EUROCALL

UCLouvain

First contact with language corpora: perspectives from students

Alex Boulton[1]

Abstract. Corpora are not the preserve of corpus linguists. In education, learners and teachers can analyse almost any collection of text for linguistic or non-linguistic purposes where regular reading is not efficient or feasible. This paper describes students' first contact with corpora in a distance master's degree where they are required to build a corpus on a topic of their choice and complete a short research report. Following a brief outline of the course, we turn to a description of 122 papers submitted over the last 5 years, with particular attention on the Personal Feedback sections of each among both high- and low-achievers. The opening sentences typically reveal bewilderment on initial encounter with corpus linguistics, which contrasts with growing mastery or sudden enlightenment. Further analysis of the 30k-word corpus suggests that a corpus approach may not be immediately easy, but most users can derive benefits with a little perseverance even in adverse conditions.

Keywords: corpus linguistics, student feedback, ESP, data-driven learning.

1. Introduction

Corpus linguistics has shown multiple uses in language education, especially perhaps in describing language use for improved materials and resources, from dictionaries and grammar books to word lists and manuals. Language is a tremendously complex object, and learning is a correspondingly complex process, such that neither reference books nor intuition can ever contain all the answers to all the questions one might have. For highly focused questions or needs, specific corpora are required, sometimes tailored to the individual level (e.g. Charles, 2012). Fortunately, as numerous studies have shown, it is not necessary to be a

1. ATILF (CNRS & Université de Lorraine), Nancy, France; alex.boulton@atilf.fr; https://orcid.org/0000-0001-6306-8158

How to cite this article: Boulton, A. (2019). First contact with language corpora: perspectives from students. In F. Meunier, J. Van de Vyver, L. Bradley & S. Thouësny (Eds), *CALL and complexity – short papers from EUROCALL 2019* (pp. 51-56). Research-publishing.net. https://doi.org/10.14705/rpnet.2019.38.985

corpus linguist to benefit from the approach. If corpus use is relatively well known in Data-Driven Learning (DDL) (Johns, 1990) and teacher education (e.g. Leńko-Szymańska, 2017), in fact corpus tools and techniques can apply to almost any field that works with large quantities of text (Adolphs, 2006, p. 11). This paper describes a course where the students are free to choose any topic and the questions they want to ask, compile a corpus, and write up their research in a paper. While not being DDL per se, it draws on similar principles of authenticity, autonomy, constructivism, and discovery-based learning. The course projects are first described in relation to the students' fields of interest, then their personal feedback is explored to gain an insight into the process of appropriation of corpus tools for their own purposes.

2. Method

The course is part of a master's degree in English in a distance teaching programme in France, and has been running in different guises since 2002 (see Boulton, 2011). The students are mainly interested in English literature and cultural studies ('civilisation') rather than linguistics, and corpora were seen as a way to involve these different interests. Currently, the students are required to define a topic and the questions they want to ask, compile a corpus of at least 20k words for analysis using AntConc, and write up their research in a template paper following the usual IMRAD[2] format (10-15 pages), thus also preparing them for academic writing. A discovery approach is adopted whereby the students play with the texts and software to understand corpora in their own way rather than imposing lengthy instructions which have been found to be counterproductive in earlier iterations. The course has been fairly stable over the last five years (2013-2018), during which time 122 papers have been submitted (discounting resubmissions and blank papers), totalling 455k tokens.

3. Results

3.1. Overview

The topics were divided into the four main disciplines of English in France, allowing for multiple themes: 22% had an overt linguistic focus and 16% a pedagogical one,

2. IMRAD stands for the 'Introduction, Methods, Results, and Discussion' organisational structure.

while 30% looked at literature and 57% were concerned with cultural studies (see sample titles below). This highlights clearly that corpus linguistics is not just a linguistic affair – language analysis can be a way in many different topics reflecting the students' interests. This is important since single-use tools are likely to be abandoned quickly, while multi-purpose tools used repeatedly for different aims are more likely to be adopted (cf. Boulton, 2011).

Literature:

- Romeo and Juliet screen adaptations since 1950's
- The importance of invented words in the Harry Potter books

Cultural studies:

- Competition between Manchester City and Manchester United
- Societal notions in the same-sex marriage debate in the United States of America

3.2. Personal feedback

The template included a Personal Feedback section where the students were encouraged to reflect on their experience of corpus linguistics (Figure 1). These sections represent 30.5k tokens (M=249.6, SD=91.9, range=37/518); there was a modest correlation between length and mark (r=.35).

Figure 1. Personal feedback instructions

This is the *only* section where you should use self-reference (*I* or *me*, etc.). Tell the reader how your project developed, what difficulties you encountered and how you dealt with them, what you think you learned from this, your perception of corpus linguistics and its applications and how your feelings may have evolved as your work proceeded, whether you think you might use some of the tools or techniques in other areas of study in the future, and so on. It might help to think of this section as tips and advice for other students who read your work in the future.

- This section will probably be about **½ to 1 page** long.

Regular reading immediately shows that the opening sentences typically highlight a certain trepidation, as seen in the following sample from the first year of the course:

- "…at first I could not understand what a language could deal with a computer analysis."

- "I was at first quite overwhelmed with the idea of entering research in this field…"

- "I was literally scared when I first read what was expected for this course."

However, these all show initial impressions which change gradually over time, or suddenly in an epiphany from a particular query. From the same students:

- "However, little by little… I got involved in the game… it made me captivated and marvelled at the fact of how information technologies can help modern linguistics in the analysis of the language… The first results after the analysis motivated me a lot as I could see myself the fruits of my work and I could understand the functioning of English in real life… This course was a very useful discovery for me."

- "…but the understanding of the field and of the possibilities it offers only really dawned on me when I started to discover AntConc for myself… I was very impressed with all its functions and the way it can in an instant sort out data which would take hours of work if it were done manually. After working on this project for some time my anxiety about Corpus Linguistics disappeared. I realized I was enjoying working on this study more and more and even got curious. I would actually really enjoy another assignment in Corpus Linguistics to 'play' a bit more with AntConc as well as explore other tools."

- "In the end, little by little, I think I found a bit of light!"

Other negative words occur quite frequently, including 11x *scar** (for *scared*, *scary*, etc.), 5x *fear**, 3x *fright**, 7x *overwhelm** and 5x *daunt**. On the other hand, there were 93 occurrences of *interest** – never preceded immediately by *not*; and although there is also one example of *uninteresting*, the File View shows this is preceded by *at first* and followed by *but then*. Indeed, a cluster analysis shows that *first* occurs 130 times in 80 papers, *at first* has 40 occurrences, and *the first time* 16. Other positive reactions include 65x *useful**, 38x *enjoy**, 14x *curio**, and even 5x *fun*. The most frequent 6-grams both include *time*: *I spent a lot of time* and *it took me some time to* (4x each), and the *time(-)consuming* (15x) nature of corpus work is clearly an issue, at least in early stages when the

students start to discover the software. The most significant adjective collocate (4L, 4R) of *AntConc* is *difficult* (6x, MI=3.9), though three include *not*; similarly, *easy* (4x, MI=3.4) was also used negatively in two occurrences; in between we have *useful* (8x, MI=3.7). A Keyword comparison of the top and bottom quartiles in terms of marks shows feedback from the most successful learners include positive items such as *curious* (f=9, LL=10.3), *clearer* (f=6, LL=6.9) and *benefit* (f=5, LL=5.7) among the top 20, while the least successful include *negative* (f=5, LL=8.3) and *complex* (f=4, LL=6.6).

4. Discussion and conclusion

This paper has attempted to show that corpus linguistics can be used for numerous purposes when dealing with text, and not just 'linguistics' or language learning. Many tools that have only a single use are likely to end up in a dusty drawer, while those that serve multiple functions are more likely to be taken out and used regularly. As such, communication between language teachers and other subject specialists is likely to be highly beneficial. In the present case, and with considerable autonomy, the students were able to build their own small corpora and analyse them in terms of their interests in literature and cultural studies in particular, all the while being exposed to considerable quantities of language in the genre that interested them. This suggests that corpus linguistics is indeed accessible even to students whose mindset is very much geared towards regular reading and qualitative interpretation of continuous text.

The objective in this course is to give the students a say in pursuing their own interests with only the basic methodology and tools imposed. Analysis of this small corpus requires the teacher to tackle the type of task required of the students and to walk in their shoes for a while. With a 30k word corpus, even free, simple corpus tools can help tremendously to gain a more objective picture from word lists (with or without a stoplist or lemmatisation), collocates, clusters, plot (for distribution) and of course concordances and file view to see the items in context. While we are not necessarily interested here in the 'aboutness' (Scott & Tribble, 2006) of the Personal Feedback section of the students' work, such tools show that initial negative reactions are outweighed by more positive views after hands-on experimentation in the majority of cases.

The course will hopefully be improved in two ways in the future. First, simply by being attentive to the views expressed in the Personal Feedback and adapting the course accordingly. Second, the intention is to make this small corpus available to

students as practice material in introductory activities in the next iteration, where they may also benefit from being able to listen to each other.

References

Adolphs, S. (2006). *Introducing electronic text analysis: a practical guide for language and literary studies*. Routledge. https://doi.org/10.4324/9780203087701

Boulton, A. (2011). Bringing corpora to the masses: free and easy tools for language learning. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: from theory to practice* (pp. 69-96). Peter Lang.

Charles, M. (2012). Proper vocabulary and juicy collocations: EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes, 31*(2), 93-102. https://doi.org/10.1016/j.esp.2011.12.003

Johns, T. (1990). From printout to handout: grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria, 10*, 14-34.

Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: tackling the challenge. *Language Learning & Technology, 21*(3), 217-241. http://llt.msu.edu/issues/october2017/lenko-szymanska.pdf

Scott, M., & Tribble, C. (2006). Textual patterns: key words and corpus analysis in language education. Amsterdam: John Benjamins. http://doi.org/10.1075/scl.22

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.