



Separating fact and fiction: The real story of corpus use in language teaching

Alex Boulton

► To cite this version:

Alex Boulton. Separating fact and fiction: The real story of corpus use in language teaching. 20 years of Eurocall: Learning from the Past, Looking to the Future, Sep 2013, Evora, Portugal. pp.51-56. halshs-00837807

HAL Id: halshs-00837807

<https://shs.hal.science/halshs-00837807>

Submitted on 27 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Separating Fact and Fiction: The Real Story of Corpus Use in Language Teaching

Alex Boulton¹

Abstract. This paper investigates uses of corpora in language learning ('data-driven learning') through analysis of a 600K-word corpus of empirical research papers in the field. The corpus can tell us much – the authors and the countries the studies are conducted in, the types of publication, and so on. The corpus investigation itself starts with frequency lists of words and clusters to detect initial themes, which are then extended (via distribution plots, collocates, concordances, etc.) to look at specific items: the researchers cited, the theoretical constructs and concepts investigated and how they are treated, and so on. The paper ends by dividing the corpus into early and more recent papers to compare evolution over time. This reveals keywords that were prevalent in earlier days as a snapshot of the past, and keywords today which may give an idea of future directions.

Keywords: corpora, language learning, DDL, data-driven learning, aboutness, empirical research.

1. Introduction

Language corpora and the tools used to investigate them are amenable to a tremendous variety of applications in many different fields. Even in language learning, there is "considerable variety in the pedagogical approaches adopted" (Johns & King, 1991, p. iii). This heterogeneity can make it rather difficult to make sense of what is really going on. Various attempts have been made elsewhere to analyse the results of empirical studies, especially in the form of a narrative synthesis (e.g. Boulton, 2010), but also more recently in a meta-analysis (Cobb & Boulton, in preparation). The aim of the present paper is not to analyse the

1. Crapel – ATILF, CNRS & University of Lorraine, Nancy, France; alex.boulton@univ-lorraine.fr

outcomes of empirical studies, but rather to identify typical themes and how they develop over time. To try to find core and peripheral areas of study, this paper investigates a corpus of published research papers in the field.

2. Methods

Given the many hundreds of papers that discuss various uses and applications of corpora in language teaching and learning, for present purposes it was decided to limit the study to published research papers which evaluate some aspect of corpus use in language learning and teaching, thus allowing a near-exhaustive collection rather than a sample. It further excludes papers in languages other than English (though only a handful of these had been identified), PhDs (which would have skewed the results given their length), duplicate papers which report the same study twice (if sometimes from different angles), and unpublished papers (e.g. conference presentations), though it did include proceedings papers (cf. the debate about including ‘fugitive literature’ for meta-analysis; e.g. [Norris & Ortega, 2006](#)). The final corpus comprises 110 papers dating from 1989 to 2012, with over half appearing in the last 6 years, which suggests a healthy growth in empirical studies in this area. Nearly half were published in Europe (52), which is perhaps unsurprising as much of the primary interest came from the UK in the work of [Johns \(e.g. 1986\)](#) in Birmingham, as well as from Europe through the TaLC conference series. Asia is also well represented (37), with 17 studies in Taiwan alone; the rest of the world (21) consists primarily of English-speaking countries. Most papers appeared in journals (68), notably *CALL* (14), *LL&T* (9) and *ReCALL* (8); a further 31 were book chapters, and 11 were in conference proceedings.

Some texts were available in electronic format, while others had to be scanned. All were then converted to plain text format (UTF8), which had to be manually checked for various conversion and other problems, a not inconsiderable task given the need to work with a suitably ‘clean’ corpus. As the main interest was in the authors’ own text, further editing was required to exclude meta-data and abstracts, tables and figures, lists of examples and corpus data, long quotes, bibliographies and appendices. However, footnotes, endnotes, embedded examples and in-text citations were retained. The final corpus comprised 615,758 tokens (mean 5,597, varying from 1,631 to 15,029; $SD = 2,068.11$).

The aim of the present study is not so much to analyse *language* per se, but rather the ideas expressed through language to arrive at an understanding of the “aboutness” of the corpus as a whole ([Scott & Tribble, 2006](#), p. 60), i.e. applying corpus linguistics tools as “a way of telling stories about texts” ([Tribble, 2012](#),

n.p.). The main software used was *AntConc* (Anthony, 2012), a free, simple, user-friendly, stable and reliable tool complete with on-line tutorials, help functions and discussion forums. *AntConc* is suitable for teachers and students (cf. Kaszubski, 2006), but also sufficiently flexible and powerful for research purposes (e.g. Charles, 2012).

3. Results and discussion

The first step was to compile a frequency list of the corpus as a whole. Inevitably, most of the high frequency items were grammar-function words (*the, of, too, and...*); though such “small words” are not without interest, they are of limited relevance for semantic purposes, so a stoplist² was applied to filter them out. The resulting list of lexical items, each with a frequency of over 1,000 occurrences per million words, allows a general overall picture to emerge of the prototypical study in this field. They can be recombined textually, if somewhat creatively, as:

A group of learners, generally students, using data from texts based on corpora for learning language or writing through concordancing. According to their level, they can look at both vocabulary (words) or grammar to gain information from the examples of actual use given in a concordance, and search for patterns to improve their knowledge in their English course. This teaching approach is known as DDL, and each research study provides analysis of results from a test.

This intuitively corresponds to the picture generally projected of data-driven learning, but the corpus-based description provides a sounder empirical basis to build on here. Frequency lists of clusters were also produced, though the results did not contribute much semantically to the present research. For example, the top three tri-grams were *the use of, of the students* and *in order to*.

Case sensitive searches enabled the compilation of a list of researchers referred to in the texts. Unsurprisingly, Tim Johns was the most frequently cited with 317 occurrences, followed by Tom Cobb (210), Angela Chambers (145) and Guy Aston (122), all tremendously influential in the field. From an analysis of authors cited at least 20 times, it is apparent that most of the 65 individuals are originally specialists in language learning and teaching who have adopted corpus linguistics techniques in their research and teaching, rather than corpus linguists moving towards language pedagogy. The implications of this require further exploration.

2. http://nlp.cs.nyu.edu/GMA_files/resources/english.stoplist

Firstly, 31% of papers make no mention of Johns at all; this may be because the authors prefer more recent references, or are simply unaware of older research. However, it might be that some researchers do not consider what they are doing to be ‘data-driven learning’, the term coined by Johns. Nonetheless, *DDL* and *data-driven learning* (± hyphen) together appear 1,106 times in 69 papers. Following up on this, the term *corpus-driven* occurred 64 times in 11 papers, compared to *corpus-based* 367 times in 70 papers (± hyphens). Given the debate about the differences between the two terms (cf. Tognini-Bonelli, 2001), it is interesting to note that only 3 of the 110 papers use both, and only 2 make an explicit distinction. However, the differences become apparent from the collocates: *corpus-driven* tends to co-occur with *research* and *study/studies*, while *corpus-based* collocates most strongly with *learning*, *activity/activities* and *approach*. Again, this seems to reflect the preoccupations among these researchers, i.e. language teaching/learning which uses corpus linguistics rather than the other way round.

The next question is to identify exactly what pedagogical aspects are developed most. A list was drawn up of likely key terms, especially the advantages frequently attributed to the approach, leading to individual analyses of 30 separate families of items. The list was topped by *context**, which occurred 1,133 times in 105 papers, and at least 10 times in 44 papers. Other items occurring at least 500 times and at least 10 times in 10 separate papers are *task**, *pattern**, *skill** and *exercise**. The bottom end of the scale is also revealing, with less than 20 occurrences of *constructiv** or *individuali** (for *individualisation*, *individualized*, etc.) for example. Even such items as *cogniti** and *autonom** are relatively infrequent: fewer than 5 papers mention them more than 10 times, suggesting they have not been the overt focus of much research. No papers feature even 10 occurrences of *collaborat** or *creativ**, again suggesting a need for further explicit work in these areas.

Pursuing the theme of this year’s EUROCALL conference (*20 years of EUROCALL: learning from the past, looking to the future*), a final stage was to divide the corpus into two roughly equal components (by publication date, with the cut-off point at 2006/2007) to compare early and more recent themes. This is achieved by a keywords analysis, which counts all items in the two sub-corpora to arrive at a measure of statistical significance (log-likelihood), and ranks them according to their positive or negative keyness (Scott & Tribble, 2006). Keywords in the earlier papers include *concordancing*, *vocabulary* and *word*, as well as specific corpora such as the *Bank (of English)*. Keywords in more recent work include *writing* (with corpora being used as reference resources as well as learning tools) and *Google* (as the internet has become ever more present for professional and study purposes

as well as in everyday life). Less significant items need careful interpretation in a corpus of this size, however, as the procedure used does not take account of distribution; so an item such as *stance*, for example, is considered key since it occurs 119 times in recent studies vs. only 2 in older ones, but 117 of these are from just one paper.

4. Conclusions

Corpus linguistics is not just for corpus linguists. The availability of high-quality yet free and simple tools opens up the methodology to teachers and learners for a tremendous variety of purposes, including language learning and teaching. In this paper, a large collection of empirical DDL-like studies was compiled and subjected to corpus analysis, revealing a picture of prototypical work in this area and suggesting themes requiring further work – especially on some of the advantages frequently attributed to a DDL approach but for which there is as yet little empirical backing. Future predictions are always delicate, but deriving them from real facts in a corpus puts them on a firmer footing than some more subjective approaches. Based on the corpus presented here, one might expect the future to hold a greater synthesis between researchers in language teaching/learning and in corpus linguistics, the continued development of corpus use as a reference tool as well as a learning aid, a bottom-up expansion from lexis to include more work at the level of text or discourse, and increasing use of the web-as-corpus and Google-as-concordancer.

References

- Anthony, L. (2012). *AntConc v3.2.4w/m*. Tokyo: Waseda University. Retrieved from <http://www.antlab.sci.waseda.ac.jp>
- Boulton, A. (2010). Learning outcomes from corpus consultation. In M. Moreno Jaén, F. Serrano Valverde, & M. Calzada Pérez (Eds.), *Exploring new paths in language pedagogy: Lexis and corpus-based language teaching* (pp. 129-144). London: Equinox. [Electronic supplement available at <http://bit.ly/STZegS>]
- Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes*, 31(2), 93-102. doi: 10.1016/j.esp.2011.12.003
- Cobb, T., & Boulton, A. (In preparation). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of corpus linguistics*. Cambridge: Cambridge University Press.
- Johns, T. (1986). Micro-Concord: A language learner's research tool. *System*, 14(2), 151-162. doi: 10.1016/0346-251X(86)90004-7

- Johns, T., & King, P. (Eds.). (1991). Classroom concordancing. *English Language Research Journal*, 4. University of Birmingham: Centre for English Language Studies.
- Kaszubski, P. (2006). Web-based concordancing and ESAP writing. *Poznan Studies in Contemporary Linguistics*, 41, 161-193.
- Norris, J. M., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.
- Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. Amsterdam: John Benjamins.
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.
- Tribble, C. (2012). Teaching and language corpora: Quo vadis? *10th Teaching and Language Corpora (TaLC) international conference*. Warsaw: Uniwersytet Warszawski, 11-14 July.

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

© 2013 by Research-publishing.net
Research-publishing.net is a not-for-profit association

20 Years of EUROCALL: Learning from the Past, Looking to the Future.
2013 EUROCALL Conference, Évora, Portugal, Proceedings
Edited by Linda Bradley and Sylvie Thouéšny

The moral right of the authors has been asserted

All articles in this book are licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 Unported License. You are free to share, copy, distribute and transmit the work under the following conditions:

- Attribution: You must attribute the work in the manner specified by the publisher.
- Noncommercial: You may not use this work for commercial purposes.
- No Derivative Works: You may not alter, transform, or build upon this work.

Research-publishing.net has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate. Moreover, Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before (except in the form of an abstract or as part of a published lecture, or thesis), or that it is not under consideration for publication elsewhere. While the advice and information in this book are believed to be true and accurate on the date of its going to press, neither the authors, the editors, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editors to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover design: © Raphaël Savina (raphael@savina.net)
Photos: © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-12-4 (Paperback, Print on Demand, Lulu.com)
ISBN13: 978-1-908416-13-1 (Ebook, PDF file, Open Access, Research-publishing.net)
ISBN13: 978-1-908416-14-8 (Ebook, Kindle Edition, Amazon Media EU S.à r.l.)
ISBN13: 978-1-908416-15-5 (Ebook, ePUB file, Open Access, Research-publishing.net)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
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: novembre 2013.