cards) which has been integrated into a new corpus tool (*The Prime Machine*) specifically designed with language teachers and language learners in mind. Feedback from teachers and students in three evaluations is reported, along with details of ongoing development and future plans.

LITERATURE REVIEW

Key Word in Context

The KWIC display provides a good way for users of a concordancer to explore multiple examples at the same time, with its vertical list of truncated sentences, typically with only a small number of words either side of the search term[1]. Since the display is very compact, it means that a large number of results can be viewed at the same time, providing a degree of "safety" as conclusions about patterns of use are drawn (Mair, 2002). KWIC results can provide a "snapshot" of how lexis is typically 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). It has also been suggested that KWIC concordance lines can free learners from getting caught up in the story or message of a text so as to be able to focus on the language (Cobb, 1999; citing a problem raised by Mondria & Wit-de Boer, 1991). Sinclair (1991) suggests this same freedom is important for researchers, as the KWIC view 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. Nevertheless, it has also long been established that wider co-text beyond the typical KWIC concordance line display may be required in order to access some information necessary for analysis or interpretation (Hunston, 2002; Sinclair, 1991).

Concordance Lines in CALL

There have been many software packages within the field of Computer Aided Language Learning (CALL) which have been designed to integrate with or incorporate concordance line data. Concordancers are a central tool for Data Driven Learning (DDL) activities, with both standalone packages such as MicroConcord (Scott & Johns, 1993), WordSmith Tools (Scott, 2010), and AntConc (Anthony, 2006), as well as web-based resources such as the Compleat Lexical Tutor (Cobb, 2000), BYU's online concordancer (Davies, 2008-) and the Sketch Engine's *SKELL* language learning site (Kilgarriff, Marcowitz, Smith, & Thomas, 2015) being used for hands on concordancing activities with language learners. Yoon (2011) gives an overview of studies into DDL. In China, studies into the use of corpora for language teaching have been promising at university level (Guan, 2013; He, 2015; Luo & Liao, 2015), and also for younger learners (Yu, Liou, Chang, & Vongpumivitch, 2011). In a recent meta-analysis of DDL research, Boulton and Cobb (2017) found that DDL approaches have been demonstrated to be effective. Their meta-analysis provides clear evidence:

… that learners seem able to perceive language patterns despite the lines chopped off the concordance output and that DDL activities are not confined to advanced learners, nor exclusively to simplified corpora or mediated data, nor to hands-off or paper-based activities, nor for learning goals limited to vocabulary and collocation. (Boulton & Cobb, 2017, pp. 39-40)

As well as studies and teaching practices based on direct use of concordancers, access to concordance lines within CALL packages has also been provided through integrated web tools (Kaur & Hegelheimer, 2005; Kuo, 2008; Yeh, Liou, & Li, 2007), and as a means for learners to extract examples to store in personal or collaborative vocabulary repositories (Cobb, 1999; Horst, Cobb, & Nicolae, 2005).

It is also worth noting that while hands on use of concordancers has maintained a steady popularity over the decades, concordance lines have also been used in a number of contexts and studies as an aid for the creation of teaching materials. This aspect of computer aided language teaching, can be understood as providing teachers with the tools needed for them to be able to access corpus data for the specific needs of a lesson or a student, and then facilitating the creation of printed or electronic materials. The use of printed handouts from corpus software extends back to the earliest days of DDL (Johns, 1994), and there have been suggestions that printed materials are an excellent way in for students to begin using concordancers and that printed materials can help students overcome difficulties using software directly (Boulton, 2010; Oghigian & Chujo, 2010). Within English for Academic Purposes, it has been suggested that when teachers get a query from a student in a class about differences between similar words or phrases, the teacher could fruitfully access a concordancer after class and present the results in the next lesson (Alexander, Argent, & Spencer, 2008). Even when students and teachers turn to a dictionary, the fact that most modern dictionaries draw directly on corpus examples, means that it could also be argued that dictionary entries provide a form of concordance line data. In a dictionary, of course, concordance lines are typically presented as individual complete sentences. Lexicographers' tools, such as the Good for Dictionary (GDEX) filter in the commercial online concordancer *Sketch Engine* (Kilgarriff, Husak, McAdam, Rundell, & Rychlý, 2008), provide a means of helping dictionary writers find corpus examples which are considered more suitable for dictionary entries, being short enough to be reproduced as complete sentences and excluding proper names, unusual symbols, low frequency vocabulary items, etc. Corpus consultation in the development of language teaching materials ("a corpus-referred approach") can be a means of both informing the reflective processes and justification of choices in a syllabus, and providing essential insights into the "nature of language and language production" (Timmis, 2003, p. 472).

Within CALL, a concordancer can be seen as a means of meeting important principles from Second Language Acquisition and language teaching pedagogy. Thomas (2015) provides an extensive set of pedagogical concordancing activities, taking language learners on a tour of *Sketch Engine* and demonstrating how a wide range of discoveries about language can be made. Frankenberg-Garcia (2014) has demonstrated that examples from corpora can serve different uses in second language learning for comprehension and production. As the author has argued previously, the SLA principles of exposure to the target language (Krashen, 1989; Nation, 1995-6), focussed attention and noticing (Doughty, 1991; Schmidt, 1990) and independent discovery (Bolitho et al., 2003; Tomlinson, 2008) can be upheld through the aims of concordancing activities (Jeaco, 2015, in press). Based on patterns and claims from Hoey's Theory of Lexical Priming (2005), *The Prime Machine* was designed to provide language learners and teachers with easy access to information about typical contextual environments of words and phrases (Jeaco, in press). Its search screen provides auto-complete functionality for words and collocations, as well as spelling support and prompts for possible words or phrases for the user to compare. The other summary data and indicators are designed to encourage language learners to look for patterns or features within the concordance lines. While the KWIC view (Lines Tab) is an important part of this software, there is also an alternative way to view concordance line data. For all the advantages of KWIC, there is little information about typical positions of the node

in terms of the paragraph or sentence. If KWIC is restricted to only showing complete sentences, some indication that a word or phrase is typically used at the beginning of a sentence may be visible because of the white space to its left, but even in such cases, paragraph breaks are usually masked. The tendencies for words to occur (or not to occur) near the beginning or end of discourse units including the sentence, the paragraph and the text are Hoey's 10th priming hypothesis, and *The Prime Machine* also includes other indicators and measures related to this (Jeaco, in press). Another aspect of the typical context for words and phrases, drawing on Lexical Priming, is the need to consider how primings are "constrained by domain and/or genre" (Hoey, 2005, p. 13). Information about the text from which a concordance line comes is hard to fit into a KWIC display.

At the time when KWIC was developed, computer screens (and printers) had a limited number of characters which could be produced on each line and using computers was essentially an experience of looking at text on screen in a form which was very different from other reading materials like magazines, etc. However, in recent years, the development of the internet and mobile technologies has meant that younger generations are regularly interacting with lists of data in different ways. Rather than scanning down a screenful of cramped data, modern apps and applications present the user with a list box where it is understood that most of the data is off screen "somewhere" and the user will scroll through the pages.

Looking at the wider context is not something new to corpus linguistics or concordancing software. Both *WordSmith Tools* and *AntConc* provide access to the original corpus texts, so it is possible to click through to the source document in each instance and look at where the concordance line occurs. In *Sketch Engine,* two ways of viewing concordance lines are provided: either KWIC or complete sentences. There is also a provision to extend the window so that more co-text for a specific concordance line can be seen. XML tags indicating paragraph breaks are visible, but these do not change the way the text is formatted. Multiple clicks forward and backwards would be needed in these software packages to explore a wider co-text for multiple concordance lines in each set of results. However, when *The Prime Machine* was being designed, a challenge was to find a way to present multiple concordance lines with a much wider context than usual in a way which also facilitated visual scanning of patterns, and at the same time could enjoy many of the benefits of KWIC. The Cards Tab was created to address this challenge.

THE DESIGN OF CARDS

The card design in *The Prime Machine* provides several important features, some of which can be seen in Figures 1, 2 and 3. Firstly, more co-text is provided as the default. This means that the learner can see not just up to 40 characters either side of the node, but the full sentence of the current concordance line plus the one before and the one afterwards. A second important feature relates to the way in which paragraphing and headings are shown on the cards. Since the theory of Lexical Priming shows that position in text is important, if the top sentence is not text initial, three dots (…) indicate more is above. Similarly, at the bottom of the concordance box, if the last sentence displayed is not the last sentence of the text, … is displayed too. Paragraphing is also an important feature of the primings of words. While concordancers like *Sketch Engine* and *WordSmith Tools* can show paragraph breaks as <p> tags, in *The Prime Machine* paragraphs are shown with line spaces and indenting. It was not thought desirable to try to implement full HTML support and be able to show text in exactly the same way as it would appear on a website or on the page of a book. Rather, the Cards provide some additional visual information about the position of words and their adjacent sentences, including how these are connected to paragraphs or headings, while providing some uniformity to the font size, colour, and layout. A range of different corpora have been processed and made available in *The Prime Machine*, including *The British National Corpus* (BNC, 2007) and academic corpora derived from *Hindawi* (2013). The way in which headings are highlighted in the raw text files of corpora such as these varies considerably, yet during the refactoring process tags indicating underlining, bold, italics and

other typeface information is discarded, and only the information about whether it was a heading or standard paragraph is retained. Therefore, if a corpus such as BAWE (BAWE, 2007) was imported, where the text formatting of headings includes simple bold or simple underlining as well as almost every combination of bold, italics and underlining possible, all the headings would actually appear in the same way when viewed in *The Prime Machine*.

One issue regarding cards is that because sentences vary considerably in length, the number of rows of text inside a card also varies considerably. Rather than having white space before and after any shorter blocks, the height of each card is not fixed, so cards with fewer lines occupy less space. A screenshot of the Cards Tab showing the paragraph layout and different heights of cards can be seen in Figure 1, which shows some of the results for *pilot* in the BNC Academic sub-corpus.

As can be seen, complete sentences are shown above and below the sentence containing the node, and the height of cards can vary considerably. One disadvantage of the card design is that since in free-flowing text the node can appear at any position horizontally (that is to say the word-wrapping system is not influenced by the position of the node), scanning to find the node on a set of cards can be quite difficult. To overcome these two options are available for gentle highlighting of the line of text on the card which contains the node. The default option is to highlight the line of text with gentle yellow colour, but it is also possible to select dotted lines or no highlighting. The effect of three options can be seen in Figure 2.

The highlighting is a way of tempting users into looking in more detail at just the words near the node, perhaps also helping to familiarize them with some of the benefits of a KWIC view. Another way in which the nearby context of the node word is highlighted is through the captions. Each of the cards on the Cards Tab view of concordance lines has a caption which highlights the relationship between the concordance line and collocations. Each caption includes the node as well as any significant items from the top 100 two word collocation log-likelihood lists[2]. When the cards are generated, the 4 word window either side of the node is checked to see whether the items are also present in this list. If they are, they are included in the caption, with ".." added between non-consecutive items. The

Figure 1. Cards of different heights on the Cards Tab with captions at the top and gentle yellow highlighting of the line containing the node
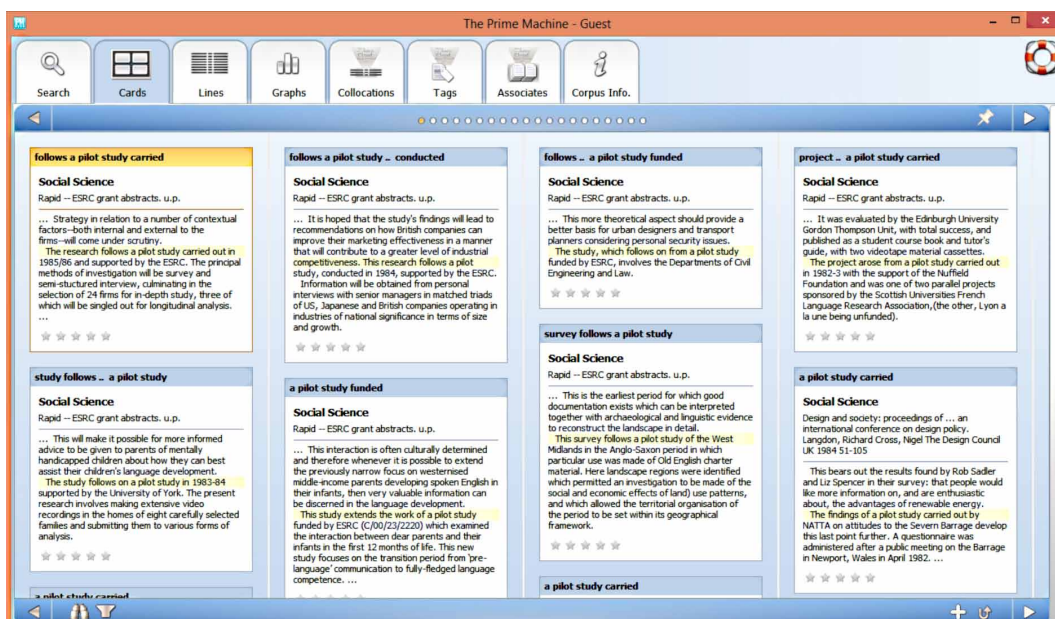
Figure 2. The same card from the Cards Tab, with no highlighting (left), dotted line highlighting (centre) and gentle yellow highlighting (right) of the node word, showing one concordance box from the BNC: Academic sub-corpus for pilot

caption also appears on the card for the currently selected row on the Lines Tab. There are two key benefits of these captions. Firstly, each provides an eye-catching snippet from the concordance line which is essentially a trimmed down KWIC of 4 words either side of the node. Secondly, they help the learner see the main use of the node in each box and should help highlight patterns so users can scan down the list of captions to see which collocations are shown. Figure 3 shows three cards with their captions for a search on the *BNC Academic* sub-corpus for *outcome*.

As can be seen, in the first card the phrase "eventual outcome" appears together on one line, while in the second card "eventual" happens to occur at the end of a line, so the phrase is broken by a line break. The third card shows how the caption may include several words within the 4 word window either side of the node. The caption, therefore, 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. Being able to scan the results quickly to find cards containing specific collocations (and longer units including grammatical words) 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

Figure 3. Collocation-based captions on cards; screenshot showing data from the BNC: Academic sub-corpus for the node outcome

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 additional information about the nearby sentences, paragraphing, and collocations, the Cards also provide information about each source. One of the most basic labels for a concordance line is the name of the text from which it comes. Typically, concordancers do not try to retrieve this information or display it as part of a citation for each concordance line. With *WordSmith Tools* and *AntConc*, since they work directly with text files on the users' hard drive, the header can be displayed as part of the full text. With *Sketch Engine*, by default a code representing the file name is displayed in a column to the left of each KWIC line[3]. The user is expected to know enough about the corpus which they are using to be able to look up information if he or she thinks it is relevant. However, users of concordancers in language learning situations are unlikely to have expert knowledge about the texts and text types which were selected when a corpus was constructed. The design of the database used for *The Prime Machine,* includes fields for a citation style label for each concordance line and a text category label to give a quick sense of the kind of text from which it is taken.

The citation and text category label can be generated in different ways, according to the way in which the corpus has been designed and the metadata available. Some corpora are created from collections of XML files where tags in the file hold suitable titles with which to identify each individual text. Others have several tags which could be used in combination in the style of an academic reference. For example, the *Hindawi* corpus and the *SpringerOpen* corpus have the title of each article, the journal name and information about the volume and issue stored within the file header. Similarly, newspaper corpora usually have information about the name and date of the publication. These fields can be easily combined to match typical referencing conventions. Other corpora, like the *BNC* and *BAWE*, have coded filenames which can be matched to a separate table of sources, giving much more detail than is contained in the file header. Corpora which are based on collections of texts from a specific group of sources can be systematically organized according to different text categories. For most of the corpora that have been loaded into *The Prime Machine*, the major category is set for each text by selecting an XML tag which holds the category name, or by entering a default category to be used for all texts in the corpus. The British National Corpus (BNC, 2007) is loaded both as one corpus, with text categories based on the main groupings from Lee (2001), and also as separate sub-corpora, using the sub-groupings as text categories. For corpora like *Hindawi* (Hindawi, 2013), decisions about how to organize texts into categories may be more complicated. However, the refactoring application which is used to import texts into *The Prime Machine* database structure allows for a lookup table to be loaded and a further option is provided to select an XML tag to use as the value for the lookup table. The lookup tables for each *Hindawi* sub-corpus contain a list of journal titles and categories[4]. Figure 4 shows one card from the *BNC* and one card from a *Hindawi* corpus with this citation information shown at the top.

Since this information does increase the height of each card substantially, not all users may want to keep the card citation information visible, and it can be turned off through clicking on a button on the task bar, or through the Options Tab. Figure 3 was created with the citation information hidden.

Since many language classrooms are equipped with data projectors and such office productivity software as *Microsoft Office* is frequently used to create slides or word-processed hand-outs, the development of the facility to generate images and text data which can be easily imported into other applications was a priority. In *The Prime Machine*, when the user double-clicks on a results page, a menu appears which provides options to copy or save the results in different formats. Copying or saving as "Picture" files allows the user to import the results as an image and is the easiest way to incorporate concordance cards, lines, tables or clouds if the size does not need to be changed dramatically. However, the MetaObject format provides an alternative way to copy the data which allows the target application to use its own drawing processes for text and vector graphics, meaning if the size of the font is increased after the image has been copied, the letters and symbols used in the text are drawn smoothly. Cards can also be saved in multiple files using the "Save all…" options. For

Figure 4. Citation information displayed at the top of a card from the Hindawi Biological Sciences corpus (left) and the complete BNC corpus (right), as shown on the Cards Tab
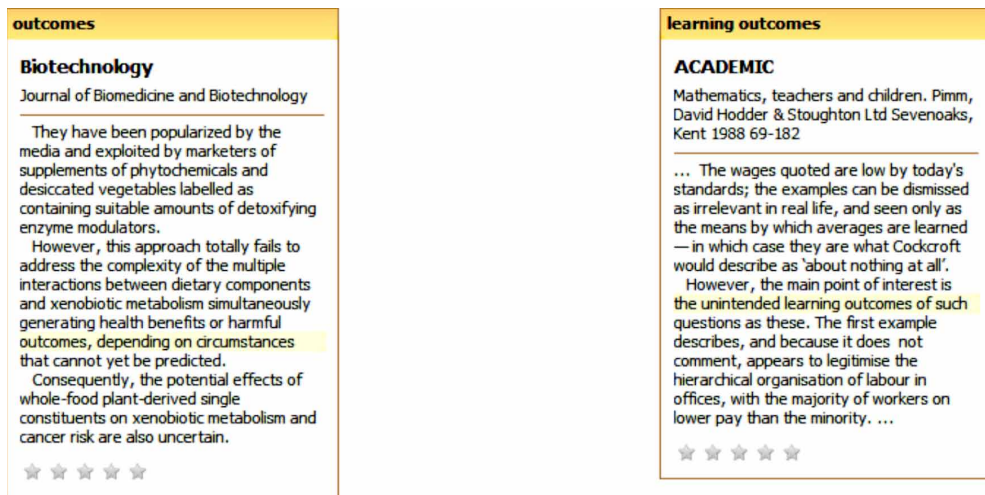
table data, a further option is to copy or save the results so they can be imported into a spreadsheet. Figure 5 shows the export options for the Cards Tab.

EVALUATION

Evaluation of the cards has taken place within three broad research projects. The first two of these studies used the same methodology, drawing on responses to a pair of questionnaires administered to participants before and after they had made direct use of the software package and complementing these data with logs from actual use of the software. The third study provided groups of language teachers and language learners with printed materials presented in both Lines and Cards format, and required participants to rate the effectiveness of several concordance line ranking methods before

Figure 5. Export options available for cards
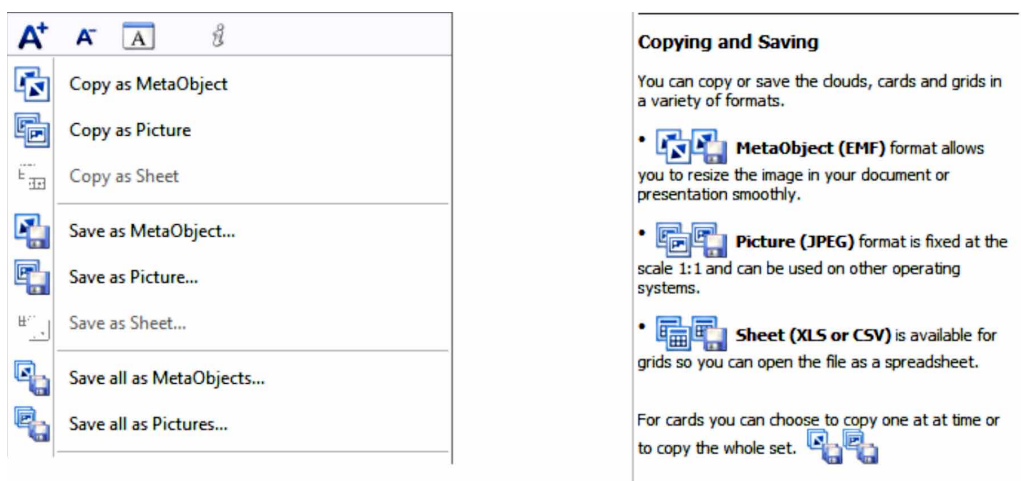
answering some questions in a questionnaire about the Cards and Lines. While the main aims of each of these studies were to look at a broader set of research questions and software features, each included key elements permitting the exploration of the following research questions:

1. How do the attitudes of potential users of *The Prime Machine* towards the Cards concordance line display compare and contrast with their attitudes to the KWIC Lines view?
2. What evidence is there (from logs and/or reported use) that Cards have provided a useful and complementary means of presenting concordance data when participants complete different kinds of task?

Study #1

Participants in the first study were volunteers from a university in Eastern China who were not studying any modules taught by the researcher, and not using the concordancing software as part of their studies. This study was the evaluation which formed part of the doctoral studies of the researcher/ developer. The materials for the evaluation were two online questionnaires, an essay prompt designed to be similar to popular language learning examinations, and *The Prime Machine* Version 1. The first questionnaire included demographic questions as well as those designed to ascertain the attitudes of the participants towards different language learning resources such as English-Chinese/Chinese-English dictionaries, monolingual dictionaries, electronic dictionaries and other reference books and their prior experience with reference tools including prior use of concordancers. After completing the first questionnaire, the software was demonstrated and the participants wrote an essay, making use of the software and other resources as they saw fit. The demonstration included how to explore collocations and colligations through looking at concordance lines on the Cards and Lines Tabs, particularly looking at differences between synonyms and differences in the use of some words across academic and non-academic corpora. After this face-to-face session, students received feedback from the researcher containing comments related to different aspects of the task, and some concordance lines presented in Lines format for at least two words or phrases related to word choice in their essay. Participants were then encouraged to continue to use the software independently, before completing the second questionnaire which asked for their views on the usefulness of the software generally, as well as specifically asking about the usefulness of different Tabs, including the Cards Tab. During the face-to-face session and over the following days, the software also automatically collected logs to record actions and time spent using a wide range of the software's features. These included the time spent looking at the Lines and Cards Tabs, and the number of concordance lines viewed. Consent for the use of questionnaire and log data was obtained, following the ethical procedures and policies at the University of Liverpool and the institution where the researcher worked. For this first study, unique identifiers were used to match logs from the software to the individual responses to the first and second questionnaires. A fuller discussion of the results of this initial evaluation is presented elsewhere (Jeaco, 2015); here, the feedback and log data related to the Cards Tab will be reported.

The number of students participating in the face-to-face study was 25, with 23 students completing the second questionnaire. They were all undergraduate students from Year 1, 2 or 3 and all were Chinese. It was evident from responses to the first questionnaire that most students (72%) claimed never to have used concordancers before. Five of the seven students who indicated they had used concordance lines selected "not often". In the second questionnaire, it was possible to compare student perceptions regarding the Cards Tab with the other tabs, as well as some aspects of their attitudes towards the software package as a whole. Figure 6 shows the responses to a question about the usefulness of the different tabs in the software from the first and second study. In the first study, the participants gave a clear indication of positive attitudes towards both the Lines and Cards tabs, with almost three-quarters of those who completed the second questionnaire indicating they were "Useful" or "Very Useful". While attitudes towards the Cards Tab are more mixed, and include

two responses of "Not Very Useful", looking at individual responses it became clear that more than a quarter of the participants gave a higher rating for the Cards Tab than they did for the Lines Tab. This seems to suggest that making both Cards and Lines displays available can accommodate different learner preferences and different uses of concordance line data. The log data was also able to confirm some of the preferences and look up habits of the participants in relation to the Cards and other views, during and after the face-to-face session. From the logs, in terms of the number of views, on average Cards were viewed more often than Lines (59:41), while in terms of time spent, on average Cards were viewed for a shorter duration than Lines (41:59). However, one limitation of this first study was that it was evident from the date stamps in the logs that the vast majority use of the software was during the face-to-face session.

Study #2

Following the completion of my doctorate, it was possible to incorporate the use of the software into two undergraduate modules for English majors, providing a new alternative for students to use for corpus-related work. For the Year 2 students, computer workshops on *The Prime Machine* were provided as part of a reflective coursework assignment where they were required to analyse word choice, collocations and other relevant features of their own timed writing or the transcription of their own recorded speech, drawing on corpus data to help them find alternative wordings and to check patterns of use. The Year 4 students were mainly using *The Prime Machine* as a way of exploring the norms and typical patterns of use within literature so as to provide concrete examples of how language use in their chosen literary passages compared or contrasted with the use of these words and phrases in other literature. For the second study, the same questionnaire and log research design could be used, but over a longer period of time, with students accessing Version 2 of the software. This project conformed to the ethical requirements of the university, with informed consent being built into the start-up page of the software and the questionnaires, and with an extra level of anonymity since the modules had assessed coursework projects requiring students to draw on corpus data from the software or other software packages. This meant that the data collected in the first questionnaire could not be matched to individual responses to the second questionnaire or to the logs, which were stamped with a temporary anonymous identifier.

For the first questionnaire, from Year 2 there were 48 participants and from Year 4 there were 19 participants, all of whom were studying a four-year undergraduate English programme. Participation in the second questionnaire was rather more limited, with 12 students from Year 2, and 14 from Year 4. As before, the feedback from the first questionnaire revealed that very few students in Year 2 had used concordancers before; 36 (75%) indicated that they did not know anything about this kind of software, four (8%) indicated they had only used it once or twice, and a further seven (15%) indicated they had only used it a few times. One student claimed to know how to use most functions on a concordancer, but named an online Chinese-English dictionary rather than a corpus tool. While seven of the Year 4 students had never used a concordancer (37%), and two had only used one once or twice, 10 had used a concordancer a few times (53%). They reported that they had used Antconc (63%), WordSmith Tools (16%) or other software packages. Responses to the second questionnaire (n=26) showed that student attitudes regarding the usefulness of the Cards and Lines tabs were again quite different. While attitudes towards the Lines tab were more positive overall, 58% of the participants rated the Cards Tab as being "Very useful" or "Useful". Well over half of the participants (62%) rated the Lines and Cards tabs equally. Although it was not possible to link individual responses from the questionnaires to the logged actions, the log data collected over this period also showed a split between the number of times the Cards and Lines Tabs were viewed. Excluding data for results viewed for less than 10 seconds, the ratio of Cards to Lines views was roughly 3:7 of the 3,277 logged concordance line views over the period of the study.

Figure 6. Evaluation of the Usefulness of the Lines and Cards tabs in Study #1 and Study #2
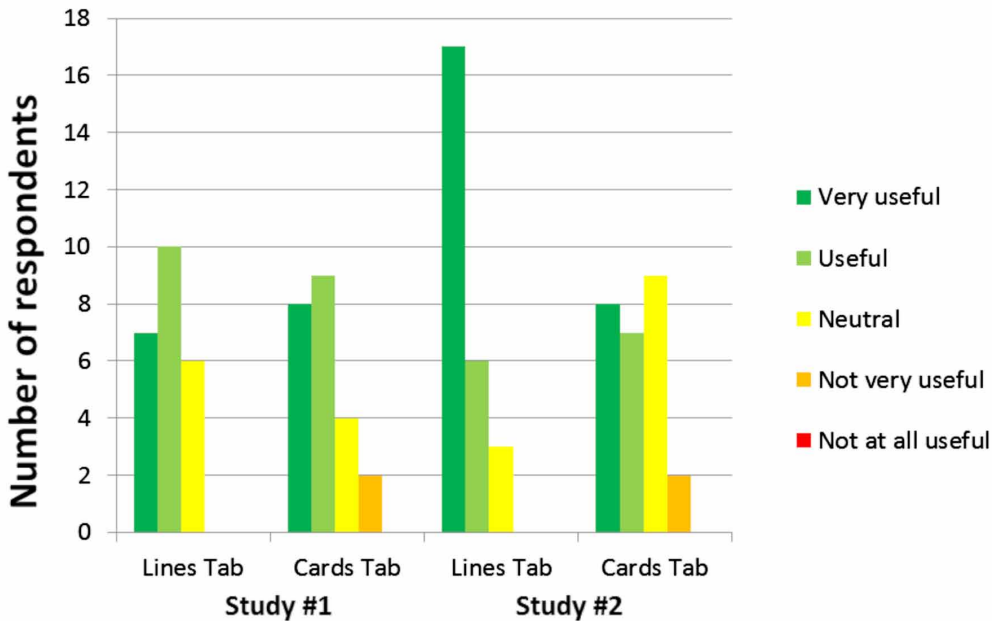
Study #3

The main focus of the third study was on attitudes of language teachers (teaching English for Academic Purposes across several programmes at the university) and language learners majoring in English towards three different concordance line ranking methods. Materials for this study were all paper-based, with a questionnaire containing demographic questions, questions about prior use of concordancers, the main concordance line rating activity, and some closing questions about perceived usefulness of the Cards and Lines printouts. 10 sets of concordance lines were laid out in a large room, with each set comprising separate printouts of Cards and Lines for the same words or collocations, sorted using four different concordance line ranking methods. The language teachers and language learners attended separate research sessions, with the questions at the beginning and end of the questionnaire being slightly reworded to reflect how, as different types of potential user of the software, they might access concordance data like these with their students (for the teachers), or for their own language study (for the students). The central task for both groups was to consider the usefulness of the Cards and Lines printouts in terms of how they would help in three different situations: general comprehension of the term, productive use of the term for a specific task and productive use of the term more generally. Thus, the questionnaire asked the respondents to consider how concordance output might be used for different purposes in an imagined language learning situation.

The results for each of the ranking methods were presented in a conference paper (Jeaco, 2016), and have led to further developments for Version 3 of the software. The research design, however, also enabled useful information to be gathered about the perceptions of both teaching staff and students towards the Lines and Cards.

Thirteen teachers participated in the teacher session. The group had a wide range of teaching experience (between less than three years to more than 10 years) and included both native and non-native speakers of English, and both male and female staff. Almost all the participants reported having little or no experience with concordancers, with only three reporting they had used a concordancer

"once or twice", one reporting they had used one "a few times", and one indicating he had could perform basic functions on a concordancer". One participant noted in the comments box that she had read about concordancers but not used them. None of the participants chose options indicating they could use "most functions" or were an "expert".

Fourteen students participated in the student session. All were Chinese, all in their second year of one of the university's English programmes, and all volunteers were female. Because some use of concordancing software was part of their undergraduate programmes, most had used a concordancer at least "once or twice" before, with nine having used *The Prime Machine.* However, only one student reported being able to perform "most functions on a concordancer", with four reporting their use "basic", a further four reporting only having used concordancers "a few times", three "once or twice" and two "never".

Differing attitudes towards the Cards and Lines were clearly evident in the participants' responses to the question "Overall, which way of looking at the concordance lines did you think was the most helpful?". As can be seen in Figure 7, only one participant from the teacher group and one participant from the student group indicated no preference.

Another interesting result was that in response to a general open-ended question at the end of the questionnaire, requesting any other comments or suggestions, a number of students and teachers commented directly on the Cards and Lines displays.

… cards were far better as lines don't have subject and journal information…
Cards more useful as students get more input by reading and also can see the subject area and citations. I like the idea of seeing citations as they need more practice in this.
I tended to look at the highlighted line on the card and the collocations; looking at the complete sentence could be distracting for students if there are many words they don't know…
At first I thought the cards were better but when my brain caught up, the lines were easier to check through.
I found it difficult to judge the sets [for the ranking task], but I found the lines easier for identifying usage

Figure 7. Attitudes of Students and Teachers Towards Cards and Lines Overall
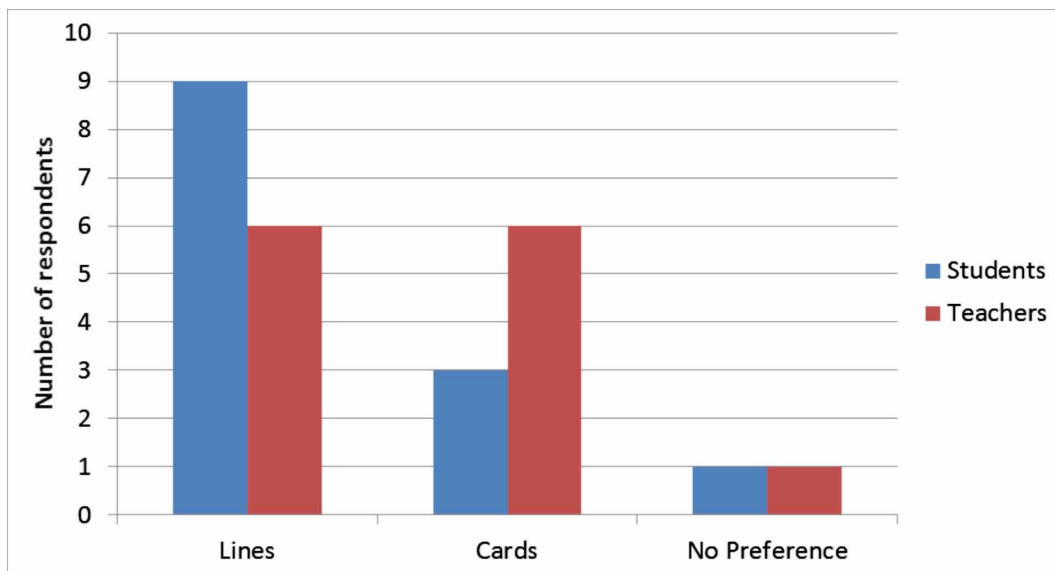
More context on the cards helps comprehension; some interesting patterns emerged with the … cards, but very often would probably not be overly beneficial; as a native speaker though, the KWIC lines were much faster to evaluate.
Lines are good for quick idea of where and how; but cards good for more context so combination of the two would be ideal.

As can be seen, some teachers commented specifically on some of the additional information available in the cards (subject areas, citations, the highlighted line and collocations), while others noted potential advantages and disadvantages of having wider contexts. Some preferences for Lines were also revealed here, and it was good to also see the comment demonstrating that Cards had been helpful at the beginning of the task. The teacher who made this comment had indicated he had used concordancers only once or twice before: both *Sketch Engine* and *WordSmith Tools*.

Comments and suggestions related to Cards or Lines from the students also revealed some differing attitudes, with one student commenting on the Lines being "too crowded", one student stating simply "Cards are better than lines", and two students making suggestions about showing complete sentences, one related to making lines longer, and the other related to making cards shorter.

In the concordance ranking evaluation task, for each set of concordances which were evaluated, participants had to indicate whether their judgement was mainly based on the Cards or KWIC Lines displays. Figure 8 shows the number of responses (from up to 10 sets of lines from each participant) where KWIC Lines or Cards were the basis of their decision on the best ranking method.

As can be seen, the Lines and Card printouts were used fairly evenly across groups. Within the teacher group, four participants selected Cards for every response, three participants selected Lines for every response except one, while the other six participants made fairly even use of each display type. In the student results, not all participants responded to all the items, but three students selected Cards for every response except one, one student selected Lines for every response except one, while the other results were more mixed. As with the previous two studies, attitudes towards the usefulness of looking at concordance lines like those presented for language learning and teaching were generally

Figure 8. Reported Use of Lines and Cards Printouts for a Rating Task
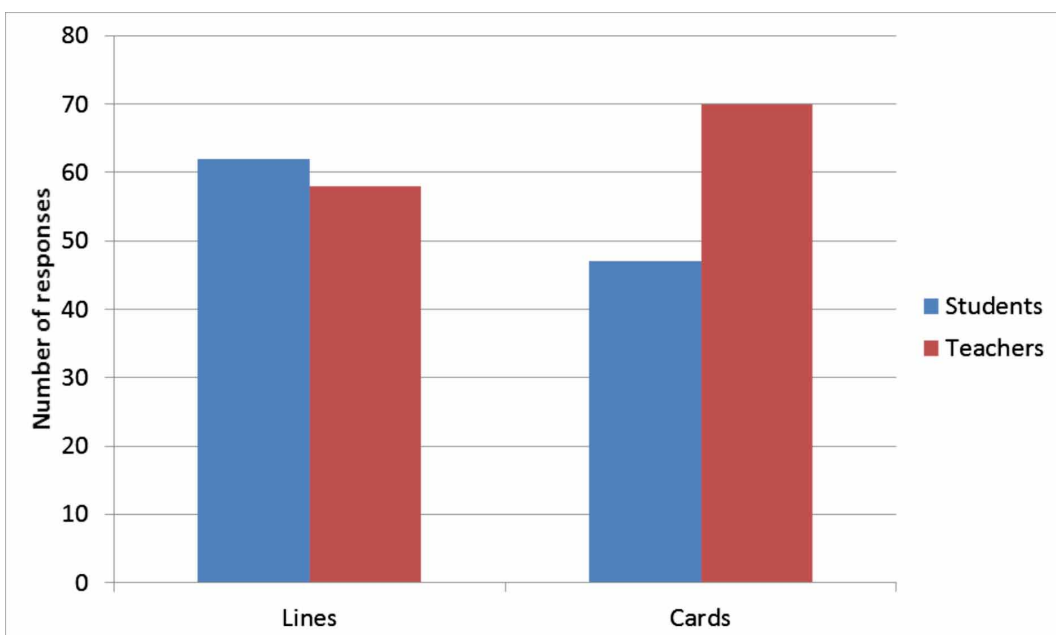
very positive. Within both groups, most participants indicated they through it would be "Quite useful" (10/13 teachers and 11/14 students), while one student and two teachers said they thought it would be "Very useful". Just one teacher selected "Not very useful", and one student selected "Not at all useful", yet both of these respondents selected "Yes, definitely" when asked whether they would be interested in hearing more about how software like this could be used in language learning and teaching.

DISCUSSION

The results of the three studies combined seem to indicate that the Cards display does provide a useful complementary way of displaying concordance line data. The three studies were based around different language learning activities. The first focused on using a concordancer as reference while drafting an essay. For the second study, the focus was either on using a concordancer to complete projects reflecting on language use and alternative wordings in the students' own writing and speech or to explore norms and patterns of use in literary texts. The third study focused on rating concordance line printouts according to potential use for comprehension or production. In response to the first research question, the fact that in all the studies different preferences and different attitudes were obtained from different participants indicates that some users would value the additional display more than others. By providing an additional format for the display of concordance lines, it can be argued that *The Prime Machine* provides an additional way of catering to the differing needs and preferences of language learners and teachers. To address the second research question, the results from the logs (studies 1 and 2) and reported use (all 3 studies) suggest that participants are finding the different ways of viewing concordance data more or less useful for different tasks. When the software is used directly, as mentioned above, the currently selected concordance line on the KWIC Lines view is also visible as a Card, and so it is not necessarily a question of having one but not the other. The KWIC Lines view was clearly a popular choice, and so it is not being suggested here that Cards should be a replacement, but there does seem to be sufficient evidence to suggest that Cards could be a useful addition.

Future Work

At the time of writing, work is progressing on a cross-platform version of *The Prime Machine* which will run on Mac OS, iOS and Android. The previous versions have been restricted to Windows only, and while the new version will bring wider access for different platforms and situations, some of the visual design is being developed to suit both large sized tablets and smaller tablets and mobile phones. One change in the design has been to allow Cards to be cropped back to show only the sentence containing the node (while retaining paragraphing information). Other developments relate to the way in which the concordance ranking mechanisms are implemented with more flexibility on the client side of the application for local sorting and re-sorting. The software will soon become available to users outside the host institution. For more details, see www.theprimemachine.com.

CONCLUSION

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 and headings. It is hoped that more users of *The Prime Machine* will discover that the Lines Tab and the Cards Tab provide different and complementary layouts of the concordance data, drawing attention to different aspects of the contextual environment and making different features more noticeable. Designers of other concordancers and CALL systems which incorporate concordance line data could consider offering their users an additional Card-like way to view results too.

ACKNOWLEDGMENT

REFERENCES

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

Anthony, L. (2006). Developing a Freeware, Multiplatform Corpus Analysis Toolkit for the Technical Writing Classroom. *IEEE Transactions on Professional Communication*, *49*(3), 275–286. doi:10.1109/TPC.2006.880753

Anthony, L. (2011). AntConc (Windows, Macintosh OS X, and Linux) Build 3.2.4 Readme file.

Baker, M. (1993). Corpus Linguistics and Translation Studies: Implications and Applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233–250). Philadelphia: Benjamins. doi:10.1075/z.64.15bak

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279. doi:10.1093/ijl/6.4.253

BAWE. (2007). British Academic Written English Corpus. Retrieved from http://ota.ahds.ac.uk/headers/2539.xml

Biber, D., & Conrad, S. M. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511814358

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

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

Bolitho, R., Carter, R., Hughes, R., Ivanič, R., Masuhara, H., & Tomlinson, B. (2003). Ten questions about Language Awareness. *ELT Journal*, *57*(3), 251–259. doi:10.1093/elt/57.3.251

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572. doi:10.1111/j.1467-9922.2010.00566.x

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

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

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

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

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

Doughty, C. (1991). Second Language Instruction Does Make a Difference. *Studies in Second Language Acquisition*, *13*(4), 431. doi:10.1017/S0272263100010287

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

Fischer-Starcke, B. (2010). *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London: Continuum.

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, *26*(2), 128–146. doi:10.1017/S0958344014000093

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319. doi:10.1016/j.system.2004.04.001

Guan, X. (2013). A Study on the Application of Data-driven Learning in Vocabulary Teaching and Leaning in Chinas EFL Class. *Journal of Language Teaching & Research*, *4*(1), 105–112. doi:10.4304/jltr.4.1.105-112

Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, *25*(4), 398–436. doi:10.1093/ijl/ecs026

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

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

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge. doi:10.4324/9780203327630

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with an interactive on-line database. *Language Learning & Technology*, *9*(2), 90–110.

Hu, K. (2016). *Introducing Corpus-Based Translation Studies*. Shanghai, Heidelberg, Berlin: Shanghai Jiao Tong University Press / Springer. doi:10.1007/978-3-662-48218-6

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

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. (2016). Get in line and show me what you've got: Exploring concordance line ranking and selection methods for different language learning goals. *Paper presented at the 12th Teaching and Language Corpora Conference*, Justus Libig University, Giessen.

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

Jin, T., Guo, K., Mak, B., & Wu, Q. (2017). Lexical Profiles of Reading Texts in High-Stakes Tests: Where are the Benchmarks? *International Journal of Computer-Assisted Language Learning and Teaching*, *7*(1), 34–49. doi:10.4018/IJCALLT.2017010103

Johns, T. (1994). From printout to hand out: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on Pedagogical Grammar* (pp. 293–313). Cambridge: Cambridge University Press. doi:10.1017/CBO9781139524605.014

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

Kaur, J., & Hegelheimer, V. (2005). ESL students use of concordance in the transfer of academic word knowledge: An exploratory study. *Computer Assisted Language Learning*, *18*(4), 287–310. doi:10.1080/09588220500280412

Kenning, M.-M. (2000). Concordancing and comprehension: Preliminary observations on using concordance output to predict pitfalls. *ReCALL*, *12*(02), 157–169. doi:10.1017/S0958344000000422

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Paper presented at the Euralex*, Barcelona.

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.

Krashen, S. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the Input Hypothesis. *Modern Language Journal*, *73*(iv), 440–464. doi:10.1111/j.1540-4781.1989.tb05325.x

Kuo, C.-H. (2008). Designing an Online Writing System: Learning with Support. *RELC Journal*, *39*(3), 285–299. doi:10.1177/0033688208096842

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 & 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.

Li, W., & Smith, S. (2015). Introduction. In B. Zou, M. Hoey, & S. Smith (Eds.), *Corpus linguistics in Chinese contexts* (pp. 1–14). Houndmills, Basingstoke, Hampshire: Palgrave Macmillan. doi:10.1057/9781137440037_1

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: A usage-based approach. *Journal of Child Language*, *30*(2), 333–370. doi:10.1017/S0305000903005592 PMID:12846301

Luo, Q., & Liao, Y. (2015). Using Corpora for Error Correction in EFL Learners Writing. *Journal of Language Teaching & Research*, *6*(6), 1333–1342. doi:10.17507/jltr.0606.22

MacWhinney, B. (2000). Childes Project: Tools for Analyzing Talk (Vol. I. *Transcription format and Programs)*. London: Routledge.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. New York: Routledge.

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. doi:10.1163/9789004334236_011

Mondria, J.-A., & Wit-de Boer, M. (1991). The effects of contextual richness on the guessability and the retention of words in a foreign language. *Applied Linguistics*, *12*(3), 249–267. doi:10.1093/applin/12.3.249

Nation, I. S. P. (1996). Best practice in vocabulary teaching and learning. *EA Journal, 3*(2), 7-15.

Oghigian, K., & Chujo, K. (2010). An effective way to use corpus exercises to learn grammar basics in English. *Language Education in Asia*, *1*(1), 200–214. doi:10.5746/LEiA/10/V1/A17/Oghigian_Chujo

Renouf, A. (2007). Corpus Development 25 Years On: From Super-Corpus to Cyber-Corpus. In R. Facchinetti (Ed.), *Corpus Linguistics 25 Years On* (pp. 27–49). Amsterdam, Netherlands: Rodopi. doi:10.1163/9789401204347_004

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158. doi:10.1093/applin/11.2.129

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

Scott, M., & Johns, T. (1993). *MicroConcord*. Oxford: Oxford University Press.

Semino, E., & Short, M. H. (2004). *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing*. Abingdon: Routledge.

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.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*. *Canadian Modern Language Review*, *64*(3), 429–458. doi:10.3138/cmlr.64.3.429

Teubert, W. (2004). Units of Meaning, Parallel Corpora, and their Implications for Language Teaching. *Language and Computers*, *52*(1), 171–189.

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

Thompson, G. (2004). *Introducing Functional Grammar* (2nd ed.). London: Arnold.

Timmis, I. (2003). Corpora and Materials: Towards a Working Relationship. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 461–474). London: Continuum.

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.

Wei, N., & Li, J. (2013). A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics*, *18*(4), 506–535. doi:10.1075/ijcl.18.4.03wei

Xia, L., Xia, Y., Zhang, Y., & Nesi, H. (2016). The Corpora of China English: Implications for an EFL Dictionary for Chinese Learners of English. *Die korpora van China-Engels: Implikasies vir 'n EVT-woordeboek vir Chinese aanleerders van Engels.*, *26*, 416-435.

Xiong, W., Xu, J., & Liang, M. (2014). An architecture for automatic opinion classification in Western online news. In *Proceedings of the 2014 IEEE Workshop on Electronics, Computer & Applications* (pp. 717-721). doi:10.1109/IWECA.2014.6845722

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. doi:10.1080/09588220701331451

Yoon, C. (2011). Concordancing in L2 writing class: An overview of research and issues. *Journal of English for Academic Purposes*, *10*(3), 130–139. doi:10.1016/j.jeap.2011.03.003

Yu, L.-T., Liou, H.-C., Chang, J. S., & Vongpumivitch, V. (2011). Integrating Concordancing into Vocabulary Learning for EFL Primary School Students. *English Teaching & Learning*, *35*(1), 47–84.

Zhang, M. (2013). The Application of Corpus Tools in the Teaching of Discipline-Specific Academic Vocabulary: A Case Study for Information Engineering Undergraduates. *International Journal of Computer-Assisted Language Learning and Teaching*, *3*(4), 33–47. doi:10.4018/ijcallt.2013100104

Zhang, W. (2015). The shifting representation of common people in Chinas news media. *Journal of Language & Politics*, *14*(2), 285–307. doi:10.1075/jlp.14.2.05zha

ENDNOTES

1. The number of letters visible to the left and right varies according to the screen width, font size and software settings. The defaults in some major concordancers are a total width of about 60 letters in *AntConc* (Anthony, 2011) and 80 letters in *WordSmith Tools* (Scott, 2010), while the default in Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004) is for 40 to 50 letters either side of the node.
2. Log-likelihood is one of the earliest measures for collocation and is built into most concordancers as a statistical measure for collocations, however the parameters used in the calculation for *The Prime Machine*, and the criteria for inclusion in a collocation list are different. More details in Jeaco (2015).
3. It is possible to change the field displayed to the left of the concordance lines using the View Options in *Sketch Engine,* so users could change the default setting to show information about the author, document title or the full header, for example.
4. For more details about how Hindawi journals were categorized, using data from its website, Intute and in one case advice from a specialist, see Jeaco (2015).

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.