Lexical list for EAP/ESP programs: Multiword sequences in computer science textbooks

By Genan Hamad, Carleton University, Canada

Abstract

Lexical Bundles (LBs) —defined by Wood (2015) as "combinations of three or more words which are identified in a corpus of natural language" (p. 45) —play a key role in the comprehension and construction of academic language (Biber & Barbieri, 2007). Despite their importance, LBs are weakly presented in second language (L2) materials (Wood & Appel, 2014). Studies show that L2 learners may misuse LBs in their production (Pérez-Llantada, 2014). With the aim of informing L2 pedagogy in the university context, this corpus study uses WordSmith Tools 6.0 (Scott, 2007) to identify 59 items that represent the most frequently occurring LBs in eight Computer Science introductory textbooks. Utilizing the functional taxonomy, suggested in Biber et al. (2004), the analysis highlights important distributional and functional patterns of LBs use in this register. Of the total bundles, two-thirds are referential, and only one-third are stance bundles and discourse organizers. Within the referential types, specification of attributes is the most common subcategory. The emergence of code reference bundles as a new subcategory is a pattern that reflects discipline specificity. Considering that "the most frequently occurring words are also the most useful items to teach" (Wood & Appel, 2014, p. 1), bundles identified by this study, may be good candidates for selecting and designing teaching content by EAP/ESP instructors. The findings may also help curriculum developers to improve the presentation of the most frequent LBs in various disciplines, such as Computer Science, in their teaching materials.

Introduction

A growing body of research has highlighted the formulaic nature of language as it is observed that native speakers prefer certain formulaic sequences over others in their oral and written production (Ellis, 1996; Erman & Warren, 2000; Wray, 1999). These sequences are defined by Wood (2006) as "fixed combinations of words that have a range of functions and uses in speech production and communication, and seems to be cognitively stored and retrieved by speakers as if they were single words" (p. 1).

Unlike creatively generated language, formulaic sequences can be processed with less time and cognitive effort because they are stored and retrieved as whole units rather than single words (Wood, 2006; Wood, 2015; Wray & Fitzpatrick, 2008). Based on this characteristic, many researchers suggest that formulaic sequences may lead to better fluency and language competence, and therefore, they can be good candidates for language instruction in English for Academic Purposes (EAP) or English for Specific Purposes (ESP) programs (Wood, 2006; Wood, 2010a; Wood, 2010b).

Lexical Bundles or multiword sequences represent a major category of formulaic language that has been the focus of much recent research. Wood (2015) defines LBs as "combinations of three or more words which are identified in a corpus of natural language by means of corpus analysis software programs" (p. 45). Examples of these bundles include *I don't know what* in spoken language, and *on the other hand* in academic writing. LBs are considered important building blocks in discourse as they serve important pragmatic functions.

Biber, Conrad, and Cortes (2004) developed a functional taxonomy in which LBs are grouped based on their pragmatic functions in discourse under three major categories; with each category having some sub-categories that serve more specific functions. The main categories include: stance expressions, discourse organizers, and referential expressions. According to Biber et al. (2004), stance bundles (e.g., *can be used to*) can be personal or impersonal, and they are often used to «express attitudes or assessments of certainty that frame some other proposition» (p. 384). Discourse organizing bundles (e.g., *in this example the*) «reflect relationships between prior and coming discourse» (p. 384), and they include two subcategories: topic introduction and topic elaboration bundles. Referential expressions (e.g., *the value of the*) "identify an entity or single out some particular attribute of an entity as especially important" (p. 393). Under this category, four subcategories are included: identification referential bundles, imprecision bundles, attribute specifying bundles, and time/place/text/ multifunctional reference bundles. Bundles specifying attributes are divided into three specific types: quantity specification, tangible framing attributes, and intangible framing attributes.

The Value of Lexical Bundles

LBs not only constitute important building blocks in academic discourse, but they are also characterized by their pervasiveness and variation across a wide range of written and spoken academic discourse (Biber & Barbieri, 2007; Hyland, 2008a, 2012). For these reasons, LBs play a key role in the comprehension and construction of written and spoken language (Biber & Barbieri, 2007). Gaining control of common LBs in a particular register can improve reading skills as they affect a reader's ability to understand and recall the main ideas provided by a text (Martinez, 2002). Nesi and Basturkmen (2006) also found that LBs aid listeners by signaling how an idea is connected to another and "help the listener predict the nature of upcoming ideas", and thus reduce the "cognitive processing demands" (p. 17). Furthermore, the mastery of LBs may facilitate successful linguistic production, as they offer ready-made sets of words to use in academic writing (Byrd & Coxhead, 2010; Schmitt, 2004). In contrast, the absence of such bundles indicates the lack of fluency of a newcomer to a particular community (Hyland, 2012).

Lexical Bundles and L2 Learners

Second language writing is characterised by the underuse, overuse, and misuse of LBs (Bychkovska & Lee, 2017; Pérez-Llantada, 2014). Therefore, the use of LBs can be a good predictor of the writer's level of proficiency in the language. Researchers who analysed LBs produced by L1 and L2 English writers found major differences across various proficiency levels in the use of these items in terms of their proportion, diversity, structures, and functions (Adel & Erman, 2012; Chen & Baker, 2010; Staples et al., 2013). According to Ping (2009), non-native learners not only underuse LBs but may also overuse them by relying on a restricted set of bundles, which they use repeatedly, due to their limited repertoire. In addition, Staples et al. (2013) reported that L2 learners may find it difficult to use LBs appropriately in their writing as a result of confusing the written register with the spoken one.

EAP/ESP Materials and University Textbooks

Proficiency levels of L2 learners can be improved by learning the most frequent LBs of their disciplines (Hyland, 2008a). Despite the fact that introductory university textbooks from each academic discipline represent an important register that novice students will frequently encounter in academia, some EAP and ESP materials are not providing L2 learners in these programs with the appropriate repertoire of LBs that they may encounter in their textbooks of introductory courses in the first year of their studies (Chen, 2010; Wood & Appel, 2014). This gap may highlight the great pedagogic value in focusing on LBs in this genre across disciplines. As such, the present study aims to bridge this gap by identifying and analyzing the most frequent LBs in introductory university textbooks from Computer Science.

Methodology

Eight university textbooks that were used as main references for two introductory courses (which focus on teaching coding using Python and Java languages) in Computer Science were selected in order to compile the Computer Science introductory textbooks corpus (CSITC) which consisted of 1.3 million words. The data then was analysed using WordSmith Tools 6.0 (Scott, 2007). Drawing on previous studies (Biber & Barbieri, 2007; Biber et al., 2004; Wood, 2015) the cluster size was set to 4-8 words and a minimum frequency cut-off of 30 times per million words was selected. Following Wood and Appel (2014), a minimum range of 2 textbooks was chosen. The analytical framework used in this study is the functional taxonomy developed by Biber et al. (2004). It was applied to classify LBs into the three main categories: stance bundles, discourse organizing bundles, and referential bundles, as well as the sub-categories of these groups based on the functions they serve in discourse.

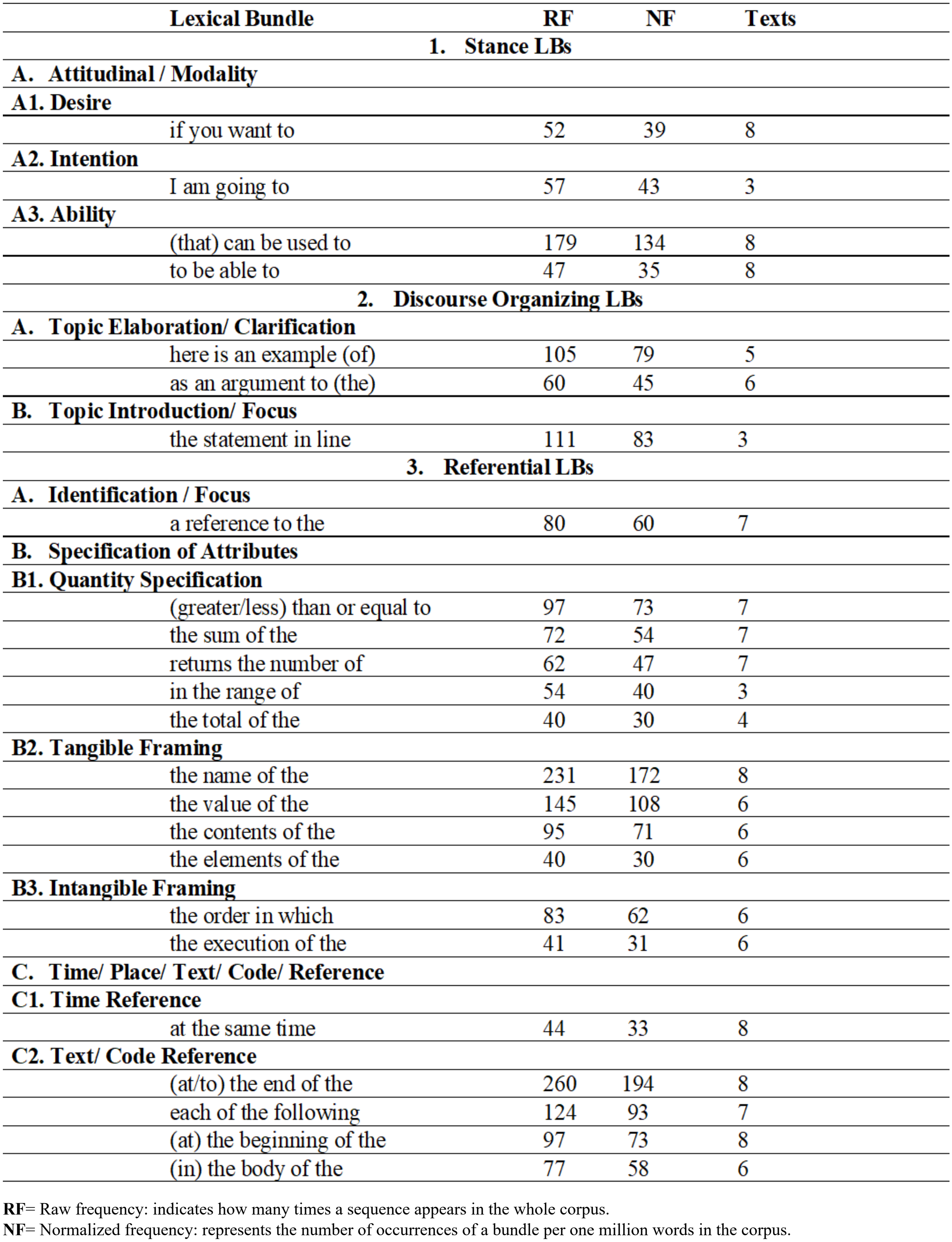
Results and Analysis

Overview of the CSITC Bundles List

In the CSITC of 1.3 million words, a total of 59 different 4-5 word LBs (the CSITC Bundles List) meet the identification criteria (frequency and range) set by the present study, with the most frequent bundle, *(at/ to) the end of the,* occurring 260 times across all the eight textbooks in the CSITC. In addition, each of the least frequent strings, *the elements of the* and *the total of the* in the list*,* appear 40 times across 4 texts in the corpus.

As can be seen in Table 1, two common structures of bundles (Biber et al., 1999) are found in the list: noun phrase + post modifier fragments (e.g., *the value of the, the contents of the*) and preposition + of phrase fragments (e.g., *in the body of the, at the beginning of the*). These structures are common in other academic discourse as they are often used to “identify quantity, place or size” as well as “to mark existence, or highlight qualities” (Hyland, 2008a, p. 10).

Table 1: Examples of LBs from the CSITC Bundles List according to their functions



Lexical Bundle	RF	NF	Texts
1. Stance LBs			
A. Attitudinal / Modality			
A1. Desire			
if you want to	52	39	8
A2. Intention			
I am going to	57	43	3
A3. Ability			
(that) can be used to	179	134	8
to be able to	47	35	8
2. Discourse Organizing LBs			
A. Topic Elaboration/ Clarification			
here is an example (of)	105	79	5
as an argument to (the)	60	45	6
B. Topic Introduction/ Focus			
the statement in line	111	83	3
3. Referential LBs			
A. Identification / Focus			
a reference to the	80	60	7
B. Specification of Attributes			
B1. Quantity Specification			
(greater/less) than or equal to	97	73	7
the sum of the	72	54	7
returns the number of	62	47	7
in the range of	54	40	3
the total of the	40	30	4
B2. Tangible Framing			
the name of the	231	172	8
the value of the	145	108	6
the contents of the	95	71	6
the elements of the	40	30	6
B3. Intangible Framing			
the order in which	83	62	6
the execution of the	41	31	6
C. Time/ Place/ Text/ Code/ Reference			
C1. Time Reference			
at the same time	44	33	8
C2. Text/ Code Reference			
(at/to) the end of the	260	194	8
each of the following	124	93	7
(at) the beginning of the	97	73	8
(in) the body of the	77	58	6

RF= Raw frequency: indicates how many times a sequence appears in the whole corpus.
NF= Normalized frequency: represents the number of occurrences of a bundle per one million words in the corpus.

Functions and Distribution of the Items in the CSITC Bundles List

Using the functional taxonomy developed by Biber et al. (2004) as an analytical framework, strings in the CSITC Bundles List can be classified under three main types based on their general functions in discourse: stance bundles, discourse organizers, and referential expressions (Biber et al., 2004). However, the distributional analysis, as presented in the pie chart in Figure 1 below, shows that referential bundles are by far the most common type. While referential bundles account for more than two-thirds of all the identified bundle types, stance expressions represent only 14%, and discourse organizers represent 15% of the total bundles in the list. Similarly, previous research on university register found that referential expressions are used more widely in textbooks and academic prose (Biber & Barbieri, 2007; Biber et al., 2004). Bundles of each main category were also grouped under different sub-categories according to their specific meanings and functions. This classification allows us to recognize the patterns of use of different bundle types in the CSITC.

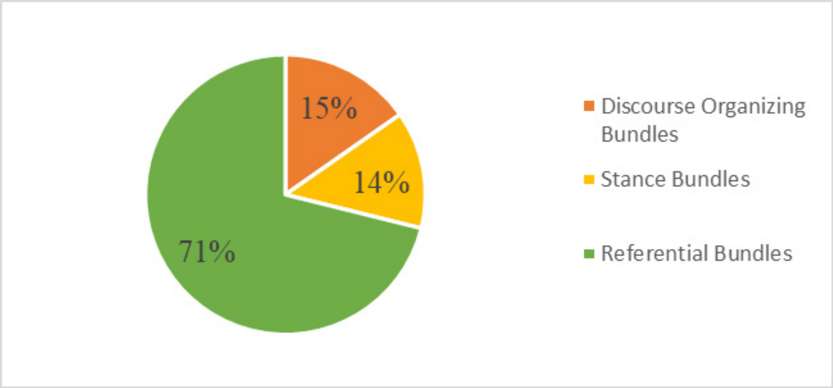


Figure 1 Distribution of LBs in the CSITC across the functional categories

Referential bundles in the CSITC

According to Biber et al. (2004), referential bundles include four sub-categories: “identification/focus, imprecision indicators, specification of attributes, and time/place/text reference” (p. 394). However, types of referential expressions, in the CSITC Bundles List, are represented in a different pattern. While imprecision indicators and place reference are completely absent from the list, code reference emerges as a new functional sub-category of time/place/text reference. Interestingly, the majority of bundles that are commonly used, in other disciplines (Biber et al., 2004), to refer to times, places or texts written by a human language (e.g., *at the end of the* and *at the beginning of the*) do not serve these functions in our corpus. Rather, their use is associated with the referral to a code constructed by a computer language (e.g., Python or Java). In this study, the term code reference is used to refer to this new function or subcategory. The following concordance lines of *at/to the end of the,* the bundle with the highest frequency in the list, illustrate the new function served by code reference bundles:

1. Notice that *at the end of the* <u>algorithm,</u> you delete the original file.
2. When the user specifies the request *at the end of the* <u>program</u>, we just need to consult the proper variable for the response.
3. During the loop, *total* is the running total, and ***at the end of the*** <u>loop</u>, *total* is the overall total of all the values in the list.
4. Sometimes complications are caused by the \n that appears ***at the end of the*** <u>strings</u> that are returned from the *readline* method.

Multi-functional reference is the subcategory under which the bundle *at the end of the* is placed in previous studies on LBs in university textbooks (Biber et al., 2004; Chen, 2010), as it is used in their corpora to refer to particular places, times, or locations in the text. However, it is clear from the prior examples that this sequence serves another specific function in our corpus. This is in line with Hyland and Tse (2007) who find that LBs behave in different ways across disciplines. The examination of the underlined words (*algorithm, program, loop* and *strings*) which follow the bundle directly in the examples above, as well as the context surrounding these words indicates that *at the end of the* is usually used when the author needs to refer to a piece of code or some parts of the programming process. For this reason, we assign this bundle to code reference as a new subcategory.

With regard to the number of bundles (types) in the functional sub-categories within the referential bundles in the list, specification of attributes is the dominant subcategory. Among specification of attributes, tangible framing records the highest number of bundles, whereas code/text reference is the dominant type across time/place/code/text reference bundles. On the other hand, the analysis of the overall frequency of specific bundle types within referential bundles suggests that while tangible framing and code/text reference are by far the most frequent bundles in the CSITC Bundles List, time reference sequences are rare and place reference bundles are absent. Accordingly, ESP instructors may need to give more attention to the most common sub-categories within the referential bundles in their classrooms.

Stance and discourse organizing bundles in the CSITC Bundles List

The pie chart in Figure 1 above demonstrates that stance bundles (e.g., *can be used to*) and discourse organizers (e.g. *here is an example of*) account for less than one-third of all the 59 bundles in the list.

This indicates that these types of LBs are less common in this register, and therefore, novice students in Computer Science may need to use them less frequently. The analysis also shows that some sub-categories of stance bundles, such as epistemic stance bundles, which are used to evaluate the level of certainty of the following information (e.g., *I do not know if*), did not occur in the CSITC Bundles List. This may suggest that epistemic bundles are less important for students in Computer Science.

Conclusion

Findings

The investigation of the Computer Science introductory university textbook corpus led to the creation of a list of 59 items, which represent the most frequent LBs that undergraduate students may encounter in their first year in Computer Science. Utilizing the functional taxonomy, suggested in Biber et al. (2004), the analysis highlights important distributional and functional patterns of LBs use in this register. The study shows that academic texts in this specialized corpus are dominated by the use of referential bundles.

The CSITC is also characterized by the dominance of bundles from two distinct subcategories: tangible framing and code/text reference. The emergence of code reference bundles as a new subcategory of time/space/text/code reference is a pattern that reflects discipline specificity. Based on the prior findings, the study concludes that the use of LBs in the CSITC is influenced by the communicative purpose of this register, namely, communicating instructions and procedures that students need to follow in order to write code that enable the computer to perform a particular task.

Implications and Future Avenues

The findings yielded by the present corpus study may have interesting pedagogical implications. Considering that “the most frequently occurring words are also the most useful items to teach” (Wood & Appel, 2014, p. 1), strings in the CSITC Bundles List may be good candidates for “selecting, sequencing, and structuring of teaching content” (Hyland, 2008b, p. 60) in EAP/ESP programs. While instructors in these courses are expected to give priority to teaching expressions that are more relevant to the academic fields of their students, those teachers do not have adequate knowledge of discourses in various disciplines. Therefore, the list developed in this study can be an important source for designing their instructional materials.

In addition, the items in the CSITC Bundles List and their functions can be invaluable for novice students planning to enroll in the Computer Science program. Familiarity with these building blocks and frames of discourse may facilitate the comprehension of and the engagement with the required textbooks for introductory courses in Computer Science. Moreover, the findings may also help curriculum developers to bridge the gaps in their teaching materials by improving the presentation of the most frequent LBs and their functions in various disciplines, such as Computer Science.

Despite the importance of the current findings, they only provide little information about LBs in a large academic genre (i.e., introductory university textbooks). As such, the present study suggests that research on LBs in introductory university textbooks from other disciplines, including nursing, health sciences, neuroscience, and other fields of hard and soft sciences, needs more attention. A comprehensive analysis of LBs in such disciplines can provide EAP/ESP programs with a complete picture about LBs in this genre which plays a key role in students' academic success.

Finally, although creating lists of the most frequent LBs in different academic registers and disciplines provides a rich source for EAP programs, these lists are available only in research articles or academic books. Instructors and learners alike may find it difficult to reach these lists and make use of them. Developing websites and software, in which learning activities are designed based on these lists, can make these important bundles more accessible and useful.

References

Adel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(1), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Philadelphia, PA: John Benjamins.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

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

Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38–52. https://doi.org/10.1016/j.jeap.2017.10.008

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5*(5), 31–64. Retrieved from https://www.semanticscholar.org/paper/On-the-other-hand%3A-Lexical-bundles-in-academic and-Byrd-Coxhead/39db7ba106e5379aaf15898d44ecf7a088a5afe7

Chen, L. (2010). An investigation of lexical bundles in Electrical Engineering introductory textbooks and ESP textbooks. In D. Wood (Ed.), *Perspectives on formulaic language* (pp. 107–125). New York: Continuum.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30–49. https://scholarspace.manoa.hawaii.edu/bitstream/10125/44213/14_02_chenbaker.pdf

Ellis, N. C. (1996). Sequencing in SLA. *Studies in Second Language Acquisition, 18*(1), 91–126. https://doi.org/10.1017/S0272263100014698

Erman, B. & Warren, B. (2000). The idiom principle and the open choice principle. *Text & Talk, 20*(1), 29–62. https://doi.org/10.1515/text.1.2000.20.1.29

Hyland, K., & Tse, P. (2007). Is there an 'academic vocabulary'? *TESOL Quarterly, 41*(2), 7–22. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*(3), 150–169. https://doi.org/10.1017/S0267190512000037

Martinez, A. C. L. (2002). Empirical examination of EFL readers' use of rhetorical information. *English for Specific Purposes, 21*(1), 81–98. https://doi.org/10.1016/S0889-4906(00)00029-6

Nesi, H. and Basturkmen, H. (2006) Lexical bundles and discourse signaling in academic lecturers. *International Journal of Corpus Linguistics, 11*(3), 283–304. https://doi.org/10.1075/ijcl.11.3.04nes

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes, 14*(1), 84–94. https://doi.org/10.1016/j.jeap.2014.01.002

Ping, P. (2009). A study on the use of four-word lexical bundles in argumentative essays by Chinese English majors: A comparative study based on WECCL and LOCNESS. *CELEA Journal, 32*(1), 25–45. http://www.celea.org.cn/teic/85/85-25.pdf

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Scott, M. (2007). *Oxford Wordsmith Tools: Version 6.0*. Released June 2007 from http://www.lexically.net

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *Canadian Modern Language Review, 63*(1), 13–33. https://doi.org/10.3138/cmlr.63.1.13

Wood, D. (2010a). *Formulaic language and second language speech fluency: Background, evidence, and classroom applications*. New York: Bloomsbury.

Wood, D. (2010b). Lexical clusters in an EAP textbook corpus. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 8–106). New York: Continuum.

Wood, D. C., & Appel, R. (2014). Multiword constructions in first-year university Engineering and Business textbooks and in EAP textbooks. *Journal of English for Academic Purposes, 15*(9), 1–13. https://doi.org/10.1016/j.jeap.2014.03.002

Wood, D. (2015). *Fundamentals of formulaic language: An introduction*. New York: Bloomsbury.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching, 32*(4), 213–231. https://doi.org/10.1017/S0261444800014154

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 123–147). Amsterdam: John Benjamins.

Author Bio

Genan Hamad obtained an MA in Applied Linguistics (TESL stream) from Carleton University, Ottawa, Canada. She has been teaching for more than 14 years and has taught L2 learners with different cultural backgrounds in Canada and abroad. She worked at UNAM-Canada, the LINC program, and the Conseil des ecoles catholiques du Centre-East (CECCE). She is interested in corpus linguistics and formulaic language.

Back