Using Concordance Software to Generate Academic Words in Applied Linguistics

Weningtyas Parama Iswari[1], Bibit Suhatmady[2], Yuni Utami Asih[3], Ida Wardani[4], Adrianto Ramadhan[5], Dynda Anastasya[6]
Faculty of Teacher Training and Education, Mulawarman University, Indonesia
[1]weningtyaspiswari@fkip.unmul.ac.id, [2]bibitsuhatmady@fkip.unmul.ac.id, [3]yuniutamiasih@fkip.unmul.ac.id, [4]idawardani@fkip.unmul.ac.id, [5]radrianto746@gmail.com, [6]dyndanastasya@gmail.com

ABSTRACT

Academic words include words that are not commonly encountered in formal circumstances and specific to particular fields of study. Undergraduate students of the English Department are required to acquire academic words in applied linguistics for academic reading and writing research articles. This paper reports on generating the academic word list for the students of the English Department by using *AntConc,* a concordance software application. In this study, corpus linguistic research was adopted, in particular the corpus-based analysis category. Data were gathered from approximately one thousand credible Applied Linguistics journal articles published from 2008 to 2021. *AntConc* software played a significant role in processing these data to get the intended corpus, which was then classified and categorized based on the frequency of occurrences. The results include an academic word list and its word family. These clusters of academic words are intended for undergraduate students of the English Department in the first up to fourth academic semesters to prepare them to participate in international academic discourse, such as writing and publishing research articles. This list can also be used as a basis for further research related to academic vocabulary.

Keywords: academic words, *AnctConc,* applied linguistics, concordance software

INTRODUCTION

Vocabulary is the core element of any language, not to mention in English. In English as a foreign language, learners need to possess an ample amount of vocabulary to communicate in the target language. Meanwhile, in academic settings, English has gained its international role as the primary language for research and publicationn. Therefore, anyone who learns at a university or college will need to acquire English in general and English academic vocabulary.

According to Coxhead and Nation (2001), the academic vocabulary is one of the English word categories, following the high-frequency words, the technical vocabulary, and the low-frequency words. It belongs to a discipline in the field of English for specific purposes. Paquot (2010) defines academic vocabulary as a term referring to "a set of lexical items that are not core words but which are relatively frequent in academic texts" (p.9). In classrooms, teachers use academic vocabulary to select words for teaching language skills, while learners refer to it as a checklist or learning goals (Coxhead & Nation, 2001).

Many researchers (e.g., Campion & Elley, 1971; Ghadessy, 1979; Xue & Nation, 1984) conducted studies to compile different academic word lists. The Academic Word List (AWL) developed by Coxhead (2000) has been most widely used by university teachers and students.

Her list is available for four faculties: Arts, Commerce, Law, and Biology, with seven subject areas for each faculty.

Academic words are commonly compiled through corpus-based research or corpus linguistics. This research method is a language study that follows procedures and uses digitalized extensive transcribed utterances or written texts (corpora) occurring in natural contexts (McEnry & Hardie, 2012). Some techniques used are to generate frequency word lists, concordance lines, collocations, and word clusters.

The corpus linguistics method was first used in the early 1960s, but corpora study began popular in the 1980s. Although corpus linguistics is relatively a new field of study, its contribution to language studies has proved significant through its new way of analyzing and describing language uses and variations (Hutson, 2002). Corpus linguistics compile and analyze corpora, extensive collections of authentic texts selected to represent a variety of languages (Sinclair, 1991).

A large number of studies have been conducted using corpus linguistics as a research methodology. Some of the previous studies are academic word lists (Coxhead 2000), corpora of sign language (Johnston & Schembri 2006), development of Academic Keyword List (Paquot, 2010), comparison between a carpentry corpus and a fiction corpus (Coxhead, Demecheleer, & McLaughlin, 2016), and a corpus-based study on metaphor (He & Young, 2017).

Realizing the importance of academic vocabulary for university students and its nature specific to any particular field of study, academic vocabulary in applied linguistics for English department students is deemed necessary. The existing Coxhead's AWL is rather general to education and linguistics, not specific to applied linguistics and the condition of the English department students at the Faculty of Teacher Training and Education, Mulawarman University. During their tertiary study, these students are required to read academic texts, such as journal articles, and write academic papers or articles as course assignments, projects, and publication purposes. Therefore, corpus-based research has been conducted to provide academic vocabulary in applied linguistics that can be used for those students in one to four semesters and can be integrated into their listening, speaking, reading, and writing courses.

Adopting corpus linguistics as a computer-based research method requires the use of concordance software applications. These tools serve to process big data or corpus collected from authentic sources stored in computers. The concordance software facilitates corpus analysis, a powerful method to analyze texts, discover data and find relationships within documents (Borhani, 2019). Other uses of the software are to analyze a large amount of data fast and accurately and help researchers detect patterns in their data, which is difficult to do manually.

AntConc software was used in this research to make corpus analysis possible, such as concordance, collocation search, word frequency, and keyword (Johnston, 2021). This free concordance program was developed by Prof. Laurence Anthony from Waseda University, Japan. *AntConc*, as a text analysis tool, is user-friendly, as it can be easily set up to run searches in corpora. It also has several features for advanced corpus analysis and performs well in working with large corpora and in other research purposes dealing with the corpus (Johnston, 2021).

This paper aims to report on the use of *AntConc* as a concordance software application in corpus-based research to compile academic words from journal articles in applied linguistics. It worked as a text analysis tool to process the corpus to generate the target academic vocabulary used, especially for university students majoring in English Language Education and generally for other interested parties and researchers in the related field of study.

METHOD

Corpus linguistics was adopted in this study based on its function to investigate language through corpus analysis (Hunston & Francis, 2000). A corpus (or corpora, in plural) is a collection of machine-readable, written, or spoken authentic texts representing a particular language or its variety (O'Keeffe, McCarthy & Carter, 2007). This study used corpora as its research data collected from approximately 1000 journal articles from seven reputable (indicating by Journal Citation Reports/JCR as Q1 & Q2, or having high impacts) journals in applied linguistics published within the period of more than ten years (2008 up to 2021). The data were then further processed electronically and manually. The electronic process used *AntConc* software, and the manual process relied on expert judgments. In this case, the researchers' expertise was in linguistics and applied linguistics.

This study combined quantitative and qualitative analyses. The quantitative analysis was used to find the frequency of word occurrence—the qualitative analysis aimed to distinguish academic words from non-academic ones and build the related word families. The list of non-academic words was referred to the New General Service List (NGSL) developed by Browne, Culligan, and Phillips (2013). The NGSL is the updated version of the original General Service List developed by Michael West initially in 1936 and its last version in 1953, considering the fast development and dynamic changes of English words (Browne 2021). This list includes core high-frequency vocabulary words for students of English as a second/foreign language.

This paper is a part of a more comprehensive study on developing academic vocabulary in applied linguistics. The quantitative analysis only focuses on using a concordance software AntConc (Anthony, 2017). *AntConc* has seven features that operate as tools to process corpora. The seven features consist of Concordance Tool, Concordance Plot Tool, File View Tool,

Clusters/N-Grams, Collocates, Word List, and Key Word List. The following table (Table 1) summarizes these critical features of *AntConc.*

Table 1. Key Features of *AntConc*

AntConc FEATURES	FUNCTIONS
Concordance Tool	To find out how words and phrases are commonly used in texts in the form of Key Word in Context (KWIC) format
Concordance Plot Tool	To find out the position of search results that occur in texts and are plotted in a barcode format.
File View Tool	To conduct further investigation towards the results of other AntConc tools.
Clusters/N-Grams	To get common expressions in a corpus.
Collocates	To examine non-sequential patterns in texts.
Word List	To find words with their frequency of occurrence in a corpus.
Key Word List	To identify specific words (related to a genre or English for specific purposes) in a corpus.

(Source: Anthony, 2019)

Three of the seven features were used for this study, including the Concordance Tool, the File View Tool, and the Word List.

RESULT AND DISCUSSION

It has been mentioned before that the primary purpose of this paper is to report on the use of *AntConc* as a concordance software to execute corpus analysis. The final result was to generate a list of academic words relevant to applied linguistics and clustered based on their frequency of occurrences. There were some research stages done, namely: collecting data (corpus), processing the data using the *AntConc* tools (i.e., the Word List, the Concordance Tool, and the File View Tool), and analyzing more deeply the *AntConc* result based on the expert judgment. The following description relates only to the use of the three tools.

The Word List

The first stage was data collection. Electronic journal articles were collected from seven high-quality journals in Applied Linguistics to guarantee correct and accurate content and language use. These journals were identified as having high impacts or known as good quality published materials. According to Journal Citation Reports or JTR (Mondragon Unibertsitatea, 2017), a tool to find out the impact indicators of leading scientific journals, the seven journals as the data sources were categorized as the first (Q1) and second (Q2) quartiles. The top 25% of journals in the JTR list are indexed in Q1, and the 25 to 50% are in Q2. The journals published in the last thirteen years were taken with approximately 1000 articles to ensure sufficient representativeness of the intended academic words. These data in Pdf format were downloaded and stored in computers for further processing using the *AntConc* software.

AntConc requires documents in a plain text format or TXT file. It is a simple text with little to no formatting that can be further used as text-based information. Therefore, the stored data in this research were first converted from Pdf to TXT files. The collected data comprised 14,733,424 word tokens.

The first feature used in this study was the Word List Tool to find words and their frequency of occurrences in the corpus. The corpus to be processed contained the journal articles converted from the Pdf to TXT file format. After *AntConc* started to run, the target corpus was selected, and the Word List Tools was clicked. The ordering option was selected. The words were ordered by frequency from the highest to the lowest frequency of occurrences. Figure 1 below shows the process of operating this tool.

Figure 1. The Process of Using the Word List Tool based on Anthony (2019)

The Word List Tools counted all the word tokens in the corpus. This process resulted in the ordered list of words to see which were the most frequent in the corpus.

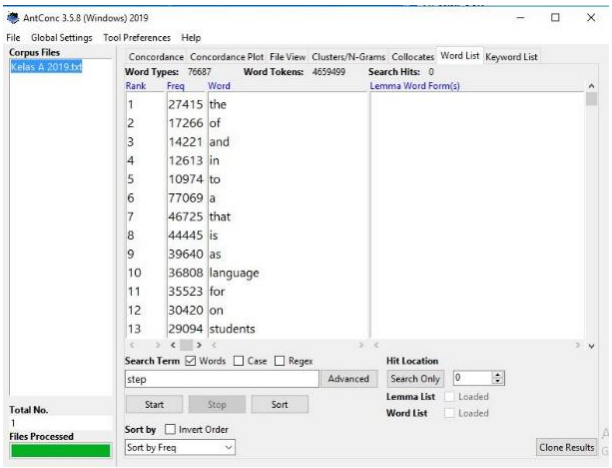

Figure 2. Screen Display of the Word List Tool

Once the 'Clone Result' button was clicked, a copy of the results was generated in a word list ordered from the most to the least frequently occurring words. In this corpus analysis using *AntConc*, the word 'the' gained the most frequent (883,848 times), followed by 'language' (117,784), 'student' (86,123), and 'learning' (68,451). Meanwhile, the lowest frequent words contained English and non-English words, such as apart (1), asteroid (1), and even words in

Bahasa Indonesia, such as *kerabat* (1) and *kerajaan* (1). All the results were copied in Excel format and listed in alphabetically ordered.

The words in the list with their frequency and ordered alphabetically were further analyzed manually to generate academic words by identifying words that did not belong to the New General Word List (NGSL). The NGSL contained words commonly found in non-academic discourse. These NGSL words were generated from corpus linguistics with the data source from a selected 273 million word subsection of the 2 billion words Cambridge English Corpus (CEC), including corpora of learner, fiction, non-fiction, journals, magazines, radio, spoken, document, and TV (Browne et al., 2013). In this stage of the research, the words generated from the Word List Tool were divided into three groups, comprising the non-NGSL group (8,709 words), the NGSL words (2800 words), and non-words, such as abbreviation, nonsense words, and symbols (14,721,915 words/characters). The generated non-NGSL words were treated as having the potential to be the targeted academic words in applied linguistics. For this purpose, different processes were taken by having a qualitative analysis through expert judgment (done manually by the researchers) and also consulting the New Academic Word List (960 words) developed by Browne, Culligan, and Phillips (2013) to crosscheck the validity of the academic words generated in this study. However, this qualitative analysis is not covered in this paper.

The Concordance Tool

The academic vocabulary generated from this research comprised the head words and their word family (affixed forms) and examples taken from the original texts to ensure their authenticity. The Concordance Tool helped to find words and phrases in context or original sentences in the corpus, as this tool's search results were in the form of keyword in context (KWIC). These words in contexts provided authentic examples of how each word was commonly used in texts.

Using the Concordance Tool started by selecting one or more files and entering a search term to generate concordance lines. The setting was adjusted by choosing the number of text characters and rows. After that, the start button was clicked to get the concordance lines. Before sorting the concordance lines, the KWIC Sort option was used to reorder them. Moving the cursor towards the highlighted search term enabled to see the searched word or phrase appearing in the original text through the Fie View Tool. Finally, a copy of the result was retrieved by clicking the Clone Result button. Figure 3 shows the summarized process of generating the concordance lines from the corpus-based on Anthony's (2019) guidance in using *AntConc* software.

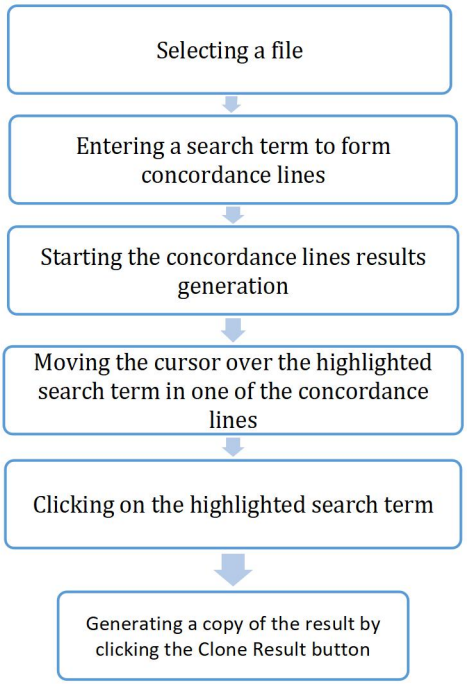

Figure 3. The Process of Using the Concordance Tool based on Anthony (2019)

The outcome of processing corpus using the Concordance Tool was in concordance lines to show how a particular word was used naturally in the original text. For instance, the words' attainment' and 'cumulative' appeared on the output of the Concordance Tool as concordance lines (see Figure 4).

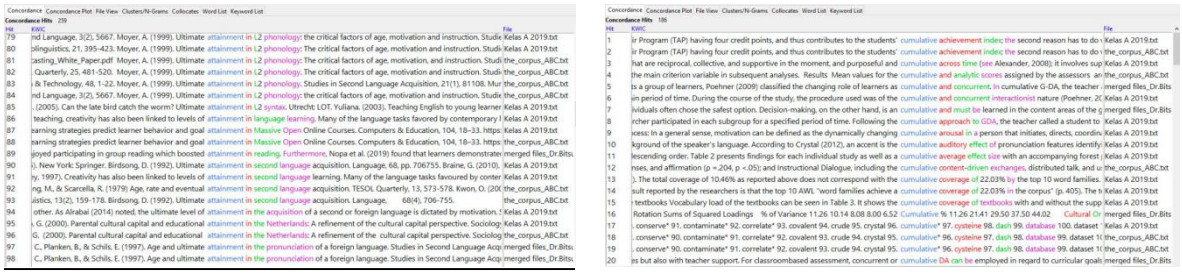

Figure 4. Screen Display of 'Attainment' and 'Cumulative' as the output of the Concordance Tool

The output of the Concordance Tools in the form of concordance lines was used in this research to complete the academic vocabulary with examples of how to use specific academic words in sentences. These examples were taken from authentic texts, which ensures their authenticity.

The File View Tool

The File View Tool presents raw texts from original files that can be used for researching more deeply the results generated from other *AntConc* tools (Anthony, 2019). The first step in

using this tool was to select a file and specify a search team in this research. The result of this step was to highlight search term hits throughout the text. The search term was changed from one to another to show other hits. When the highlighted text was chosen as the search term, the tool presented a group of KWIC lines. This process can be seen in Figure 5.

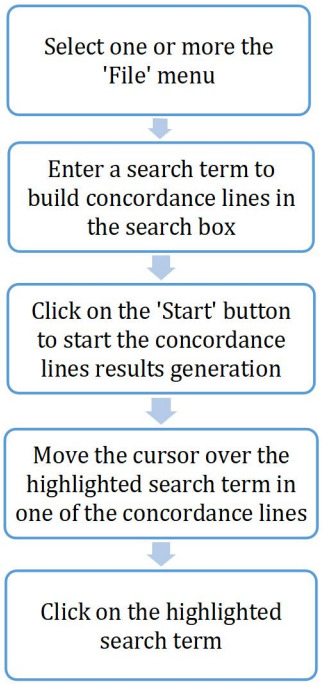

Figure 5. The Process of Using the File View Tool

Using the File View Tool in this research was to identify the individual text where a particular academic word originated. It is helpful for coding purposes in the research. Figure 6 presents the screen display of the word 'attainment' and 'cumulative as sample output of the tool.

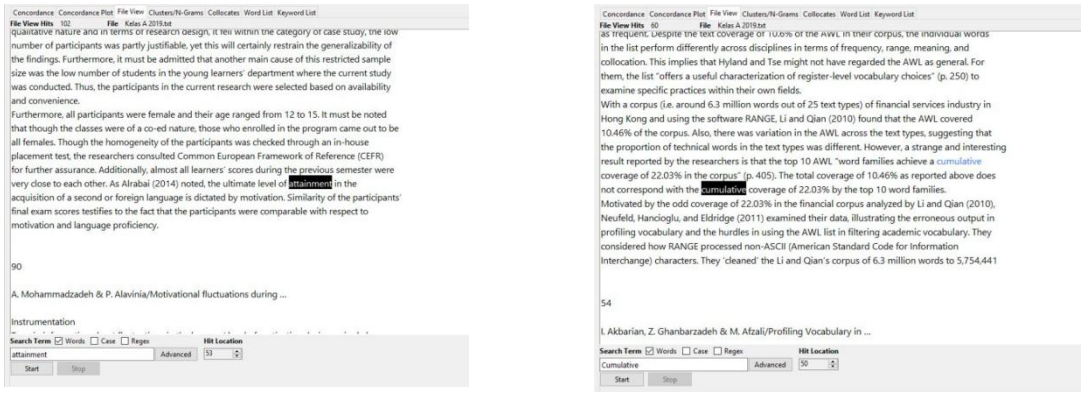

Figure 6. Screen Display of "attainment' and 'cumulative' as the output of the File View Tool

The result of executing the File View Tool, such as for the words' attainment' and 'cumulative' presented in Figure 6, can be used for further corpus analysis resulting from the other *AntConc* tools.

Pedagogical Implications

The use of concordance tools in conducting corpus linguistics is essential. In this research, *AntConc* software played a significant role, especially in generating word frequency and concordance lines for further corpus analysis. Corpus linguistics has given considerable contributions to the study of language, as it gives new perspectives in language studies, provides powerful tools for language analysis, and offers new ways of doing linguistic research (Szurdaski, 2018). Some contributions of corpus-based studies are producing dictionaries, finding patterns in grammar, analyzing the stylistics in the language of literature, developing learning materials, and compiling words for many different purposes (O'Keeffe, McCarthy & Carter, 2007). The ability to use corpus or concordance tools is necessary to carry out different corpus-based studies. As there are many concordance tools available, educators and researchers may choose one suitable to their condition and need.

The outcome of this corpus linguistics research is a list of academic words in applied linguistics from the related journal articles that can be useful for university students majoring in English language education whose primary field of study is in applied linguistics. Learning the academic vocabulary (and word families) will enrich students' knowledge, especially for reading academic texts and writing academic papers in their discipline. Therefore, it is recommended that this academic vocabulary be used as a pedagogical tool for both teachers and students.

The students of the English Department usually have to take English courses, including language skills (i.e., listening, speaking, reading, and writing) and language components (i.e., vocabulary, pronunciation, spelling) in their first four semesters. They can use this list as learning materials for self-study and as teachers' instructional materials embedded in the English courses. However, it is also suggested that the students do not exclusively learn the academic words but also need to learn general words that might be useful to discuss specialized academic discourses. Academic and general words are necessary for students to survive during their study, since they can be a shortcut for students to learn the essential words in their learning context (Browne 2021).

CONCLUSION

Concordance tools play a significant role in corpus linguistics since they facilitate researchers in analyzing big data or corpus, as the essential part of this kind of research.

AntConc, a free, user-friendly concordance tool, was employed in this research to generate data for further qualitative analysis. The sources of data were reputable journals in applied linguistics, as it was in line with the research purpose. *AntConc* provided the list of words ordered by frequency of occurrence in the corpus and concordance lines. The quantitative analysis part of the fundamental research is continued with the qualitative analysis to sort out the academic and non-academic word list. Reporting the process of using the concordance tool can be evidence of this tool in providing rich data for different researchers related to language and language use.

ACKNOWLEDGEMENT

This research on generating academic vocabulary in applied linguistics was financed under the Faculty of Teacher Training and Education grant, Mulawarman University, Indonesia. The researchers would like to thank the Dean and Vice Dean for Academic Affairs of the faculty who have approved to provide the financial support to this research. Hopefully, this research results in valuable academic vocabulary for the students of the English Department, other related parties, and future researchers.

REFERENCES

Anthony, L. (2019). Laurence Anthony's Website. https://www.laurenceanthony.net

Anthony, L. (2017). *AntConc (Version 3.5.0) [Computer Software].* Tokyo, Japan: Waseda University. http://www.antlab.sci.waseda.ac.jp/

Borhani, M. (2019). Corpus Analysis Using Relaxed Conjugate Gradient Neural Network Training Algorithm. *Neural Process Lett* **50,** 839–849. https://doi.org/10.1007/s11063-018-9948-8.

Browne, C. (2021). The NGSL Project: Building Wordlists and Resources to help EFL Learners (and Teachers) to Succeed. *TEACHING with2020*, 1.

Browne, C., Culligan, B., & Phillips, J. (2013). New General Service List. http://www.newgeneralservicelist.org

Campion, M. E., & Elley, W. B. (1971). *An Academic Vocabulary List*. Wellington, New Zealand: New Zealand Council for Educational Research.

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

Coxhead, A., Demecheleer, M. & McLaughlin, E. (2016). The technical vocabulary of Carpentry: "Loads, lists and bearings. *TESOLANZ Journal, 24: 38–71*.

Coxhead, A., & Nation, P. (2001). The Specialized Vocabulary of English for Academic Purposes. In J. Flowerdew, & M. Peacock (Eds.), *Research Perspectives on English for Academic Purposes, Chapter: The Specialised Vocabulary of English for Academic Purposes* (pp. 252-267). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524766.020

Ghadessy, P. (1979). Frequency counts, word lists, material preparation: A new approach. *English Teaching Forum, 17,* 24–27.

He, Q. & Yang, B. (2017). A corpus-based study of the correlation between text technicality and ideational metaphor in English*. Lingua, vol. 203/2018, pp. 51-65.* doi:10.1016/j.lingua.2017.10.005

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

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Johnston, B. (2021, March 4). *Educational & classroom technologies*. https://mcgrawect.princeton.edu/tool/antconc/

Johnston, T. and Schembri, A. (2006). 'Issues in the creation is a digital archive of a signed language,' in L. Barwick and N. Thieburger (Eds.) *Sustainable Data from Digital Fieldwork*, pp. 7–16. University of Sydney Press.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: method, theory, and practice.* Cambridge University Press.

Mondragon Unibertsitatea. (2017). *Publication Impact Indexes*. https://www.mondragon.edu/en/web/biblioteka/publications-impact-indexes

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

Paquot, M. (2010*). Academic vocabulary in learner writing*. London: Continuum.

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics,* 32: 130–149*.*

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

Szudarski, P. (2018). *Corpus linguistics for vocabulary. A guide for research.* New York: Routledge.

Xue, G. & Nation, I. S. P. (1984). *A university word list.* Language Learning and Communication, 3(2): 215-229.