Review

EDITORS: Cortes, Vivian & Csomay, Eniko
TITLE: Corpus-based Research in Applied Linguistics
SUBTITLE: Studies in Honor of Doug Biber
SERIES TITLE: Studies in Corpus Linguistics (66)
PUBLISHER: John Benjamins
YEAR: 2015

SUMMARY:

The importance of Douglas Biber's work to corpus linguistics is phenomenal, and this volume is a tribute to his ongoing legacy in the form of nine contributions from those who have benefitted from his tutelage at the Northern Arizona University, or 'the Flagstaff School' of corpus-based research. Biber's work on register analysis and variation in the late 1980's and 90's (notably Biber's 1988 *Variation across Speech and Writing*, his 1995 *Dimensions of Register Variation*, and his work on 1999 *Longman Grammar of Spoken English*) heralded a new quantitative tradition of corpus-based research in applied linguistics. In the introduction, penned by the volumes' editors Viviana Cortes and Eniko Csomay, Biber's Flagstaff heritage regarding the notions of corpus size, representativeness, sampling, systematic analysis, programming, statistical procedures and qualitative functional analyses are recognised as distinguishing features or 'fundamental pillars' of Flagstaff-like research, leaving other programs - in the words of the volume's editors - as simply 'doing' corpus-based research (xvi). All authors in the present volume were, at one point or another over the last twenty years, part of that heritage, and now seek to further that heritage under Biber's guiding principles over a diverse range of corpus-based research studies. These studies include further work on multidimensional analyses of written and spoken registers (Chapters 1,2), the analysis of certain features within written discourse (Chapters 3-4), getting teachers involved with the construction of learner corpora (Chapter 5), the creation of word lists (Chapter 6), new varieties of English (Chapter 7), textual borrowing (Chapter 8) and lexical bundles (Chapter 9). In particular, the selection of contributors are intended to represent a coming of age for corpus-based research in applied linguistics, built upon the backs of 'old codgers' – to use Michael McCarthy' parlance in his foreword to the volume (ix) - for a new generation of corpus linguists.

The volume begins with the previously-mentioned foreword by Michael McCarthy, in which he heralds the work of Biber and his colleagues at the Flagstaff School as a 'productive and internationally respected body of research and publication' (ix). McCarthy reminisces on the historical challenges faced by the early generation of corpus linguists who had to defend the value of their research in the face of rejections of corpus data as decontextualized, and therefore limited in value (following Widdowson's 2000 *On the Limitations of Linguistics Applied*, a criticism again revisited in Chapter 5). This gives way to an appraisal of the valuable contribution that corpus linguistics has given to the field of applied linguistics, with Biber's work a central part of that contribution. He then goes on to discuss the work of the Flagstaff School specifically regarding multi-word strings, and on the interaction between registers and discrete linguistic items. McCarthy then touches on the quantitative/qualitative divide in corpus linguistics, and how both can work to ratify the findings of the other, and finishes by wishing for reconciliation between corpus-based and non-corpus based approaches to the study of genre and register. A brief introduction to the volume by Cortes and Csomay follows, briefly setting out the respective studies.

Chapter 1 (Csomay) presents a study in the tradition of which Biber is perhaps best known, that of a multi-dimensional analysis of variation. Focusing on the differences between teacher and student presentations in a corpus of 271,500 words, the results suggest that teachers produce more features associated with oral and content-focused discourse as well as devices to present stance, while students do not produce any linguistic feature of stance, and produce language that is more typical of literate and procedural discourse. Importantly, as the first work in the volume, some useful background is provided regarding what register variation is and how it is calculated in a multi-dimensional framework across four dimensions, and useful examples from both teacher and student discourse are provided. Also of note is the claim in the conclusion that classroom discourse is not characterised by difficult vocabulary and complex grammar, but that a continuum of literate and oral discourse is present.

Chapter 2 (Friginal) presents a multi-dimensional comparison of telephone interactions in customer service transactions, conversations between friends and family, and participants discussing topics on fixed prompts, compared with corpus data from face-to-face English conversation. The results here suggest that the use of the telephone and the task at hand are major determinants of linguistic variation. An impressive set of spoken data corpora was collected for the study, totalling many hundreds of hours of discourse, which showed clear differences in the levels of politeness, addressee focus and elaboration required, with the call centre conversations exhibiting the most variation among the other two types of telephone interaction. Given the vast amount of data, the analysis taken in this study is technically impressive and is a great example of the Biber tradition in action.

Chapter 3 (Gray) looks at phrasal compression and clausal elaboration structures across six academic disciplines with a view to teasing out inter-disciplinary differences in written structural complexity. Building upon earlier work by Biber and Gray (2010, 2013) regarding the observation of a dense nominal style in academic writing, particularly in science writing, the study investigates the aforementioned structures used in the humanities, social sciences, and natural sciences, using a corpus of just under 2,000,000 words. The results suggest that the ‘soft’ subjects (i.e. philosophy, history, political science and linguistics) entail a high frequency of elaboration devices such as complement clauses, while ‘hard’ subjects make more frequent use of compression features such as complex NPs. Detailed charts and qualitative examples are provided in support of the claims made, giving this study a sense of purpose, rather than just presenting a list of facts and figures.

Chapter 4 (Albakry) studies hedging and negative evaluations in academic letters of recommendation, from a corpus of 114 letters spanning 46,000 words. The study is interesting given the rarity of the dataset, as such letters are not usually available in the public domain. The results suggest patterned uses of hedging in the form of particular modals, evaluative adjectives and mitigation strategies. The author stresses the need for a qualitative approach when analysing such quantitative data given the careful pragmatic nuances evident in the writing, particularly when a writer wishes to be negative about the letters' subject – to quote, 'letters of recommendations, for better or worse, are telling in both commission and omission' (p.95). This negates the author's criticism regarding the small corpus size, given the level of manual annotation needed for such data, and the rarity of the dataset as a whole.

Chapter 5 (Urzúa) presents the construction of a learner corpus with a rare variable – language teachers, in a bid to bridge the gap between corpus linguistics and L2 pedagogy by fully contextualising learner corpus data through teacher-mediated design. There is a real need for studies involving language teachers in corpus-related research with some promising new guides finally becoming available in the literature (e.g. Quinn, 2015), and so this research refreshingly documents the involvement of teachers in the process of constructing a quasi- (i.e. by level) and fully-longitudinal (i.e. across time) corpus of students' L2 writing. Specifically, teachers were invited to information, planning and research-oriented sessions during construction of the corpus, and given hands-on workshops in text processing and concordance software after construction. Doing so increased teachers' level of involvement with the project as a whole and helped define the areas the corpus analysis would explore. Findings from the corpus related to reference tracking helped shape future curriculum design. As a researcher involved in a similar project, the description of teacher involvement in this study should stand as a call to action for corpus linguists who are also involved with language pedagogy.

Chapter 6 (Miller) explores difficulties encountered in the generation of a discipline-specific wordlist for psychology, based on the construction of a corpus from introductory psychology textbooks totalling 3.1 million words. This study is placed between existing literature on wordlists that talks of the usefulness of such lists for vocabulary learning and teaching (such as Coxhead, 2000) and studies that criticise 'general' lists (e.g. Hyland & Tse, 2007) for failing to incorporate discipline-specific concerns regarding vocabulary, and asks whether the corpora on which wordlists are based 'reliably represent the lexical variability in their domains of interest' (P.125). The study compared the words meeting a criteria of 'importance' based on occurrences in 50% of the chapters in a single sample textbook or across the whole corpus of 10 textbooks, with words meeting the importance criteria in both labelled as truly important. The results show that having one, three, five or even nine textbooks included in the sample was not enough to produce a wordlist that was comparable to that of the whole corpus, suggesting that the amount of lexical variability within individual samples makes it difficult to trust the wordlist generated from the whole corpus as representative of the field in general. The chapter concludes with a warning that wordlist users should not generalise the findings of wordlists to other texts, even in the same domain, which might come as a disappointment to those who use such information to design curricula.

Chapter 7 (Balasubramanian) explores the use of wh- and adverbials 'also' and 'only' in a corpus of spoken and written Indian English taken from news and academic registers, in a bid to demonstrate that Indian English exhibits a similar level of internal variation as native varieties.

Indian English contains elements of subject-auxiliary inversion in wh-questions such as ‘Where you are going?’ or ‘Who you are going out with?’, as well as positions ‘also’ and ‘only’ in initial, medial or final positions according different registers, unlike the medial position preferred in English. The results highlight the variation in Indian English use of these constructions according to register, with conversational Indian English exhibiting the highest frequency of ‘Indian’ features, such as ‘also’ in initial or final position, or wh-questions lacking subject-auxiliary inversion. While perhaps unsurprising, the finding that Indian English features were present in the corpus’ sample of written academic English, with an added age effect, suggesting younger users of Indian English were more likely to use such features. While there are perhaps too many large tables included in the chapter, the chapter is quite accessible to those with little experience in corpus-based research on variation.

Chapter 8 (Keck) presents an interesting study on how corpora can be used to analyse paraphrasing (known in this study as ‘textual borrowing’) in L1/L2 student academic writing, which usually falls under investigations of plagiarism or copying. The paper sets out to challenge three ‘beliefs’ in studies on textual borrowing, that 1) L2 writers borrow more from source texts than L1 writers, 2) students copy because they do not understand what they are reading, and 3) students should be taught to paraphrase to avoid suspicion of plagiarism. Based on a small corpus of elicited summaries of three 1,000 word source texts from 124 L1 and 103 L2 writers collected by the researcher, the analysis involved annotating the data from a taxonomy of four paraphrase types (‘near copy’, ‘minimal revision’, ‘moderate revision’ and ‘substantial revision’). The results suggest that novice writers borrow more than experienced writers regardless of L1/L2 group, that both L1/L2 writers judge which words are technical in the source texts and paraphrase surrounding words accordingly, and that both L1/L2 writers used paraphrase to accomplish rhetorical moves appropriately. This main benefit of the study is the methodology used for the study of textual borrowing (although it would be necessary to have access to the same programs Keck developed), and the notion that corpus-based research can be used to challenge typically-held assumptions about L1/L2 writers.

Chapter 9 (Cortes) looks at lexical bundles, providing an overview of corpus-based studies of bundles leading to a detailed description of treatments by Biber on this topic (e.g. Biber et. al. 1999) and Cortes’ own work. The chapter examines the internal structure of lexical bundles along three major categories incorporating verb phrase fragments (‘is going to be’), dependent clause fragments (‘if you want to’) and NP/prepositional fragments (‘the end of the’), then goes on to examine their functional usage in expressions of stance, discourse organisation, and referring expressions. The chapter finishes with a description of bundles used as a rhetorical move devices in academic prose, built largely on Cortes’ own work. This is the only chapter in the book that is not present new research data, and while a very useful resource for those looking for literature on how lexical bundles are defined and used, the chapter feels out of place among the others in the volume.

EVALUATION:

The main contribution of the volume is that of promoting the legacy of the work of Douglas Biber and the Flagstaff school and to act as ‘a passing of the torch’ from one generation of corpus linguists to the next. Certainly, given the quality of the individual papers in this volume, it appears as though the field of corpus linguistics is in very safe hands, and one would hope that these authors are training the next generation of corpus linguists to uncover new trends in new data using new techniques.

The first issue with the volume is one of scope. The title ‘Corpus-based Research in Applied Linguistics’ is rather broad, and unless one is aware of Biber’s work, the subtitle ‘Studies in honor of Doug Biber’ does little to narrow the potential range of studies to the Flagstaff tradition, at least without reading the information on the back cover. On the other hand, the first four chapters are largely in the tradition of Biber’s work on register variation, but this leaves the other corpus-based research areas (learner corpora, wordlists and work on new Englishes in particular) with only a single chapter, so while those looking for a very general volume on corpus-based research in applied linguistics may be pleased, those looking specifically for studies on wordlists, for example, may likely feel little need to purchase the volume. While both points mentioned above may seem contradictory, if a volume is neither particularly general nor particularly specific, the target readership may be hard to determine.

The second issue with the volume is one of accessibility for readers without a) much knowledge of Biber’s approach to corpus linguistics and b) much technical knowledge as an aspiring corpus linguist.

Regarding point a), it would have been useful to have a more user-friendly overview of how multi-dimensional analysis is conducted prior to the studies, in particular information regarding the nature of dimensions themselves and how the results of a multidimensional analysis (i.e. z-scores) may be interpreted. For newcomers to Biber’s approach (and for quantitative linguistics in general), details regarding factor analysis, for example, should be framed with prior explanations for what this does, how it was done, and what it means. Chapter 2 (Friginal) is particularly dense in terms of terminology in this regard, particularly on P29 when mentioning such features as’ Kaiser-Meyer-Olkin Measures for Sampling Adequacy’ or ‘34.29 cumulative percentage of Initial Eigenvalues’, which while useful information for those in the know, may be difficult for those who are not.

Regarding point b), while reference is made in the introduction to this volume with regards to the importance of programming to those involved in the Flagstaff tradition, this makes a number of studies (Chapter 1, Chapter 2) largely unreplicable by those without access to such programs or without the ability to make their own, which I would argue places these authors in a minority of corpus linguistics, rather than the majority. For example, in Chapter 1, Csomay mentions that she ‘developed computer programs with Delphi Pascal to count the various linguistic features for the study' (p9). Chapter 3 (Gray) introduces two computer programs, one for automatically re-tagging systematic errors, and the other for identifying 'elaborating' and 'compression' features, while Chapter 6 (Miller) introduces a 'vocabulary analysis program' based on lemma, and Chapter 8 (Keck) developed a series of programs in Delphi software to extract n-grams, classify paraphrases into a taxonomy, and analyse each summary from the corpus. While certainly impressive in technical terms, given my earlier comments about the book's scope, even those who do have an interested in Flagstaff themed research may feel left out should they wish to attempt any of these studies in their own context (although I am sure that the authors would gladly provide said programs and assistance with them if contacted!).

However, despite the general issues raised above, the quality of the individual chapters shines through and the volume as a whole should work as an inspiration to those of us currently working in corpus linguistics, whatever your particular speciality may be.

References:

Biber, D. (1988). *Variation across Speech and Writing*, Cambridge: CUP.

Biber, D. (1995). *Dimensions of Register Variation*. Cambridge: CUP.

Biber, D. & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes*, 9:2-20. DOI:10.1016:j.jeap.2010.01.001

Biber, D. & Gray, B. (2013). Discourse characteristics of writing and speaking task types on the TOEFL iBT Test: A lexico-grammatical analysis [TOEFLiBT Research Report (TOEFL eBT-19)]. Princeton, NJ: Educational Testing Service.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. London: Longman.

Cortes, V. & Csomay, E. (eds.)(2015) *Corpus-based Research in Applied Linguistics: Studies in Honor of Doug Biber*. Amsterdam/Philadelphia: John Benjamins.

Coxhead, A. (2000). A new academic wordlist. *TESOL Quarterly*, 34(2): 213-238. DOI:10.2307/3587951

Hyland, K & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2):235-253.

Quinn, C. (2015). Training L2 writers to reference corpora as a self-correction tool. *ELT Journal*. 69(2):165-177. DOI:10.1093/elt/ccuo62

Widdowson, H. (2000). On the limitations of linguistics applied. *Applied Linguistics*, 21(1): 3-25. DOI:10.1093/applin/21.1.3.