Using Word Frequencies to Introduce Corpora in the Classroom

Jonathan Ferries

Bukkyo University

Classroom use of corpora has long been advocated but remains rare in practice. In this paper, I summarize the obstacles to corpus use, describe a simple frequency-based method of analysis that avoids these problems, and offer a gateway to a more sophisticated use of corpora to promote more active, independent learning.

In this paper, I describe a two-stage method of using word frequency counts from online corpora to provide a simple introduction to using online corpora in the classroom. Researchers have long advocated using corpora to enable learners to explore and make their own inductive language discoveries through data-driven learning (see Johns, 1986). In practice, however, corpora remain rarely used in the classroom (Boulton, 2017; Breyer, 2009; Ma et al., 2022; O'Keefe & Farr, 2003). This is attributed on the teacher side to lack of training in corpus skills (Boulton, 2017; Breyer, 2009; Leńko-Szymańska, 2014), and on the student side to insufficient knowledge of the metalanguage required to perform corpus searches (Chang, 2014; Yeh et al., 2007). Studies have also drawn attention to the problems faced by both teachers and students in analyzing and interpreting concordance lines (Breyer, 2009; Ma et al., 2022).

Concordance lines are lines of text centered around a nodal word or phrase, and producing them is one of the basic functions of corpus analysis software. However, they can be challenging to read for the inexperienced (O'Keefe et al., 2007) and are considered more suitable for use with more advanced learners. This poses a challenge for practitioners in locations such as Japan, where many learners do not advance beyond the lower intermediate level (Hadley, 2001).

To address these issues, the method presented below makes use of another basic function of corpus software: calculation of word frequencies. These calculations are relatively easy to perform, and they produce results that are visually easy to grasp. I illustrate how this method can be applied with an example drawn from the domain of English for academic purposes (EAP) that I have used with Japanese L1 learners of English majoring in sciences at the undergraduate level. With appropriate modifications, this method may be used with learners of lower intermediate level and above in other domains of English for specific purposes (ESP).

However, word frequencies cannot tell us how and in what context a word or phrase is used. For this, concordance lines showing a certain degree of co-text are required. The method described below serves only as a relatively accessible gateway to more sophisticated use of corpora, including concordances, in the classroom.

Use of “I” in academic writing

My example concerns use of the first-person pronoun in academic writing. This is a topic of clear relevance to EAP students, and one regarding which students may have received advice in the past. For example, both lower-intermediate textbooks (e.g., Ackert et al., 2014) and writing guides for advanced students (e.g., Bailey, 2011) from major publishing houses recommend avoidance of the first person in academic writing. Several corpus-based studies have demonstrated that this advice is not always followed in actual academic writing. Hyland (2001) and Dobakhti and Hassan (2017), for example, assembled their own large corpora to reveal extensive use of the first person in research articles and considerable variation between disciplines.

However, approaches like these that use purpose-built corpora are unfeasible for ESP classroom use. This is because they use corpora that take time to prepare and specialized software that takes more time to master. With time in the language curriculum often scarce, students and teachers need ready-made corpora that are cheap, accessible, and relevant. Several large corpora and corpus analysis tools are now accessible online that meet these requirements. These include the British National Corpus (Davies, 2004), the Corpus of Contemporary American English (COCA) (Davies, 2008), and Lextutor (Cobb, 2017). I chose COCA because of its size and breadth in content, the simplicity of its search interface, the availability of instructions on its use in Japanese, and its cost (students can perform up to 50 searches per day for free).

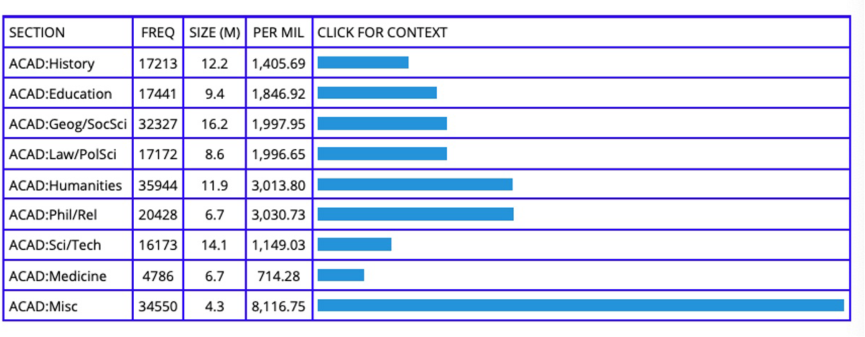
Stage 1: Searching the corpus

The first stage consists of searching the chosen corpus for the features under investigation. In my example, the aim was to investigate how academic writers refer to themselves and their actions in order to determine whether they avoid using "I" and instead use structures such as the passive voice. This was operationalized by searching for occurrences of the pronoun *I* (as a proxy for first-person authorial references) and *be* + past participle (as a proxy for passive-voice authorial references) in each of the academic disciplines contained in COCA. This is potentially the most challenging step for teachers and students unfamiliar with corpora. Corpora are normally annotated by "tagging" to indicate the part of speech of each word. The purpose of this is to allow users to search for instances of, for example, *will* used as a modal verb but not as a noun. Users require some knowledge of these tags in order to perform effective searches. For simplicity, I used the query syntax `i_p*` to search for instances of the pronoun *I*, and the query syntax `are|were  [vvn*]` to search for passive *be* + past participle constructions.[1]

A search of this kind produces results indicating the frequency of occurrence in each discipline of the features searched for (Figure 1).

These may be converted by students or the teacher to chart form (Figure 2). This shows, in a visually easy-to-grasp manner, that the advice not to use "I" in academic writing is not always followed by academics in practice; the first person appears to be used to some degree in all disciplines, and it is used more frequently than the passive voice in the Humanities and Philosophy/Religion.[2]

Owing to the crudity of the proxies and query syntax used, these results provide only a rough indication of the relative frequencies of use of the first



SECTION	FREQ	SIZE (M)	PER MIL	CLICK FOR CONTEXT
ACAD:History	17213	12.2	1,405.69	
ACAD:Education	17441	9.4	1,846.92	
ACAD:Geog/SocSci	32327	16.2	1,997.95	
ACAD:Law/PolSci	17172	8.6	1,996.65	
ACAD:Humanities	35944	11.9	3,013.80	
ACAD:Phil/Rel	20428	6.7	3,030.73	
ACAD:Sci/Tech	16173	14.1	1,149.03	
ACAD:Medicine	4786	6.7	714.28	
ACAD:Misc	34550	4.3	8,116.75	

Figure 1. Selected raw outputs (Source: COCA).

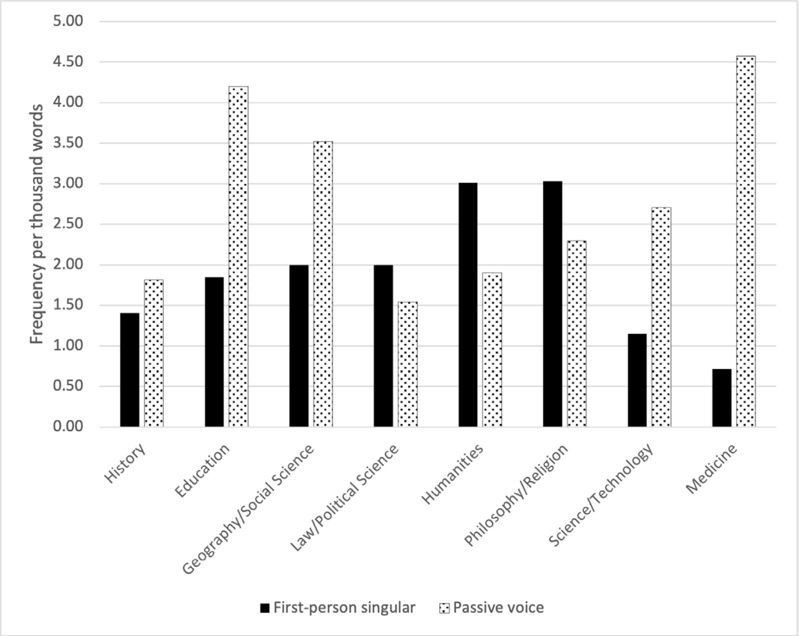


Figure 2. First person and passive voice frequencies by discipline

person and passive voice in different disciplines. Use of pronoun *I* as a proxy excluded first-person pronouns in other cases (*me*, *my*, *us*, *our*), which means that the results do not accurately reflect the frequency of use of all forms of first-person expression. The search for passive *be* + past participle using the syntax `are|were [vvn*]` is likewise limited in scope. It fails to capture, for example, constructions using the more colloquial passive auxiliary *get* and instances where an adverb occurs between the auxiliary and past participle (e.g., *were **frequently** referred to*). A second, more in-depth stage of analysis is therefore required.

Stage 2: Manual search

For the second stage, a manual search is made of a smaller sample of texts. In the case of my example, a manual search was made for all instances of pronoun *I* referring to the writer (excluding, for example, instances in quotations and appended questionnaires) and all instances of the passive voice used to describe actions (including judgments and other mental processes) performed by the author.

Given time and other classroom constraints, the manual analysis was limited to the two disciplines found to exhibit the greatest relative differences in frequency of use of the first person and passive voice, namely Education and the Humanities (Figure 2). Three research articles (RAs) were randomly selected from the most recent issue of a journal from each of these disciplines—*ELT Journal* and *The Art Bulletin*—chosen based on online availability to students and the advice of expert informants in each discipline. However, alternative methods of selection, such as a journal rating metric or students' own knowledge of the field of interest, could be used.

This analysis, performed manually by the teacher and students, revealed that passive-voice use exceeded first-person use in all three Education RAs (Figure 3). Conversely, first-person use exceeded passive-voice use in all three Humanities RAs (Figure 4). These results corroborate the results of the cruder computerized analysis at the first stage.

Implications for the classroom

The first stage of the above analysis introduces newcomers to corpora to the tags and metalanguage needed to perform more sophisticated corpus analyses. The second stage introduces learners to complete texts. This stage exposes them to the organization and constitution of texts in their fields of interest. Thus, in the case of the example described above, students noted, unprompted, the different writing strategies employed in the acknowledgement sections compared to the

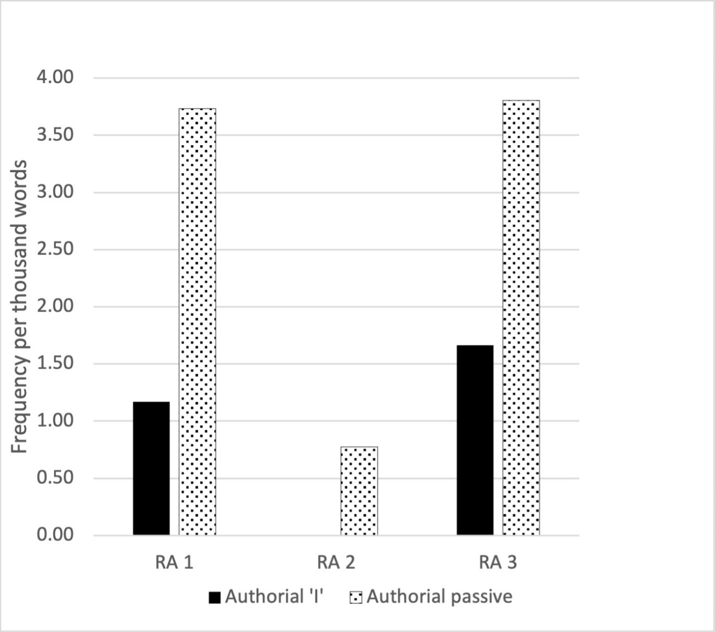


Figure 3. Education sample.

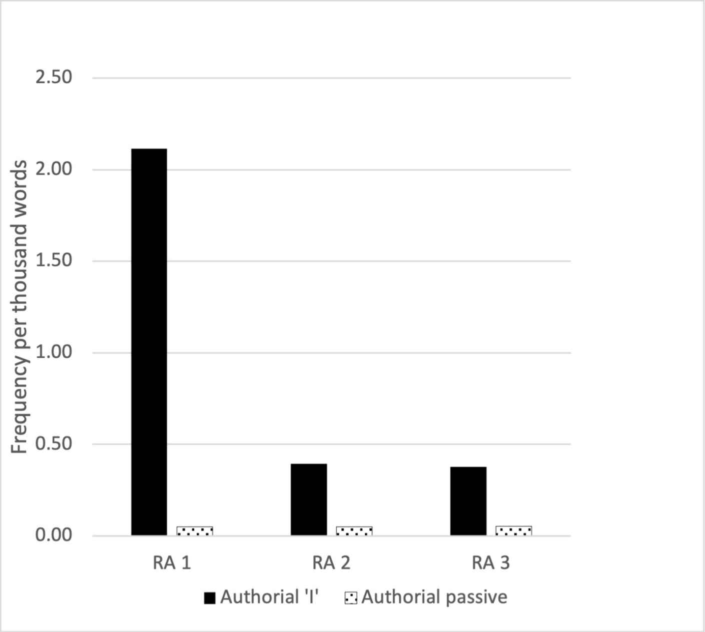


Figure 4. Humanities sample.

other parts of academic articles. Importantly, from the point of view of data-driven learning, the second stage raises students' awareness of how the words that they searched for at the first stage are used in context. This serves as a gateway for introducing concordances once students have mastered the necessary basic corpus skills. Concordances, and the tools used to obtain them, are considered a mainstay of data-driven learning, and the simple frequency-based corpus analysis proposed here offers one way of overcoming the obstacles to corpus use outlined at the outset to equip students to become more active, independent learners (Chen, 2011).

Notes

1. A full list and descriptions of these and other tags used by many online corpora, including COCA, can be found at the CLAWS part-of-speech tagger for English website. https://ucrel.lancs.ac.uk/claws
2. The disciplinary categories cited here are those used in COCA.

References

Ackert, P., Lee, L., Haynes, H., & Beck, J. (2014). *New reading and vocabulary development 2: Thoughts and notions (teacher's edition)*. Cengage Learning.

Bailey, S. (2011). *Academic writing: A handbook for international students*. Routledge.

Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, *50*(4), 483–506. https://doi.org/10.1017/S0261444817000167

Breyer, Y. (2009). Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning*, *22*(2), 153–172. https://doi.org/10.1080/09588220902778328

Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, *26*(2), 243–259. https://doi.org/10.1017/S0958344014000056

Chen, H.-J. H. (2011). Developing and evaluating a web-based collocation retrieval tool for EFL students and teachers. *Computer Assisted Language Learning*, *24*(1), 59–76. https://doi.org/10.1080/09588221.2010.526945

Cobb, T. (2017). *Range for texts v.3* [computer software]. Lextutor. https://www.lextutor.ca/cgi-bin/range/texts/index.pl

Davies, M. (2004). *British national corpus* (from Oxford University Press). https://www.english-corpora.org/bnc/

Davies, M. (2008). The corpus of contemporary American English (COCA): 560 million words, 1990-present. https://corpus.byu.edu/coca

Dobakhti, L., & Hassan, N. (2017). A corpus-based study of writer identity in qualitative and quantitative research articles. *The Southeast Asian Journal of English Language Studies, 23*(1), 1–14. https://doi.org:10.17576/3L-2017-2301-01

Hadley, G. (2001). Concordancing in Japanese TEFL: Unlocking the power of data-driven learning. In K. Gray, M. Ansell, S. Cardew, & M. Leedham (Eds.). *The Japanese learner: Context, culture and classroom practice* (pp. 138–144). Oxford University Department for Continuing Education.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes, 20*, 207–226. https://doi:10.1016/S0889-4906(00)00012-000012-0 "10")

Johns, T. (1986). Micro-concord: A language learner's research tool. *System*, *14*(2), 151–162. https://doi.org/10.1016/0346-251X(86)90004-790004-7 "11")

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, *26*(2), 260–278. https://doi.org/10.1017/S095834401400010X

Ma, Q., Chiu, M. M., Lin, S., & Mendoza, N. B. (2023). Teachers' perceived corpus literacy and their intention to integrate corpora into classroom teaching: A survey study. *ReCALL*, *35*(1), 19–39. https://doi.org/10.1017/S0958344022000180

O'Keefe, A., & Farr, F. (2003). Using language corpora in initial teacher education: Pedagogic issues and practical applications. *TESOL Quarterly*, *37*(3), 389–418. https://doi.org/10.2307/3588397

O'Keefe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom:*

Language use and language teaching. Cambridge University Press.

Yeh, Y., Liou, H. C., & Li, Y. H. (2007). Online synonym materials and concordancing for EFL college writing. *Computer Assisted Language Learning*, *20*(2), 131–152. https://doi.org/10.1080/09588220701331451

Author bio

Jonathan Ferries *is a contract lecturer at Bukkyo University in Kyoto, Japan. f-paul@bukkyo-u.ac.jp*

Received: January 20, 2023

Accepted: May 18, 2023