RiCL Research in Corpus Linguistics

Review of Moskovich, Isabel, Inés Lareo and Gonzalo Camiña. 2021. *"All Families and Genera": Exploring the Corpus of English Life Sciences Texts*. Amsterdam: John Benjamins. ISBN: 978-9-027-20924-5. https://doi.org/10.1075/z.237

Stefania Degaetano-Ortlieb
Saarland University / Germany

Historical linguistics is witnessing a major shift since the early twenty-first century towards the integration of quantitative approaches in the methodological repertoire of the discipline. As Jenset and McGillivray (2017) observe, while the shift towards quantitative methods has penetrated many subfields of linguistics already, historical linguistics has only recently boosted interdisciplinary collaboration with quantitative linguistics. Indispensable for this endeavor is corpus-based work, which has been taken very seriously in this book by Isabel Moskowich, Inés Lareo and Gonzalo Camiña in building the *Corpus of English Life Sciences Texts*[1] (CELiST). From the collection of papers of this book, the interdisciplinary work becomes evident, including not only transparency in corpus-building decisions accounting for the history of science, but also evaluation procedures on the corpus and its representativeness in light of the integration of sound knowledge on the history of science, as well as insights gained from working with the corpus. The chapters of the book fall roughly in the above mentioned three parts. In Chapters 1 to 4, the making of the corpus, the editorial policy adopted to select and encode material and a detailed description of the eighteenth and nineteenth century samples are presented. Chapter 5 evaluates in a detailed manner the representativeness of the corpus, whereas Chapters 6 to 15 present corpus-based studies on various linguistic aspects of the CELiST texts ranging from lexical variation to discourse

[1] https://varieng.helsinki.fi/CoRD/corpora/CELiST/

matters up to register-internal shifts, most of the analytical contributions focusing on evaluative language.

Considering the first part of the book, the first chapter introduces the CELiST corpus in terms of the fields of natural science covered and how the authors decided to group these under the broader term 'Life Sciences'. The authors are very transparent about their selection procedure (something that is faithfully continued throughout the first part of the book). The corpus amounts at 10,000-word samples per decade with a total of 400,305 words. Given the endeavors put into corpus quality, this is quite an achievement. Moreover, the corpus is rich in metadata reflecting the socio-historical context of both time periods covered (eighteenth and nineteenth centuries). The metadata encompasses authors, authors' gender, genres, as well as geographical information, that is, information about the English-speaking country in which the authors were educated and acquired their linguistic habit. For the latter, on page 12 there is a mismatch between the description of increase of Ireland-related authors when comparing Figure 4 to 5: in fact, it decreases, while percentage of authors educated in North America increases from the eighteenth to the nineteenth centuries. Also missing in this chapter is the description of authors' age, a variable used in some of the analyses. The second chapter describes, in detail, the editorial decisions taken by the book authors to encode the corpus material. The challenges described, which are well-known by historical linguists working with digital material, are remarkably dealt with. The effort was not limited to one aspect but went into OCR-error correction up to visualizing the texts as authentically as possible, and particularly highlighted should be the endeavor to truthfully reflect the authors language giving the possibility to exclude language of others present in the texts, such as quotes. Chapters 3 and 4 are summarizations of the content of the texts represented in the corpus, for the eighteenth and nineteenth centuries respectively, however with details on the texts' length and document structure and most importantly on the socio-historical context, which is especially relevant to the historical linguist. Collecting this kind of information would mean quite some work and having that covered in the book is of great value.

In the second part, Chapter 5 presents what is an often-missed contribution in books about corpora, namely corpus evaluation. It provides a detailed computational evaluation of the representativeness of CELiST, confirming the corpus adequacy to represent the Late Modern English scientific discourse in a satisfactory way.

Analytically, the chapter presents results on how CELiST shows a constant lexical growth in line with specialization processes shaping scientific English at that time.

The third part of the book, dealing with analyses of the corpus, is introduced by looking at the lexical fixedness within CELiST considering binomials (such as *more or less*), their distributions across time and genres, as well as semantic relations between the components of the binomials. Nominal pairs are by far the most frequent binominals and unsurprisingly especially those joined by *and*. Diachronically, the distribution of binominals seems to stay relatively equally distributed after a higher usage in the early eighteenth century. For the genre analysis, the unbalanced representation of genres (a choice taken to represent the socio-historical context) hinders a valid diachronic analysis. In terms of differences in semantic relations, synonymy, antonymy, hyponymy and complementation haven been looked at, with the latter being the most prevalent semantic relation. The choice of excluding single occurrences due to the qualitative amount of data to be inspected lead to possibly excluding synonymy relations which the author expected to be much more prominent in the scientific discourse based on previous related work. On the other hand, the wide range of topics covered in CELiST is also considered a possible source of bias.

Chapter 7 analyzes female English scientific writing in CELiST. Botany writing has a large female tradition, which is clearly reflected in this corpus. The focus of the analysis is on directives as engagement features within scientific writing, that is, how the reader is engaged into the discourse. Female writers are compared to contemporary male writers. Knowledge about the surnames of each writer is essential to best comprehend Section 3. For the unknowledgeable reader, it is advisable to have the list of authors from pages 5–9 in Chapter 1 at hand, in order to know the gender of the author. A better audience design would have been advisable. An introductory section to the history of botany in the eighteenth and nineteenth centuries leads the ground to ask whether, in pragmatic terms, there is variation between female and male writers in how they render their discourse authoritative. For this, the use of directives is analyzed. The author well describes the challenges of finding directives properly in the corpus and how essential qualitative methods are to address this in a corpus-based fashion. Results show that male writers use directives more prominently than women in the eighteenth century and that in the nineteenth century women's engagement with the reader is

marked by delicate forms of engagement such as first-person plural combined with modal verbs.

The topic of female writers is continued in Chapter 8, where linguistic indicators of persuasion are analyzed. In focus are prefaces and dedications said to have a persuasive nature as evidenced in Section 2, which are then compared to main texts. The linguistic features analyzed are taken from the literature and encompass various forms of stance features. Again, the historical context is nicely introduced and valuable for any historical linguist. The results clearly show a prevalence of *to*-infinitives and first-person pronouns in prefaces by women writers as opposed to main texts. Wishful would have been a more qualitative analysis of the *to*-infinitive, as it is the most frequently used feature. In fact, from the few examples presented for the *to*-infinitive (cf. pp. 156–157) a tendency of its usage being an evaluative one seems to be quite evident, such as the *it-be*-ADJ-*to* pattern, a usual evaluative feature in academic writing, which increased its usage over time (cf. Hunston and Francis 2000 or Degaetano-Ortlieb 2015). It would have been interesting to see an evaluation of the possible different contexts the *to*-infinitive was used in.

Chapter 9 focuses on suasive verbs and compares CELiST with a corpus of non-fiction texts from the twentieth and twenty-first centuries. While comparatively this seems an odd selection, the diachronic insights gained taking this perspective are indeed valuable. The author shows how suasive verbs as a persuasion strategy are increasingly used in more contemporary non-fiction texts and seem to promote audience design in terms of the involvement of the reader. Moreover, women are more prominently using this persuasive strategy than men over time. A small typo in Figure 9 (*snd* to *and*) has found its way into the text.

Chapter 10 slightly changes the focus to the evolution of scientific practice within the scientific register, while accounting for the authorial presence of the author in the text through the analysis of conditionals and citation sequences. Yet, it still eludes at the evaluative character of the texts, specifically the authors stance towards the texts content. The author shows how epistemic evaluations (certainty-uncertainty) through conditionals and quotations are used within CELiST, and how these usages are complemented by attitudinal or stylistic ones.

The epistemic nature of the CELiST corpus is further investigated in Chapter 11, where epistemic adverbs are considered. The focus is on how writers of the life sciences

persuaded their readers to believe in the truth value of their statements. The comparison with the *Corpus of Historical English Texts* (CHET; cf. Moskowich *et al.* 2019) would have profited from either including evidence from the corpus or at least introducing reference to the respective work. This chapter also makes extensive use of the metadata provided in the CELiST corpus (authors' age, gender, time) showing how mid-career authors most frequently used epistemic adverbs to promote the truthfulness of their statements, how women underused them —possibly due to the descriptive register they were publishing for (botany)— and that their usage was most prominent in articles, lectures and essays.

Chapter 11 presents a very detailed analysis of *that* complement clauses in CELiST accounting for gender differences. Methodologically, the authors would have profited from a linguistically annotated version of CELiST such as part-of-speech annotation. Great effort is taken to extract *that* complement clauses relevant to the analysis. Detailed inspection of various variables combined with statistical evaluation allow the authors to draw valuable insights on the evaluative use of *that*-structures in CELiST. While the distributions provide evidence of no differences in use between female and male writing, by considering the evaluative functions and local contextual settings of evaluative *that* complement clauses the authors arrive at insightful conclusions towards a preference for a cognitive way of expression of scientific claims by female as opposed to a procedural way adopted by male writers. Some typos should be corrected (for instance, the word *attitudinal* is incorrectly written in the graphs on page 232 and *human-subjective* lacks a space on page 237).

Chapter 13 considers authority and deontic modals in CELiST. After a definition and a methodology section, a quantitative analysis of deontic modals follows. The quantitative analysis shows significant differences in the use of particular deontic modals between female and male, as well as regarding distributions across modal use. The qualitative section elaborates on the functions these modals fulfill, considering various functions which provide insights on the discursive patterns. It would have been nice to include a qualitative inspection here as well or, at least, explain why that might not have been possible.

Chapter 14 introduces a different aspect from evaluation and looks at coherence relations by the use of conjunctions in CELiST. It also uses the metadata provided in a fashionable way closely connected to the history of science of the field and categorizing

each metadata into meaningful categories (genres into specialized and non-specialized texts). The results show a steady increase in the use of the analyzed conjunctions over time. While all types of conjunctions rise, adversative and causal ones are definitely the most frequent ones. However, given the high frequency of *and*, which is excluded from the analysis, the reader might wonder whether the picture would have been different. The most important finding is that of a higher use of conjunctions in non-specialized texts as opposed to specialized ones, possibly favoring ease of processing in that genre. The argument of explicit mentioning is not quite straightforward as there is no comparison to implicit relations, so statements in this direction should be made with caution.

Chapter 15 is most quantitative in nature using multidimensional analysis to inspect register-internal variation of CELiST, including comparison to the *Corpus of English Texts on Astronomy* (CETA; cf. Moskowich and Crespo 2012) and the *Corpus of English Philosophy Texts* (CEPhiT; cf. Moskowich 2016). The focus is on the dimension of variation of descriptive and argumentative style. By comparison to the other disciplines, Life Science (CELiST) is fundamentally descriptive as opposed to Astronomy (CETA). Genre and gender differences round the picture nicely up also in light of the studies preceding this chapter.

Overall, the book is a great complement to the preceding series of the *Coruña Corpus of Early Scientific Writing*.[2] Two general remarks relate to (1) the cohesiveness of the single contributions and intended audience and (2) the contextualization of the studies to the international endeavors of historical corpus-based work as well as computational historical linguistics. As for (1), the chapters would have profited from more intersectional reference, especially because most of the analytical chapters engage in the topic of evaluative language. An introductory chapter to the books' single contributions would have been of great value to the interested reader. Here, the particular foci of the book could have been highlighted especially for the analytical chapters, such as the set of papers on evaluative language in CELiST as well as the more quantitative parts as opposed to more qualitative work. The gender aspect is taken up in the preface very nicely boosting interest in this direction. As for (2), while extensive related work has been considered in all contributions, the more international and more contemporary work of various historical corpus-based work has been rather

[2] https://varieng.helsinki.fi/CoRD/corpora/Coruna/

neglected. For example, the work around Tanja Säily's group[3] (Helsinki) on English female writing in the Helsinki corpora of correspondences faces similar challenges in corpus building and seems relevant considering the contributions on female writing. The work by Elke Teich's group[4] (Saarbrücken) on English scientific writing of the *Royal Society of London* provides a huge corpus and various linguistic annotations compared to the CELiST corpus, whose focus is on providing a qualitatively high resource truthful to the originals and authors' language. A smaller remark is directed at the poor linguistic annotation of the CELiST corpus, which would have profited from part-of-speech tagging or a least an explanation in the first chapters of why linguistic annotations have not been integrated. This said, the book presents a great endeavor taken to create the CELiST corpus, with a lot of effort put into beautifully enriching the corpus with valuable metadata and the aim to achieve a highly qualitative corpus resource reflecting the socio-historical setting of the time. Especially the transparency of the decisions made is to be highlighted.

References

Jenset, Gard B. and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford: Oxford University Press.

Hunston, Susan and Gill Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Degaetano-Ortlieb, Stefania. 2015. *Evaluative Meaning in Scientific Writing: Macro- and Micro-analytic Perspectives Using Data Mining*. Saarland: Saarland University dissertation.

Moskowich, Isabel. 2016. Philosophers and scientists from the modern age: Compiling the *Corpus of English Philosophy Texts* (CEPhiT). In Isabel Moskowich Gonzalo Camiña Rioboó, Inés Lareo and Begoña Crespo eds. *'The Conditioned and the Unconditioned': Late Modern English Texts on Philosophy*. Amsterdam: John Benjamins, 1–23.

Moskowich, Isabel and Begoña Crespo eds. 2012. *'Playne and simple': The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins.

Moskowich, Isabel, Luis Puente-Castelo, Begoña Crespo and Leida María Monaco. 2019. *Writing History in Late Modern English. Explorations of the Coruña Corpus*. Amsterdam: John Benjamins.

[3] https://www2.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence
[4] https://sfb1102.uni-saarland.de/projects/information-density-in-english-scientific-writing/

Reviewed by
Stefania Degaetano-Ortlieb
Saarland University
Department of Language Science and Technology
Campus A2.2
66123. Saarbrücken
Germany
E-mail: s.degaetano@mx.uni-saarland.de