REVIEW OF THREE SOFTWARE PROGRAMS DESIGNED TO IDENTIFY LEXICAL BUNDLES

Title	*KfNgram 1.2.03*	*N-Gram Phrase Extractor (Compleat Lexical Tutor 4.0)*	*Wordsmith Tools 3*
Platform	PC (download)	PC (use on Web site)	PC (download)
Minimum hardware requirements	No information provided	Windows or Linux	Windows 98, 2000, and XP
Publisher	William H. Fletcher http://kwicfinder.com/kfNgram/kfNgramHelp.html	Tom Cobb http://www.lextutor.ca/	Mike Scott http://www.lexically.net/wordsmith/index.html
Support offered	Brief manual provided on the software's web site	Directions provided on each screen. Contact: http://www.lextutor.ca/mailer	Online manual provided on the website
Target language	English	English and French	English
Target audience	Beginning to advanced users	Beginning to advanced users	Beginning to advanced users
Price	Free	Free	License for a single user is currently around £50 (approx. US$92 or €75); a license for up to 10 users is around £250 (US$460, €376) and for up to 50 users around £500 (US$919 ,€753).

Review by Omer Ari, Georgia State University

OVERVIEW

Three software programs--N-Gram Phrase Extractor, kfNgram, and Wordsmith Tools--are reviewed in terms of their user-friendliness and efficiency for searching for lexical bundles, which are recurring chunks of words in text. User-friendliness is defined as the ease in operating the interface of the program; efficiency is defined as fulfilling the criteria by which word combinations qualify as lexical bundles, such as frequency and multi-text occurrence (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber, Conrad, & Cortes, 2004).

Of the three software programs, N-Gram Phrase Extractor is the most user-friendly program and could be used by language teachers and learners for information on raw frequency of lexical bundles. kfNgram has an easy-to-use main interface and could also be useful to language teachers and learners. kfNgram and Wordsmith Tools additionally provide information on raw frequency and are more efficient than N-Gram Phrase Extractor. kfNgram and Wordsmith Tools could be used by researchers and others interested in multi-text occurrence as well as raw frequency information. Specific details are explained further below.

BRIEF BACKGROUND ON LEXICAL BUNDLES

Corpus investigations of natural language data have resulted in major changes in the way language is viewed. Using specially developed software, researchers have discovered frequently recurring multiword lexical chunks in texts or corpora (Biber et al., 1999; Cortes, 2004; Sinclair, 1991; Stubbs & Barth, 2003), indicating that language is more repetitive than has been assumed. What is more, these chunks have been shown to vary across registers, i.e., conversation, academic prose, newspapers, fiction, etc. (Pawley & Syder, 1983; Stubbs & Barth, 2003). Although findings regarding frequency and variation have gained consensus among researchers, defining what counts as a chunk has met with broad disagreement. As a result, the field has seen a plethora of labels for chunks, such as lexical bundles (Biber et al., 1999), prefabs or lexical phrases (Nattinger & DeCarrico, 1992), formulaic sequences (Schmitt & Carter, 2004), and sentence stems (Pawley & Syder, 1983).

Biber and his colleagues (1999, 2004) postulated a set of defining criteria to identify register-bound lexical bundles. Accordingly, there are two fundamental criteria for a multi-word combination to be considered as a lexical bundle: (a) it must occur frequently in a register, and (b) it must occur in multiple texts in that register. Frequency cut-off points for both criteria have usually been determined based on the researchers' goals. For example, Biber et al. (1999) set out their register-based research with a very flexible cut-off point of ten in one million words. Biber, Conrad, & Cortes(2004), however, were more conservative in their search for lexical bundles, using the criteria of 40 in one million words. Cortes (2004), on the other hand, opted to set the cut-off point in her data at 20 in one million words.

The second criterion of multi-text occurrence was intended by Biber et al. (1999) to avoid idiosyncratic uses of lexical bundles by individual speakers or writers in a given register. Multi-text occurrence thus assumes that a lexical bundle is shared by other members of the discourse community who communicate in that register. Working with small corpora may make it difficult to apply this criterion due to limited availability of different texts. This criterion has largely been ignored in the search for lexical bundles, mainly because raw frequencies satisfied researchers' purposes or because researchers did not have access to large corpora.

DESCRIPTION OF SOFTWARE PROGRAMS AND COMPARATIVE EVALUATION

The three software programs designed to help researchers and teachers search for lexical bundles that are reviewed here are: kfNgram, a free downloadable software program; N-Gram Phrase Extractor, part of the online corpus tool Compleat Lexical Tutor; and Wordsmith Tools, a downloadable software program available for purchase. The programs are reviewed for their efficiency and user-friendliness (see Table 1). To reiterate, software efficiency is defined as a program's capability to identify lexical bundles in running text by frequency and multi-text occurrence; and user-friendliness is defined as the ease with which the program can be used by a user who may have little experience in using computers.

Table 1. The Software Programs Rated for their Ability to Perform Various Tasks

	kfNgram	N-Gram Phrase Extractor	Wordsmith Tools
analyzing a long text	+	-	+
analyzing multiple texts	+	-	+
reporting frequency	+	+	+
determining multi-text occurrence	+	-	+
user-friendly	+	+	-
efficient (frequency and multi-text occurrence)	+	-	+

kfNgram

kfNgram is a user-friendly tool. After the user adds a text file into the input field, he or she has to select only the desired length of particular lexical bundles and the floor, the minimum frequency of occurrence, in the corpus or text. The search takes place on kfNgram's single interface and does not require additional page viewing or operations. The status of the operation is reported in the output field on the main interface and the results are displayed in a new window for each file with frequency numbers aligned to the right. The following are screenshots of the software: Figure 1 shows the main interface; Figure 2 shows an outcome window displaying 4-word lexical bundles occurring at least three times in a section of *Alice's Adventures in Wonderland* by Lewis Carroll (Carroll, 1994).

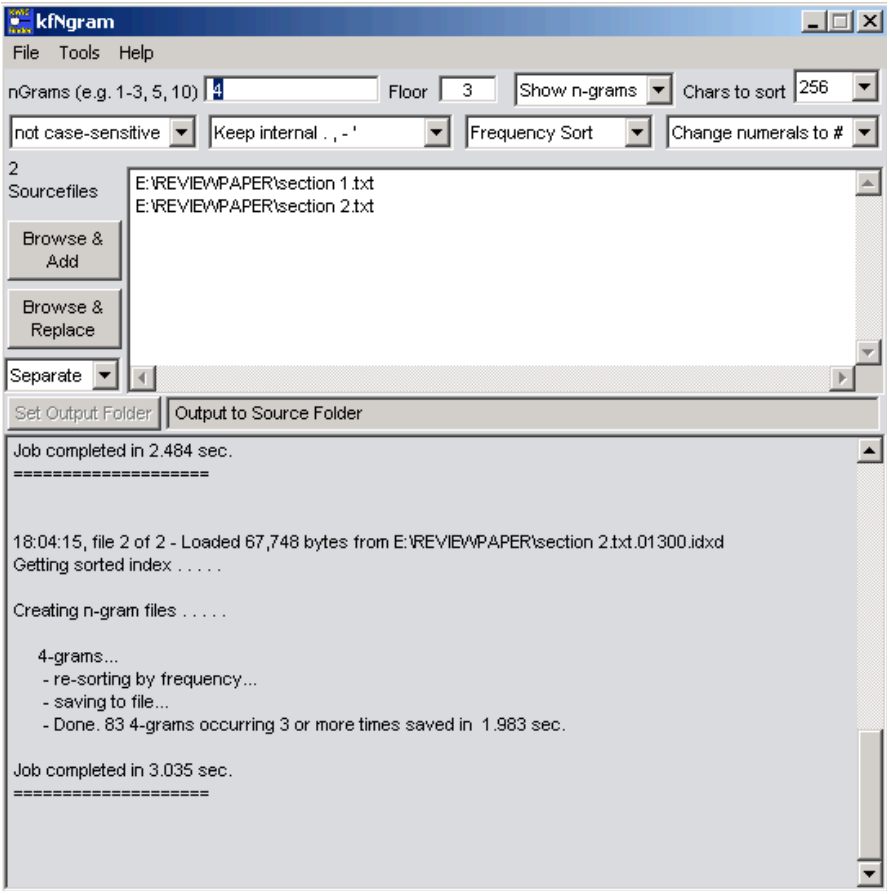


Figure 1. The main interface of kfNgram

Although a very effective tool in extracting repeating lexical bundles with frequency numbers, kfNgram's functions are limited. To determine multi-text occurrences of a lexical bundle, the user/researcher has to enter the texts separately and compare lexical bundles across outcome windows for each text. Therefore, determining multi-text occurrences requires entering the text in subtexts or a corpus in a number of sections.

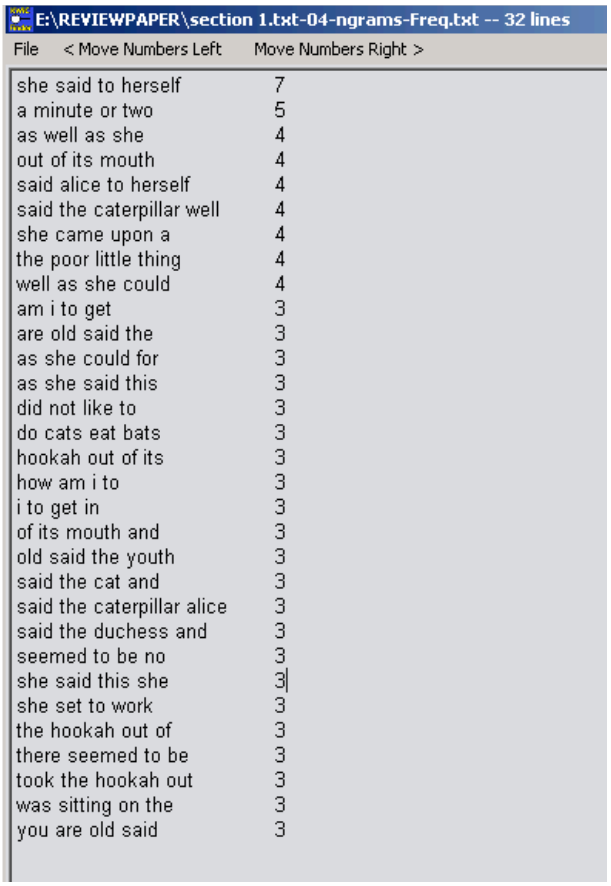


Figure 2. Output window displaying 4-word lexical bundles in the first half of Alice's Adventures in Wonderland using kfNgram

N-Gram Phrase Extractor

This software is available as as part of the website Compleat Lexical Tutor. N-Gram Phrase Extractor analyzes a given text - the shorter, the better, extracting recurring phrases and displaying the output in varying spans of co-text (usually 17-20 words) with the phrases centered and listed in alphabetical order. Information about how many times a phrase occurs in the text is reported to the left of the page with phrases listed alphabetically. There is no information about multi-text occurrences of phrases, however, or in how many different texts a phrase appears (see Figure 3).

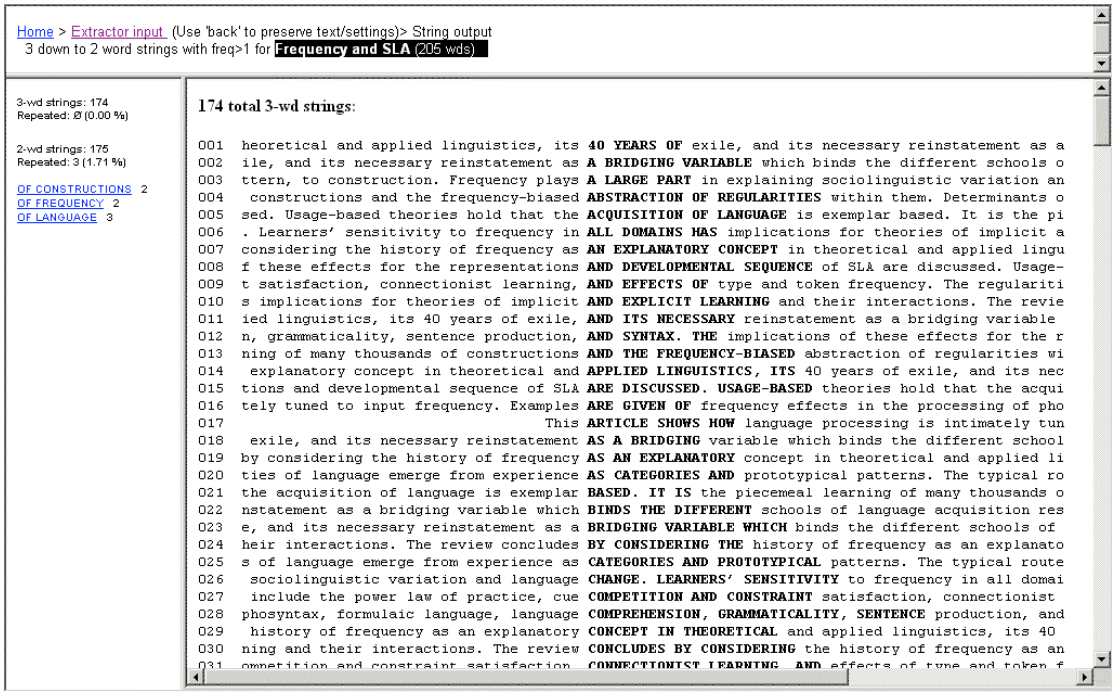


Figure 3. The output display of N-Gram Phrase Extractor for 4-word phrase size

Since the software cannot process texts larger than 1MB, it may not be as useful for a researcher as it may be for a teacher interested in his/her students' use of lexical bundles in their writing. The interface, however, merits more praise, for all that needs to be done is to paste a text into the input field or upload a plain text file, click the desired length of lexical bundles (i.e., 2, 3, 4, 5), and hit the submit button (see Figure 4). However, it is not possible to instruct the software to look for specific target lexical bundles in the text or text file. The program reports only those lexical bundles that recur in the text or file submitted by the user, which constitutes another limitation for N-Gram Phrase Extractor.

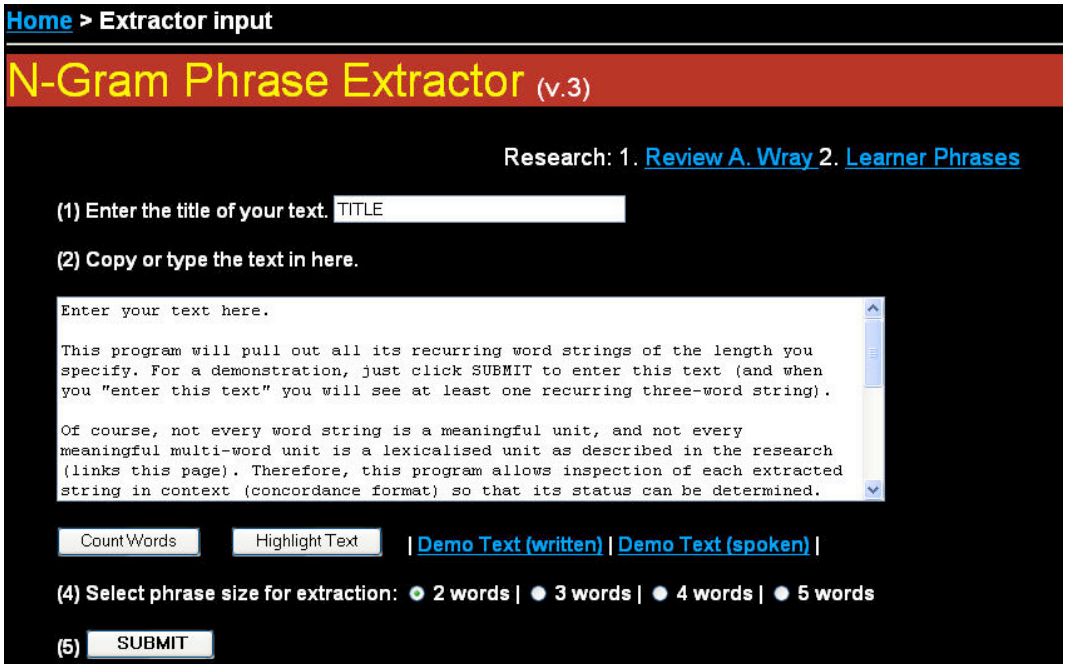
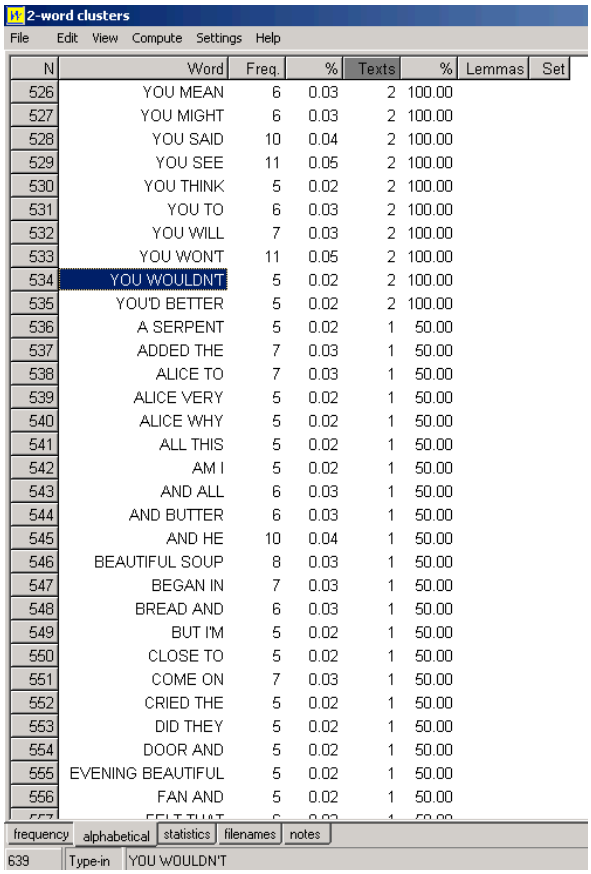


Figure 4. The main interface of N-Gram Phrase Extractor

Therefore, this software makes for a useful tool for teachers who are interested in the kinds of word combinations that repeat in single texts, most likely those of their students. Using this program, teachers can examine the extent and type of lexical bundles that their students are learning as reflected in their writings. This program also allows teachers to analyze the phrasal structure of class readings that their students engage with. Teachers can find out the lexical bundles that occur in the readings assigned to students, and they can make these lexical bundles noticeable to students using a variety of activities. Researchers who seek more detail (e.g., multi-text occurrence) in the analysis of lexical bundles should consider using either or both of the other programs examined in this review: kfNgram and Wordsmith Tools.

Wordsmith Tools

Among the software reviewed here, Wordsmith Tools is the most efficient in its search for lexical bundles. The program satisfies the two criteria for lexical bundles as outlined by Biber et al. (1999). Lexical bundles are reported in the output window in order of frequency, and to the far right of this output window users find information about the different texts from which the lexical bundles were extracted. In order to look for text files in this manner, the user needs to enter texts separately and give each an easily identifiable name, such as *text 1*. One advantage this program has over kfNgram is that users do not have to read through multiple output windows when searching for multi-text occurrences of lexical bundles; they can scroll up and down on only one screen that lists all lexical bundles according to frequency and text. Clicking on the text tab at the top of the page groups the lexical bundles into their source texts, thus allowing the user to compare and/or tally texts for the occurrence of a lexical bundle or lexical bundles in texts (see Figure 5).



N	Word	Freq.	%	Texts	%	Lemmas	Set
526	YOU MEAN	6	0.03	2	100.00		
527	YOU MIGHT	6	0.03	2	100.00		
528	YOU SAID	10	0.04	2	100.00		
529	YOU SEE	11	0.05	2	100.00		
530	YOU THINK	5	0.02	2	100.00		
531	YOU TO	6	0.03	2	100.00		
532	YOU WILL	7	0.03	2	100.00		
533	YOU WON'T	11	0.05	2	100.00		
534	YOU WOULDN'T	5	0.02	2	100.00		
535	YOU'D BETTER	5	0.02	2	100.00		
536	A SERPENT	5	0.02	1	50.00		
537	ADDED THE	7	0.03	1	50.00		
538	ALICE TO	7	0.03	1	50.00		
539	ALICE VERY	5	0.02	1	50.00		
540	ALICE WHY	5	0.02	1	50.00		
541	ALL THIS	5	0.02	1	50.00		
542	AM I	5	0.02	1	50.00		
543	AND ALL	6	0.03	1	50.00		
544	AND BUTTER	6	0.03	1	50.00		
545	AND HE	10	0.04	1	50.00		
546	BEAUTIFUL SOUP	8	0.03	1	50.00		
547	BEGAN IN	7	0.03	1	50.00		
548	BREAD AND	6	0.03	1	50.00		
549	BUT I'M	5	0.02	1	50.00		
550	CLOSE TO	5	0.02	1	50.00		
551	COME ON	7	0.03	1	50.00		
552	CRIED THE	5	0.02	1	50.00		
553	DID THEY	5	0.02	1	50.00		
554	DOOR AND	5	0.02	1	50.00		
555	EVENING BEAUTIFUL	5	0.02	1	50.00		
556	FAN AND	5	0.02	1	50.00		

Figure 5. Display of 2-word clusters after cluster search in Wordsmith Tools

The procedure to arrive at lexical bundles is not as easy in Wordsmith Tools as in kfNgram or N-Gram Phrase Extractor. This makes Wordsmith Tools less user-friendly than the other two programs. The procedure starts with the user inputting the text using the wordlist function—the program is a composite of three functions: concordance, keywords, and wordlist. After the addition of text, the user has to create or add to an index file of the text, which needs to be saved. The same index file then has to be opened from the wordlist window, resulting in the wordlist of the corpus with frequency order of words. In this window, the user has to click *compute* and *cluster* on the drop-down menu to extract the lexical bundles in the corpus. On the next, smaller window, the desired minimum frequency and lexical bundle size needs to be specified, upon which a new window with lexical bundles extracted will open. The lexical bundles are reported in frequency order. Although the program does not have a separate function for finding out multi-text occurrences, this information can be gathered by examining the text information to the right of the reported lexical bundles.

CONCLUSION

To summarize, kfNgram and N-Gram Phrase Extractor provide user-friendly and effective tools for teachers who are interested in the frequency of recurrent word combinations in their students' writing. For researchers who are more interested in word combinations that are used by the discourse community of a specific register or registers, Wordsmith Tools and kfNgram are the best tools to use. Although a very efficient tool, Wordsmith Tools is less user-friendly due to its somewhat tedious interface and complicated operations. To satisfy multitext occurrences (one of the defining criteria for lexical bundles) text files have to be entered separately in Wordsmith Tools and kfNgram, except N-Gram Phrase Extractor, which works only with single text files. If this is not done, the software programs yield only raw frequency information about lexical bundles. Raw frequency information itself is not sufficient for a word combination to qualify as a lexical bundle. The second criterion of multi-text occurrence also has to be satisfied.

Finally, the procedure to obtain multi-text occurrences, the second criterion for software efficiency, is quite laborious and time consuming in all programs except N-Gram Phrase Extractor, which provides only raw frequencies of lexical bundles. Users either have to view separate output windows and compare lexical bundles from one output window to the other as in kfNgram, or follow a very similar procedure and read through text file names for multi-occurrence of lexical bundles in the same window in Wordsmith Tools. Therefore, there is need for new software that reports the frequency of texts in which lexical bundles occur, in addition to raw frequencies of lexical bundles in an entire corpus. In their current format, however, these three software programs can be useful to obtain raw frequencies of lexical bundles. Teachers and language learners can benefit from raw frequency information using all three software programs reviewed here. Researchers and others who are interested in multi-text occurrence can benefit only from two of the software packages(i.e., Wordsmith Tools and kfNgram).

ABOUT THE REVIEWER

Omer Ari is a doctoral student in Middle /Secondary Education and Instructional Technology (MSIT) at Georgia State University. His research interests include college reading, corpus linguistics, and second language acquisition.

Email: **ariomer@hotmail.com**

REFERENCES

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25,* 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Carroll, L. (1994). *Alice's adventures in Wonderland*. Retrieved April 29, 2005, from **http://birrell.org/andrew/alice/Alice.pdf**.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23,* 397-423.

Nattinger, J. R., & DeCarrico, J. (1992). *Lexical phrases and language teaching.* Oxford: Oxford University Press.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). London: Longman.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1-22). Amsterdam: John Benjamins.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language, 10*(1), 61-104.