Slipping through the cracks of e-lexicography: lessons from ColloCaid

Rees, Geraint; Frankenberg-Garcia, Ana; Lew, Robert; Roberts, Jonathan C.; Sharma, Nirwan; Butcher, Peter

Published: 01/10/2019

Early version, also known as pre-print

Slipping Through the Cracks of e-Lexicography

Geraint Rees[1], Ana Frankenberg-Garcia[1], Robert Lew[2], Jonathan C. Roberts[3], Nirwan Sharma[3], Peter Butcher[3]

1 University of Surrey, Guildford, UK
2 Adam Mickiewicz University, Poznań, Poland
3 Bangor University, Bangor, UK

E-mail: g.rees@surrey.ac.uk, a.frankenberg-garcia@surrey.ac.uk, rlew@amu.edu.pl, j.c.roberts@bangor.ac.uk, n.sharma@bangor.ac.uk, p.butcher@bangor.ac.uk

Thanks to the corpus revolution which underpins e-lexicography, headword lists and defining vocabularies can now be adjusted to better reflect current language use. Definitions can be enhanced with information that goes beyond introspection alone. Syntactic patterning, lexical collocations and phraseology in general can now be given much more comprehensive coverage. Good dictionary examples can be more easily found (Kilgarriff, Husak, McAdam, Rundell, & Rychlý, 2008). Yet despite these unparalleled and undeniable advantages, there are elements of word usage that appear to be slipping through the cracks of corpus analyses and corpus-based lexicographic resources. Issues encountered during the development of the ColloCaid project, particularly its base word list, provide examples of such slips and a valuable opportunity to reflect on why they come about and how they might be addressed in future e-lexicography projects.

The ColloCaid tool is an integrated text editor and writing assistant which aims to help academic writers choose the most appropriate collocations in a way that does not distract them from the writing task at hand. The ColloCaid base list comprises circa 500 trigger lemmas which when typed into the text editor offer the user the possibility of selecting a collocate typically found in academic writing along with examples of the collocation in use. Three respected published sources—the AVL-BAWE list (Durrant, 2016), the Academic Keyword List (Paquot, 2010), and the list of bases from the Academic Collocation List (Ackermann & Chen, 2013) as presented in Mayor (2013)—were used in an attempt to ensure that the tool covered words that academic writers really need (Frankenberg-Garcia, Lew, Roberts, Rees, & Sharma, 2019). However, in the process of developing the tool it has become apparent that there are key words—denoting important interdisciplinary academic concepts (e.g., the nouns *omission, paper, step*)—which are not present in the base list. It is hypothesised that these omissions are artefacts of the methods used to compile the sources from which the base list was derived.

In a critical examination of the methods employed to create the source lists, the present paper posits a number of possible accounts for these omissions. These include differences in vocabulary use between student texts written for assessment purposes and professionally published academic writing, the limitations of distributional approaches when dealing with synonymy, particularly those base words which exhibit discipline-specific meanings (c.f. *code* in computing versus biology), and the decontextualized use of qualitative criteria such as judgements about pedagogical relevancy to delimit collocations. In order to test these explanations an expert informed procedure is carried out with the aim of revealing those key words which are missing from the base list. Cross-referencing these missing words with the accounts resulting from the critical examination reveals that they have considerable explanatory power.

These findings not only serve as a reminder of the dangers of the uncritical reuse of lexicographical sources for purposes beyond their original design, but also highlight the limits of many e-lexicographical methods and suggestions for their improvement.

Acknowledgements

This research is funded by the Arts and Humanities Research Council (AHRC), UK, ref. AH/P003508/1

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocations List (ACL) – A Corpus-driven and Expert-judged Approach. *Journal of English for Academic Purposes*, *12*, 235–247.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, *43*, 49–61. https://doi.org/10.1016/j.esp.2016.01.004

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(01), 23–39. https://doi.org/10.1017/S0958344018000150

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425–432). Barcelona: Universitat Pompeu Fabra.

Mayor, M. (2013). *Longman Collocations Dictionary and Thesaurus*. Harlow: Pearson Education.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.