ColloCaid: a tool to help academic English writers find the words they need

Ana Frankenberg-Garcia[1], Geraint Rees[2], Robert Lew[3], Jonathan Roberts[4], Nirwan Sharma[5], and Peter Butcher[6]

Abstract. This short paper summarizes the development of ColloCaid (www.collocaid.uk), a text editor that supports writers with academic English collocations. After a brief introduction, the paper summarizes how the lexicographic database underlying ColloCaid was compiled, how text editor integration was achieved, and results from initial user studies. The paper concludes by outlining future developments.

Keywords: collocation, EAP, writing, e-lexicography.

1. Introduction

Research has shown that less experienced users of academic English have a limited repertoire of collocations (Frankenberg-Garcia, 2018). Indeed, collocations like *REACH+conclusion* are among the most frequent look-ups among novice users of written academic English (Yoon, 2016).

There are a number of tools and resources that academic writers can use to search for such idiomatic combinations of words. These include general English dictionaries and more targeted ones like the *Longman Collocations Dictionary and Thesaurus* (Mayor, 2013) or the *Oxford Learner's Dictionary of Academic English* (Lea, 2014). Writers familiar with corpora can also consult general English corpora

1. University of Surrey, Guildford, United Kingdom; a.frankenberg-garcia@surrey.ac.uk; https://orcid.org/0000-0001-9623-7990
2. University of Surrey, Guildford, United Kingdom; g.rees@surrey.ac.uk; https://orcid.org/0000-0002-9204-8073
3. Adam Mickiewicz University, Poznań, Poland; rlew@amu.edu.pl; https://orcid.org/0000-0002-6772-210X
4. Bangor University, Bangor, Wales, United Kingdom; j.c.roberts@bangor.ac.uk; https://orcid.org/0000-0001-7718-3181
5. Bangor University, Bangor, Wales, United Kingdom; n.sharma@bangor.ac.uk; https://orcid.org/0000-0002-6576-3848
6. Bangor University, Bangor, Wales, United Kingdom; p.butcher@bangor.ac.uk; https://orcid.org/0000-0002-3361-627X

How to cite this article: Frankenberg-Garcia, A., Rees, G., Lew, R., Roberts, J., Sharma, N., & Butcher, P. (2019). ColloCaid: a tool to help academic English writers find the words they need. In F. Meunier, J. Van de Vyver, L. Bradley & S. Thouësny (Eds), *CALL and complexity – short papers from EUROCALL 2019* (pp. 144-150). Research-publishing.net. https://doi.org/10.14705/rpnet.2019.38.1000

like the BNC and COCA, and corpora of student papers like BAWE (Nesi, 2011) and MICUSP (Romer & Swales, 2010). Other useful tools include SkELL (Baisa & Suchomel, 2014), arguably the easiest to use English corpus available, FlaxLC (Wu, Fitzgerald, Yu, & Witten, 2019), a learner-friendly corpus-based collocation tool, and LEAD (Granger & Paquot, 2015), an academic English dictionary-cum-corpus.

However, writers may not know where or how to look up collocations (Frankenberg-Garcia, 2011), or may simply not realize that their emerging texts could be made more idiomatic (Frankenberg-Garcia, 2014; Laufer, 2011). Moreover, even when writers realize they need help, looking up collocations while writing can be distracting and disruptive (Yoon, 2016).

To address this challenge, we are developing a text editor that assists writers with academic English collocations (Frankenberg-Garcia et al., 2019a). ColloCaid provides writers with collocation suggestions as they write, helping them find idiomatic combinations of words and expand their collocational repertoire. ColloCaid can also be used to revise collocations in existing drafts.

2. Lexicographic database

The ColloCaid lexicographic database aims to address core collocations used across disciplines in general academic English. As detailed in Frankenberg-Garcia et al. (2019a), it draws on the noun, verb and adjective lemmas that occur in at least two of three well-known academic vocabulary lists: the Academic Keyword List (Paquot, 2010), the Academic Collocation List (Ackermann & Chen, 2013), and the Durrant (2016) subset of the Gardner and Davies (2014) Academic Vocabulary List.

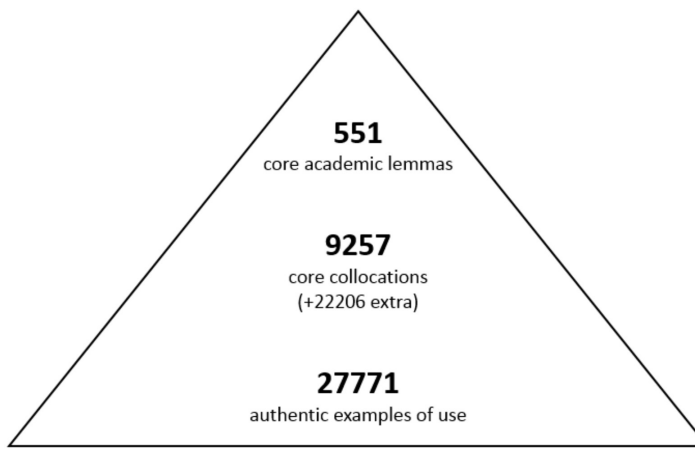
The original selection of lemmas has been revised to (1) disambiguate polysemy (e.g. *figure* as image, as number and as person); (2) include homographs used in academic contexts (e.g. *aim* was initially only listed as a noun, but its less frequent verbal lemma was added to avoid the impression that only the noun was idiomatic); (3) discard lemmas that are not collocationally productive (e.g. *actual*); and (4) add high-frequency interdisciplinary academic lemmas like *paper* and *table*, which slipped through initial selection thresholds (Rees et al., 2019).

The database was populated with interdisciplinary collocates pertaining to the above lemmas extracted from corpora of expert academic English writing. As

detailed in Frankenberg-Garcia et al. (2019a), this was undertaken using Sketch Engine (Kilgarriff et al., 2014), which automatically summarizes the main collocations of a lemma in a corpus. Issues with the extraction have been dealt with using lexicographic judgment on a case by case basis. This included, for example, overruling the classification of *regard* as a verb, since its primary use in academic texts is preposition-like, in contexts such as *decisions regarding safety*, or in prepositional phrases like *with regard to* (Rees et al., 2019).

The database was further populated with authentic examples of collocations in use, selected according to typicality, informativity, and intelligibility. Examples were also curated to address language production needs and maximize their potential for data-driven learning, as explained in Frankenberg-Garcia (2014). Figure 1 summarizes the lexical coverage of ColloCaid in its current 0.4 version (20 September 2019).

Figure 1. ColloCaid 0.4 lexicographic database



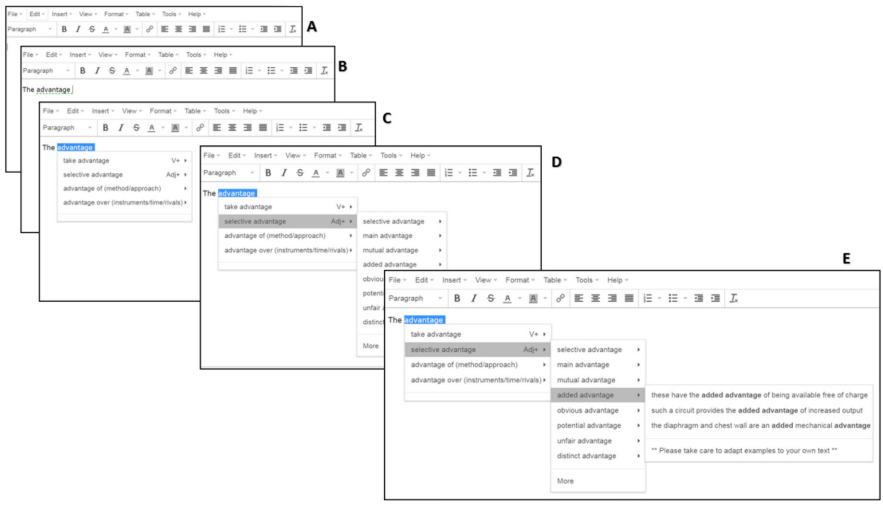
3. Text editor integration

Academic writers from different disciplines have their own preferred operating systems and text editors. In our interdisciplinary research team, for example, papers initiated by the linguists are normally drafted in a Windows environment using Microsoft Word, whereas the computer scientists prefer to use Macs and LaTeX editors. For developing a prototype and testing it with different users, we opted for an online editor that can be accessed from a normal browser compatible with

multiple devices and operating systems, without the need to download additional software. TinyMCE (https://www.tiny.cloud/), a widely used open-source editor that looks like any regular editor was selected for this purpose (Figure 2: A).

We adopted a dynamic, data-driven learning approach to the integration of the lexicographic data into the editor. It is data-driven because collocations suggestions are shown rather than explained. It is dynamic because collocations are displayed only when wanted, and in as much detail as desired, via progressive interactive menus (Figure 2: B-E).

Figure 2. ColloCaid editor

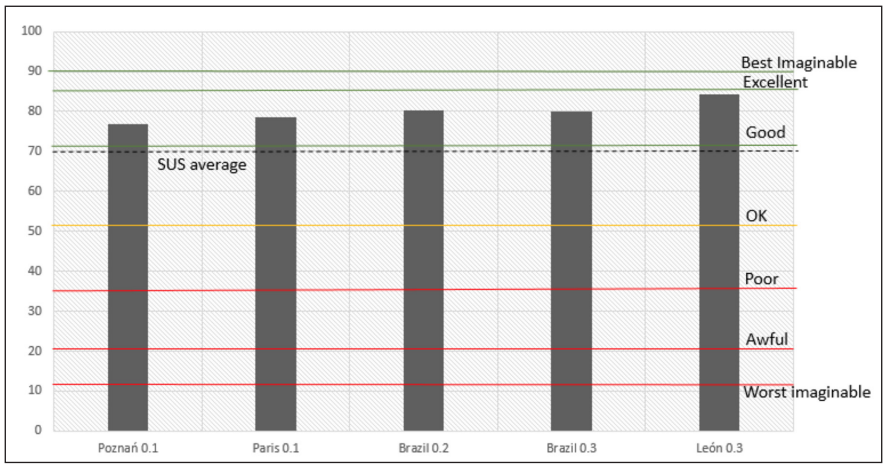


4. Initial user studies

Development versions of ColloCaid have been tested during university writing workshops and seminars in Brazil, France, Poland, and Spain (Frankenberg-Garcia et al., 2019b). Participants (N=122) included novice and expert L2 English writers from a wide range of disciplines. Due to space restrictions, we are only able to present here the scores obtained on the Brooke (2013) System Usability Scale (SUS). The SUS is a standard for measuring the usability of systems (hardware, software, websites, etc.), with the advantage that its results can be compared on the same scale with hundreds of other systems. It comprises ten alternating positive and negative statements about system usability which users rate with a Likert-type

scale. As shown in Figure 3, the SUS scores obtained for ColloCaid are between good and excellent (and above the SUS average of around 70), despite known bugs and minor issues with the lexicographic database.

Figure 3. Usability scores of ColloCaid v0.1 to v0.3 and interpretation of SUS values (right) according to Bangor, Kortum, and Miller (2009)



5. Conclusion and future work

Previous studies on academic writing needs and dictionary use have led us to develop a text editor integrated with a large, lexical database of general academic English collocation suggestions, enriched with corpus examples of collocations in use. Our prototype, which draws on the principle of dynamic data-driven learning, has been well received by L2 users of academic English, scoring between good and excellent on the SUS. Future development of ColloCaid includes adjustments to the lexical database (i.e. expanding and proofreading current coverage), experimenting with new ways of visualizing collocations, and further user testing with think-aloud and diary studies.

6. Acknowledgments

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

References

Ackermann, K., & Chen, Y. (2013). Developing the academic collocations list (ACL) – a corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247. https://doi.org/10.1016/j.jeap.2013.08.002

Baisa, V., & Suchomel, V. (2014). SkELL: web interface for English language learning. In P. Rychlý (Ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing* (pp. 63-70).

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114-123.

Brooke, J. (2013). SUS: a retrospective. *Journal of Usability Studies*, *8*(2), 29-40.

Durrant, P. (2016). To what extent is the academic vocabulary list relevant to university student writing? *English for Specific Puroposes, 43*, 49-61. https://doi.org/10.1016/j.esp.2016.01.004

Frankenberg-Garcia, A. (2011). Beyond L1-L2 equivalents: where do users of English as a foreign language turn for help? *International Journal of Lexicography, 24*(1), 97-123. https://doi.org/10.1093/ijl/ecq038

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL, 26*(2), 128-146. https://doi.org/10.1017/s0958344014000093

Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes, 35*, 93-104. https://doi.org/10.1016/j.jeap.2018.07.003

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G., & Sharma, N. (2019a). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL, 31*(1), 23-39. https://doi.org/10.1017/s0958344018000150

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G., Sharma, N., & Butcher, P. (2019b). Collocations in e-lexicography: lessons from Human Computer Interaction research. Paper at *Collocations in Lexicography: existing solutions and future challenges*, 30 September 2019, Sintra, Portugal.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: design and structures of a needs-driven online academic writing aid. *Lexicographica - International Annual for Lexicography, 31*(1), 118-141. https://doi.org/10.1515/lexi-2015-0007

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7-36. https://doi.org/10.1007/s40607-014-0009-9

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography, 24*(1), 29-49. https://doi.org/10.1093/ijl/ecq039

Lea, D. (Ed.). (2014). *Oxford learner's dictionary of academic English*. Oxford University Press.

Mayor, M. (Ed.). (2013). *Longman collocations dictionary and thesaurus.* Pearson Education.

Nesi, H. (2011). BAWE: an introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (Eds), *New trends in corpora and language learning* (pp. 213-228). Continuum. https://doi.org/10.5040/9781474211925.ch-013

Paquot, M. (2010). *Academic vocabulary in learner writing: from extraction to analysis*. Continuum.

Rees, G., Frankenberg-Garcia, A., Lew, R., Roberts, J., Sharma, N., & Butcher, P. (2019). Slipping through the cracks of e-lexicography: lessons from ColloCaid. Paper at *eLex 2019: Smart Lexicography*, 1-3 October 2019, Sintra, Portugal.

Romer, U., & Swales, J. (2010). The Michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes, 9*(3), 249-249. https://doi.org/10.1016/j.jeap.2010.04.002

Wu, S., Fitzgerald, A., Yu, A., & Witten, I. (2019). Developing and evaluating a learner-friendly collocation system with user query data. *International Journal of Computer-Assisted Language Learning and Teaching, 9*(2), 53-78. https://doi.org/10.4018/ijcallt.2019040104

Yoon, C. (2016). Concordancers and dictionaries as problem-solving tools for ESL academic writing. *Language Learning and Technology, 20*(1), 209-229.

Published by Research-publishing.net, a not-for-profit association
Contact: info@research-publishing.net

CALL and complexity – short papers from EUROCALL 2019
Edited by Fanny Meunier, Julie Van de Vyver, Linda Bradley, and Sylvie Thouësny

Publication date: 2019/12/09

Typeset by Research-publishing.net
Cover theme by © 2019 Frédéric Verolleman
Cover layout by © 2019 Raphaël Savina (raphael@savina.net)

Fonts used are licensed under a SIL Open Font License

ISBN13: 978-2-490057-54-2 (Ebook, PDF, colour)
ISBN13: 978-2-490057-55-9 (Ebook, EPUB, colour)
ISBN13: 978-2-490057-53-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2019.