


Please cite the Published Version

Pérez-Paredes, Pascual and Curry, Niall  (2024) Epistemologies of corpus linguistics across disciplines. *Research Methods in Applied Linguistics*, 3 (3). 100141 ISSN 2772-7661

DOI: <https://doi.org/10.1016/j.rmal.2024.100141>

Publisher: Elsevier

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/635166/>

Usage rights:  [Creative Commons: Attribution-Noncommercial 4.0](https://creativecommons.org/licenses/by-nc/4.0/)

Additional Information: This is an open access article which appeared in *Research Methods in Applied Linguistics*, published by Elsevier.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



Epistemologies of corpus linguistics across disciplines

Pascual Pérez-Paredes^{a,*}, Niall Curry^b

^a Universidad de Murcia, Spain

^b Manchester Metropolitan University, UK

ARTICLE INFO

Keywords:

Corpus linguistics
Corpus linguistics for education
Corpus linguistics for philosophy
Corpus linguistics for sociology
Corpus methods
Epistemology

ABSTRACT

Despite the growing use of corpus linguistics across an ever-growing range of disciplines such as sociology, sports studies, journalism, media discourse or education, there is a dearth of research that examines the epistemological foundations of corpus methods in these disciplines. This paper builds on well-established conceptualisations about research methodology and the role of methods in the wider literature. Drawing on existing discussions about the use of research methods in objectivist and subjectivist conceptualisations of social reality, we seek to bring to the fore the underlying methodological tensions found in the use of corpus linguistics in the application of corpus methods in research that lies outside the interest of major linguistic disciplines. Through this process, we explore how the notions of natural language use and data elicitation are interpreted by current research in order to advance our understanding of how experts from different research camps engage with and epistemologically localise corpus linguistics.

1. Introduction

The boundaries and emergence of disciplinary knowledge are demarcated by a consensus drawn between the scientific community, the institutions and scientific associations in which scientists carry out their research and the ultimate social utility of the object of the discipline (Guy & Small, 1993). Since the late 19th century, the social utility and specialisation rather than the prestige of researchers have been essential for the establishment of new disciplinary knowledge. In this context, specialisation usually involved high standards of explanatory rigour that could potentially maintain the status of a new discipline (Guy & Small, 1993).

In this view, corpus linguistics (CL) has gradually established itself as a distinct discipline over the last five decades (Engwall & Hedmo, 2016). In the early days, the field that shaped CL was computational linguistics, with the Association of Computational Linguistics (ACL)¹ being founded in 1962, and the International Computer Archive of Modern and Medieval English (ICAME) following in suit, in 1969. This break away from computational linguistics and the establishment of corpus linguistics as “an independent discipline with its own theoretical background” (Teubert, 2001, p.127) was, for Teubert, ignited by the need for corpus data to support the empirical analysis of speech. The standards established and maintained by the intellectual authority that stemmed from computational linguistics contributed greatly to the methodological rigour that was attached to the use of corpora in linguistics research.

As the field evolved independently of computational linguistics, there emerged a sustained need to maintain methodological rigour, as the boundaries of the field expanded. This may explain why much of the activity of corpus linguists in the 20th century has been dominated by introspection, with foci on issues such as the compilation of data, the definition of the field and its parameters, the

* Corresponding author.

E-mail address: pascualf@um.es (P. Pérez-Paredes).

¹ Originally its name was Association for Machine Translation and Computational Linguistics (AMTCL)

encoding of textual data, and automatic annotation, among others (Teubert, 2001). While important, these are methodological issues that also limit our capacity to look beyond the confines of our field, and truly engage with the goal of many a corpus linguist i.e., conducting research with the view that “the analysis of corpora may contribute to our knowledge of language” (Teubert, 2001, p.126). Looking outward in this way may not only serve to enhance the reach and remit of corpus linguistics, but also inform and develop corpus linguistics by incorporating diverse epistemological perspectives from a range of disciplinary contexts.

Recognising the potential for such an outcome, this paper seeks to bring to the fore the underlying methodological tensions found in the use of corpus linguistics in what Hunston (2022) has designated as outward-facing corpus studies. That is, the application of corpus methods in research that lies outside the interest of major linguistic disciplines. This paper discusses the scientific impact of corpus linguistics methodology and methods by paying attention to the selection, analysis and reporting in a range of disciplines including, philosophy, sociology and, most notably here, education.

Initially, a broad focus is offered in Section 2, with a more detailed reflection on the use of corpus linguistics methods in education, in Section 3. The focus on education is owing to the relatively expansive use of CL methods in education when compared to the likes of philosophy and sociology – disciplines in which its use is also apparent. Throughout these sections, the focus rests on the distinction between the use of corpus linguistics as a main methodology and corpus linguistics as a set of research methods. Section 4 endeavours to operationalise the reflections generated in Sections 2 and 3. Through this, we explore how experts from different research camps engage with corpus linguistics, offering a critical appraisal of the methods and approaches used in their research and a pathway for future transdisciplinary collaborations, exchanges, and enrichment.

2. Corpus linguistics across the disciplines

Epistemological frameworks shape how scholars within a field interpret evidence and construct knowledge. Although there is room for alternative interpretations, corpus linguistics can be conceptualised as a critical realist approach to social science (McEnery & Brezina, 2022) that embraces an empirical epistemology (Teubert, 2005) and studies the statistical properties of language (Stefanowitsch, 2020) through a variety of forms of analysis (McEnery & Hardie, 2012). Critical realism promotes epistemological relativism (Bhaskar, 1975; Stutchbury, 2022), acknowledging that reality is mediated by (social) structures and understood through different research methods capable of generating knowledge. Critical realists embrace the concept that multiple perspectives contribute to the generation of knowledge within an ontological realism paradigm. In CL, the prevailing ontological presupposition is that reality exists, and researchers can interact with it (McEnery & Brezina, 2022).

CL has attracted the attention of other disciplines as, we argue, corpus methods offer to critical realism practitioners ways for gathering empirical data and analysing how “observable events arise as a result of activities within social structures” (Stutchbury, 2022, pp. 114–115). CL methods offer concrete ways into analysing such events within a realist ontology that presupposes² a real world and a linguistic reality, which is represented in language data (Teubert, 2005). Notably, CL methods are used in academic disciplines not strictly related to linguistics where language or the analyses of discourse(s) play a significant role, spanning an ever-growing range of disciplines that includes, among others, sociology (Rutbcova, 2015), sports studies (Brooke, 2020), journalism (Touri & Koteyko, 2015), media discourse (O’Halloran, 2010) and historical research (Tumbe, 2019). Though the impact of corpus linguistics in the wider social sciences can be impeded by a lack of epistemological alignment (McEnery & Brookes, 2024), these contributions to disciplinary knowledge generally highlight the affordances of using corpus methods in the fields mentioned above, usually by complementing the standard, disciplinary methods and adding some sort of powerful knowledge to the otherwise well-established epistemologies in these fields.

We argue that the research that has used CL methods across disciplines are often loosely attached to core epistemological methodological foundations of CL itself. In particular, the notion of representativeness as formulated in linguistics as the prerequisite for accurate data collection that reflects the linguistic diversity of the target language variety or community (McEnery & Brezina, 2022) is challenging in fields of study beyond linguistics, and potentially less relevant for other disciplines less concerned with linguistic description or generalisability. In this context, the construction or selection of the corpus influences the validity, reliability, and applicability of the findings (McEnery & Brezina, 2022). Therefore, understanding the impact (or lack thereof) of this relationship with representativeness is important for epistemological exchange and disciplinary development. In the following paragraphs, we offer an inevitably simplified overview of how some disciplines have addressed the use of corpus methods to exemplify how CL methods are localised in other disciplines.

CL methods are often used as complementary data analysis techniques. For Bluhm (2016), the use of corpus linguistics in philosophy offers controlled access to empirical linguistic data, that is, access to how natural language represents reality. In this regard, a corpus approach was employed to research the uses of “hoffen” (hope as a verb) and “Hoffnung” (hope as a noun) in German as means to understanding the social construction of emotions. Bluhm argues that the use of corpora can overcome some of the shortcomings present in the analysis of language and, particularly, they can offer independent and unbiased evidence of the use of language.. This approach echoes Baker’s (2023) call to unbiased scrutiny of language data in a corpus in critical discourse analysis, avoiding subjectivism (where appropriate) and embracing accountability of the data (McEnery & Hardie, 2012).

² According to McEnery & Brezina (2022, p.15), the ontological presuppositions of CL “include the existence of reality and our ability to interact with it; [...] the existence of an individual mind (brain) and the social reality of human interactions, with language being an important part of both. We thus assume language as having some form of existence in the brain of an individual (e.g. in the form of a mental lexicon) as well as the social existence as a shared entity in a society.”

Since 2007, the journal *Synthese*, (h5-index = 54), the top 1 journal in philosophy according to Google Scholar, has published 12 articles where corpus linguistics is used. *Synthese* publishes high-quality research in the areas of philosophy, epistemology, logic, metaphysics, philosophy of science and philosophy of language, which may facilitate the uptake of CL methods in philosophy research. Except for one of those papers, the remaining 11 have been published since 2021. Despite the interest of corpus linguistics in epistemological foundations of science-making and knowing (McEnery & Brezina, 2022), only 0.8 % of the research published in *Synthese*³ makes use of corpora or corpus methods for its inquiry. These papers all share an interest in the philosophy of science and in the analysis of language in different domains and disciplines. Operating from this perspective, Mizrahi (2021) used CL methods to offer empirical substantiation for both the epistemic and noetic (i.e., when research focus is placed on the mental processes involved in acquiring knowledge) paradigms in contrast to the semantic framework (where scientific progress is construed in terms of truth) concerning multi-disciplinary scientific advancement. This evidence implies that scientists exhibit a notable predilection towards employing the concepts of *knowledge* and *understanding* as opposed to *truth* when articulating the objectives of scientific inquiry within their scholarly publications. Mizrahi maintains that CL methods such as frequency analysis and n-grams can offer valuable insights about science by studying “what practicing scientists say and do, specifically, what they say and do in their scholarly publications”. For Chapman (2023), CL has gained popularity thanks to its “possibilities of big data analysis [...] moving on from a central concern with language itself to focus on more social and ideological issues”. Systma and Fischer (2023) conducted a comparative corpus analysis of *experience* talk in philosophical, academic, and non-academic discourse, which involved the analysis of both publicly available and purpose-built corpora in conjunction with the analysis of a randomised selections of texts through manual, qualitative analysis. For non-academic discourse, the authors extracted randomised instances of use from the Corpus of Contemporary American English (COCA); for academic discourse, the authors extracted randomised instances from a selection of 10 top-rated philosophy journals available through JStor. The authors claim that the experimental data, analysed through CL methods, contribute to critical ordinary language philosophy, helping explain an illusion of sense.

Generally speaking, claims to objectivity are central to the uptake of corpus methods across disciplines. In Library and Information Science (LIS), Bowker (2018) has suggested that CL offers researchers a “high degree of objectivity” that allows them to approach texts “free from any preconceived or existing notions regarding their, linguistic, semantic or pragmatic content” (p. 20). Bowker argues that LIS researchers could use keywords to identify recommender systems and the analysis of lexical patterns for knowledge discovery systems – two key areas of research and practice in LIS. In the field of journalism and communication research, Bednarek and Carr (2019) designed and collected a corpus of Australian news about diabetes. They put together a series of tips for journalists in terms of the language to use but also on the absences, including issues such as equity of access to specialists or lack of references to Aboriginal and Torres Strait Islander people. The data in the corpus corroborated the need to avoid both blaming individuals and contributing to stigma associated with diabetes. Bednarek and Carr (2021) have argued that corpus linguistics offers a type of computer-assisted linguistic analysis that requires little technical expertise, e.g., programming skills are not required. In public health, Millar and Budgell (2008) have studied lexical and syntactical features of the public health research literature in a corpus of 554 research papers of around 2 million words published in 4 professional journals. The authors set out to identify words particular to their public health corpus by means of log-likelihood, dispersion and comparison to other word lists. To do this, they put together a wordlist that included items in the Academic Word List (AWL), the General Service List (GSL) and words in neither of them such as epidemiological terms (e.g., prevalence), abbreviations (e.g., BMI) and names of diseases and medical conditions (e.g., cancer). In health communication, Adolphs et al. (2004) and Crawford, Brown and Harvey (2014), among others, have highlighted the use of CL methods in flexible ways “to explore topics of mutual interest [to the healthcare research community] and reach conclusions that lead to tangible benefits in terms that make sense to policymakers, patient groups, practitioners and commercial partners” (Crawford, Brown & Harvey, 2014, p. 85).

In sociology, Rubtcova (2015) and Rubtcova et al. (2017) have argued that representative corpora allow researchers to understand and operationalise social categories. Their research about the notions of *altruism* and *mercy* in the Russian National Corpus allowed them to isolate institutional contexts in which the terms are used and interpret their uses in communities of users. A quantitative analysis of these contexts lends evidence to the conclusion that their use as synonyms in sociology should be questioned and revisited. Some sociological research has studied diachronic social change using corpus methods. Zinn (2020), for example, examined the use of the prepositional phrase *at risk* in *The Times* from the 19th century to the 21st century. The author combines frequency analyses with sociological interpretations to understand how the term has come to be mainly associated with people (e.g. jobs, lives or children) rather than with institutions or economy. This research design makes use of well-designed ad-hoc corpus method where the collection, digitalisation and access to *The Times* corpus was key. The texts published in this newspaper model the public sphere as a space in which risks are selected and social meaning is shaped and negotiated. The corpus consists of 23 subcorpora where each of them contains all the articles published in one of the decades. In terms of corpus methods, Zinn used collocation and concordance analyses. For the author, corpus linguistics is experiencing rapid growth and holds great potential for enriching the study of social dynamics and change through the development of a corpus sociology. As such, Zinn argues that the gap between quantitative big data approaches and qualitative analysis of extensive data sets can be bridged by employing a mixed-method design, such as that typically espoused in CL methods.

In the next section, we analyse the uptake of CL methods in education, a discipline that has shown an interest in methodological pluralism and where CL have made a modest yet promising impact (e.g., Pérez-Paredes, 2020).

³ 1,431 papers since 2007.

3. Corpus linguistics in education research: epistemological boundaries, methodology and methods

One of the disciplines beyond linguistics in which CL methods have been moderately used is education. There are two factors that explain this wider proliferation. Firstly, education research is a field that encompasses “deliberative, complex, subtle, challenging thoughtful activity” (Cohen, Manion & Morrison, 2018, p.3) which draws on a plethora of both methods and methodologies across a wide variety of research designs (Hedges & Hanis-Martin, 2009). Second, education research is not established upon one single epistemological view of research and reality (Gray, 2004; Cohen, Manion & Morrison, 2018), revealing the tensions between post-positivism and constructionism as related to an objective-subjective binary that conceptualises knowledge dependent or independent from knowers (Taylor & Raykov, 2020). Thus, education research designs show substantial diversity in terms of methods and methodologies (Cohen, Manion & Morrison, 2018), with mixed methods designs gaining momentum in the field (Cohen, Manion & Morrison, 2018). In this section we pay attention to the broader field of education research (Pérez-Paredes, 2020) rather than to the specific field of English Language Teaching (ELT), where corpus tools (Limberg, 2022) and corpus methodology have been widely tested (e.g., - Crosthwaite, 2024; Curry & Riordan, 2021; Curry et al., 2022).

In education research, the top 1 research journal according to Google Scholar at the time of writing is *Education and Information Technologies*, h5-index=91. In the 2000–2024 period, 10 papers, 8 of those published after 2020, included corpus linguistics either in their methodology or in their scope of analysis, often in the field of language education and language learning. The top 2 journal, *Teaching and Teacher Education*, h5-index=88, included 7 papers featuring corpus linguistics, some of them dealing with the potential of corpus methods for reflective practice in teacher education. The top 3 journal, *British Journal of Educational Technology*, h5-index=86, has published 4 papers in the 2000–2024 period where CL is used.

Corpus methods have been used in education research in two distinct scenarios: (1) as a complement to text analysis in mixed-methods methodology; (2) as the main methodology, including the design, collection and query of a representative corpus.

3.1. Corpus methods as complementing textual data analysis in a mixed-methods methodology

The scenario where corpus methods have been relatively frequently used in education research is as complementary methods to analyse textual data, commonly using a mixed-methods methodology (Tashakkori & Creswell, 2007; Cohen, Manion & Morrison, 2018). In this scenario, researchers look at individuals’ textual data in the form of transcribed interviews, focus groups or, alternatively, they look at educationally relevant documents, including legislation or, among other types of text, official institutional websites.

Corpus methods contribute to the overall aim of mixed-methods methodology by allowing researchers “greater certainty in inferences, conclusions or statements which formulate its findings” and allow for the production of more reliable research (Ponce & Pagán-Maldonado, 2015, p.114). This is the case for studies in which content analysis is complemented with corpus methods. For example, Fest (2015) examined 6 h of audio-only interviews with 14 German speaking teenagers about an online self-assessment tool that offers secondary school students the chance to evaluate future study and career options. The interviews, conducted in German, were designed to elicit their opinions about the usability of the tool both in a mentoring context and as a stand-alone tool. First, the interviews were classified into topics using 18 emerging themes. These themes were reduced to five major categories such as affordances of the tool or usability of the tool. Then the researcher quantified specific linguistic features such as personal pronouns, modal verbs (wollen, sollen, etc.) and qualifiers (etwas or bisschen vs. sehr or absolut) across topics. In this exploration, frequencies, concordance lines and collocations were combined. For the researcher, the combination of content analysis and the actual linguistic units used for conveying opinion offered a more nuanced analysis of the students’ stance and was instrumental in identifying the complexity involved in the phenomenon analysed.

Sometimes a combination of more than two data analysis methods seeks to reinforce the validity of the results. Pérez-Paredes and Curry (2022) examined interviews and focus groups with English as Medium of Education (EME) lecturers in a Spanish university. Instead of combining content analysis with corpus methods, they merged an existing conceptual framework, ROAD- MAPPING (Dafouz & Smit, 2016), with a critical grounded theory analytical framework (Hadley, 2017) using keyword analysis (Curry & Pérez-Paredes, 2023). The researchers combined bottom-up and top-down coding in their analyses of the interviews and focus groups as a way to identify salient aspects in the lecturers’ data. The number of codes applied illustrates their interest in accomplishing a comprehensive analysis of the data by examining every single word in the data, reflecting thus an interest in empirical knowledge (Teubert, 2001, 2005). One of the main contributions of this research was the identification of 115 distinct words and terms used by the lecturers when discursively constructing EME. These terms would most likely have not been isolated without a computational approach. Table 1 shows just the top 30 single keywords in the data.

Figs. 1–3 show an implementation of top-down codes pertaining to different areas of analysis in the ROAD-MAPPING framework (Dafouz & Smit, 2016): Roles of English (RO), Academic Discipline (AD), Language Management (M), Agents (A), Practices and Processes (PP), and Internationalisation and Globalisation (ING). These Figures show the spread of the codes across three of the interviews analysed, offering a unique snapshot of the codes analysed and how EME was constructed by each lecturer.

This approach has several advantages. In addition to making explicit the methodology followed for the quantification and analysis of codes, it offers the possibility of reducing the complexity inherent in the qualitative analysis of complex phenomena in the field of education. The Figures above show simple yet powerful visualisations of how the different interviews display preferences for different codes and topics, facilitating interpretations about the nature of the structure of the data that are not always offered in corpus software (Anthony, 2018). Using this CL methodology to complement the mixed-methods study, the analysis demonstrates how corpus analysis techniques, designed to provide empirical linguistic description, can also support text analysis of the type of interview data that is frequent in education research (Cohen, Manion & Morrison, 2018). In this way, CL methods, unconcerned with issues of

Table 1
Top 30 single keywords in Pérez-Paredes and Curry (2022).

Rank	Item	Keyness	Rank	Item	Keyness
1	bilingual	219.7	16	faculties	30.1
2	Spanish	122.2	17	teach	30.
3	Erasmus	121.3	18	vocabulary	29.7
4	English	79.1	19	slides	27.2
5	lingua	62.2	20	speaking	26.7
6	EMI	60.4	21	grammar	24.5
7	blackboard	54.3	22	motivate	24.2
8	franca	51.5	23	translator	23.3
9	pedagogy	40.4	24	B2	23.2
10	cocoon	39.4	25	Cambridge	23.1
11	transactional	35.0	26	translate	22.8
12	teaching	33.5	27	mistakes	21.1
13	vehicular	32.1	28	face-to-face	20.7
14	fluency	31.1	29	B1	19.9
15	language	30.5	30	accent	19.0

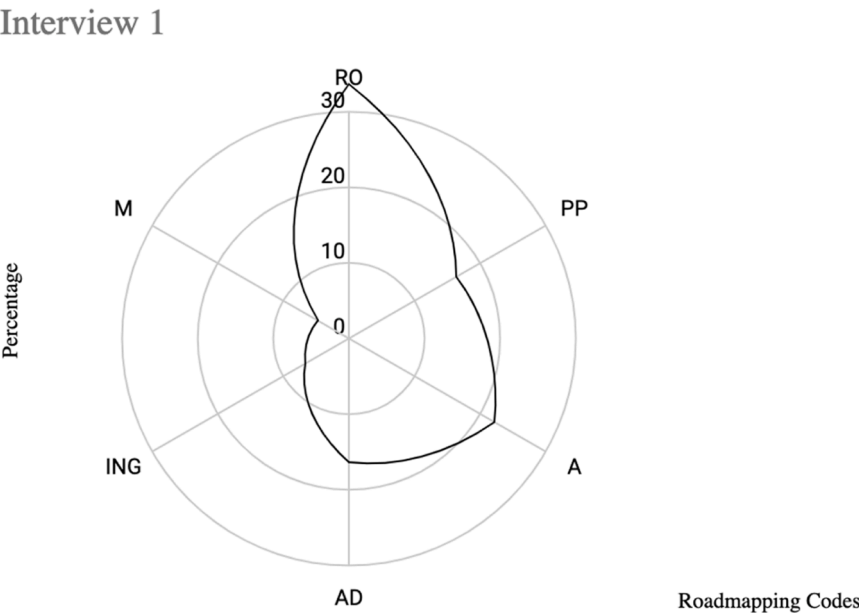


Fig. 1. Interview 1 profile in Pérez-Paredes and Curry (2022).

representativeness and sampling, can be employed more broadly when the analysis is concerned with what people being studied say, not how they say it, necessarily.⁴

Other education studies have used official documents as their primary source of analysis. Villares (2019) researched how the language policies of universities in Spain interpret and articulate the role of languages, in particular English, as part of the University mission. She constructed a corpus of official language policies ($n = 37$) across 29 Spanish universities from 2001 to 2018. As in Fest (2015), a mixed methods approach was followed (although the researcher did not describe her methodology as such). The researcher first used content analysis to identify the main themes and strategic areas in the documents. Villares looked at how the institutions presented themselves, their roles and their responsibilities in the creation of language policies. A comparative analysis examined how the English language was used in contrast to the local languages. This research situates the corpus findings as a means of revealing of the strategies adopted by the universities analysed. Villares argues that the use of a corpus conveys credibility and rigour to the findings: “The corpus results established a direct relationship between accreditation, language competence, and language requirements” (p.3) or “the corpus findings indicated [...] that universities follow a similar line to the one established by the national LP framework” (p.7). These reveal that the validity and credibility of the findings rest upon the chosen corpus methods. Pérez-Paredes (2017) used CL to examine one text: the 2015 UK “Higher education (HE): teaching excellence, social mobility and student choice

⁴ A detailed description and reflection on this approach is presented in Curry and Pérez-Paredes (2023).

Interview 2

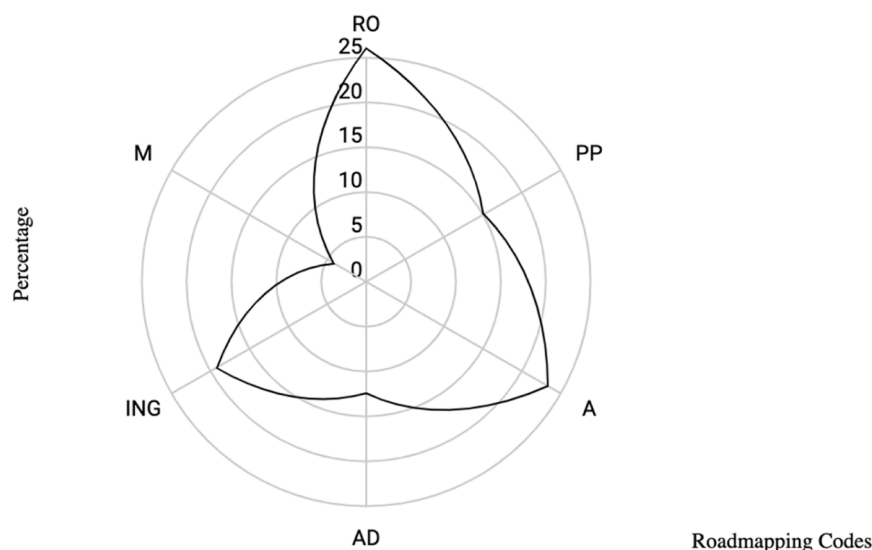


Fig. 2. Interview 2 profile in Pérez-Paredes and Curry (2022).

Interview 5

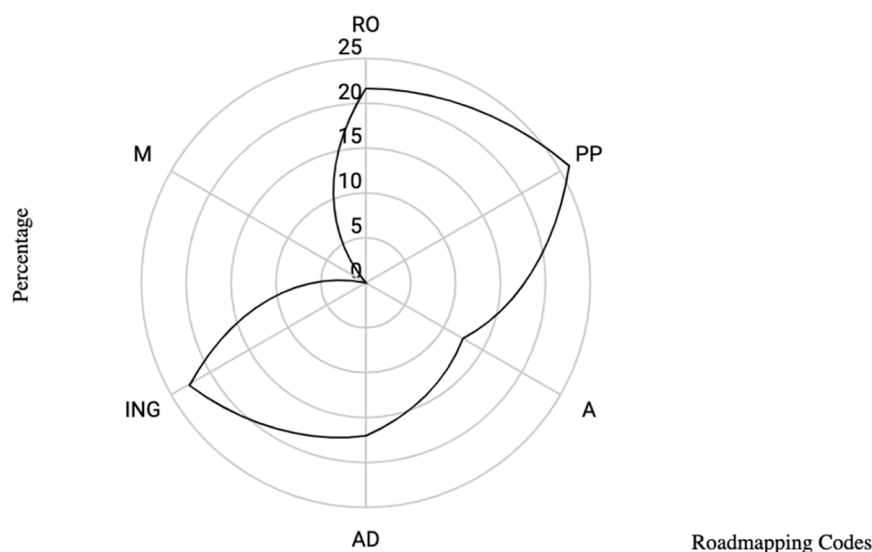


Fig. 3. Interview 5 profile in Pérez-Paredes and Curry (2022).

Green Paper” (HEGP). The Green Paper showed the implementation of the Teaching Excellence Framework (TEF), which would give the UK government the power to monitor and assess the quality of teaching in English universities. For the study of the 33,099-word HE Green Paper, the author used part of speech (POS) keyword analysis (Rayson, 2008), a corpus method that is capable of revealing the role that some specific word categories play in the document when compared with a reference corpus, providing an automatic profile of the POS tags that are significantly more frequent in the HEGP. As a result of this analysis, Pérez-Paredes (2017) exposed the strategies used to represent Higher Education providers as idealised business models capable of delivering excellence through noun phrases that represent minorities and disadvantaged students in need of support. Also, modality played an important role in the HEGP, hiding personal agency and highlighting the power structure of the speech act situation present in the Green Paper. The corpus analysis established a dichotomic tension epistemic modality, used to represent the UK Administration as an open interlocutor, and deontic modality, used to represent the Government as a firm defender of the society. The use of corpus methods facilitated an understanding of the discursive strategies employed to legitimise neoliberal policies in HE and the construction of UK universities as a market.

3.2. Corpus linguistics as the main methodology

Few studies in education research have utilised corpus linguistics methodology in the way it is conceptualised in McEnergy and Hardie (2012) in terms of corpus design that meets both representativeness and balance criteria (Pérez-Paredes, 2020). A substantial part of these efforts are focused on language education research. Jäkel (2022) put together the 56,000-word Flensburg English Classroom Corpus (FLECC) to analyse interaction in primary education. The corpus contains 39 lessons where trainee English language teachers taught a selection of different primary education students across 7 years. He examined the Initiation-Response-Follow up (IRF) patterns in real classrooms, offering language evidence of how interaction is shaped in Northern German primary and secondary schools. A diverse range of age groups of EFL learners is represented in the data, from primary school grade 3 to the end of secondary level grade 10. For Jäkel, the corpus offers future English language teachers opportunities to develop their analytical skills in dealing with authentic classroom discourses and gain a profound awareness of the linguistic and communicative patterns of real, classroom-based teaching discourse. The corpus was envisaged as a proxy for the reality of classroom dialogue in the micro and meso contexts where the data was collected, allowing researchers generalise about the nature of such interactions.

Durrant and Brechly (2019) also designed a corpus to capture the range of writing that students in England produce during different key stages of the school system. Thus their corpus included texts ($n = 2901$) sampled at the ends of Key Stage 1 (Year 2), Key Stage 2 (Year 6), Key Stage 3 (Year 9), and Key Stage 4 (Year 11). The texts were part of children's regular schoolwork and accordingly they were educationally authentic. Various disciplines were represented, including English, Humanities and Science, ensuring a multi-discipline scope in the way texts were analysed. The authors examined the development of vocabulary sophistication in children's writing across the aforementioned stages of education. The corpus in this research was conceptualised as a proxy for writing development in the school years before university education. Through the analysis of the corpus, the authors argued that vocabulary sophistication does not necessarily increase as children progress through schooling. Their study suggests that vocabulary sophistication may not be solely related to word frequency and that the measures of frequency may not be sensitive enough to capture changes in vocabulary sophistication. This research offers implications for curriculum design, writing instruction, assessment practices, support for diverse learners, and the promotion of evidence-based practices in education. Methodologically, the study draws on widely accepted principles in corpus linguistics such as an empirical approach to the study of language and the use of quantitative methods to analyse language data systematically (McEnergy & Brezina, 2022).

One may wonder what the implications are of using CL methods to study corpora that do not meet the established requirements of representativeness and sampling that are characteristic of research in linguistics. We argue that to answer this, one must first understand the motivations of any such study outside of corpus linguistics that employs its methods. In education, for example, research is typically localised to specific institutions, courses, and classrooms. Researchers acknowledge the variable nature of education research sites, which, far from reflecting lab-like qualities associated with scientific thinking, can change from day to day, teacher to teacher, and class to class. In education studies, there has been a rise in participatory and action research to address this potential variability. While such methodological variability may be a problem in corpus linguistics, multiple methodologies are widely embraced by educators and education researchers alike. As such, the notion of corpus representativeness, corpus repeatability, and corpus generalisability that preoccupy many a linguist and a corpus linguist do not necessarily map onto the concerns of those conducting research in education. Many studies in this domain are not concerned with studying language. However, most of them arguably explore interpretations of phenomena through education or situated in educational institutions by means of analyses of subjective meanings constructed through language in texts (Cohen, Manion & Morrison, 2018). Adopting a critical realist perspective, they see language as an observable phenomenon that offers a way into the data. From that point, the interests of education researchers and corpus linguists will likely diverge. Therefore, a reflexive approach to the use of CL methods in disciplines outside of linguistics is imperative when acknowledging the social and cultural factors that shape the use of methods that are novel or are not mainstream in a discipline such as education.

4. The (promising) future of corpus linguistics

The impact of corpus linguistics on some of the disciplines we have discussed in this article demonstrates an interest in recent years in adopting more quantitative methodologies in areas customarily alien to quantitative methodologies or experimental designs, such as philosophy or sociology. This fact could explain the introductory tone of some of the contributions to the field in which corpus linguistics is presented as a viable methodology, although still distant from the epistemological and social conventions of some of these disciplines (McEnergy & Brookes, 2024). Corpus linguistics could be about to enter a new era of influence in disciplinary fields such as education or sociology where there is an epistemological plurality that is no stranger to the impact of data science and big data on the generation of knowledge.

In their book on academic writing and science, Hyland and Jiang (2019) highlight two opposing trends in science writing in English today. While *hard sciences* seem to be interested in adopting a tone that is closer to the reader, a tone where the researcher adopts a more interpretative role and where the transmission of scientific facts and findings is not the only protagonist of academic writing, the social sciences and humanities are moving in the opposite direction. Disciplines such as sociology or applied linguistics have increased, according to Hyland and Jiang (2019), their interest in exploring a more data-driven approach, experimental designs and the use of research methodologies that make use of an increasing amount of empirical data. Our own modest analysis of top philosophy and education journals suggests an increase in the use of corpus methods in the last few years.

Such increase in empirical designs and the use of data found in the social and human sciences may be due to various motivations. Scientific knowledge is built around communities of knowledge that share assumptions about how research should be conducted, and

therefore about how scientific knowledge should be constructed. The need to publish research in high-impact international journals entails the adoption of rigorous research designs and methodologies in accordance with ontological and epistemological principles accepted in the scientific community in which one wishes to develop a scientific career (Hyland, 1999). Some of the most prestigious journals embrace some form of implicit normativism that validates the prestige associated to an “epistemology based on a detached attitude to an external reality of objective facts” (Hyland, 2004, p. 17). This epistemological framing provides credibility, influence and prestige to CL methods for researchers by aligning them with a disciplinary standard of rhetorical inquiry (Hyland, 1999, 2004).

4.1. Factors that contribute to the emergence of corpus linguistics methods across the disciplines

There are, despite the challenges, exogenous variables that have seen corpus linguistics acquire a more relevant status in disciplines such as education research or sociology. Firstly, the interest in discursive practices in these fields (Schwarze, 2022; Urmina et al., 2022) has generated an interest in the analysis of the texts in which we can situate these discursive practices. The combination of corpus approaches with qualitative methods widely used in education, such as interviews or ethnography, complement each other, under a mixed-methods methodology (Adolphs et al., 2004). This approach offers a window into the communicative practices and conventions of various user communities and in this context, and the use of ad-hoc corpora can serve as the main methodology to model widely adopted discursive practices. The emphasis for disciplines such as sociology, philosophy or education is typically on the social activity enacted through the texts in the corpus rather than on the linguistic properties of the text itself. While the implementation of corpus linguistics in wider social sciences still faces some epistemological barriers (see McEnery & Brookes, 2024), a form of corpus-linguistics *light* emerges wherein the corpus design principles that, at least in linguistics, affect the types of texts represented in the corpus, do not appear to hold as much value.

While a lack of concern for representativeness or sampling may not hamper studies unconcerned with communities or samples of populations, these fields within social sciences address social groups and defined communities in many ways. As such, we argue that there is an opportunity for corpus linguists and researchers in the above-mentioned disciplines to work together in the design of corpora that can capture discursive practices across target communities. In such an exchange, researchers from other disciplines could gain a more nuanced appreciation of how language contributes to the production of knowledge. Likewise, corpus linguists may need to shift from genre-aware design principles to community and activity-aware principles that can facilitate new avenues for design of ad hoc corpora. Such an approach would help to localise corpus linguistics epistemologies in different disciplines, while continuing to respect the principles of representativeness and balance central to the corpus linguistics methodology (McEnery & Brezina, 2022). This too could shed light on the rationale for the long-standing importance of these concepts in corpus linguistics for those adopting its methods.

A second variable that can explain the adoption and the ever-increasing interest in CL across disciplines is the significance of big data and data analysis in research (Suhr, Nevalainen & Taavitsainen, 2019). Using Bhaskar’s (1975) terminology, the necessary conditions for scientific research in the 21st century require the use quantitative approaches to data analysis. In this context, corpus linguistics methods can be attractive to social sciences and humanities disciplines. The 2020 Alan Turing Institute *White Paper on the Challenges and Prospects of the Intersection of Humanities and Data Science* outlines some potential areas of impact for corpus linguistics, including improved research methodologies, enhanced data analysis, data visualisation, predictive modelling and interdisciplinary collaboration. While some of the assertions in the white paper are open to debate and different epistemological interpretations (e.g., data analysis methodologies lead to more rigorous and comprehensive research), we can agree that the impact of data science on research in 2024 and beyond is unstoppable. Corpus linguistics as outlined in Bednarek and Carr (2021) offers an entry door to the analysis of big datasets that is user friendly and does not necessarily require programming skills. Curiously enough, the Alan Turing White Paper represents Humanities data as “often unstructured, fragmentary, ambiguous, contradictory, multilingual, heterogeneous and bounded by the subjectivities of their data collection” (p. 11), offering data science approaches as a way to overcome these shortcomings. Corpora, we argue, are highly structured, annotated and searchable, offering a principled design and, oftentimes, an interface where data and data analysis methods are easily accessible for the researcher. Effective corpus design and analysis is possibly the best response to harmful, incidental subjectivities in research, as corpus linguistics can help us understand bias in our findings (McEnery & Brezina, 2022) by employing systematic approaches to data collection, analysis, and critical and self-reflective interpretation.

As suggested by McEnery and Brezina (2022), corpus linguistics embraces an interdisciplinary perspective that draws on insights and methodologies from linguistics, computer science, statistics, and other fields. By integrating diverse approaches, corpus linguists and researchers from other disciplines can address complex research questions and explore language-related phenomena from multiple angles. As demonstrated in Section 2, the use of corpus methods in mixed-methods research designs is relatively widely spread across the disciplines. This is a fertile territory for conversations about methodological tensions found in the use of corpus linguistics in linguistics and elsewhere in other disciplines. These tensions can be located in at least 2 continua. The first continuum, corpus representativeness vs. specificity, exposed the limitations for most disciplines to embrace the notion of linguistic representativeness. Striving for a corpus that is specific to the research question at hand can unleash tensions in the way corpus methods are used (e.g. the use of reference corpora in keyword analysis). While researchers must carefully select texts to ensure their corpus meets their analytical needs, this may create further tensions with notions such as replication and objectivity. A second continuum, corpus generalisability vs. contextualisation, is also likely to expose ontological tensions across disciplines. Corpus linguists strive to draw broad conclusions about language from corpus data, but finding the right balance between generalisability and contextualisation in concrete research projects can be challenging.

4.2. Subjectivity and objectivity

Given their centrality to this discussion, we must, raise some notes of caution on the notions of subjectivity and objectivity. While corpus linguistics is often positioned as an objective approach to a language analysis, this is not necessarily always the case. CL's preoccupation with frequency reflects an epistemological and theoretical perspective that quantifying language use offers a meaningful insight into language in use. Likewise, the embedded computations within tools or the use of statistical methods drawn from other disciplines also bring with them epistemological and theoretical perspectives that converge around shared perspectives on counting, what is counted, and how it counted. These perspectives are subject to debate in corpus linguistics literature (Teubert, 2001, 2005). Therefore, it is important be cautious about the degree to which corpus linguistics can promise objectivity, as ultimately, it is through the interpretative process involved in data collection, design and interpretation that objectivity and subjectivity emerge. Crucially, adopting an empirical perspective, such as that discussed by Hyland and Jiang (2019) does not necessarily lead to objectivity – this arrival at some degree of objectivity is entirely in the hands of the researcher.

On this matter of objectivity and subjectivity, a second point is worth making. While in empirical design, objectivity is often seen as a pinnacle of research excellence, there are domains in which subjectivity is not only admissible, but preferable. In fields like education, philosophy, and sociology this is often the case. So too is it in subfields of corpus linguistics itself, in which ontological positioning and reflexivity impact and shape hermeneutical approaches to language analysis. Corpora can be seen as operationalisations of discourse and the language used in such discourses that facilitate an empirical epistemology. The value of the subject in corpus research can be found in the likes of corpus-assisted discourse studies (Baker, 2023), corpus approaches to translation studies (Curry et al., 2021), corpus-based contrastive linguistics (Curry, 2022, 2023), and corpus approaches to ethnography (Harrington, 2018), for example. Subjectivity is, in this sense, a part of corpus linguistics research and, with this in mind, objectivity should not necessarily be seen as superior to subjectivity. It is the research aim and design that dictates the importance of one or the other. Recognising this and the evident potential for corpus linguistics to continue to branch into other fields of study, it will be important for corpus linguists to approach the notions of objectivity and subjectivity critically, reflecting on their relevance for data design, construction, and analysis across disciplines.

A third note of caution for corpus linguists pertains to the notion that CL is built on one critical-realist epistemology. While much of the cannon reflects this view, it is important to recognise that epistemologies are not simply disciplinary, they are also culturally bound (Curry & Pérez-Paredes, 2021). This means that the movement of CL methods across disciplinary areas and knowledge cultures will require a reflexivity and openness to new permutations and applications. Within linguistics, corpus linguistics has already shaped and been shaped by other subfields. In translation studies, for example, corpus linguistics has shaped the field and its practices in many ways, offering valuable contributions in terms of data design, collection, and analysis, through the development of multilingual corpora. Yet, corpus linguistics has also been shaped by translation studies, broadening the notions of representativeness and sampling (Joahsson, 2007) and ushering in technological advances through the development of parallel concordancing tools, specifically designed for the analysis of translated texts. This example demonstrates the openness of CL to transform in response to the epistemological paradigm of another discipline. In broadening its reach and remit, corpus linguists will need to extend this same openness to fields of study beyond corpus linguistics. Moreover, core field views such as Teubert's (2005) statement that “the linguist is a specialist in investigating texts, not in analysing the real world” merit further reflection and debate in terms of the interdisciplinary nature of corpus methods.

Overall, the kind of research that makes use of corpus methods across some of the disciplines discussed in Sections 2 and 3 is either interested in the use of concrete corpus methods to complement text analysis (e.g., collocation analysis, keywords analysis, frequency analysis), or in the power behind representative corpora to capture how language users discursively construct ideas and objects in the Foucauldian sense of the term. Both pose interesting and concrete methodological challenges, including:

- How reliable are keyword or collocation analyses of individual texts across projects?
- How reliable are general purpose corpora such as enTenTen corpora in keyword analysis where they are selected as reference corpora?
- To what extent can existing general purpose corpora model thought across populations?
- Are there alternative ways to design representative corpora that capture the broader research needs of disciplines such as sociology and education?

These and other questions can only be explored in the context of individual projects. As Egbert, Larsson and Biber (2020) note, querying a corpus entails so much more than obtaining frequencies and opaque measures. Researchers need to make sure that corpus data “is evaluated, appropriately analysed, and interpreted, transforming the data into linguistically meaningful information” (p. 72). This is perhaps a challenging task that will require collaboration between corpus linguists and experts across the disciplines in order to extend the type of interdisciplinary academic conversation that has begun in some of the research discussed in Sections 2 and 3. We hope that in future analyses, bibliometrics studies and systematic reviews can offer a comprehensive overview about the use of corpus methods in the disciplines discussed in this paper as well as in other disciplines where the analysis of language and discourse informs the main research methodologies.

As in foreign language teacher education, corpus literacy initiatives should reach out to practitioners across disciplines and facilitate conversations around the fundamental principles behind corpus linguistics methods and methodology. Similarly, corpus linguists need to understand how other disciplines view the role of discursive practices and how the “functionalist principle of purposefulness of communication” (McEnery & Brezina, 2022, p. 182) can be extended and applied therein. This must be an iterative

process where CL and other disciplines establish a fruitful dialogue on how to streamline methodological decisions for research that does not necessarily show an interest in linguistic analysis, but which uses language as evidence of cultural, psychological, communicative, or discursive practices central to disciplinary praxis and researcher ontologies.

CRedit authorship contribution statement

Pascual Pérez-Paredes: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Niall Curry:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguistics in a health care context. *Journal of Applied Linguistics*, 1(1), 9–28.
- Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor, & A. Marchi (Eds.), *Corpus approaches to discourse: A critical review* (pp. 197–224). Routledge.
- Baker, P. (2023). *Using corpora in discourse analysis*. Bloomsbury Publishing.
- Bednarek, M., & Carr, G. (2019). *Guide to the diabetes news corpus (DNC)*. URL: <https://osf.io/jrhx2/>.
- Bednarek, M., & Carr, G. (2021). Computer-assisted digital text analysis for journalism and communications research: Introducing corpus linguistic techniques that do not require programming. *Media International Australia*, 181(1), 131–151.
- Bhaskar, R. (1975). Forms of realism. *Philosophica*, 15, 99–127.
- Blumh, R. (2016). Corpus analysis in philosophy. In H. Martin (Ed.), *Evidence, experiment and argument in linguistics and the philosophy of language* (pp. 91–109). Peter Lang.
- Bowker, L. (2018). Corpus linguistics is not just for linguists: Considering the potential of computer-based corpus methods for library and information science research. *Library Hi Tech*, 36(2), 358–371.
- Brooke, M. (2020). 'Feminist' in the sociology of sport: An analysis using legitimation code theory and corpus linguistics. *Ampersand*, 7, Article 100068. <https://doi.org/10.1016/j.amper.2020.100068>
- Chapman, S. (2023). When Arne met JL: Attitudes to scientific method in empirical semantics, ordinary language philosophy and linguistics. *Synthese*, 201(4), 146.
- Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education*. Routledge.
- Crawford, P., Brown, B., & Harvey, K. (2014). Corpus linguistics and evidence-based health communication. In H. Hamilton, & W. Y. S. Chou (Eds.), *The Routledge handbook of language and health communication* (pp. 75–90). Routledge.
- Crosthwaite, P. (Ed.). (2024). *Corpora for language learning: Bridging the research-practice divide*. Routledge.
- Curry, N. (2022). On contrastive analysis and language pedagogy: Reimagining applications for contemporary English language teaching. In L. McCallum (Ed.), *English language teaching in the European Union: Theory and practice across the region* (pp. 239–256). Springer. https://doi.org/10.1007/978-981-19-2152-0_14.
- Curry, N. (2023). Question illocutionary force indicating devices in academic writing: A corpus-pragmatic and contrastive approach to identifying and analysing direct and indirect questions in English, French, and Spanish. *International Journal of Corpus Linguistics*, 28(1), 91–119. <https://doi.org/10.1075/ijcl.20065.cur>
- Curry, N., Clarke, J., & Vincent, B. (2021). Ponyming the slovos: A parallel linguistic analysis of translations of A Clockwork Orange. In I. Campbell (Ed.), *Science fiction in translation* (pp. 165–188). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-84208-6_9.
- Curry, N., Love, R., & Goodman, O. (2022). Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora*, 17(1), 1–38. <https://doi.org/10.3366/cor.2022.0233>
- Curry, N., & Riordan, E. (2021). Intelligent CALL systems for writing development: Investigating the use of Write and Improve for developing written language and writing skill. In K. Kelch, P. Byun, S. Cervantes, & S. Safavi (Eds.), *CALL theory applications for online Tesol education* (pp. 252–273). IGI Global. <https://doi.org/10.4018/978-1-7998-6609-1.ch011>.
- Curry, N., & Pérez-Paredes, P. (2021). Stance nouns in COVID-19 related blog posts. A contrastive analysis of blog posts published in The Conversation in Spain and the UK. *International Journal of Corpus Linguistics*, 26(4), 469–497.
- Curry, N., & Pérez-Paredes, P. (2023). Using corpus linguistics and grounded theory to explore EMI stakeholders' discourse. In S. Curle, & J. K. H. Pun (Eds.), *Qualitative research methods in English medium instruction for emerging researchers: Theory and case studies of contemporary research* (pp. 45–61). Routledge. <https://doi.org/10.4324/9781003375531-5>.
- Dafouz, E., & Smit, U. (2016). Towards a dynamic conceptual framework for English-medium education in multilingual university settings. *Applied Linguistics*, 37(3), 397–415.
- Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing*, 32(8), 1927–1953.
- Egbert, J., Larsson, T., & Biber, D. (2020). *Doing linguistics with a corpus: Methodological considerations for the everyday user*. Cambridge University Press.
- Engwall, L., & Hedmo, T. (2016). The organizing of scientific fields: The case of corpus linguistics. *European Review*, 24(4), 568–591.
- Fest, J. (2015). Corpora in the social sciences. How corpus-based approaches can support qualitative interview analyses. *Revista de Linguas para Fines Especificos*, 21(2), 48–69.
- Gray, D. (2004). *Doing research in the real world*. Sage.
- Guy, J. M., & Small, I. (1993). *Politics and value in English studies: A discipline in crisis?* Cambridge University Press.
- Hadley, G. (2017). *Grounded theory in applied linguistics research: A practical guide*. Routledge.
- Harrington, K. (2018). *The role of corpus linguistics in the ethnography of a closed community: Survival communication*. Routledge.
- Hedges, L., & Hanis-Martin, J. (2009). Can non-randomized studies provided evidence of causal effects? A case study using the regression discontinuity design. In P. B. Walter, A. Lareau, & S. H. Ranis (Eds.), *Education research on trial: Policy reform and the call for scientific rigor*. Routledge.
- Hunston, S. (2022). *Corpora in applied linguistics*. Cambridge University Press.
- Hyland, K. (1999). Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics*, 20(3), 341–367.
- Hyland, K. (2004). *Disciplinary discourses, Michigan classics ed.: Social interactions in academic writing*. University of Michigan Press.
- Hyland, K., & Jiang, F. K. (2019). *Academic discourse and global publishing: Disciplinary persuasion in changing times*. Routledge.
- Jäkel, O. (2022). Revisiting an old acquaintance: Exploring the IRF pattern in corpus data from 5th grade EFL lessons. In K. Thomson (Ed.), *Classroom discourse competence: Current issues in language teaching and teacher education* (pp. 205–220). Narr Francke Attempto.
- Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. John Benjamins.
- Limberg, H. (2022). Classroom corpora as tools for reflective practice in pre-service teacher education. In K. Thomson (Ed.), *Classroom discourse competence: Current issues in language teaching and teacher education* (pp. 189–204). Narr Francke Attempto.

- McEnery, T., & Brezina, V. (2022). *Fundamental principles of corpus linguistics*. Cambridge University Press.
- McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- McEnery, T., & Brookes, G. (2024). Corpus linguistics and the social sciences. *Corpus linguistics and linguistic theory*. <https://doi.org/10.1515/cllt-2024-0036>
- Millar, N., & Budgell, B. S. (2008). The language of public health—A corpus-based analysis. *Journal of Public Health*, 16, 369–374.
- Mizrahi, M. (2021). Conceptions of scientific progress in scientific practice: An empirical study. *Synthese*, 199, 2375–2394. <https://doi.org/10.1007/s11229-020-02889-5>
- O'Halloran, K. (2010). How to use corpus linguistics in the study of media discourse. In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 563–577). Routledge.
- Pérez-Paredes, P. (2017). A Keyword Analysis of the 2015 UK Higher Education Green Paper and the Twitter Debate. In M. A. O. Llopis, R. Breeze, & M. Gotti (Eds.), *Power, persuasion and manipulation in specialised genres: providing keys to the rhetoric of professional communities* (pp. 161–191). Peter Lang.
- Pérez-Paredes, P. (2020). *Corpus linguistics for education. A guide for research*. Routledge.
- Pérez-Paredes, P., & Curry, N. (2022). Exploring the internationalization and globalization constructs in EMEMUS lecturers' interviews and focus groups. In E. Dafouz, & U. Smit (Eds.), *English-medium education across multilingual university settings: Applications and critical evaluations of the road-mapping* (pp. 92–116). Routledge.
- Ponce, O. A., & Pagán-Maldonado, N. (2015). Mixed methods research in education: Capturing the complexity of the profession. *International Journal of Educational Excellence*, 1(1), 111–135.
- Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, 13(4), 519–549.
- Rubtcova, M. (2015). *Corpus linguistics in sociological research*. URL: <https://ssrn.com/abstract=2577167>.
- Rubtcova, M., Vasilieva, E., Pavenkov, V., & Pavenkov, O. (2017). Corpus-based conceptualization in sociology: Possibilities and limits. *Espacio Abierto*, 26(2), 187–199.
- Schwarze, T. (2022). Discursive practices of territorial stigmatization: How newspapers frame violence and crime in a Chicago community. *Urban Geography*, 43(9), 1415–1436.
- Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press. <https://doi.org/10.5281/zenodo.3735822>
- Stutchbury, K. (2022). Critical realism: An explanatory framework for small-scale qualitative studies or an 'unhelpful edifice'? *International Journal of Research & Method in Education*, 45, 113–128. <https://doi.org/10.1080/1743727X.2021.1966623>, 2.
- Suhr, C., Nevalainen, T., & Taavitsainen, I. (2019). Corpus linguistics as digital scholarship: Big data, rich data and uncharted data. In C. Suhr, T. Nevalainen, & I. Taavitsainen (Eds.), *From data to evidence in English language research* (pp. 1–350). Brill.
- Sytsma, J., & Fischer, E. (2023). 'Experience', ordinary and philosophical: A corpus study. *Synthese*, 201(6), 210. <https://doi.org/10.1007/s11229-023-04190-7>
- Tashakkori, A., & Creswell, J. W. (2007). Exploring the nature of research questions in mixed methods research. *Journal of mixed methods research*, 1(3), 207–211.
- Taylor, A., & Raykov, M. (2020). Towards Critical and Dialogical Mixed Methods Research: Reflections on Our Journey. In B. Grummell, & F. Finnegan (Eds.), *Doing critical and creative research in adult education: Case studies in methodology and theory* (pp. 127–137). Brill.
- Teubert, W. (2001). Corpus linguistics and lexicography. *International Journal of Corpus Linguistics*, 6, 125–153.
- Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics*, 10(1), 1–13.
- Touri, M., & Koteyko, N. (2015). Using corpus linguistic software in the extraction of news frames: Towards a dynamic process of frame analysis in journalistic texts. *International Journal of Social Research Methodology*, 18, 601–616, 6.
- Tumbe, C. (2019). Corpus linguistics, newspaper archives and historical research methods. *Journal of Management History*, 25(4), 533–549. <https://doi.org/10.1108/JMH-01-2018-0009>
- Urmina, I., Onuchina, K., Irza, N., Korsakova, I., Chernikov, I., & Yushchenko, N. (2022). Communicative and discursive practices in the 21st century: Culturological analysis of the educational process in higher education. *Revista Conrado*, 18(87), 34–43.
- Villares, R. (2019). The role of language policy documents in the internationalisation of multilingual higher education: An exploratory corpus-based study. *Languages*, 4(3), 56.
- Zinn, J. O. (2020). *The UK at risk. A corpus approach to social change 1785-2009. (Critical studies in risk and uncertainty)*. Palgrave Macmillan.