Replicating corpus-based research in English for academic purposes: Proposed replication of Cortes (2013) and Biber and Gray (2010)

Taha Omidian1*, Oliver James Ballance2 and Anna Siyanova-Chanturia1,3
1Te Herenga Waka - Victoria University of Wellington, Wellington, New Zealand, 2Massey University, Palmerston North, New Zealand and 3Ocean University of China, Qingdao, China

*Corresponding author. Email: taha.omidian@vuw.ac.nz

Abstract

Accurate description of language use is central to English for academic purposes (EAP) practice. Thanks to the development of corpus tools, it has been possible to undertake systematic studies of language in academic contexts. This line of research aims to provide detailed and accurate characterization of academic communication and to ultimately inform EAP practice. Very few studies, however, have attempted to ascertain whether, and to what extent, corpus-based findings have achieved such goals. The diverse nature of EAP, and the unique methodological challenges involved in compiling and using corpora, provide sufficient incentive for replication research in this area. The present article makes a case for replication of corpus-based studies in the field of EAP. It argues that replication research not only enhances the credibility of corpus linguistics for EAP pedagogy and research but also provides practical advice for EAP teachers and materials designers. It then looks at how two key corpus-based studies on the topic, Cortes (2013) and Biber and Gray (2010), can be replicated with respect to replication approaches described in Porte (2012).

1. Introduction

As an area of English for specific purposes, English for academic purposes (EAP) seeks to provide a better understanding of the nature of communication in English-medium academic contexts, as well as to highlight the literacy skills required to use language for academic purposes. Thanks to advances in corpus methodologies since the late 1970s, it has been possible to undertake large-scale and systematic studies of academic communication. Academic communication can be thought of as the act of developing and transmitting academic meaning through the use of written or spoken language (e.g., essay writing, grant writing, writing for research publication, delivering lectures, presenting seminars). Corpus-based research in this area has shown that: (i) academic communication is characterized by highly specialized, field-specific discourses that are disparate in nature (see e.g., Durrant, 2017; Omidian & Siyanova-Chanturia, 2021), (ii) linguistic characteristics of academic communication are likely to change as the use of English in academic contexts continues to globalize (see e.g., Biber & Gray, 2010), and (iii) the description of different types of language use in academic contexts is central to EAP practice (see e.g., Durrant, 2014; Omidian et al., 2021). Findings from this line of research indicate that EAP is an eclectic field of inquiry that encompasses a wide range of educational and linguistic topics, requiring a broad array of empirical evidence that could provide insights into various aspects of academic communication in English (see Flowerdew, 2014; Hyland & Hamp-Lyons, 2002).

Due to the diverse and complex nature of EAP, it is important to ascertain the extent to which empirical research in this area has provided findings generalizable to communication in various academic contexts and whether they can be effectively used to inform EAP practice. These reasons...
highlight the importance of replication research in EAP and provide sufficient incentive for replication of corpus-based studies that have made key contributions to the development of the field. The primary objective of replication research in this area would not be to only assess the falsifiability of observations; rather, the principal aim would be to examine whether, and to what extent, a particular finding represents generalizable empirical patterns, rather than an artifact of the method (or the nature of texts) used to explore the topic under investigation (see Porte & McManus, 2018, Chapter 1; also see McEnery & Hardie, 2012, Chapter 1, for a detailed discussion of the difference between falsifiability and replicability in corpus-based research). It should, however, be noted that different types of replication do not necessarily share the same aims mentioned above and might vary as a result of their operational definitions (more on this later).

Replication-based research is particularly important in corpus linguistics as corpus data are highly sensitive to factors such as corpus design, sampling methods, and corpus techniques (see e.g., Baker & Egbert, 2016). Since corpus-based inquiries in the field of EAP aim to provide a detailed and accurate description of the nature of academic texts (spoken or written) and to ultimately inform EAP practice, it is important to assess their replicability to further investigate the impact of methodological choices on their findings.

The current article presents an argument for replication of corpus-based research in the field of EAP. It proposes a number of ways in which corpus-based studies in this area could be replicated with reference to the methodological approaches described in Porte (2012). The studies selected for the purpose of this article are Cortes (2013) and Biber and Gray (2010). These studies have been published in a high-profile journal in the field – Journal of English for Academic Purposes – and represent systematic corpus-based analysis of academic communication. The selected studies address key questions related to EAP and academic discourse (more on this in Section 3). In addition, they provide sufficiently detailed descriptions of their corpus-based methods and procedures, which allows for systematic replication of the original analyses.

The remainder of the present article is organized into three sections. Section 2 provides background information about corpus-based research in the field of EAP. In Section 3, we describe the two original studies, their significance and contribution to EAP, and make suggestions for possible approaches to replication. Section 4 concludes the paper.

2. Background

Since its inception in the 1960s, the field of EAP has expanded to encompass a wide range of methodological approaches. Within this broad and diverse research area, two approaches in particular have been widely adopted in EAP studies: genre analysis and corpus analysis. The genre analysis approach, which was spearheaded by John Swales’ seminal work on academic genres, focuses on the discursive and textual conventions that characterize academic discourse (e.g., Swales, 1990). Applying the genre analysis approach, numerous EAP studies have explored the ways in which discourse is constructed in academic texts. The main output from this line of research has been the development of rhetorical frameworks for the main sections of research articles (RAs) (i.e., abstract, introduction, methods, results, and discussion). These frameworks are important as they show how the textualization of discourse is carried out based on rhetorical conventions within academic genres. Due to the labor-intensive and predominantly qualitative nature of the genre analysis approach, however, analysts often resort to analyzing a small number of academic texts, which can curtail the generalizability of genre-based characterizations of academic communication (see Moreno & Swales, 2018 for further discussion).

In contrast, a corpus-based approach to the analysis of academic discourse requires large-scale samples of naturally occurring cases of language use in academic contexts. Corpus analysis has become a highly influential approach in various fields of inquiry within applied linguistics, including EAP (see Flowerdew, 2014; Hamp-Lyons, 2011 for a detailed discussion of corpus-based analyses in EAP). This approach, which was made possible by advances in computer technology in the late 1970s, relies on
computer-assisted methods and procedures for the identification, quantification, and systematic analysis of patterns of language use in a principled collection of machine-readable texts. Using this approach, many studies have explored academic communication by looking at its characteristics at different linguistic levels, such as single words (e.g., Durrant, 2016; Omidian & Siyanova-Chanturia, 2021), multi-word expressions (e.g., Cortes, 2013; Durrant, 2017; Omidian et al., 2018; Omidian et al., 2021), and lexico-grammatical structures (e.g., Biber, 2006; Biber & Gray, 2010; Omidian et al., 2021). The attractive feature of such studies is that, unlike manual analysis of relatively small data sets, the automated, corpus-based analysis allows analysis of much larger data sets. Thus, it can be argued that corpus-based studies are in a position to lay claim to greater generalizability compared to linguistic investigations that are primarily small scale and reliant on manual analysis (for further discussion of the advantages of using corpus-based techniques, see Baker, 2006, Chapter 1).

It is important to note that the corpus-based approach does not represent a single set of tools, techniques and procedures that are universally agreed upon as the most appropriate methodology for corpus-based research (see e.g., Baker & Egbert, 2016; Gries, 2015; McEnery & Hardie, 2012). In general, corpus linguistics allows the researcher to use the statistical procedures, observational units, computer programs, and corpus resources that they deem suitable for describing language use in the target domain. This autonomy in using various resources and analytical methods in corpus linguistics offers both opportunities and challenges for corpus-based studies. On the one hand, this provides the opportunity for the analyst to use more than one method to validate and verify corpus findings and arrive at more robust interpretations than any single method could achieve on its own (see e.g., Baker & Egbert, 2016; Omidian & Siyanova-Chanturia, 2021). On the other hand, the variety of methods in corpus linguistics has the potential to pose unique challenges to research design in a corpus-based study. Such challenges are primarily related to the principled decisions that should be made regarding the corpus design, methodological orientation, and linguistic variables that are well-suited for a corpus-based analysis of language use in a particular context (for further discussion of methodological issues concerning corpus-based studies, see e.g., Biber et al 1998, p. 243; also see Siyanova-Chanturia & Omidian, 2020). Such decisions and methodological considerations can substantially impact the findings obtained from a corpus analysis and determine whether the emerged patterns are representative of language use in the target domain. This is particularly important in the context of corpus-based research in EAP, in that the ultimate goal of research findings in this area is to inform EAP practice by offering scientific and evidence-based advice to EAP teachers and materials designers (also see Omidian et al., 2017). These considerations highlight the importance of replication research in establishing the generalizability of corpus findings and their ecological validity for EAP practice.

The following section discusses two specific corpus-based studies (i.e., Biber & Gray, 2010; Cortes, 2013), explains their significance and contribution to EAP, and provides specific suggestions for possible approaches to replication. As was mentioned earlier, these studies are based on systematic corpus-based analyses of academic texts and offer important insights into linguistic characteristics of academic communication. Replication of these studies has the potential to corroborate findings, strengthen claims, and verify whether certain observations represent generalizable empirical patterns, or artifacts of the methodological decisions made in the process of conducting the studies.

3. The original studies and suggested approaches to replication


As was discussed in the previous section, two approaches that have been widely adopted in EAP studies are genre analysis and corpus analysis. While the genre analysis approach mainly focuses on the rhetorical strategies that characterize academic genres, corpus analysis techniques are often used in EAP research to identify and describe the lexical and grammatical characteristics of language use in academic contexts. Combining the methodological advantages of these two approaches, Cortes (2013) developed a hybrid research approach through which the predominance of linguistic items
in a given academic genre is systematically characterized with respect to the rhetorical strategies commonly adopted in that genre.

More specifically, Cortes conducted a corpus-driven analysis of the use of multi-word expressions (i.e., lexical bundles of various lengths) in the rhetorical ‘moves’ of RA introductions. In the genre analysis approach, a move is described as ‘a discoursal/rhetorical unit that performs coherent communicative functions’ and often comprises a number of obligatory and/or optional ‘steps’ (Swales, 2004, p. 29). Cortes (2013) first extracted lexical bundles of 4+ words from a large-scale corpus of RA introductions. The corpus used in Cortes (2013) comprises RA introductions from the Published Research Article Corpus (PRAC). The PRAC is a multidisciplinary corpus of 1,020 RAs (c. 5 million words) from journals selected by scholars in more than 20 academic fields (for a detailed description of the corpus, see Cortes, 2007). Following the extraction procedure, Cortes (2013) classified the identified bundles based on their structures and functions using taxonomies developed in previous corpus-based research (e.g., Biber et al., 2004). As the final step, Cortes used the move model for RA introductions proposed by Swales (1990) to connect the identified bundles to the moves in which they occurred. Cortes’s analyses revealed a systematic functional correspondence between lexical bundles and the rhetorical aims of the moves and steps commonly employed in RA introductions. She also found that this functional connection was stronger for certain bundles compared with others. Specifically, it was observed that while some of the identified bundles were exclusively linked to one move/step (e.g., PLAY AN IMPORTANT ROLE IN THE; Move 1 Step 1: claiming relevance of the field), others occurred across several moves and steps (e.g., HAVE SHOWN TO BE). In addition, Cortes found that certain lexical bundles are typically employed to trigger the communicative functions of their corresponding move/step (e.g., LITTLE IS KNOWN ABOUT THE; Move 2 Step 1A: indicating a gap). Taken together, Cortes’s (2013) observations provide a thorough and systematic account of the links between lexical bundles and rhetorical moves.

Cortes’s findings and the methodological synergy of genre analysis and corpus linguistics have been particularly influential in addressing what has recently been labeled as the ‘function-form’ gap in EAP (Moreno & Swales, 2018). The significance of Cortes (2013) for EAP research is reflected in the growing number of studies it has inspired since its publication (e.g., Gray et al., 2020; Khany & Malmeer, 2020; Le & Harrington, 2015; Li et al., 2020; Lu et al., 2020; Moreno & Swales, 2018; Omidian et al., 2018; Yoon & Casal, 2020). Cortes’s findings make a significant contribution to genre-based and corpus-based research in the EAP tradition by providing a detailed and systematic description of function-form connection in academic communication.

### 3.2. Approach to replication

Cortes’s findings leave open the question of whether such a function-form connection between rhetorical strategies and their linguistic realizations is present in other academic genres. As was mentioned above, Cortes’s (2013) findings are based on an analysis of frequently used rhetorical strategies and multi-word items in RA introductions. Generally speaking, RAs are a highly conventionalized genre characterized by conventionally accepted rhetorical and linguistic features (see e.g., Bhatia, 1993; Hyland, 2001; Swales, 1990). Such linguistic norms and conventions have come into being for functional purposes, allowing researchers to communicate the outcomes of their scholarly activity in a way that is retrievable and understandable by their target readership. Thus, it is possible that the conventionalization of rhetorical strategies and their linguistic realization in this genre has engendered their strong function-form connection. In other words, there is a case to be made about whether the strong function-form link observed in Cortes (2013) holds true for less conventionalized forms of academic communication (e.g., promotional genres of academic writing including personal statements, grant application; for further explanation and examples see Bhatia, 2005; Shaw, 2006; also see Casal & Kessler, 2020). This suggests interesting avenues for replication research in relation to Cortes’s findings. For instance, an ‘approximate’ replication of Cortes (2013) would provide important insights into the generalizability of the close form-function link observed in her results. A replication of this
type involves repeating the original study but modifying certain variables of its research methodology with the aim of comparing the original results with those of the replication study (Porte, 2012). The primary purpose of approximate replications is to assess the extent to which the findings of the original study can be generalized to other contexts and whether they have a wide range of applications (Porte, 2012). In the case of Cortes (2013), such a replication study could examine the extent to which conventionality in an academic genre can potentially give rise to systematic function-form relations by exploring such connections in a corpus of texts from a less conventionalized academic genre compared to the one investigated in the original study (i.e., RA introductions). Specifically, a replication of this kind could use the research approach employed in Cortes (2013) to first identify all instances of meaningful relations between function and form in an ‘occluded’ genre (e.g., promotional writing, grant application writing) and then compare the pervasiveness of the identified patterns with those emerging from a highly conventionalized genre, such as RA introductions. The results from such a study will allow us to ascertain the extent to which the strong function-form connection found in Cortes (2013) is generalizable to other genres of academic writing.

There is also great value in a ‘conceptual’ replication of Cortes (2013). Conceptual replications are conducted with similar research questions as those in the original study, but with a new research design, such as the use of qualitative methods in addition to the quantitative analysis employed in the original study (Porte, 2012). Conceptual replications are valuable in that they examine the reliability and validity of outcomes, and can, thus, potentially provide further evidence in support of the original findings. In the context of Cortes (2013), such a replication could investigate function-form connection in RA introductions, using a different research method. More specifically, a replication of this kind could use the same procedure as the one employed in the original study, but ask raters (e.g., experienced EAP instructors) to judge the extent to which the relationship between the identified forms and their rhetorical functions is perceptually salient (also see Simpson-Vlach & Ellis, 2010). Many instances of the functional connections uncovered in Cortes (2013) may not represent a direct inferential link between function and its linguistic realization. For example, it may not be easy to infer that the use of the multiword item AN UNDERSTANDING OF THE is connected to fulfilling the rhetorical function of MAKING TOPIC GENERALIZATIONS, although the corpus analysis conducted in the study proves the existence of such a connection (see Cortes, 2013, p. 39). Consequently, the pedagogical implications of highlighting such function-form mappings may not be immediately self-evident. Thus, a conceptual replication that uses qualitative survey-based methods, in addition to the quantitative analysis employed in the original study, can provide important insights into both perceptual salience and pedagogical relevance of the function-form connections found in Cortes (2013), highlighting their ecological validity for EAP teaching purposes.


Accurate description of the target domain of academic discourse is a priority for the field of EAP. However, it is not entirely clear whether the majority of EAP practice is informed by research, as opposed to introspection or tradition. Biber and Gray (2010) engage with this important issue. The paper explores and challenges stereotypes about the linguistic features that are typical of academic writing. It analyses grammatical features related to complexity, elaboration and explicitness in two corpora: a corpus of academic writing and a corpus of conversation. The corpus of academic writing used in Biber and Gray (2010) is composed of RAs (c. 3 million words), university textbooks (760,000 words), and course syllabi/assignments (c. 52,000 words). The research article sub-corpus represents RAs in four academic fields (science/medicine, education, psychology, and history) published in years 1965, 1985, and 2005. University textbooks and course syllabi/assignments are part of the T2K-SWAL Corpus (see Biber, 2006). The conversation corpus used in the study is part of the Longman Spoken and Written Corpus (see Biber et al., 1999). The corpus comprises c. 42 million words of face-to-face conversation in English.

The above two corpora were annotated using the tagger developed for the purposes of Biber et al. (1999). Specialized computer programs were also developed for linguistic analyses of the specific
grammatical features associated with structural elaboration and compression. The features associated with elaboration included five different types of dependent clauses: finite complement, non-finite complement, finite adverbial, finite relative, and non-finite relative clauses. The features associated with structural compression included five types of phrasal structures: adjective as nominal pre-modifier, noun as nominal pre-modifier, prepositional phrase as nominal post-modifier, prepositional phrase as adverbial, and appositive noun phrase as noun post-modifier (for linguistic examples of elaboration and compression features, see Biber & Gray, 2010, Tables 3–4).

Biber and Gray’s (2010) analyses of these features showed that, due to the informational purpose and compact discourse style of academic prose, academic writing tends to rely on the use of nouns, nominal phrases, and embedded phrasal structures (e.g., prepositional phrases as post-modifiers, THE OBSERVED POST-TEST CRACK PATTERNS FOR THE CONCRETE BLOCKWORK AND CLAY BRICKWORK WALLS). Biber and Gray (2010) findings also demonstrated that the nominal and phrasal discourse style of academic writing stands in sharp contrast to the clausal discourse style of real-time conversation, which tends to heavily rely on clausal structures (especially THAT-and WH-complement clauses and IF- and BECAUSE-clauses functioning as adverbial clauses) and verbs. These findings show that, contrary to many claims in the field, current academic writing is often structurally unelaborated, compressed, and implicit, and that, on the basis of a longitudinal analysis, this discourse style is a fairly recent development. These findings are then related back to a discussion of how the current features of academic writing reflect the current features of academic life (e.g., the preeminence of knowledge communication through academic writing in the modern academic system).

Taken together, Biber and Gray (2010) provide important insights into the linguistic characteristics of academic writing and contest the long-held belief that this linguistic mode of communication is elaborated and explicit. The study by Biber and Gray (2010) has been influential, as indicated by 453 citations listed on Google Scholar (2021). However, although frequently cited, the full implications of this research are far from commonly acknowledged by practitioners. This alone is one reason why this paper would benefit from replication, which will help to draw attention to these important findings. However, more significantly, decisions made in the original study design limit the generalizability of its findings, and as such conceptual replications of this study would not only increase awareness of these important findings but also improve its interpretability.

3.4. Approach to replication

There is value in a conceptual replication of Biber and Gray (2010) that addresses the same research questions as those of the original study but uses a different yet related methodology, making different assumptions about complexity, elaboration, and explicitness. Biber and Gray’s study defines these terms in relation to linguistic features, concluding that divergence in their data shows previous claims were based on stereotypes. However, there is a degree of assumption here: assumption that these linguistic features are indeed salient in judgements about complexity, elaboration, and explicitness. Hence, one useful type of conceptual replication would be to follow the same research design, but to ask raters (e.g., EAP teachers) to score texts for complexity, elaboration, and explicitness. If the results converge, we can be confident both of the original findings and also that the linguistic features that measured elaboration and explicitness are salient in judgements about these qualities of academic texts. If, however, the judgements diverge from the scores on linguistic features, we might consider revising the relationships between linguistic features and these three qualities, perhaps, leading to a re-analysis of the original data but with different mappings between features and qualities. If successful, conceptual replications can strengthen the case for dispelling the stereotypes regarding these features of academic texts. If not, the mismatch between analyses based on formal descriptions and judgements made by raters would encourage a revaluation of which features underlie which types of judgement.

We might also question the decision to compare conversation with academic writing. Such a comparison can potentially confound modality with domain: that is, differences observed in relation to one
variable (a spoken modality as compared with a written modality) may be confounded with differences pertaining to a second (academic discourse as compared with conversational discourse). The original study could be usefully replicated but with systematic control of variables by comparing, for example, non-academic writing with academic writing and conversation with academic speech. Results from this kind of approximate replication would resolve the potential confound issue and could provide a richer account that sheds light on the persistence of the stereotype. That is, potentially, the difference between Biber and Gray’s (2010) findings and common perceptions might be explained by variations related to differences between discourse domains having more perceptual salience than variations pertaining to modality differences. Replication with systematic variance of these variables might shed light on whether this could be the case: if results for non-academic writing versus academic writing are consistent with the stereotype (even if less pronounced), it might be that perceptual saliency is more strongly influenced by domain than by modality. It may be that comparisons across modality are counter intuitive or lack perceptual saliency.

4. Conclusion

The field of EAP has long been concerned with accurate depiction of its target domain of language use, and corpus linguistics has provided tools that allow far more ambitious surveys than were previously possible. The diverse and complex nature of EAP, together with the unique methodological challenges involved in conducting corpus-based studies, provide sufficient incentive for replication research in this area. In this article, we highlighted the rationales for replication of corpus-based research on EAP and proposed specific approaches to replication of two key corpus-based studies in the field.

In conclusion, we argue that replication research plays an important role in establishing the generalizability of corpus findings and their potential for informing EAP practice. Such research can not only enhance the credibility of corpus linguistics for EAP pedagogy and research but also provide practical advice for EAP teachers and materials designers.

References

Taha Omidian is a PhD student at Victoria University of Wellington, New Zealand. Taha’s research interests include corpus-based and computational approaches to language research, quantitative research methods, language for specific purposes, phraseology, vocabulary, semantics, and multilingualism. His recent articles have been published in English for Specific Purposes, TESOL Quarterly, and the Journal of English for Academic Purposes.
Oliver James Ballance teaches at Victoria University of Wellington, New Zealand, where he earned his PhD. He works on the English Language Training for Officials Programme, an ESP-focused, capacity-building programme for government officials from developing countries in the Asia-Pacific region. Before this, he worked in a variety of EFL, ESOL, and EAP settings. He occasionally lectures in various fields of applied linguistics.

Anna Siyanova-Chanturia is Senior Lecturer in Applied Linguistics at Victoria University of Wellington, New Zealand. Anna’s research interests include psychological aspects of second language acquisition, bilingualism, usage-based approaches to language acquisition, processing and use, vocabulary and multi-word expressions, and quantitative research methods (corpora, eye movements, EEG/ERPs). She has published in Studies in Second Language Acquisition, Language Learning, Applied Linguistics, Journal of Experimental Psychology: Learning, Memory and Cognition, and other venues.