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

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

09. Oct. 2020

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, 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 like the BNC and COCA, and corpora of student papers like BAWE (Nesi, 2011) and MICUSP (Romer and Swales, 2010). Other useful tools include SkELL (Baisa and Suchomel, 2014), arguably the easiest to use English corpus available, FlaxLC (Wu et al., 2019), a learner-friendly corpus-based collocation tool, and LEAD (Granger and 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 whilst 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 and 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

[1] University of Surrey, UK ; a.frankenberg-garcia@surrey.ac.uk
[2] University of Surrey, UK; g.rees@surrey.ac.uk
[3] Adam Mickiewicz University, Poland; rlew@amu.edu.pl
[4] Bangor University, UK; j.c.roberts@bangor.ac.uk
[5] Bangor University, UK; n.sharma@bangor.ac.uk
[6] Bangor University, UK; p.butcher@bangor.ac.uk

as 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 (Frankenberg-Garcia et al., 2019b).

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 judgement 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* (Frankenberg-Garcia et al. 2019b).

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 maximise 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

551
core academic lemmas

9257
core collocations
(+22206 extra)

27771
authentic examples of use
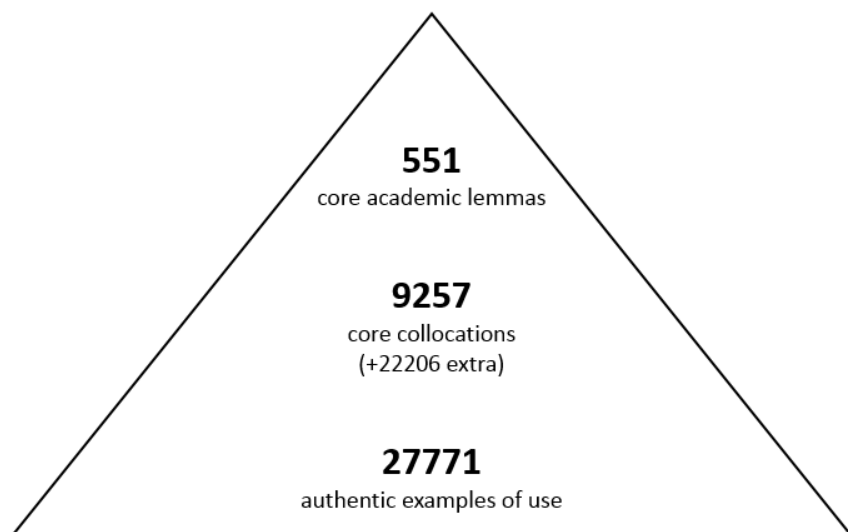
3.   Text editor integration

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, a widely used open-source editor that looks like any regular editor was selected for this purpose (Figure 2A).

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 2B-E).

Figure 2. ColloCaid editor
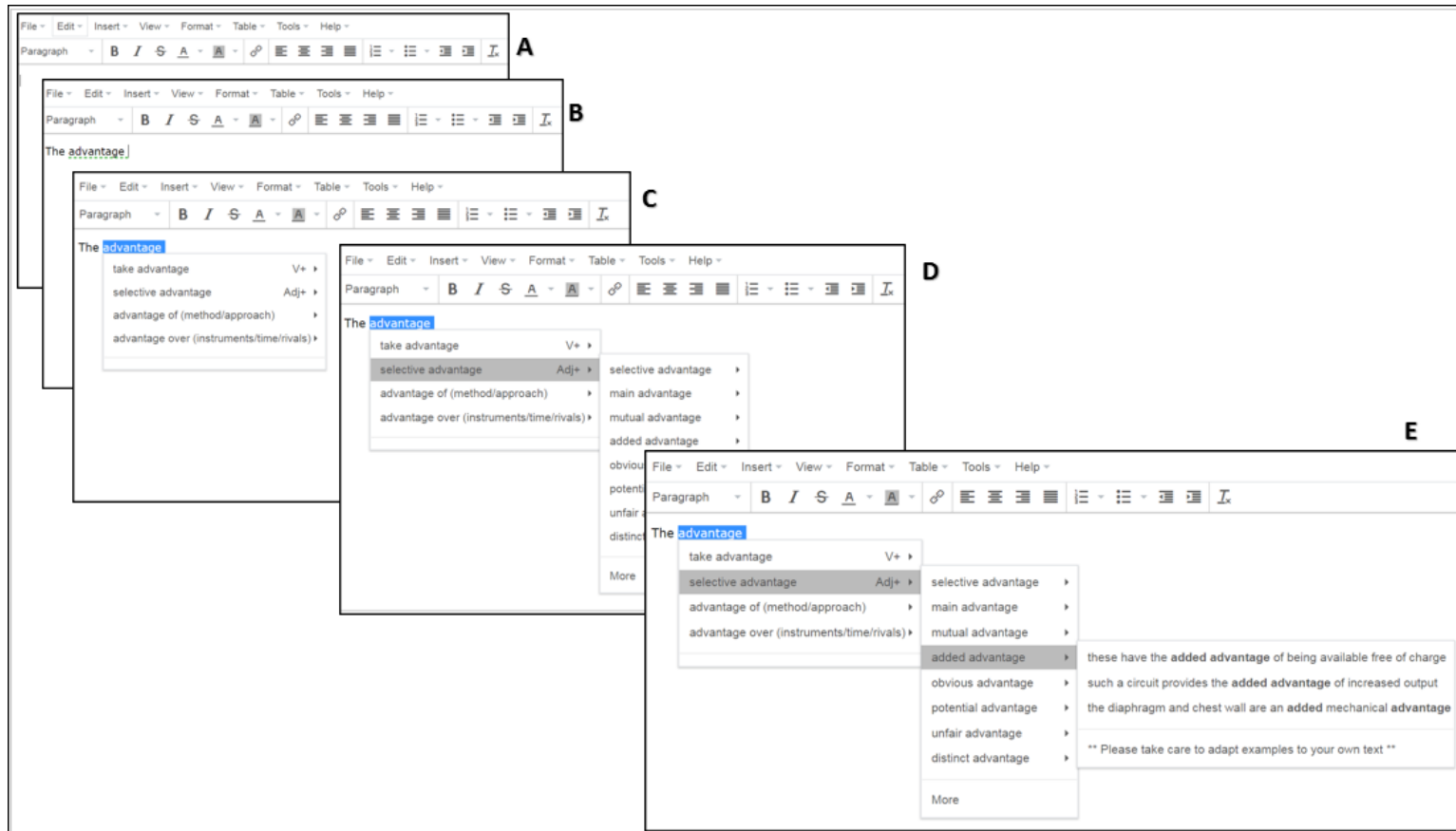
4.   Initial user studies

Development versions of ColloCaid have been tested during university writing workshops and seminars in Brazil, France, Poland and Spain (Frankenberg-Garcia et al., 2019c). 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)
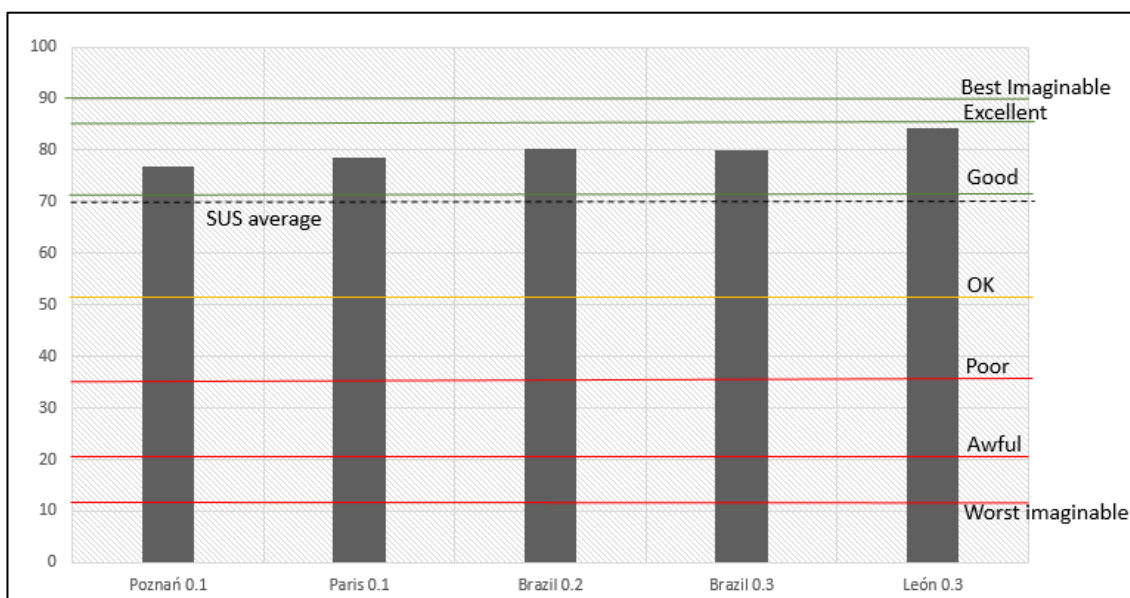
5.   Conclusion and future work

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 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 System Usability Scale. Future development of ColloCaid includes adjustments to the lexical database (i.e., expanding and proofreading current coverage), experimenting with new ways of visualising collocations, and further user testing with think-aloud and diary studies.

6.   Acknowledgements

References

Ackermann, K. and Y. Chen (2013). Developing the Academic Collocations List (ACL) – a corpus-driven and expert-judged approach, *Journal of English for Academic Purposes* 12, 235-247.

Baisa, V. and Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In P. Rychlý (ed.), *Proceedings of Recent Advances in Slavonic Natural Language Processing*, 63-70.

Bangor, A., Kortum, P. and 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.

ColloCaid (n.d.) Online: www.collocaid.uk (11/01/2019)

Durrant, P. (2016) To What Extent is the Academic Vocabulary List Relevant to University Student Writing? *English for Specific Puroposes,* 43: 49-61.

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.

Frankenberg-Garcia, A. (2014).The use of corpus examples for language comprehension and production. *ReCALL* 26(2), 128-146.

Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, 35, 93-104.

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G. and Sharma, N. (2019a) Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL* 31(2), 23-39.

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

Frankenberg-Garcia, A., Lew, R., Roberts, J., Rees, G., Sharma, N. and Butcher, P. (2019c) 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. and Davies, M. (2014) A New Academic Vocabulary List. *Applied Linguistics,* 35(3): 305-327.

Granger, S. and 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.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J. and Suchomel, V. (2014) The Sketch Engine: Ten Years On. *Lexicography*, 1: 7-36.

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.

Lea, D. (ed.) (2014). *Oxford Learner's Dictionary of Academic English*. Oxford: Oxford University Press.

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

Nesi, H. (2011). BAWE: an introduction to a new resource. In: Frankenberg-Garcia, A., L. Flowerdew and G. Aston (eds.) *New Trends in Corpora and Language Learning*. London: Continuum. 213-228.
Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis.* London: Continuum.

Romer, U. & Swales, J. (2010). The Michigan Corpus of Upper-level Student Papers (MICUSP) *Journal of English for Academic Purposes*, 9(3):249-249.

TinyMCE (n.d.). Online: https://www.tiny.cloud/ (15/08/2019).

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

Wu, S., Fitzgerald, A., Yu, A., and 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.