# Selecting Lexical Bundles for Writing Courses

Daniel Parsons
*International University of Japan*

Master's degree students transitioning from reading textbooks to writing a thesis face a language gap of primarily instructional language to reporting language, but this gap is not well known for economics, politics, and international relations fields. This study aims to examine differences in lexical bundles between textbooks and research papers in these fields. This report outlines a preliminary analysis carried out on stance lexical bundles which occur much more frequently in research papers.

Lexical bundles play an important role in managing the discourse of academic texts, but there is little information for practitioners about which specific lexical bundles can best help students to write a master's thesis. These bundles come in three varieties: referential, stance, and meta-discursive (Biber, Conrad, & Cortes, 2004), though Simpson-Vlach and Ellis (2010) modified the categories to include cause-effect and comparison bundles.

However, Cortes (2004) has shown that the English lexical bundles students encounter in their academic reading do not make it into their academic writing. This should not be surprising as Ellis, Simpson-Vlach, and Maynard (2008) have shown that non-native speakers' ability to process lexical bundles accurately and fluently is dependent on the frequency with which such bundles are encountered. Wulff (2019) explained that for acquisition of language to begin, students require a great deal of input in the early stages, but as Wood and Appel (2014) have demonstrated, current English for Academic Purposes (EAP) textbooks do not offer any pedagogical focus on lexical bundles. In addition, one of the defining characteristics of lexical bundles is that they are evenly

dispersed throughout a corpus with cut-off frequencies of 40 occurrences per million words (Biber, Conrad & Cortes, 2004). This means that the bundles are spread out across the millions of words in student textbooks, not all of which the students will read. Given this lack of salience, and given the important discursive role played by lexical bundles, there is a need for teachers to know which lexical bundles students need to focus on in a thesis writing course.

This study aims to explore the lexical bundles within the fields of international development, international relations, politics, public management and policy, and business, which represent the fields studied within the various programs at International University of Japan (IUJ). Cortes (2013) listed lexical bundles for a wide range of disciplines, but only for those occurring in introductions, whereas Mizumoto, Hamatani, and Imao (2017) analyzed lexical bundles used in all the sections of a corpus of research articles, but only in the field of applied linguistics. Consequently, there is little knowledge about what lexical bundles might be important to teach students in a thesis writing course for the subjects at IUJ, where students are expected to read textbooks, research papers, and eventually write their own thesis in English. The purpose of the study is, therefore, to isolate those lexical bundles which students need to write their thesis, and those lexical bundles that might be useful for students to read their textbooks. To clarify, the research reported here focuses only on stance lexical bundles which Biber, Johannsson, Leech, Conrad, and Finegan (1999) defined as those bundles which either attribute knowledge to other sources, hedge ideas, or evaluate ideas.

## Methods

To construct the corpus, six of IUJ's international development program (IDP) course textbooks were first scanned as PDF files, then converted in Adobe Acrobat to text files through optical character recognition. The token count for this corpus, according to AntConc (Anthony, 2019a) was 1.35 million. Similarly, AntCorGen (Anthony, 2019b) was used to create the corpus of research papers. The following fields were selected: development economics, economic models, macroeconomics, mathematical economics, and microeconomics. These fields

collectively represent the core courses during the first year in the IDP, and the token count for this corpus was 1.9 million.

Stance lexical bundles were selected from two sources: The Academic Formulas list (Simpson-Vlach & Ellis, 2010) and The Longman Grammar of Spoken and Written English (Biber et al., 1999). Biber et al. (1999) also reported that certain nouns, such as *fact* and *possibility* can be used with prepositional phrases and complement clauses to express stance. Examples include *the fact that* and *the possibility to*. Stance noun lexical bundles were constructed based on this information and added to the lists provided in the above two sources. In total, a list of 149 stance lexical bundles was created.

The list of stance lexical bundles was then subject to a test of salience. Following Simpson-Vlach and Ellis (2010), a cut-off point was set at 10 occurrences per million. Any bundle that occurred at a relative frequency lower than this in either corpora was excluded from further analysis. The second stage of the test was to calculate the ratio of the relative frequency in research papers to the relative frequency in textbooks. A ratio higher than 2 indicated that a bundle occurred more than twice as frequently in research papers in comparison to textbooks. This, therefore, allowed the bundles to be ranked in accordance with how much more frequently they occurred in papers as compared to textbooks.

## Results

The test of salience described in the previous section revealed that of the 149 stance lexical bundles selected for this study, 64 were above the cut-off point in either one or both of the corpora. In addition, 18 other bundles were below the cut-off point in textbooks but present in research papers, while five bundles were below the cutoff-point in papers but present in textbooks. The remaining 85 items were either below the cut-off point in both corpora, or did not occur in either corpora. Table 1 shows those bundles which occurred in only one of the two corpora; Table 2 ranks those bundles which were more salient in papers; and Table 3 ranks those bundles which were more salient in the textbooks.

Table 1
*Stance Lexical Bundles That Occur Above the Cut-Off Point Only One Corpus*

| Bundles | Corpus |
|---|---|
| have shown that<br>be explained by<br>been shown to<br>can be considered<br>be seen as<br>be considered as<br>important role in<br>the notion of<br>it is clear that<br>it should be noted<br>it has been shown<br>it can be seen<br>it should be noted that<br>it is not possible<br>it has been shown that<br>at least in<br>it is not possible to<br>the tendency to | Occur above the cut-off point only in the corpus of research papers. Either below the cut-off point or do not occur in the corpus of textbooks. |
| we have seen<br>it was found that<br>the sense that<br>it can be shown that | Occur above the cut-off point only in the corpus of textbooks. Either below the cut-off point or do not occur in the corpus of papers. |

Table 2
*Stance Lexical Bundles Occurring More Frequently in Research Papers and Ranked According to Ratio of Frequency in Papers to Frequency in Textbooks*

| Stance lexical bundles | Ratio of Papers Relative Frequency to Textbooks Relative Frequency | Comments |
|---|---|---|
| is consistent with | 5.01 | Occur more than twice as frequently in research papers than textbooks. |
| appears to be | 3.40 | |
| the importance of | 2.46 | |

| Stance lexical bundles | Ratio of Papers Relative Frequency to Textbooks Relative Frequency | Comments |
|---|---|---|
| more likely to be | 2.33 | |
| less likely to | 2.10 | |
| it is difficult to | 2.07 | |
| are likely to | 1.91 | Occur slightly more frequently in research papers than textbooks. |
| it is possible that | 1.68 | |
| it is necessary to | 1.57 | |
| it is worth | 1.55 | |
| the fact that | 1.48 | |
| the possibility that | 1.42 | |
| it is possible to | 1.37 | |
| may not be | 1.36 | |
| according to the | 1.33 | |
| be regarded as | 1.32 | |
| the most important | 1.32 | |
| it appears that | 1.29 | |
| to some extent | 1.27 | |
| is likely to be | 1.27 | |
| the idea that | 1.14 | |
| it is important to | 1.09 | |

Table 3
*Stance Lexical Bundles Occurring More Frequently in Textbooks and Ranked According to Ratio of Frequency in Papers to Frequency in Textbooks*

| Stance lexical bundles | Ratio of Papers Relative Frequency to Textbooks Relative Frequency | Comments |
|---|---|---|
| the possibility of | 0.98 | Occur slightly more frequently in textbooks than research papers. |
| be the case | 0.92 | |
| assumed to be | 0.91 | |
| the assumption that | 0.91 | |
| there may be | 0.89 | |
| it is likely that | 0.84 | |
| we assume that | 0.73 | |
| the problem of | 0.70 | |
| the idea of | 0.65 | |
| to show that | 0.63 | |
| it may be | 0.60 | |
| the hypothesis that | 0.57 | |
| we can see | 0.52 | |
| the assumption of | 0.44 | Occur more than twice as frequently in textbooks than research papers. |
| as a whole | 0.43 | |
| if they are | 0.42 | |
| assume that the | 0.40 | |
| it is easy to | 0.29 | |
| is determined by | 0.26 | |

## Discussion

It is interesting to note that the phrase *it can be shown that* is much more prominent (in terms of frequency) in textbooks, whereas *have shown that* is much more prominent in research papers. This justifies the comparative methodology chosen here, as it allows the language to be selected with more focus for the specific language course being designed. A thesis writing course would certainly need to focus on the top 18 lexical bundles.

Analysis of the concordance lines reveals further pedagogically valuable features of these lexical bundles that could be useful for instructors to know. Due to limitations of space only a few are outlined here. For example, *been shown to* was often used with citations in the introductions of research papers, where it usually refers to established facts from previous research, such as: "nighttime lights have also been shown to be a good predictor of local wealth as measured by the DHS wealth index" (Bruederle & Hodler, 2018, p. 2). This function is especially unnecessary in textbooks which are instructional in nature. Similarly, *can be considered* was used frequently in results and discussion sections where they are used to attribute a quality to the data, thus helping to evaluate the data. For example, "the results obtained can be considered to be robust" (Meier et al., 2015, p. 16), and "From the results of our study, we hypothesize that the organization of centres can be considered a valuable economic resource for metropolitan areas" (Khaili-Miab, van Strien, Axhausen & Grêt-Regamey, 2019, p.13). In addition, the bundle *important role in* was frequently used in introductions to establish the relationship between the research context and the research topic. For example, "Many studies have shown that entrepreneurship plays an important role in stimulating economic growth" (Naminse & Zhuang, 2018, p. 2). Examining the wider context of this concordance line reveals that *stimulating economic growth* refers to the context of poverty, and entrepreneurship was the topic of the whole paper.

The bundle *it is important to* occurred in the introductions to research papers and appeared to have two main functions. The first was to help distinguish between two ideas or to narrow the focus of an idea to be more specific, as in "it is important to recognize improvements in gentrified neighborhoods may not be

associated with city-wide health inequalities" (Gibbons, Barton & Brault, 2018, p. 3) The second function was to provide a specific context that rationalizes the research, as in "It is important to ensure that health systems are responsive to women's and men's needs yet this requires a robust evidence base" (Hosseinpoor et al., 2012, p. 1). In the methodology section, this bundle is used to explain a necessary aspect of the methodology, as in "Therefore, it is important to use appropriate weights to make the estimates representative and comparable over the two survey rounds" (Pathak, Singh & Subramanian, 2010, p. 4). This bundle occurs most frequently in the results and discussion sections and appears to provide some evaluation of some aspect of the methodology, as in "However, it is important to note that there was limited variation in the types of satellite data used in the studies" (Seto, Fragkias, Günerlap & Reilly, 2011, p. 8). This can also include mentioning limitations. The use of this bundle in the results and discussion section also includes reiterating the rationale of the research.

## Conclusion

As can be seen, ranking lexical bundles according to the ratio of relative frequencies facilitates the extraction of salient lexical bundles. They are salient due to the fact that they play a critical role in managing the typical discourse features of the research paper. They can therefore be selected to be taught to students in their thesis writing courses through the specific functions that they have.

There are, however, a number of limitations to this study. First, the current corpora only represent one program within the university, whereas the thesis writing course caters for all students. Additionally, selecting lexical bundles from pre-determined lists limits the scope of the study since it is possible that these corpora contain other lexical bundles in the same categories. Finally, the analysis of the concordance lines has only been carried out by this researcher. A second rater would improve the reliability of the interpretations.

One dimension that remains unexplored in this research is the situation where lexical bundles may have a higher (e.g., *the fact that*) or lower frequency (e.g., *the possibility that*) simultaneously in both corpora and what impact this can have for selecting items for a thesis writing syllabus. Going forward, the

corpora will continue to grow, lexical bundles will be identified directly from the corpora, and the analyses that inform the curriculum will also continue. Software written in R is being developed to help identify lexical bundles more quickly, and this is due to be completed during 2020.

## References

Anthony, L. (2019a). *AntConc*. Retrieved from https://www.laurenceanthony.net/software/antconc/

Anthony, L. (2019b). *AntCorGen*. Retrieved from https://www.laurenceanthony.net/software/antcorgen/

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., Johannsson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English* (9th ed.). Longman.

Bruederle, A., & Hodler, R. (2018). Nighttime lights as a proxy for human development at the local level. *PLoS ONE*, *13*(9). https://doi.org/10.1371/journal.pone.0202231

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Cortes, V. (2013). The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes*, *12*, 33–43. https://doi.org/10.1016/j.jeap.2012.11.002

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). The processing of formulas in native and second language speakers: Psycholinguistic, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x

Gibbons, J., Barton, M., & Brault, E. (2018). Evaluating gentrification's relation to neighborhood and city health. *PLoS ONE, 13*(11), e0207432. https://doi.org/10.1371/journal.pone.0207432

Hosseinpoor, A. R., Stewart Williams, J., Amin, A., Araujo de Carvalho, A.,

Beard, J., Boerma, T., Kowal, P., Naidoo, N., & Chatterji, S. (2012). Social determinants of self-reported health in women and men: Understanding the role of gender in population health. *PLoS ONE*, *7*(4). https://doi.org/10.1371/journal.pone.0034779

Khiali-Miab, A., van Strien, M. J., Axhausen, K. W., & Grêt-Regamey, A. (2019). Combining urban scaling and polycentricity to explain socio-economic status of urban regions. *PLoS ONE 14*(6). https://doi.org/10.1371/journal.pone.0218022

Meier, T., Senftleben, K., Deumelandt, P., Christen, O., Riedel, K., & Langer, M. (2015). Healthcare costs associated with an inadequate intake of sugars, salt and saturated fat in Germany: A health econometrical analysis. *PLoS ONE 10*(9). https://doi.org/10.1371/journal.pone.0135990

Mizumoto, A., Hamatani, S., & Imao, Y. (2017). Applying the bundle–move connection approach to the development of an online writing support tool for research articles. *Language Learning, 67*(4), 885–921.

Naminse, E. Y., & Zhuang, J. (2018). Does farmer entrepreneurship alleviate rural poverty in China? Evidence from Guangxi province. *PLoS ONE 13*(3). https://doi.org/10.1371/journal.pone.0194912

Pathak, P. K., Singh, A., & Subramanian, S. V. (2010). Economic inequalities in maternal health care: Prenatal care and skilled birth attendance in India, 1992-2006. *PLoS ONE 5*(10). https://doi.org/10.1371/journal.pone.0013593

Seto, K. C., Fragkias, M., Güneralp, B., & Reilly, M. K. (2011). A meta-analysis of global urban land expansion. *PLoS ONE 6*(8). https://doi.org/10.1371/journal.pone.0023777

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512. https://doi.org/10.1093/applin/amp058

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

Wulff, S. (2019). Acquisition of formulaic language from a usage-based perspective. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 19–37). Routledge. https://doi.org/10.4324/9781315206615-2

## Author Bio

***Daniel Parsons*** *teaches English for Academic Purposes to master's degree students of international relations and business at IUJ. His primary research interest is in corpus linguistics. <dparsons@iuj.ac.jp>*

**Received:** October 20, 2019
**Accepted:** September 9, 2020

# Appendix

## Textbook corpus

Bradley, T., & Patton, P. (2002). *Essential mathematics for economics and business.* (2nd ed.). Wiley.

Gans, J., King, S., Stonecash, R., Byford, M., Libich, J., & Mankiw, N. G. (2015). *Principles of Economics.* (6th ed.). Cengage Learning.

Mankiw, N. G. (2010). *Macroeconomics.* (7th ed.). Worth Publishers

Newbold, P., Carlson, W. L., & Thorne, B. M. (2013). *Statistics for business and economics*. (8th ed.). Pearson.

Varian, H. R. (2006). *Intermediate microeconomics: A modern approach* (7th ed.: International student edition). W. W. Norton & Company.

Wooldridge, J. M. (2013). *Introductory econometrics: A modern approach* (5th ed.). South-Western Cengage Learning.