An Innovative (and Easy) Approach to Corpus Analysis

By Julia Williams, University of Waterloo, Canada

Abstract

Lee and Swales (2006) suggest that using corpus analysis activities in the classroom provides students with pragmatic tools they can use to identify patterns of language use without relying on native-speaker expertise. In addition, subsequent research on the use of corpus analysis, or data driven learning (DDL) (Boulton & Cobb, 2017), demonstrates that substantial benefits accrue to students who work with corpora (Bridle, 2019; Charles, 2012; 2014). However, the complexity of existing corpus analysis technologies may deter instructors from implementing existing corpora (e.g. COCA) or corpus toolkits (e.g. AntConc) which may require significant time investments to learn and transform into effective pedagogy.

This article describes an easy, innovative approach to harnessing the benefits of corpus analysis using technology with which teachers and students are already familiar. Students build corpora in Word and use the search function to identify grammatical, lexico-grammatical, collocational, and even organizational patterns. This teaching innovation was effective with a class of graduate students. Using easily accessible corpus analysis activities in the classroom encourages students to rely less on teacher expertise and develop skills that support greater learner autonomy.

The benefits of using corpus analysis in EAP classes has been well documented (Anthony, 2017; Boulton & Cobb, 2017; Charles, 2012, 2014, 2018; Lee & Swales, 2006). Lee and Swales (2006) suggest that using corpus analysis activities in the classroom provides students with pragmatic tools they can use to identify patterns of language use without relying on native-speaker expertise. In addition, subsequent research on the use of corpus analysis, or data driven learning (DDL) (Boulton & Cobb, 2017), demonstrates that substantial benefits, in particular enhanced learner autonomy, accrue to students who work with corpora (Bridle, 2019; Charles, 2012; 2014). And Lawrence (2017) advocates for the use of corpus analysis is a useful way for students to build discipline-specific knowledge related to vocabulary and linguistic features, as

well as encourage collaborative, problem-solving skills that students can apply to their own composition processes.

However, the complexity of existing corpus analysis technologies may deter instructors from implementing existing corpora (freely available online, e.g. the Corpus of Contemporary American English [COCA]) or corpus toolkits (freely downloadable for use offline with self-constructed corpora, e.g. AntConc) which may require significant time investments to learn and transform into effective pedagogy. With regards to time requirements, two thirds of students in Bridle's (2019) study felt that learning to use corpus analysis techniques required "too much time and effort" (p. 65). And students are not the only ones concerned about time requirements; instructors are also cautious of new technologies that require significant time to learn and implement. With regards to technical issues, Anthony (2017) himself notes that students and instructors using corpora may encounter "mundane" (p. 165) technical challenges related to slow Internet connections and system failures. In addition, instructors may experience unanticipated technical glitches, such as one this author faced when introducing the COCA to her graduate students. In class, students used the university's wifi to search the freely accessible COCA website for a specific linguistic feature. Free access to the COCA website allows users 50 queries per day. Unfortunately, the COCA website considered the students in class that day as a single user as the university's wifi had assigned the students the same identical Internet protocol (IP) address. The 24 students were only able to complete approximately 2 searches each before exceeding their allowable queries, resulting in wasted class time, and a hastily revised lesson. This incident was sufficient to encourage the author to search for alternative ways to use corpus analysis in her classroom that would neither be susceptible to technical challenges, nor consume unreasonable amounts of limited class time to orient students to the benefits of corpus analysis.

Using Word and its search function, a technology both students and instructors are familiar with, this author developed a series of corpus analysis assignments that allowed graduate students to identify linguistic features common to academic English, and those that were specific to their disciplines. Student responses to the assignments reveal insights into their analytical and composing processes.

The class consisted of 24 graduate students from disciplines as varied as accounting, computer science, earth sciences, engineering (civil, chemical, electrical, and systems design), geography, kinesiology, and statistics. The students attended a graduate-level writing course called Scholarly Writing in English taught by this author. The course textbook was the third edition of Swales & Feak (2012) *Academic Writing for Graduate Students*, and elements of academic style and cohesion identified in this textbook were used as searchable linguistic features in the corpus analysis assignments. Prior to the start of the course, the author developed

a model corpus, established objectives for three assignments, and developed assignment descriptions and rubrics.

The Corpus Analysis Assignments

The objective of the first assignment was simply for students to build a discipline-specific corpus for use in the subsequent corpus analysis assignments. Initial discussions on how to find and identify legitimate and reliable journal articles within the students' disciplines served to raise awareness that the quality of query output would be dependent on the quality of texts used in the corpus. To create their own corpora, students identified academic journal articles relevant to their fields of study and cut and pasted the text (minus tables, figures, author biographies, and references which do not maintain format when transferred to Word) into a Word document. An accurate citation for each paper was included based on the citation format appropriate in the student's discipline. Students then used the colour highlighting feature in Word to colour code the abstract, introduction, methods, results, and discussion sections in each paper. This colour-coding later allowed students to easily identify the sections of the journal articles in which linguistic features were most frequently found.

The resulting corpora were required to be at least 20 pages (or approximately 150,000 words) in length. This resulted in relatively short (or small) corpora, but the 20-page recommendation was suggested to limit the time and effort required to cut and paste text from a journal article pdf. to a Word document. Some students were able to create their corpus by using advanced features of Adobe Acrobat that allowed them to either convert pdf. files directly to Word, or to combine pdf. files into a single searchable document. However, these useful features are not freely available; therefore, to keep costs at zero and avoid technical issues, most students cut and pasted text into a Word document to create no-cost, low-tech corpora.

The Second Corpus Analysis Assignment: Searching for Elements of Academic Style

The second corpus analysis assignment was designed to focus student attention on how the elements of academic style, as identified in Unit 1 of *Academic Writing for Graduate Students* (Swales & Feak, 2012), were used in their disciplines. These elements include the use of formal one-word verbs to replace verb + preposition combinations (e.g. *implement* rather than *put in place*), use of first person pronouns (e.g. I, we), contractions (e.g. can't), formal negative forms (e.g. *no* instead of *not any*), vague expressions (e.g. etc. and so forth), direct questions (e.g. Why has antibiotic resistance increased?), adverbs in mid-position (e.g....was originally developed...), split infinitives (e.g. ...to adequately meet...), passive voice (e.g. ...was determined...), and the second person pronoun (e.g. you) to address the reader. Students were asked to

search their corpora for these elements, represent the results in a table, and, in some cases, create a second column to comment on the results in the first column. After each search, students were asked to briefly explain what they learned from the query.

Figure 1 displays a typical student response to the verb + preposition search in the second assignment. In this example, students searched for the prepositions *up* and *on*, identified the instances where the preposition followed a verb, and represented the results in a table. In the second column, students were asked to find a 'more academic' one-word verb synonym to replace the verb + preposition combination.

Figure 1: Example of Student Searches for verb + up and verb + on

Table 1: search ***up***

#	**Verb + Preposition combination**	**More academic single-word verb**
1	which ***takes up*** valuable land	occupies
2	if WWTPs' generation ***scales up***	This is an accepted "coined" combination in my discipline

There is only a single occurrence of the verb + up combination in each paper. This indicates how rare it is used in my discipline.

Table 2: search ***on***

#	Verb + Preposition combination	More academic single-word verb
1	***quantified on*** a sector by sector basis	
2	energy ***expended on*** drying	
3	converted to ***run on*** biogas in 2005	
4	wild algae ***cultured on*** oxidation ponds	
5	***based on*** actual fitted probability distributions	
6	This paper ***focuses on*** the trade-off	
7	the WWTP is selected ***depending on*** whether the	
8	***running on*** digested biogas from solid end products	
…		

The use of verb + on combination is found to be much more frequent compared to verb + up. No equivalent singles-word verbs were found as a substitute.

In this example, the student discovers that the verb + up combination is not frequently used in their disciplinary corpus, that the *takes + up* combination is replaceable with a more academic one-word verb (*occupies*), and decides that the second occurrence of this linguistic feature (*scales up*) is an acceptable exception to the recommended avoidance of the verb + preposition rule. Interestingly, in Table 2, the

student represents the results of the verb + on search, identifies that verb + on is much more frequently used in their discipline than verb + up, and decides that there are no easily identified one-word verbs to replace these frequently used combinations. As an instructor who often encourages students not to use verb + preposition combinations in their academic writing, I was struck by the difficulty of finding one-word verbs that would eliminate the use of on. For example, we might reasonably replace *based on* with *premised on*, but that does not eliminate the preposition *on*. This seemed to be a useful discovery, not only for the students, but also for instructors, like this author, who might encourage the elimination of verb + preposition combinations in principle, without making a distinction amongst specific prepositions, which may be more or less replaceable. The colours represented in the table reflect the sections in which the examples were found: blue for introduction, pink for methodology, red for results, and grey for discussion. Although the colour coding was not necessary for the analysis of this linguistic feature, it was helpful when, for example, students searched for passive voice verbs, which, they discovered, were mostly found in the methods (or pink) sections of their corpora.

The following quotes are taken from student responses to the 'What have you learned from this search?' question in the second corpus analysis assignment that focused on searches related to the elements of academic style. The quotes reflect what the students learned about written academic English, their disciplines, and in some cases, reflect their sense of humour.

In response to the verb + preposition (*up* and *on*) search, students commented:

- I learned that "up" preceded by a verb is not a word combination that is commonly used in [my discipline]. Even when I look in other papers excluded to my corpus, this combination seems to be way less used than the verb + on. I guess that "on" is more used in general both in written and English speaking.

- I learned that a verb followed by "on" is a common sentence structure in English. Unfortunately for me, the rule to select the most suitable preposition remains a mystery and I will probably continue to use a single word when possible. In my corpus, we could replace the verb + preposition structure with a single word in most of the cases. Nevertheless, there is some structure where the 'on' seems required in the sentence even when using synonyms. The most popular combination is "based on" and appears in every paper while "focus on" appears in ¾ papers. This means that these verbs + prepositions are broadly accepted.

In response to the first-person pronoun searches (search *I*, *we*), students responded:

- Among the papers selected for my corpus, there was no first-person pronoun used. The community prefers to use the first-person plural pronoun especially when they are describing or referring to their own experiences. This thought can be extended to computer science paper. …computer science papers are usually written in groups and therefore, the first-person [singular] pronoun is not adopted. As demonstrated by the colour code, the first-person plural pronoun can be used in every section of the paper.

- Personal pronouns should not be used in academic writing. I just found one personal pronoun we in the corpus, in the discussion section. In my discipline active voice is not common.

These two comments were from students in different disciplines. In class, students were interested in the differences in how their disciplines took up the use of specific linguistic features, and there was often cross-talk in the class as one disciplinary group of students asked students in other disciplines the results of their queries. In the second comment here, the student has learned from the search for first person personal pronouns not only that the first person pronouns were not frequently used in her discipline, but that the avoidance of first person pronouns translates to the frequent use of the passive voice. Both events – the interdisciplinary discussions that occurred in class, and the individual student's insight into how the avoidance of personal pronouns connects with frequent use of the passive voice in her discipline – were indicators that the corpus analysis activities were relevant and useful for the students.

In response to the searches for formal negative forms (search *no*, *few*, and *little*), students commented:

- I learn that the words "few" and "little" are quite uncommon to behaviour planner and computer science papers since they do not quantify properly the claims. Nevertheless, "no" seems to be a word accepted by the community since it can be found in 3/4 of the paper in my corpus. Sometimes, a more academic substitution exists like "irrelevant" instead of "no longer relevant", but in most case the word "no" is used to make the claim less wordy.

In response to the searches for vague expressions which should be avoided (search *etc.* and *so forth*), students noted:

- It [etc.] was only found once in the whole corpus, which indicates that it is uncommon to use them.

- I learned that "so forth" is uncommon to computer science paper. Nevertheless, "etc." is profusely used in my corpus. I do not believe that this is generally the case in computer science papers, but it is rather a particularity of my field of application. The behaviour planning problem is scenario-

based which means that there exist many special cases that a system needs to cover. Thus, the community exemplifies some of these cases based on their properties and simply mention that the reader must extrapolate the remainder.

In response to searches for passive voice verb use, students searched 'was' and 'were' followed by a past participle. A student responded to the searches as follows:

- The passive voice is extensively used in all sections of the corpus. "was used" and "was selected" are the most common passive verbs present in the corpus due to the nature of my research area where several parameters need to be set for modelling purposes.

Unfortunately, space constraints prevent the inclusion of student responses from each search. In each case, students' comments demonstrated their ability to think critically about the results of their searches and extrapolate their discoveries to broader understandings of their disciplines.

Searching for Elements of Cohesion

In addition to searching for elements of academic style, students were also asked to search for elements of textual cohesion as identified in the textbook. Included in this category was repetition of key words in various forms (i.e. parts of speech), use of it as a pronoun that refers to an antecedent, use of this/these + noun as a summary phrase, and use of connectors such as coordinate and subordinate conjunctions.

In response to searches for repetition of key words in various forms (parts of speech), students identified key words in their research areas and searched for the root form of the key words. For example, a student researching the recovery of resources from wastewater noted that recovery was a key word in his discipline. For this query, he searched *recover* (root word) to identify frequencies of use of *recover* as well as *recovery*, *recovering*, and *recovered*. The students had frequently been told to use synonyms to avoid repetition of key words. They were astonished to see how textual cohesion was achieved through repetition of key words, and how shifting key words to new parts of speech was a strategy to reduce the appearance of repetition.

A student responded to this search for key word repetition using various parts of speech as follows:

- The ability to connect ideas by means of repetition of key words and phrases sometimes meets a natural resistance based on the fear of being repetitive. We've been trained to loathe redundancy. Now we must learn that catching a word or phrase that's important to a reader's comprehension of a piece and replaying that word or phrase creates a musical motif in that reader's head. Unless it is overworked and obtrusive, repetition lends itself to a sense of coherence (or at least to the illusion of coherence).

The Third Corpus Analysis Assignment: Observing Characteristics of Academic Writing in English

The objective of the third corpus analysis assignment was to draw students' attention to some common features of academic writing in English. Swales and Feak (2012) identify these features in the *Language Focus* sections of their textbook. These features included attention to verbs used in definitions (e.g. *known as, defined as*), prepositions used before *which* to start adjective clauses (e.g. *at which*), -ing phrases of cause and effect (e.g. *resulting in*), word order in indirect questions, linking *as* clauses (e.g. *as can be seen*), indicators of strength of claim (e.g. *we think, likely, clearly*), and claim modification (e.g. *may, tends to, based on limited data*).

To identify the frequency of definitions in academic texts, students searched for verbs commonly used in definitions such as *called, known as, defined (as), denoted,* and *referred to*. Students commented on the results of these searches as follows:

- I think that the papers in my corpus may be focused on an engineering audience which results in fewer definitions. However, definitions and even short definitions are very important when the writer is introducing equation variables, acronyms, or specific terms.
- The definition verbs are common in academic texts. The most common verb in my corpus is defined/ defined as that can mostly be found in the methodology sections.

To identify how prepositions are used before which when adjective clauses are objects of a preposition, students searched for *at which, for which, in which,* and *of which*. A student responded to the results of these queries as follows:

- I would say I found another reason why it is not good style to end a sentence with a preposition. In my examples, it would be odd to have the preposition at the end.

To identify the frequency of using *–ing* clauses of cause and effect, students searched for *thus + verb-ing, resulting in + verb-ing, leading to + verb-ing, and causing + verb-ing*. Post-search, students explained what they had learned.

- I think that the "ing" clauses helps to reduce the wordiness of the writing while explaining the cause and effect.
- These -ing clauses are used to introduce the result of an action within a single sentence. Although these clauses are very common in formal written English, I only found 1 in my corpus, indicating

it's not common in my discipline.

To identify the frequency of linking as clauses in their corpora, students searched for *as can be seen, as seen, as such, as a result, as a consequence, as noted, as determined, as expected*. In response to the question 'What did you learn from this search?' students stated:

- I could not find instances of "as seen" in my corpus; however, "as shown" is extensively used with 30 search results. Also, I think that *as* clauses are very useful to introduce informative statements.
- Linking *as* clauses are a nice, short and efficient way to refer to a figure or table in the method or result section.

To identify how researchers modified their claims, students searched their corpora for hedging indicators (e.g. *may, might, could*) as well as *it seems* and *it appears*. Once finished, students reported the following:

- The word "may" is used many times to moderate a claim in my corpus. Additionally, I would say there are many ways to moderate or qualify a claim.
- I found hedging an important feature of academic writing, because academic writers need to clearly indicate whether they think claims are certain, likely, unlikely, or just false. On the other hand, I realized writing that contains too many qualifiers can sound unclear and wordy.

To determine researchers' verb tense usage when writing about research, students searched for *(* or *[* as indicators of citations that would reveal when researchers were paraphrasing or summarizing others' research. Students then verified the verb tense and voice used to write about research. A student responded to the search results as follows:

- The researchers in my corpus used present, present perfect and past tense verbs to refer to past studies. It seems that when referring to current research, there is no rule about the use of the verb tense. However, it looks like the present tense is often used when referring to a process of the research, and past tense when referring to the research itself or the researchers.

Evidence of Success

Although it can be difficult to determine the success of new pedagogical intervention, the author noted the following outcomes that suggested students were attentive to the corpus analysis searches. First, several students independently expanded their corpora to achieve more robust results. They quickly noted that a larger corpus would provide more reliable data, and without being prompted, they added papers to their corpora to achieve more consistent results. Second, two students who were studying in interdisciplinary

fields became aware of the disciplinary differences of writing expectations in their fields. In both cases, they developed two new corpora, one corpus for each discipline. They enjoyed informing their classmates of the differences between the disciplines and attempting to discern which patterns of use they should follow. A further positive outcome was demonstrated by students' in-class behaviour. Students rapidly habituated to completing searches of their corpora, and when the author drew her students' attention to specific linguistic features, without prompting, they would open their corpora files and search for the feature immediately. They would then enjoy informing the class whether the feature was present in their corpora, and to what extent. They were also interested in knowing the results of the searches in corpora constructed from journal articles in different disciplines. And significantly, as can be seen from their comments, students were successful at making connections between the linguistic features for which they searched and academic writing conventions in their disciplines.

The final indicator of success came in the form of student comments on their course evaluations. Several students' comments related to the corpus analysis assignments.

- First, the corpus analysis allowed me to discover the expectations of my target audience.
- The introduction of corpus assignments was very helpful, and I think it achieved more than regular grammar classes.
- Democratization of knowledge. Let students gain knowledge of academic style, grammar and vocabulary based on corpus search in their own disciplinary.

Conclusion

The experience of creating self-compiled, low-tech corpora in Word and integrating corpus analysis assignments in a graduate writing class seemed to be successful at stimulating student awareness of how linguistic features (as identified by Swales and Feak in their 2012 textbook *Academic Writing for Graduate Students*) were used in the students' disciplines. In addition, the assignments were created by the instructor and completed by the students at no additional cost, without encountering the technical challenges that may affect corpus analysis work with existing large corpora or corpus toolkits, and with minimal class time devoted to the explanation of the process of corpus analysis. It would appear that some form of search can be completed in Word to identify each of the textually significant linguistic features as enumerated by Swales and Feak. (For example, although it would be possible to search the COCA directly for a comprehensive list of verb + preposition combinations, the results of this search can be replicated in Word by searching for common prepositions and asking students to identify which of the occurrences are preceded by a verb.) Furthermore, students appeared to learn from their search queries about how linguistic features are used in

their disciplines, and how different disciplines may employ these features to varying degrees.

Paralleling the corpus analysis assignments, students were also writing discipline-specific assignments such as a problem-solution text, a data commentary, and an article summary (all addressed in Swales and Feak, 2012). Students were encouraged to integrate the knowledge they learned through their corpus searches into their writing assignments. Further research could be done to determine the extent to which students were successfully able to integrate the linguistic features they searched for in their corpora into their writing.

The author wishes to thank her students for their in-class work and the generous permission to use their comments in this paper.

References

Anthony, L. (2017). Introducing corpora and corpus tools into the technical writing classroom through data-driven learning (DDL). In J. Flowerdew & T. Costley (Eds.). *Discipline-Specific Writing* (162–180). New York; Routledge.

Antony, L. (2019). *Laurence Anthony's website*. Retrieved from https://www.lawrenceanthony.net/

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning: A Journal of Research in Language Studies, 67*(2), 348–393.

Bridle, M. (2019). Learner use of a corpus as a reference tool in error correction: Factors influencing consultation and success. *Journal of English for Academic Purposes, 37,* 52–69.

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes, 31*(2), 93–102.

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes, 35,* 30–40.

Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes, 36,* 15–25.

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1), 56–75.

Nesi, H. (2016). Corpus Studies in EAP. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes* (206-217). New York: Routledge.

Swales, J. & Feak, C. (2012). *Academic writing for graduate students (3rd ed.).* U.S.A.: Michigan University Press.

Tribble, C. & Wingate, U. (2013). From text to corpus – A genre-based approach to academic literacy instruction. *System 41/1*(2), 307–321.

Author Bio

Julia Williams is an experienced EAP instructor with over 30 years of teaching in second language contexts. She is the author of LEAP Reading and Writing, levels 3 and 4, and the Director of English Language Studies at Renison University College, University of Waterloo. She attempts to translate theory into effective pedagogy and teaching materials that are easy to implement.