RiCL

Research in Corpus Linguistics

RiCL 10/2 (2022)

Special Issue

"Register in understudied academic contexts"

edited by Larissa Goulart, Randi Reppen and Douglas Biber

aelinco
Asociación Española de Lingüística de Corpus

RiCL 10/2 (2022)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

10/2 (2022)

Articles

Book Reviews

RiCL Research in Corpus Linguistics

Register variation in understudied academic contexts

Larissa Goulart
Montclair State University / USA

Abstract – A major focus of register research has been language variation in academic discourse. These studies describe patterns of language use in spoken and written academic texts. Although there have been numerous studies of this type, most have focused on academic registers in English and on descriptions of published academic registers (e.g. textbooks, research articles, and abstracts). Much less work has been caried out on academic registers in other languages or unpublished academic registers. This special issue presents five studies describing the language patterns of understudied academic discourse in English (learners' writing and statutory law), as well as descriptions of published academic registers in languages other than English (Russian, Portuguese, and Arabic). We hope that the papers in this special issue will pave the way for future research in other understudied academic contexts.

Keywords – register studies; academic writing, understudied academic context

From a text linguistic perspective, registers are described as "text varieties that are defined by the situational characteristics of a text" (Gray and Egbert 2019: 1), with situational characteristics being used to refer to attributes, such as communicative purpose, mode of communication, addressor, etc. (see Biber and Conrad 2019: 40). Register studies seek to describe the linguistic profile of a group of texts and explain the use of these linguistic features in terms of their function in these texts. That is, register researchers believe that there is an intrinsic relationship between the situational characteristics and the linguistic profile of a text (or register). Therefore, a register investigation usually includes a situational analysis, a linguistic analysis, and a functional interpretation. This three-pronged analysis is usually referred to as the 'register triangle'.

To date, the register approach has been applied to describe several discourse domains, from face-to-face conversations (e.g. Biber 2006; Quaglio and Biber 2006) to the language of popular science (e.g. Egbert 2016), to academic discourse (e.g. Gray 2015). In fact, Goulart and Wood (2021) show that academic registers are the most common types of discourse domain investigated using multidimensional analysis, a

method often used in register studies. Nevertheless, most of this research has focused on published academic registers (e.g. research articles, textbooks, and abstracts), especially on research articles. In Goulart and Wood's (2021) survey, the authors find 31 studies describing the language patterns of research articles alone. In addition, register studies have focused primarily on academic registers written in English. That is, few studies have examined the language patterns of academic discourse in languages other than English or in unpublished academic registers such as student writing and grant proposals, among others.

The articles in this special issue of *Research in Corpus Linguistics* seek to shed some light into the language patterns of understudied academic contexts from two different perspectives. First, we have a set of two papers that discuss language variation in understudied academic contexts in English student writing and in English legal language. Second, we have a set of three papers that describe the language of research articles in languages other than English (Portuguese, Russian, and Arabic).

In the first paper**, Magali Paquot, Damien De Meyere, Hilde Hasselgård, Tove Larsson, Signe Oksefjell Ebeling, Natalia Judith Laso, Hubert Naets, Larry Valentin, Isabel Verdaguer, and Sanne van Vuuren** describe the compilation process of the *Varieties of English for Specific Purposes dAtabase* corpus (VESPA) and discuss possible applications of this corpus to explore register variation in learner writing. VESPA is a collection of texts written by learners of English with different first languages across three disciplinary components: linguistics, business communication, and literature. Texts in the corpus are classified into the same register categories as texts in the *Michigan Corpus of Upper-level Student Papers* (MICUSP; Römer and O'Donnell 2011), allowing for comparisons between these two corpora. The authors conclude their paper with a case study illustrating the applications of VESPA to investigate register variation in learner writing. In the case study, the authors report on the results of a multidimensional analysis that compares the language profiles of argumentative texts and research papers in different corpora. Finally, they argue that VESPA can help us further understand the language patterns of a somewhat understudied academic register: that of university writing produced by English learners.

In the second paper, **Margaret Wood** discusses the language patterns of the understudied register of English statutory law. Her study examines the extent to which the language of statutory law differs from plain language. To this end, Wood conducts a

key-feature analysis comparing the use of several lexico-grammatical features between states statutes and a corpus of popular written language comprising news, sports reports, encyclopedia articles, historical and magazine articles. It is worth noting that the corpus of state statutory law is part of the *United States State Statutes* (CorUSSS) Corpus (Egbert and Wood, in preparation), which comprises the state codes for each of the 50 states in the United States. The results show that statutory language is denser in terms of clausal embedding, with more passive voice, prepositions, and *wh-* and *that* relative clauses than plain language. Such findings provide a broad overview of the language patterns encountered in state statutes.

In the third paper, **Marine Matte and Elisa Stumpf** examine the use of reporting verbs in Portuguese research articles. To date, few studies have investigated the language patterns of academic Portuguese (Hoffnagel 2010; Bessa 2011). The authors seek to bridge this gap by examining how authors in hard and soft sciences include sources in their research articles. With this goal in mind, Matte and Stumpf first retrieve occurrences of the word *autor* 'author' followed by verbs from the *Corpus of Portuguese for Academic Purposes* (CoPEP; Kuhn and Ferreira 2020). Secondly, they retrieve the most frequent verbs in both hard and soft sciences and analyze them in terms of their structural use (tense, mood, and aspect) and their semantic use (research, cognition, and discourse) in the corpus. The findings show that there is a considerable overlap between the verbs that introduce citations in hard and soft sciences and further suggest that sources are usually incorporated with verbs in the simple present or simple past tenses. Matte and Stumpf's study is innovative in its approach to identify reporting verbs. In the conclusion, the authors highlight the need for further research dealing with the language patterns of academic Portuguese.

Examining the use of lexical bundles in research articles, **Maria Kostromitina** analyzes cross-linguistic transfer in writing between English as a second language (L2) by Russian native speakers and Russian as a first language (L1). To this end, the author compiles a corpus of research articles in educational psychology by L1 Russian writers, L2 English writers, and L1 English writers. Kostromitina retrieves lexical bundles from these three corpora and investigates the L2 English bundles that are mirrored in the L1 Russian and L1 English research articles. The results show that bundles produced by L2 English writers who are native speakers of Russian have a considerable overlap in form with the bundles extracted in the L1 Russian corpus. More importantly, the study shows

one possible approach to investigating language transfer when examining the language patterns of English learners.

In the last paper, **Mai Zaki** investigates the use of metadiscourse and rhetorical features in Arabic academic abstracts. The author compiles a corpus of 400 abstracts in Arabic from research articles and dissertations. Zaki analyzes the extent to which the use of metadiscourse features varies across 1) types of abstracts (dissertation or research articles) and 2) authors' gender (male, female, or mixed). The abstracts are annotated for their use of transition markers, frame markers, evidentials, endophorics, and code glosses, following Hyland's (2019) framework. The results show that engagement markers are quite frequent in Arabic abstracts. Interestingly, the study also finds that transition and frame markers are the most frequent types of metadiscourse within textual markers. This study provides insights into how Arabic academic writers use language features that can engage their readers with the text.

This collection of papers displays a range of different methods (key-features, multidimensional analysis, lexical bundles, etc.) used to describe the language patterns of understudied academic registers. We hope that these studies will motivate further research on other understudied academic registers that are central to academic life but rarely published, such as grant proposals, personal statements, or fellowship applications, among others.

REFERENCES

Bessa, Jose Cezinaldo Rocha. 2011. (Re)pensando a citação em textos acadêmico-científicos. *Signum: Estudos da Linguagem* 14/2: 421–439.

Biber, Douglas. 2006. *University Language*. Amsterdam: John Benjamins.

Biber, Douglas and Susan Conrad. 2019. *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Egbert, Jesse. 2016. Stylistic perception. In Paul Baker and Jesse Egbert eds. *Triangulating Methodological Approaches in Corpus-linguistic Research*, 167–182. London: Routledge.

Egbert, Jesse and Margaret Wood. (In preparation). *Constructing and Designing a Specialized Corpus of Statutory Law* (CorUSSS).

Goulart, Larissa and Margaret Wood. 2021. Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science* 6/2: 107–137.

Gray, Bethany. 2015. *Linguistic Variation in Research Articles*. Amsterdam: John Benjamins.

Gray, Bethany and Jesse Egbert. 2019. Register and register variation. *Register Studies* 1/1: 1–9.

Hoffnagel, Judith C. 2010. A prática de citação em trabalhos acadêmicos. *Cadernos de Linguagem e Sociedade* 10/1: 71–88.

Hyland, Ken. 2019. *Metadiscourse: Exploring Interaction in Writing*. London: Bloomsbury.

Kuhn, Tanara Zingano and José Pedro Ferreira. 2020. O Corpus de Português Escrito em Periódicos-CoPEP. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada* 36/2: 1–42.

Quaglio, Paulo and Douglas Biber. 2006. The grammar of conversation. In Bas Aarts and April McMahon eds. *The Handbook of English linguistics*, 692–723. Oxford: Blackwell.

Römer, Ute and Matthew Brook O'Donnell. 2011. From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora*, 6/2: 159–177.

Corresponding author
Larissa Goulart
Montclair State University
Department of Linguistics
1 Normal Ave.
NJ 07043 Montclair
USA
e-mail: goulartl@montclair.edu

RiCL Research in Corpus Linguistics

The *Varieties for Specific Purposes dAtabase* (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing

Magali Paquot[a] – Tove Larsson[b] – Hilde Hasselgård[c] – Signe O. Ebeling[c] – Damien De Meyere[a] – Larry Valentin[a] – Natalia J. Laso[d] – Isabel Verdaguer[d] – Sanne van Vuuren[e]

Université catholique de Louvain[a] / Belgium
Northern Arizona University[b] / USA
University of Oslo[c] / Norway
University of Barcelona[d] / Spain
Radboud University[e] / The Netherlands

Abstract – The *Varieties of English for Specific Purposes dAtabase* (VESPA first release) is the result of an international corpus compilation project that aims to address the lack of large-scale, open access, multi-L1, multi-discipline and multi-register learner corpora. This corpus report provides a detailed description of VESPA and illustrates possible uses of the corpus for register exploration of learner data. Specifically, it first offers an overview of the makeup of the corpus and the online interface that can be used to search and download the corpus. It then gives an illustrative example of a study where multi-dimensional analysis was used to investigate the relative importance of register vis-à-vis other factors in learner academic writing. In the concluding remarks, we identify priorities for future developments in the VESPA project, including the addition of more L1 components, more disciplines and more registers, as well as the compilation of a comparable corpus of native student writing.

Keywords – learner corpus; learner corpus research; English as a Foreign Language; academic writing, register variation; student writing

1. INTRODUCTION[1]

The main objectives of this corpus report are to provide a detailed description of the *Varieties of English for Specific Purposes dAtabase* (VESPA first release) and to illustrate how the corpus can be used to facilitate exploration of learner languages across registers

[1] We are most grateful to Paul Rayson (Lancaster University, UK) for giving us access to the CLAWS7 POS-tagger. We also thank Hubert Naets (UCLouvain, Belgium), main developer of the corpor@uclouvain.be platform, for his help at the initial stages of the project.

and different first-language (L1) backgrounds. As outlined below, corpora enabling large-scale, multi-L1, multi-discipline and multi-register investigations of learner language have previously not been available to researchers in the field. In making VESPA publicly available, we hope to help facilitate such studies, thus contributing one among many resources needed in order to provide a more accurate and nuanced picture of learner language.

Traditionally, the vast majority of written learner corpora available to the research community have included general argumentative or narrative texts produced by foreign language learners in the context of foreign/second language courses for general purposes (e.g. the *International Corpus of Learner English* (ICLE), 3rd edition, Granger *et al*. 2020). More recently, a number of learner corpora that comprise official language tests have also been released (e.g. *ETS Corpus of Non-Native Written English*, Blanchard *et al.* 2013; the *Open Cambridge Learner Corpus* 2017). By focusing almost exclusively on these contexts of use (and associated tasks), however, the field of learner corpus research has arguably developed a somewhat narrow perspective on what learner languages typically are. For example, overuse of first person pronouns, pragmatic inappropriateness and overstatements are linguistic features commonly reported in the literature to be typical of English as a Foreign Language (EFL) (e.g. Paquot 2010). This is somewhat problematic given that a growing body of research (e.g. Paquot *et al*. 2013; Larsson and Kaatari 2019) has noted that learners' use of many of these features (most particularly features related to writer-reader visibility) are often register-specific, thereby demonstrating the importance of including a broader range of registers in studies of learner language.

Further, in the context of English for Academic Purposes (EAP), the scope of registers analyzed to identify (i) typical characteristics of learner writing (development) and (ii) learners' difficulties remains overly restricted, meaning that the results of such studies often are of limited utility for EAP pedagogy. As stated by Biber *et al*. (2020: 49)

> university students are expected to produce a bewildering array of different registers, associated with the expectations of different disciplines, at different levels of study, and associated with the particular tasks required by their academic programs.

Therefore, there is a need for EAP researchers and practitioners to broaden their empirical basis. Corpora of EFL learner academic writing have been, or are being, compiled, but for different reasons, they are rarely available (Granger and Paquot 2013). Examples

include the *Corpus of Academic Learner English* (Callies and Zaytseya 2013) and the corpus of L2 disciplinary writing used in recent studies by Biber and colleagues (Staples *et al*. 2018; Biber *et al*. 2020). In addition, they often represent the writing of just one L1 population (e.g. German EFL learners in the *Aachen Corpus of Academic Writing*; Ströbel *et al*. 2020) or one register with a focus on dissertations (e.g. *Chinese Academic Written English Corpus*; Lee and Chen 2009). In that sense, the situation has not evolved much since Alsop and Nesi's (2009: 72) remark that discipline-specific student writing "has tended to be collected for individual scholarly purposes rather than as part of formal corpus-building projects."

While recently compiled open access corpora of academic writing such as the *British Academic Written English* corpus (BAWE; Nesi *et al*. 2008) and the *Michigan Corpus of Upper-level Student Papers* (MICUSP; Römer and O'Donnell 2011) include some texts by L2 writers, they were not compiled with a view to studying learner writing and/or learner writing development. Rather, the main objective of their collection is to investigate register and disciplinary differences in academic writing through a record of highly proficient university-level (mostly native-speaker) student writing. This means that only a limited number of learner texts per discipline or register are included; for example, there are only 39 EFL learner texts written in the field of linguistics in BAWE, with a variety of first languages represented (Bulgarian, Chinese, French, German, Greek, Italian, Japanese and Portuguese).

Given this lack of large-scale, open access, multi-L1, multi-discipline and multi-register corpora of learner academic writing, the VESPA learner corpus compilation project was initiated by Dr. Magali Paquot at the *Centre for English Corpus Linguistics* (CECL, UCLouvain, Belgium) with the aim to build a large collection of disciplinary writing by L2 English university students across registers and disciplines. Like other CECL corpora, VESPA is a corpus compilation project that involves collaborative work among several universities internationally. Partners have joined at different times and the corpus is still under compilation, with new components (e.g. new L1 backgrounds and more disciplines) continuously being added. The compilation process is described in detail in Section 2 together with an overview of the makeup of the corpus and the online interface.

While still work-in-progress, VESPA has already been used in a variety of studies to analyze linguistic features of EFL learners' academic writing in content courses (e.g.

Hasselgård 2014; Larsson 2019; Paquot 2019; Larsson *et al*. 2020), and to compare learners and native speakers' use of recurrent word combinations across disciplines (Ebeling and Hasselgård 2015). VESPA has also been used to complement data from other learner corpora such as ICLE: used together, the two learner corpora enable large-scale, multi-L1, multi-register explorations of learner data (Paquot *et al*. 2013; Larsson *et al*. 2021). With more subcorpora being added (especially subcorpora representing more disciplines) in the future, VESPA will also allow researchers to compare learner academic writing across registers and disciplines. In Section 3, we illustrate one of the many possible uses of VESPA by providing a brief overview of a recent study that made use of multi-dimensional analysis to investigate the relative importance of register vis-à-vis other factors in learner academic writing (Larsson *et al*. 2021). Finally, in Section 4, we make some concluding remarks.

2. VESPA: CORPUS COMPILATION, CORPUS PROCESSING AND ACCESS

In its current form (first release), VESPA comprises 941 texts (over 2 million words) produced by university students at the Bachelor's and Master's levels and collected by VESPA partners from five European universities (Radboud University, The Netherlands; UCLouvain, Belgium; University of Barcelona, Spain; University of Oslo, Norway; Uppsala University, Sweden), as shown in Table 1. The majority of the texts were written by students who have one of the official languages of the partner institutions (Dutch, French, Norwegian, Spanish, and Swedish, respectively) as their first language. Given the cultural diversity of some of the cities where the partner institutions are situated and the internationalization of higher education, however, 26 per cent of the collected texts across the various institutions represent academic writing by EFL learners with other L1 backgrounds than the official language of the respective institutions (examples of these other L1 backgrounds include Chinese, Czech, German, Greek, Italian, Polish, Russian, Turkish, and Vietnamese). 23 per cent of the students also report that they speak two languages or more at home.

Institution	Main L1 language represented	Number of texts	Total number of words	Number of words per text (median [Q1 – Q3])
Radboud University (The Netherlands)	Dutch	118	310,099	2,616 [1,992 – 3,152]
UCLouvain (Belgium)	French	154	648,483	4,072 [3,295 – 4,816]
University of Barcelona (Spain)	Spanish	85	57,323	575 [525 – 755]
University of Oslo (Norway)	Norwegian	515	772,964	1,180 [738 – 2,005]
Uppsala University (Sweden)	Swedish	69	399,352	6,038 [2,894 – 7,634]
Total		**941**	**2,188,221**	**1,809 [822 – 3,224]**

Table 1: Corpus size per institution and main L1 language represented

With regard to the types of text included, VESPA comprises assignments that students submitted for course credit in disciplinary content courses. In that sense, the corpus answers repeated calls for greater ecological validity in L2 writing research (Polio 2017; Biber *et al*. 2020). As shown in Table 2, the large majority of the texts (79%) were collected in linguistic courses (taught by VESPA partners or colleagues in the same department) but some VESPA partners have also started compiling sub-corpora in literature and business communication.

Discipline	**Number of texts**
Linguistics	741
Business communication	126
Literature	74
Total	941

Table 2: Disciplines represented in VESPA

To classify the VESPA texts into register categories, we used the classification system from MICUSP (Römer and O'Donnell 2011: 170–171), which has two main advantages: the number of text categories is limited to seven, and each category comes with a set of defining linguistic features that can serve as simple guidelines. Table 3 provides an overview of the texts across the five register categories currently represented (critique/evaluation, proposal, report, research paper and response paper). This categorization is the result of an annotation procedure where each text was coded either using the register category identified by looking at the course requirements or, for the cases where we did not have access to the course requirements or could not obtain the information from the course instructor, texts were double coded by two VESPA partners (or a VESPA partner and a trained research assistant). Any disagreements were discussed and resolved with the VESPA coordinator. As shown in Table 3, the majority of texts (78%) fall into one of two categories: reports and research papers. However, given that texts were collected in different courses with different requirements at different institutions, the corpus is not balanced in terms of register by L1.

Institution	Radboud University (The Netherlands)	UCLouvain (Belgium)	University of Oslo (Norway)	University of Barcelona (Spain)	Uppsala University (Sweden)	
Main L1 represented	*Dutch*	*French*	*Norwegian*	*Spanish*	*Swedish*	
Registers						**Total**
Critique / evaluation	5	3	129	0	0	**137**
Proposal	45	0	0	0	0	**45**
Report	26	36	268	85	0	**415**
Research paper	42	115	93	0	69	**319**
Response paper	0	0	25	0	0	**25**
Total	**118**	**154**	**515**	**85**	**69**	**941**

Table 3: Registers represented in VESPA

Table 4 provides information about the main rhetorical purpose of each register, its defining features and examples as detailed in Römer and O'Donnell (2011).

Register	Rhetorical purpose	Defining features	Example
Critique/evaluation	Presents a positive or negative assessment of an outside source/project/text	- The text is driven by an in-depth assessment of a product/policy/procedure/text (although often interwoven with a description or observation of the product/policy/procedure/text) - Gauges the effectiveness, validity, or usefulness of something - Recommendations for improvement may be offered	Evaluation of business practices, problem-solution, literary critique, operations report
Proposal	Puts forth a research question, a theory or a model that the author feels should be explored in order to further the understanding of a given topic	- Formulates a research question or model, or proposes a potential study - Usually does not collect or synthesize new data, but may include projected results; any collected data will be to support the proposal - Justifies the need for data collection or data verification - Critiques relevant literature and/or prior studies	Research proposal
Report	Describes the state or gives an account of a problem/issue/text, or describes the carrying out of a procedure (demonstrates the ability to gather data and summarize)	- Most space is devoted to description, rather than critical assessment - Not driven by an original thesis or research question - Author's opinion/evaluation may be present, but is not foregrounded and does not appear to drive the text	Lab report, literature review, article review, annotated bibliography, compare/contrast paper
Research paper	Presents original research in the field	- Entire text serves to answer a clearly stated research question - Contains original data, or compiles existing data for the purpose of providing a new interpretation - Structured into predictable sections (usually with subheadings) - Includes most of the following: abstract, literature review, methods, results, discussion, conclusion	Research paper, replication study
Response paper	Short piece of writing responding to a given prompt or question, although prompt may not be explicit in the text	- Short in length (typically 1-2 pages) - Style tends to be informal (e.g. expressions of emotional response; frequent references to mental processes, such as 'I was confused', 'I was surprised') - May lack a formal introduction/'jumps right in' to content of paper, because author assumes reader's familiarity with the given topic (shared knowledge or in-group knowledge) - The text provokes new questions for the author that may not be thoroughly answered	Solution to a homework problem, personal response to a text

Table 4: VESPA text categories and definitions for text classification (adapted from MICUSP paper categories, Table 5 in Römer and O'Donnell 2011: 170–171)

The VESPA corpus compilation followed the same procedure across all institutions; this procedure aimed to maximize homogeneity of texts by applying the same inclusion criteria for all the texts across all institutions. First, we recruited students in specific content courses via their instructors.[2] The students filled out a questionnaire that is used to collect a set of learner and task variables (e.g. first language, level of study, number of years studying English at university, and content course for which the text was written) as well as a permission form. Both files are available in paper format and as an online survey. Second, the VESPA partner(s) at each institution collected the student work in electronic format, typically as Microsoft Word documents, and then annotated and processed the files with a series of tools developed or adapted for the project. These steps resulted in marked-up .xml files that are then ready for inclusion into VESPA. More specifically, following the procedure used in the BAWE corpus (Ebeling and Heuboeck 2007; Heuboeck *et al*. 2008), the texts were first processed using Word macros to annotate main sections (e.g. abstract, introduction), block quotes and so-called mentioned items (e.g. cited works, foreign words, linguistic examples). Next, the annotated texts were processed by means of Perl scripts to produce .xml files that include both the text and the metadata.[3] The complete corpus compilation procedure is described in the VESPA manual (Paquot *et al*. 2015).[4]

VESPA is available open access for non-profit educational and/or linguistic research purposes from the corpor@uclouvain.be platform, an online catalogue of corpora compiled at UCLouvain.[5] The platform can be used to search or download the corpus, in parts or in whole. Users first select texts by ticking variables of interest (e.g., all texts written in linguistics courses by French EFL learners) in the first three tabs of the 'Text selection' menu (Learner variables I, Learner variables II, and Task variables). Figure 1 shows the 'Task variables' page. The distribution of texts for each variable is dynamic; for example, in VESPA as a whole, there are more texts at the Bachelor's level than at the Master's level. However, if Radboud University is the only university that is

[2] Note that this is the main reason why each partner started with the collection of papers written in linguistic courses. Most of the time, VESPA partners were also the instructors for these courses and had direct access to the students and their writing.
[3] The Word macros and Perl scripts were developed by Alois Heuboeck (Reading University, UK); they are largely based on what was developed for the *British Academic Written English* (BAWE) corpus (cf. Ebeling and Heuboeck 2007; Heuboeck *et al*. 2008).
[4] The corpus collection guidelines and all associated material (student questionnaire, permission form, and Word macros) are available at https://tinyurl.com/VESPAguidelines.
[5] https://corpora.uclouvain.be/cecl/vespa/

ticked in the institution variable, the figures are recomputed for that particular institution, and we see that, in this subcorpus, the majority of texts were collected at the Master's level. As shown of Figure 2, the distribution of texts can also be explored graphically.

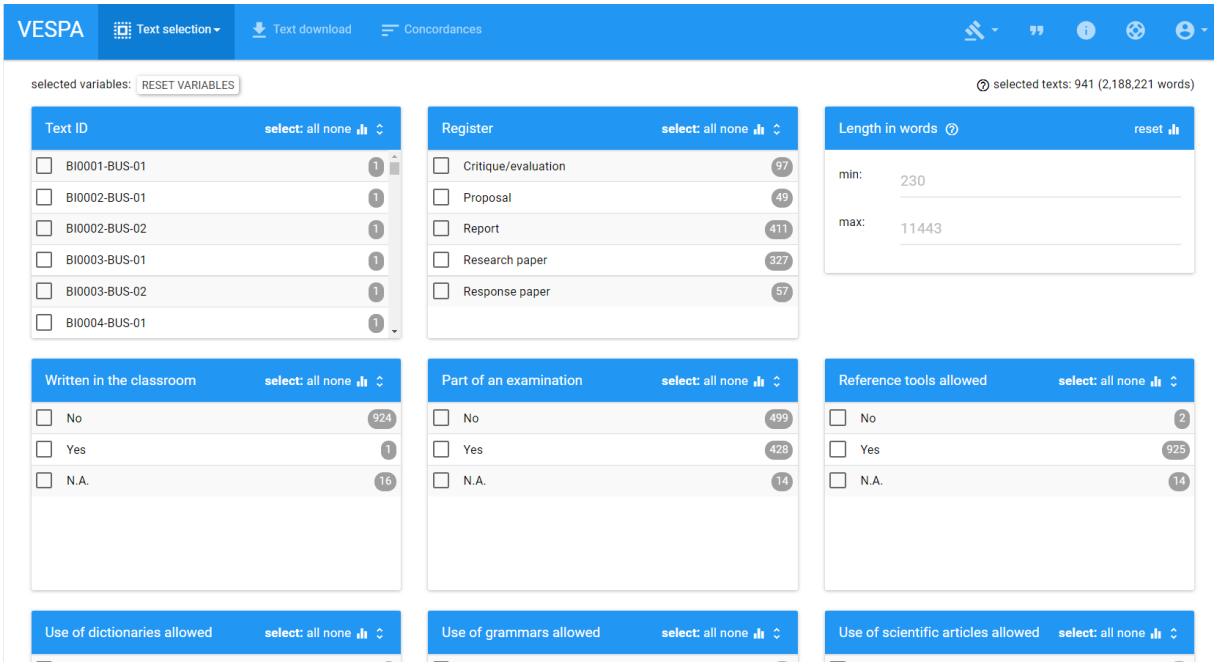


Figure 1: Selecting VESPA texts

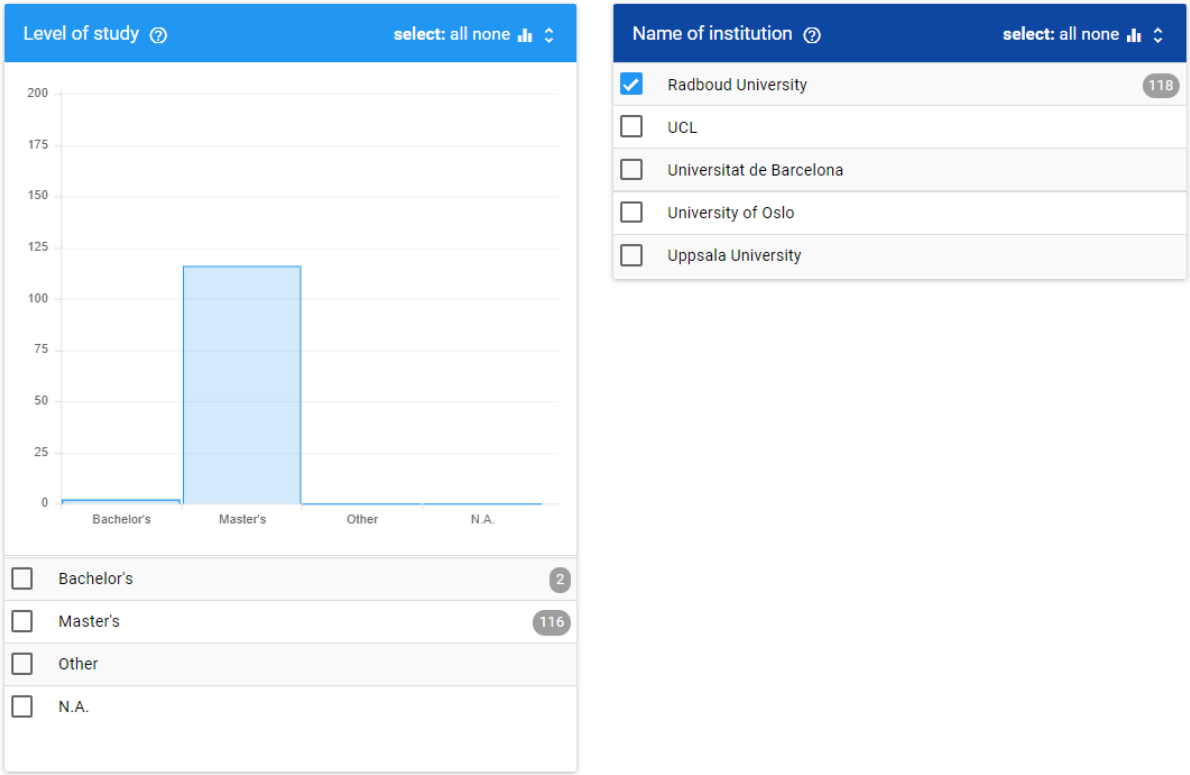


Figure 2: Exploring the corpus with the https://corpora.uclouvain.be/catalog/ platform

When a set of texts has been selected, the user can download it as a .zip file that will contain:

- A folder containing separate txt files for each text in the corpus (in UTF-8 format, no header);
- A file grouping all the texts in the corpus in a single txt file (in UTF-8 format);
- A database containing the learner profile information (metadata) for each text in the corpus in both .csv and .xlsx formats.

Alternatively, the selected texts can be explored online with a built-in concordancer that was initially developed for the third version of ICLE (Granger *et al*. 2020). One major improvement to the system is that it is configured to only search for linguistic items produced by EFL learners. Thus, if a user searches for the connector *however*, occurrences found in block quotes and mentioned items (see above) will not be retrieved.

All texts in VESPA are lemmatized and part-of-speech (POS) tagged with CLAWS7.[6] The concordance therefore makes it possible to search for word forms, lemmas, POS tags as well as combinations of word forms and lemmas with POS tags (see Part IV of Granger *et al*. 2020 for more details). Note, however, that the results of the automatic annotation were not manually checked and users of the corpor@uclouvain.be platform should check their accuracy when conducting a linguistic study that relies on lemma- and/or POS-based queries. Figure 3 shows the results of a search for the sequence *it* + modal verb + *be* + past participle in the whole corpus. Such concordances can then be exported in .xlsx or .csv format together with associated metadata, thus facilitating further analysis and treatment of the data outside the interface.

[6] https://ucrel.lancs.ac.uk/claws7tags.html

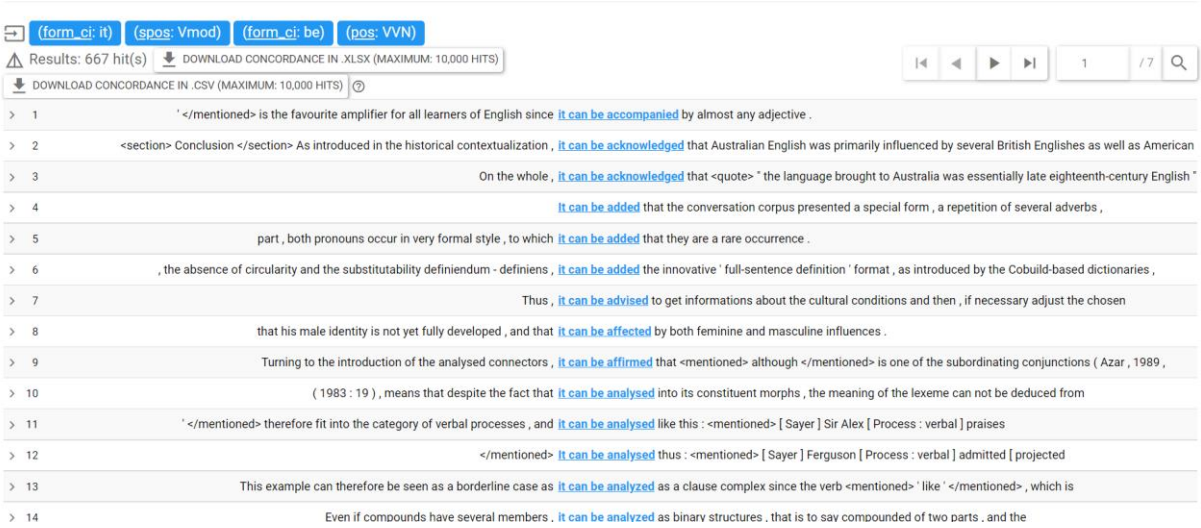


Figure 3: Searching VESPA with the corpor@uclouvain in-built concordancer

3. MAKING USE OF VESPA TO EXPLORE REGISTER VARIATION

As mentioned in Section 1, VESPA can be used for many different kinds of multi-L1 register comparisons, especially as a complement to the widely used ICLE (which almost exclusively includes argumentative essays). We will here illustrate this line of research by means of a recent study that made use of multi-dimensional (MD) analysis (Biber 1988) to examine learner and native-speaker student writing from two registers (argumentative essays and research papers) and published scientific articles, with the aim of investigating possible register effects in EFL learner writing. MD analysis is an approach used to describe and compare registers employing a wide range of linguistic co-occurrence patterns reduced to a few underlying 'dimensions' of variation that are then interpreted functionally (for a more detailed account of MD analysis, see Biber 1988, 1992). As such, the approach is ideally suited to investigate the extent to which features commonly attributed to EFL learner writing should be seen as more general characteristics of learner writing, as indicated in previous studies, or whether they may instead be prompted by (or at least moderated by) the register investigated. As shown in Table 5, the selection of corpora included in this study allowed for several different comparisons:

- Argumentative essays vs. research papers[7] vs. scientific articles (ICLE + LOCNESS vs. VESPA + BAWE + MICUSP vs. LOCRA)

[7] It is important to note that when the study reported on in Larsson *et al*. (2021) was conducted, the more detailed register categorization of VESPA texts had not been conducted yet. In that study, the term 'research paper' was used in a broader sense, as a superordinate category to refer to any piece of academic disciplinary

- Non-native vs. native speakers of English (ICLE + VESPA vs. LOCNESS + BAWE)
- L1 background (French, Spanish, Norwegian, Swedish and Dutch)

Corpus	L1	Register	Number of words	Number of texts
ICLE	French, Spanish, Norwegian, Swedish and Dutch	Argumentative essays	708,541	1,073
LOCNESS	English	Argumentative essays	99,520	88
VESPA	French, Spanish, Norwegian, Swedish and Dutch	Research papers in linguistics	1,303,278	584
BAWE	(British) English	Research papers in linguistics	167,482	76
MICUSP	(American) English	Research papers in linguistics	313,785	34
LOCRA	NA	Scientific articles in linguistics	956,761	109
Total			**3,549,367**	**1,964**

Table 5: Overview of the corpora used in Larsson *et al*. (2021)

The results of the multi-dimensional analysis showed that the features investigated vary along two dimensions in the texts: 'Personal vs. topic-focused style' (Dimension 1) and 'Evaluative style vs. factual descriptions' (Dimension 2). While the study also reported certain differences across native vs. non-native status or L1 groups, the main differences were found between the registers, stressing its importance as a moderating variable. With both dimensions taken together, the novice writers' research papers (natives and non-natives) and the experts' scientific articles were found to be characterized by topic-focused and factual descriptions, the scientific articles significantly more so than the research papers. By contrast, the argumentative essays were shown to be personal and evaluative (L2 learners) or personal and topic-focused (English L1 students). Only very limited evidence was found to support claims made in previous studies about learner-specific characteristics such as a more involved style.

Larsson *et al*.'s (2021) results provide empirical evidence to support the increasingly more accepted view that "if we limit our investigations to argumentative writing only, the findings are likely to reflect that register and the results cannot (and should not) be used to make general claims about 'learner writing'" (Larsson *et al*. 2021: 254). The release of VESPA and its newly developed register classification will enable further explorations of learner (disciplinary) writing across more varied and specific

writing that provides analysis, interpretation, and/or argument based on independent research work. As such, the different register categories represented in VESPA are subsumed under this broader category (see Table 3).

registers than have often been the focus of previous research. With its focus on specialized registers in academic writing, VESPA can help answer (sometimes widely debated) questions such as (i) What are the main difficulties L2 writers face in an academic setting?; (ii) Are EFL learners' needs the same across disciplines and registers?; (iii) Does it make sense to provide general EAP courses?; and (iv) To what extent are L2 learners' needs the same as those of novice L1 students in an academic setting? (e.g. Gilquin *et al*. 2007; Römer 2009).

4. Conclusion

This corpus report has served to introduce VESPA and illustrate some of its many uses. While the corpus in its current form has already proven useful for describing linguistic features typical of specific types of disciplinary writing (mostly linguistics), and comparing learner features across registers, it is our belief that the following developments will make the corpus even more useful for the research community in the future. First, more partners have joined the project and corpora of disciplinary writing by Czech, Filipino and Turkish students are currently under development. Second, VESPA will soon also include comparable data in the discipline of linguistics by English-speaking L1 students. Third, we are also exploring avenues to collect data in other disciplines than linguistics, literature, and business.

It is our hope that the release of VESPA coupled with the publication of this corpus report will serve to inspire more research on learner languages across registers and disciplines.

References

Alsop, Sian and Hilary Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora* 4 /1: 71–83

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1992. The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities* 26: 331–345.

Biber, Douglas, Randi Reppen, Shelley Staples and Jesse Egbert. 2020. Exploring the longitudinal development of grammatical complexity in the disciplinary writing of L2-English university students. *International Journal of Learner Corpus Research* 6/1: 38–71.

Blanchard, Daniel, Joel Tetreault, Derrick Higgins, Aoife Cahill and Martin Chodorow. 2013. TOEFL11: A corpus of non-native English. *ETS Research Report Series*, 2013/2: i–15. https://doi.org/10.1002/j.2333-8504.2013.tb02331.x (29 September, 2021.)

Callies, Marcus and Ekaterina Zaytseva. 2013. The Corpus of Academic Learner English (CALE) – A new resource for the assessment of writing proficiency in the academic register. *Dutch Journal of Applied Linguistics* 2/1: 126–132.

Ebeling, Signe O. and Hilde Hasselgård. 2015. Learners' and native speakers' use of recurrent word-combinations across disciplines. In Ann-Kristin H. Gujord, Susan Nacey, Silje Ragnhildstveit eds. *Learner Corpus Research: LCR2013 Conference Proceedings* (Bergen Language and Linguistics Studies 6), 87–106.

Ebeling, Signe O. and Alois Heuboeck. 2007. Encoding document information in a corpus of student writing: The British Academic Written English Corpus. *Corpora* 2/2: 241–256.

Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. In Paul Thompson ed. *Corpus-based EAP Pedagogy. Special issue of the Journal of English for Academic Purposes* 6/4: 319–335.

Granger, Sylviane, Maïté Dupont, Fanny Meunier, Hubert Naets and Magali Paquot. 2020. *The International Corpus of Learner English* (version 3). Louvain-la-Neuve: Presses universitaires de Louvain.

Granger, Sylviane and Magali Paquot. 2013. Language for specific purposes learner corpora. In Carol A. Chapelle ed. *The Encyclopedia of Applied Linguistics*. Oxford: Blackwell-Wiley.

Hasselgård, Hilde. 2014. *It*-clefts in English L1 and L2 academic writing. In Kristin Davidse, Caroline Gentens, Lobke Ghesquière and Lieven Vandelanotte eds. *Corpus Interrogation and Grammatical Patterns*. Amsterdam: John Benjamins, 295–320.

Heuboeck, Alois, Jasper Holmes and Hilary Nesi. 2008. The BAWE Corpus Manual. http://www.reading.ac.uk/AcaDepts/ll/app_ling/internal/bawe/BAWE.documentation.pdf (29 September, 2021.)

Larsson, Tove. 2019. Grammatical stance marking in student and expert production: Revisiting the informal-formal dichotomy. *Register Studies* 1/2: 243–268.

Larsson, Tove, Marcus Callies, Hilde Hasselgård, Natalia J. Laso, Magali Paquot, Sanne van Vuuren and Isabel Verdaguer. 2020. Adverb placement in EFL academic writing: Going beyond syntactic transfer. *International Journal of Corpus Linguistics* 25/2: 155–184.

Larsson, Tove and Henrik Kaatari. 2019. Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. *International Journal of Learner Corpus Research* 5/1: 33–62.

Larsson, Tove, Magali Paquot and Douglas Biber. 2021. On the importance of register in learner writing: A multi-dimensional approach. In Elena Seoane and Douglas Biber eds. *Corpus-based Approaches to Register Variation*. Amsterdam: John Benjamins, 235–258.

Lee, David Y. W. and Sylvia Xiao Chen. 2009. Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18/3: 149–165.

Nesi, Hilary, Sheena Gardner, Paul Thompson and Paul Wickens. 2008. *British Academic Written English Corpus*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2539

Open Cambridge Learner Corpus (v1). 2017. Distributed by Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment.

Paquot, Magali. 2010. *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London: Continuum.

Paquot, Magali. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research* 35/1: 121–145.

Paquot, Magali, Hilde Hasselgård and Signe O. Ebeling. 2013. Writer/reader visibility in learner writing across genres: A comparison of the French and Norwegian components of the ICLE and VESPA learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Louvain-la-Neuve: Presses universitaires de Louvain, 377–387.

Paquot, Magali, Signe O. Ebeling, Alois Heuboeck and Larry Valentin. 2015. *The VESPA Tagging Manual* (version 2.3). Louvain-la-Neuve: Centre for English Corpus Linguistics.

Polio, Charlene. 2017. Second language writing development: A research agenda. *Language Teaching* 50/2: 261–275.

Römer, Ute. 2009. English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies* 20/2: 89–100.

Römer, Ute and Matthew Brook O'Donnell. 2011. From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan Corpus of Upper-level Student Papers (MICUSP). *Corpora* 6/2: 159–177.

Staples, Shelley, Douglas Biber and Randi Reppen. 2018. Using corpus-based register analysis to explore authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Modern Language Journal* 102/2: 310–332.

Ströbel, Marcus, Elma Kerz and Daniel Wiechmann. 2020. The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning* 70/3: 732–767.

Corresponding author
Magali Paquot
SSH/ILC
Collège Erasme
Place Blaise Pascal 1, bte L3.03.31
1348 Louvain-la-Neuve
Belgium
e-mail: magali.paquot@uclouvain.be

received: October 2021
accepted: June 2022

RiCL Research in Corpus Linguistics

A lexico-grammatical comparison of statutory law and popular written language

Margaret Wood
Northern Arizona University / United States

Abstract – While the plain language movement has shed light on the lack of readability of statutory texts for the lay person, there has been a lack of empirical methodology employed to determine the ways in which statutory language differs lexico-grammatically from forms of popular language that are familiar to the lay person. With this in mind, the present study conducts a comparative analysis of statutory language and other forms of popular written language (i.e., a corpus of news reports, sports reports, encyclopedia articles, and historical articles) with two goals: 1) to provide a detailed lexico-grammatical description of statutory law independent from other forms of legal writing, and 2) to identify pervasive lexico-grammatical features of statutory language that the lay person has relatively less exposure to in comparison to other written registers. Following a bottom-up selection of lexico-grammatical features for analysis, a key feature analysis is used to identify linguistic features that are more pervasive in statutory law relative to other forms of popular written language as measured through Cohen's *d* effect sizes. Results reveal the pervasive use of the passive voice, prepositions, a variety of coordinating conjunctions, the pied-piping *wh*-relative clause construction, and non-finite *-ing* and *-ed* clause constructions in statutory language. These results complement previous research regarding the features that are characteristic of statutory language and help to identify features that potentially contribute to the lack of readability of statutory law.

Keywords – statutory law; register variation; readability; popular language; key feature analysis

1. INTRODUCTION

For years, people have bemoaned the lack of readability of written legal documents, in particular for those outside of the profession or without detailed knowledge of the law. The plain language movement, which has its roots in the 1970s, calls for legal language that is accessible and readable for the lay person. With this has come numerous attempts to describe the language of written legal documents and identify the features that are detrimental to the readability of the texts.

While these linguistic descriptions have concerned a variety of written legal texts (all of which pose readability challenges for the lay person), there are comparatively fewer empirical descriptions of statutory language independent from other forms of legal writing (e.g., contracts, agreements, treaties). The current lack of independent

focus on statutory law is problematic, as the domain carries an extraordinary amount of power over the lay person; explicitly creating, modifying, and terminating legal rights and obligations of everyday individuals (Tiersma 1999: 1). Because a long history of register variation studies tells us that linguistic characteristics of a text will differ in relation to the situational context in which they occur (Biber and Conrad 2009), it is important for the discussion of readability of legal texts to identify the lexico-grammatical features that are pervasive in the register of statutory law as an independent form of written legal language.[1]

Claims are frequently made about the pervasiveness of certain features of legal writing based on simple frequency counts within a register or across multiple combined registers. While this may tell us which features are more common in the register relative to other features in that same register, if we wish to identify features that are uniquely characteristic of statutory law and aim to make claims about pervasiveness, the register must be described in relation to a different text variety or domain. Using non-legal language as domain for comparison stands to contribute to the discussion of readability as it allows for the identification of pervasive lexico-grammatical features in statutory law that the lay person has relatively less exposure to on an everyday basis. The value of this lies in the assumption that a lack of exposure to the characteristic linguistic structures of a specific text variety has the potential to impede one's understanding of it. We see evidence of this in the fact that the typical 'audience' of statutes, or those who interact with them on a daily basis (i.e., lawyers and judges), seem to be able to make sense of the texts more readily than the lay person.

With this in mind, the present study aims to provide a linguistic description of codified state statutory law in relation to other forms of non-legal, popular written language, with two goals: 1) to provide a detailed lexico-grammatical description of the features that are characteristic of state statutory law (as a text variety that holds great power over the lay person), and 2) to identify pervasive lexico-grammatical features of statutory law that the lay person has markedly less exposure to in comparison to other forms of popular written language. The present study proposes that the bottom-up (rather than top-down) identification of features that are pervasive in statutory law and relatively less common in other forms of popular written language will allow for the

[1] The present study uses the term 'register' to refer to culturally-recognized text varieties (Biber and Conrad 2009: 6).

removal of personal intuition concerning which features are relevant in the conversation of readability. Through a linguistic comparison of a corpus of state statutory law (i.e., bills proposed by an elected member of a state house or senate, drafted by a draftsperson, passed through various committees, and signed into law by the governor of the state), and a corpus of popular written language comprising online news reports, sports reports, encyclopedia articles (i.e., *Wikipedia*), and historical articles, the present study aims to contribute to the discussion concerning the language that poses a threat to the readability of statutes for the lay person.

2. LITERATURE REVIEW

2.1. Linguistic descriptions of written legal language

Previous linguistic descriptions of written legal language have concerned a variety of registers including decisions, directives, regulations, law journals, commercial law documents, case law, contracts, law reports, and legislation. Linguistic studies of these registers have most commonly focused on lexis: in particular, lexical bundles and keyword analysis (Caliendo *et al*. 2005; Trebits 2009; Jablonkai 2010; Breeze 2013; Biel 2017; Alasmary 2019; Serachini 2020), phraseology (Biel 2009, 2014; Pontrandolfo 2015) and on single features such as modals (Foley 2002; Andersson 2007; Gibova 2011) and personal pronouns (Rodríguez-Puente 2019).

Only a select number of studies that have described forms of written legal language in terms of their lexico-grammatical characteristics, though these have largely been undertaken without the use of a reference register (notable exceptions include Goźdź-Roszkowski 2011, and Biber and Gray 2019). Studies that have focused on the lexico-grammatical characterization of legislative writing have described it as both structurally compressed and structurally elaborated. The frequent use of nominalization (nouns that have been morphologically derived from verbs or adjectives), which are features often associated with structurally compressed written language (Biber 1988; Biber and Gray 2016), are considered highly characteristic of legislation (Goźdź-Roszkowski 2011; Sun and Cheng 2017). Williams (2013: 354) similarly characterized legislative language as structurally compressed, noting in particular its reliance on nouns, including the frequent use of nominalization and high density of noun phrases.

Others describe structural elaboration of legislative language through the density of clausal embedding, which is frequently considered one of the most detrimental features to readability (Williams 2007). Charrow and Charrow (1979: 1329) specifically attribute readability issues to central embedding, in which there are two subordinate clauses; one enclosed within the other. Bhatia (1983: 50) also noted that legislation displays a high degree of subordination, citing adverbials and non-finite prepositional constructions as particularly common. Embedded clauses have been referred to as 'qualification inserts', which are used to flesh out main ideas of a clause and directly contribute to the syntactic complexity of legislative language (Bhatia 1993). Goźdź-Roszkowski (2011: 136) found that legislative language made particularly frequent use of different types of post-nominal clauses, including *wh*-relative clauses, *that* relative clauses and the pied-piping construction. Tiersma (1999: 62) also noted that legislation frequently makes use of coordinating conjunctions *and* and *or* to combine multiple clauses, contributing to the 'wordy' nature of the texts, and states that "the possibilities of creating tremendously long phrases and sentences by use of conjunctions like *and* and *or* are virtually limitless."

Use of the passive voice is also considered highly characteristic of legislative writing. According to Williams (2004: 231), approximately one quarter of all verbal constructions in prescriptive legal English are in the passive voice. Bulatović (2013: 103) found that of the verb phrases counted in a corpus of acts, around 65 per cent were in the active voice and 35 per cent were in the passive voice. Of those passives, around 24 per cent served as post-nominal modifiers in the form of past participles (Bulatović 2013: 104).

However, as previously noted, a majority of the studies above describe legislative writing without comparison to other registers. It is difficult to know, for example, how notable it is to have a text with 35 per cent of its verbal constructions in the passive voice, if there is nothing to compare this percentage too. For this reason, the present study aims to test these claims about pervasiveness through empirical, comparative means.

2.2. Linguistic descriptions of popular written varieties

The present study focuses on written language that is 'popular', that is, on language that is written specifically for the lay person as its audience and is easily accessible to them. For this reason, the study investigates the online popular written registers of news reports, sports reports, encyclopedia articles, and historical articles as registers that fit these criteria (see Section 3.1.2).

The most prominent large-scale linguistic description of forms of popular language was undertaken by Biber *et al.* (1999) in the *Longman Grammar of Spoken and Written English*. Using the *Longman Spoken and Written English Corpus* (LSWE), which comprises over 40 million words representing six registers, Biber *et al.* (1999: 5) compiled a "descriptive and explanatory account of English grammar." Four core registers were used in their analysis: conversation (British), fiction (American and British), news (British), and academic prose (American and British). Biber *et al.* (1999: 25) also included two other sets of texts for dialect comparison (American conversation and American news), and two supplementary registers (British non-conversational speech, and British and American general prose). Biber *et al.* were able to investigate structural descriptions of the features and patterns of use, and comment on the pervasiveness of the features in comparison to other registers. They undertook extensive functional interpretation of the quantitative data, in particular in terms of three functional associations: the work that a feature performed in discourse, the processing constraints that it reflected, and the situational and social distinctions that it conventionally indexed (Biber *et al.* 1999: 41). Of particular interest to the present study is the lexico-grammatical comparison of formal academic prose to other non-academic registers, as academic prose generally shares much in common with previous descriptions of legislative writing, namely, the tendency towards dense, informational, compressed language.

The popular written register of news has frequently been the subject of investigation, largely studied through discourse analysis (e.g., Davies 2012; Fowler 2013; Bednarek and Caple 2014; Scollon 2014; Xie 2018). These studies have often focused on highly specific contexts; for example, political posts in the Jakarta post newspaper (Yana 2015) and socio-political influences on lexico-grammatical features in Ecuadorian Spanish news (Tapia and Biber 2014). However, select others have had a broader focus. In a multi-dimensional analysis of registers on the searchable web, Biber

and Egbert (2016: 109) found that news reports were characterized by a set of co-occurring features frequently associated with written informational language; largely, a variety of nominal modifiers. Biber and Egbert (2016) also found, however, that news was characterized by the co-occurrence of features such as complement clauses and *that* deletion, which are often associated with oral language varieties. Through a later key feature analysis in *Register Variation on the Web*, Biber and Egbert (2018) found that when set aside a reference corpus of other web registers, news reports displayed a relatively higher use of communication verbs, proper nouns, common nouns, perfect aspect, pre-modifying nouns, and prepositions.

In both studies, Biber and Egbert (2016, 2018) provided linguistic descriptions of a variety of other web registers, including encyclopedia articles, historical articles, and sports reports (registers of analysis in the present study). Biber and Egbert (2016) found that encyclopedia articles were characterized by the co-occurrence of features associated with literate-informational language (prepositional phrases, passive non-finite relative clauses, relative clauses). In the later key feature analysis, Biber and Egbert (2018:162) found that passives, prepositions, longer word length, and nominalizations were key in the register. They found that historical articles had similar key features, though with the notable added use of the past tense, which was the most key feature in the register with a large effect size of $d > 1.0$ (Biber and Egbert 2018: 95). On the other hand, sports reports made pervasive use of features associated with narrative and oral varieties when set aside a reference corpus of the web registers, including proper nouns, third-person pronouns, activity verbs, past tense, perfect aspect, contractions, and adverbs of place (Biber and Egbert 2018: 90). Notably, while proper nouns were also key for news reports, the effect size was more than two times larger in sport reports (Biber and Egbert 2018: 91).

Largely influenced by the work of Biber and Egbert (2016, 2018), the present study combines several of these web registers in order to build a reference corpus representing popular written language as a whole. This has been done in order to increase coverage of the various types of online language that individuals have frequent exposure to.

2.3. Comparisons of legislative language and non-legal language

While the literature discussed in Section 2.1 has constituted a great contribution to our knowledge of legislative language, the prevailing gaps remain: 1) a focus on legislation independent from other written legal language, and 2) a lexico-grammatical description of statutory law in reference to other types of language. To the best of the researcher's knowledge, only two studies have made empirical lexico-grammatical comparisons of legislative writing and non-legal registers. Goźdź-Roszkowski's (2011) register variation study of legal language was undertaken with the goal of comparing a variety of legal registers to one another, including academic journals, briefs, contracts, legislation, opinions, professional articles, and textbooks. While the primary goal of Goźdź-Roszkowski's study was to examine lexico-grammatical variation between legal registers, he briefly compares the seven legal registers to select forms of non-legal language (i.e., fiction, textbooks, conversion, research articles, academic prose) through an additive multi-dimensional analysis on Biber's (1988) dimensions. In doing so, Goźdź-Roszkowski characterized legislation as comparatively informational, non-narrative, explicit (as opposed to situation-dependent), and lacking overt persuasion. Also of importance for the present study is the considerable amount of variation that Goźdź-Roszkowski found *between* legal registers, lending further support for the argument that for a clear and accurate description of a particular type of legal language, one must study it as a unique, independent register.

The other study that has undertaken a comparative lexico-grammatical analysis of legal and non-legal language was conducted by Özyildirim (2011), who investigated Turkish legislation in relation to other forms of non-legal language, including Turkish scientific research articles, newspaper articles, television commercials, men's/women's magazines, and stand-up comedy shows. Özyildirim (2011:78) made use of an additive multi-dimensional analysis on Biber's (1988) dimensions as Goźdź-Roszkowski did, but focused only on the narrative vs. non-narrative dimension, similarly characterizing legislation as highly non-narrative.

In some ways, this analysis follows in the footsteps of Goźdź-Roszkowski (2011) and Özyildirim (2011), though the present study differs both in methodology and research aims. First, both Goźdź-Roszkowski and Özyildirim made use of a multi-dimensional analysis for their register comparisons, which is used to characterize a number of individual registers in terms of the co-occurrence patterns of lexico-

grammatical features. In contrast, the analysis here focuses on identifying features that are markedly pervasive in one register relative to a combined reference corpus of other registers and does not concern feature co-occurrence. Finally, the selection of non-legal registers is targeted specifically for the purposes of investigating readability. While both Goźdź-Roszkowski and Özyildirim used a mixture of spoken and written registers as well as academic registers (i.e., textbooks and research articles, which are not considered 'popular' in the present study due to the restricted audience), this study makes use of a much more narrowly defined group of texts, specifically representing language that is both accessible and familiar to a lay audience.

3. Methodology

3.1. Corpora

The present study makes use of two corpora for analysis: a corpus of state statutory law, and a corpus representing other forms of popular written language. The following sections will describe the motivation for selection of the text varieties in the two corpora and the compilation processes.

3.1.1. Corpus of state statutory law

The corpus of state statutory law used for the present study was sampled from the larger *Corpus of United States State Statutes* (CorUSSS) (Egbert and Wood under review), which comprises the state codes for each of the 50 states in the United States. CorUSSS was compiled using a *Python* script to web-scrape texts located on https://www.justia.com. Statutes were initially scraped and aggregated at the top level by title, each of which contains a set of statutes representing specific topical content (e.g., Agriculture, Criminal Code, Businesses, Corporations). Text files were cleaned through a second *Python* script that removed all boiler-plate text and inserted brackets into the text files to denote meta-data, including the name of statute, year, and universal citation. A secondary cleaning process was undertaken through the regular expression program *Sublime Text* in order to remove extraneous boiler-plate text leftover following the initial cleaning process.[2]

[2] https://www.sublimetext.com/

To compile the corpus of statutory law used in the present study, a sample of eight state codes was selected from the 50 states. This smaller selection was made for logistical reasons, namely, any linguistic analysis on a corpus of such size (the totality of CorUSSS consists of over 420 million words and almost 8 million texts) would be challenging to conduct with existing corpus analysis tools. Additionally, because the compilation process of a corpus this large was fairly time-consuming, only a limited number of the state codes were available for use (web-scraped, cleaned, and tagged) at the time the present study was carried out. However, during the design and construction of CorUSSS, exploratory frequency counts of a variety of linguistic features in these states revealed very little variation between the codes from state to state, providing a high level of confidence that even if a complete corpus of all 50 state codes was used, there would not be substantial changes to the results. Still, in the selection process of state codes available at the time of the study, care was taken to select state codes that represented a variety of geographical regions in the United States in order to control for representativeness of the country as closely as possible. The final selection of states resulted in a corpus of state statutory law comprising 670 texts and 90,388,372 words. The final composition of the corpus is presented below in Table 1.

Codes	**Number of texts**	**Number of words**
Rhode Island	155	6,190,952
West Virginia	133	6,952,846
Kansas	85	5,795,347
Connecticut	72	7,798,889
New Jersey	68	10,855,203
South Dakota	68	4,210,208
South Carolina	63	5,993,304
Alaska	43	873,860
Total	**670**	**90,388,372**

Table 1: The statutory law texts

The texts were tagged for lexico-grammatical features using the *Biber Tagger*, which identifies a larger set of characteristics than other existing taggers (over 150 features) and is able to identify these features at a more fine-grained level, for example, the identification of the gap position for *wh*-relative clauses (Biber and Egbert 2018: 22). Staples *et al.* (2016) reported that the tagger tagged at 90 per cent accuracy for formal writing.

3.1.2. *Popular Written Language* corpus

The *Popular Written Language* corpus (PWL) used for the present study comprised a selection of web registers. This decision was made based on the criteria that language needed to be written for an audience of the general public, and easily accessible to that population. Because the Internet is highly accessible to the general public in the United States (whether personally or in public establishments) and reaches a wide audience, registers selected to represent popular written language were sampled from the *Corpus of Online Registers of English* (CORE). CORE is a corpus compiled by Biber and Egbert (2016) sampled from the larger *Corpus of Global Web-based English* (GloWbE; Davies 2013). The entirety of CORE holds 48,571 documents and nearly 54 million words (Biber and Egbert 2016: 14). Using CORE was also beneficial as Biber and Egbert (2016) had previously removed any texts from the sample that had fewer than 75 words, which is undesirable for studies of lexico-grammatical characteristics (Biber and Egbert 2018: 13).

Popular written registers were selected from CORE with the aim of keeping the PWL corpus as cohesive as possible in terms of situational characteristics. To be included in the PWL corpus, registers needed to be written by an author that has formal expertise or insider knowledge of the topic about which they are writing (i.e., news, sports, history, etc.). Registers were not selected for the corpus if they varied in mode (i.e., spoken language), were not written for a lay audience (academic research articles), did not represent real-world topical content, or were highly stylistically varied (i.e., fiction). To be included in the corpus for the present study, registers also needed to be originally in the written mode and be non-interactive (categorized as such by Biber and Egbert 2018).

Appendix 1 provides an overview of situational characteristics for all five registers used in the present study, demonstrating the relative similarity in most of their characteristics. The situational difference between these registers lies predominantly in topic, with a small range of variation in communicative purpose.

Extensive consideration was given to blogs, which were selected for the corpus in the early stages of the project due to the popularity of the text type. However, Biber and Egbert (2018) identified this text type as one that does not seem to clearly fit a register, as topic and blog type are highly variable. In the end, blogs were excluded from consideration with the exception of two types: sports blogs and news blogs. This

decision was made for two reasons. First, these two types of blogs are infrequently written by the lay person, but rather individuals with relatively specialized knowledge of the topic. Along with this, they infrequently concern personal experience, instead reporting on outside stories or occurrences. This is in contrast to other blog types identified by Biber and Egbert (2018), such as personal narrative blogs, travel blogs, and opinion blogs, all of which were excluded from the PWL corpus. Second, Biber and Egbert (2018: 42) chose to incorporate news blogs and sports blogs to their respective registers based on the finding that often these blogs were "virtually indistinguishable from published reports." The final composition of the PWL corpus is presented in Table 2.

Codes	**Number of texts**	**Number of words**
News	600	498,780
Sports reports	600	472,795
Encyclopedia articles	430	1,291,380
Historical articles	206	413,537
Total	**1,871**	**2,756,389**

Table 2: The *Popular Written Language* corpus

3.2. Linguistic analysis

3.2.1. Key feature analysis

To identify pervasive lexico-grammatical features in statutory law, the present study makes use of a key feature analysis. Key feature analysis makes use of a reference corpus in order to identify features that are markedly more frequent in a target corpus, which are considered 'key'.

Key feature analysis makes use of the mean rate of occurrence and standard deviations of linguistic features to calculate Cohen's *d* effect sizes (Cohen 1977). Large positive Cohen's *d* values indicate that the feature is markedly more frequent in the target corpus than in the reference corpus, while large negative Cohen's *d* values indicate that the feature is markedly less frequent. In accordance with Cohen (1977), *d* values will be interpreted as small (> +/- 0.20), medium (> +/- 0.50) and large (> +/- 0.80).

In the present study, features with large positive effect sizes in the corpus of statutory law are considered pervasive linguistic features of statutory language that the

lay population is expected to have less exposure to on a daily basis. Cohen's *d* values approaching zero are an indication of a similar frequency of use in the two corpora.

3.2.2. Feature selection

The lexico-grammatical features selected for analysis were generated through bottom-up means in order to remove the influence of personal intuition concerning the pervasiveness of certain features. The general tag count generated from the *Biber Tagge*r was used, which provides normed frequency counts per 1,000 words for over 150 lexico-grammatical features. A normed frequency count for one additional lexico-grammatical feature —the non-finite post-nominal *-ing* clause— was manually added.

Once normed frequency counts for all features were obtained, a dispersion threshold of 90 per cent was established, meaning that the feature had to appear in at least 90 per cent of the texts in either of the two corpora in order to be retained for analysis. This narrowed the list of features for analysis to 81 features. An additional 19 features were then eliminated from the analysis due to overlap. This included the removal of several 'all' features (e.g., 'all adjectives'), in favor of more specific types of that feature (e.g., 'predicative adjectives' and 'attributive adjectives'). Specific semantic domains were later removed if they included similar lexical items; for example, private verbs and mental verbs, which include words such as *think* and *believe.* In such cases, the semantic domain with the larger effect size (positive or negative) was retained for analysis. This resulted in a final list of 62 linguistic features, which are presented in Table 3.

Verbs	**Dependent clauses**
Present tense	Non-finite *-ing* clauses
Past tense	Non-finite *-ed* clauses
Perfect aspect	*To* complement clause controlled by verbs of modality, causation
Progressive aspect	*To* complement clauses controlled by stance nouns
Passive + *by*	*That* complement clause controlled by verbs
Passive post-nominal modifier	*That* relative clauses
Short passives	*Wh* relative clause, object position
Infinitive	*Wh* relative clause, subject position
Split auxiliary	*Wh* relative clause, prepositional fronting (pied-piping)
Be as main verb	
Have as main verb	**Other**
Modals of prediction	Stranded preposition
Modals of possibility	Prepositions
Mental verbs	Clausal coordinating conjunction
Communication verbs	Phrasal coordinating conjunction
Activity verbs	Subordinating conjunction – conditional
Suasive verbs	Subordinating adverbial - other
Aspectual verbs	Attributive adjectives
Verbs of likelihood	Predicative adjectives
Verbs of existence	Linking adverbials
Verbs of causation	1st person pronouns
Verbs of occurrence	3rd person pronouns
	Indefinite pronouns
Nouns	Pronoun *it*
Process nouns	Indefinite articles
Abstract nouns	Definite articles
Human nouns	Contractions
Place nouns	Topical adjectives
Technical nouns	Adverb of time
Cognitive nouns	Adverb of place
Quantity nouns	Downtoner
Group noun	Type/token ratio
Proper nouns	
Pre-modifying nouns	

Table 3: Features for keyword analysis

4. RESULTS

Results from the key feature analysis are presented in Figure 1 and Tables 4 and 5 below. Figure 1 shows an oral/literate divide between the PWL corpus and the corpus of statutory law which, as suggested by Biber (2014), is a universal dimension in multi-dimensional analysis studies. More specifically, popular written language displays the lexico-grammatical characteristics that are highly typical of narrative language (i.e., first and third-person pronouns, past tense, perfect aspect, progressive aspect, contractions, verbs), while statutory language can be characterized as highly detail-oriented, dense, and topically narrow (type/token ratio was key with a large effect size in the PWL, indicating high lexical diversity relative to statutory language).

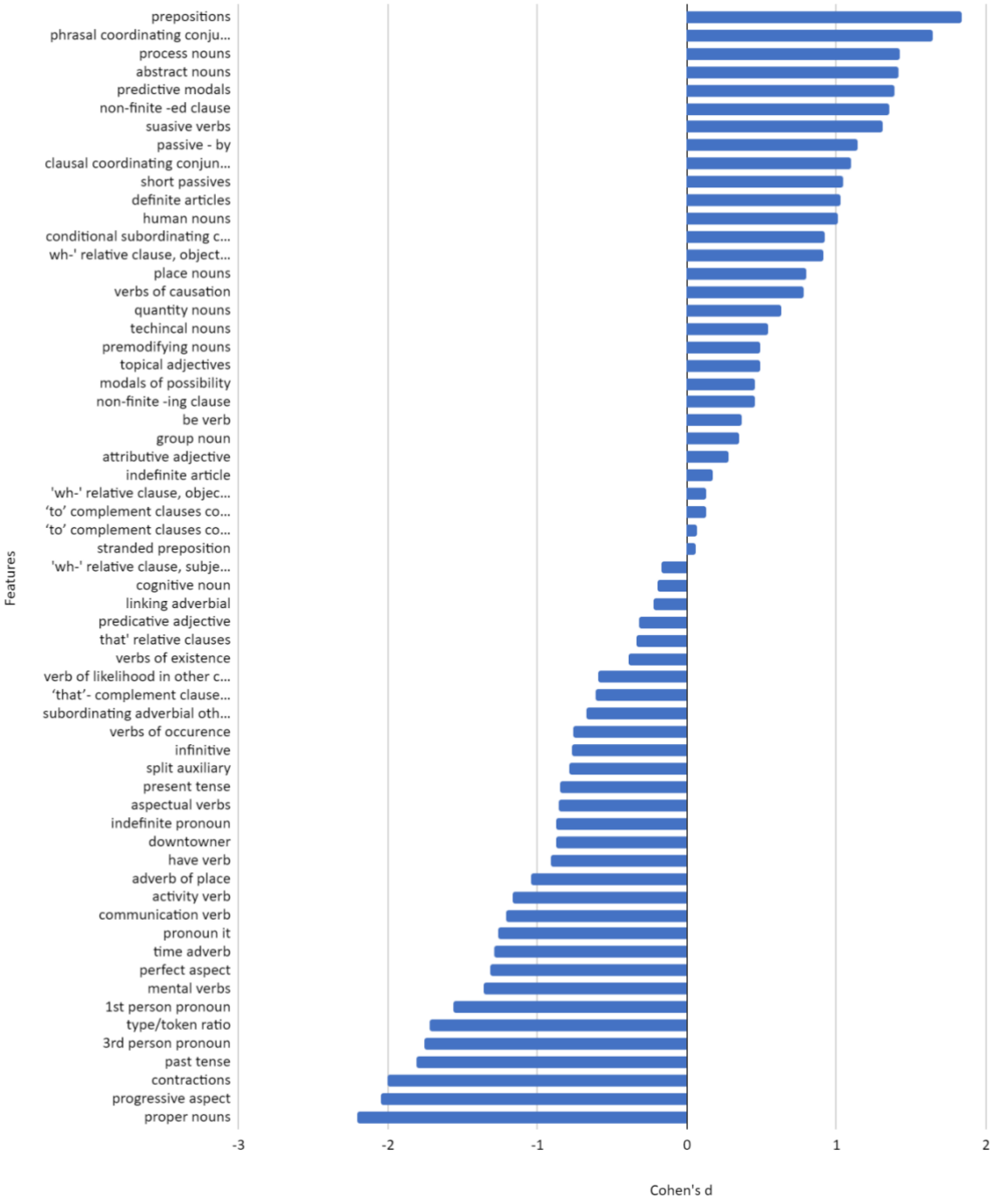


Figure 1: Key feature analysis results

Key features of statutory law indicate the pervasive use of both phrasal language (contributing to the dense, literate nature of statutes) and clausal language (contributing to the long-winded, detail-oriented nature of statutes). The corpus of statutory law has 15 features with large effect sizes over d=0.90. Of these, six features are typically associated with literate language (Table 4). Two passive constructions are key in statutory law with large effect sizes (*by* passives, d=1.14; short passives, d=1.04), as are

prepositions (with the highest keyness score in the corpus of *d*=1.84). Statutory law also demonstrated frequent use of phrasal coordinating conjunctions, which had the second largest effect size in the corpus (*d*=1.64). Nouns of several different semantic domains were also key in statutory law, including process nouns, abstract nouns, human nouns, quantity nouns, place nouns, and technical nouns (with medium to large effect sizes). Key clausal language in the corpus included the use of clausal coordinating conjunctions (*d*=1.10), conditional subordinating conjunctions (*d*=0.92), non-finite *-ed* clauses (passive post-nominal modifier) (*d*=1.35), *wh*-relative clauses with the pronoun in the object position and prepositional fronting (*d*=0.91), and non-finite *-ing* clauses (*d*=0.45). This mixture of both phrasal and clausal features is consistent with past research of legal language previously discussed in Section 2.1. The notable keyness of conjunctions —phrasal, clausal, and subordinating— has been tied to the characteristically long, drawn out sentences found in statutory language (Tiersma 1999).

Effect size	***d***	**Feature**
Large	1.84	prepositions
	1.64	phrasal coordinating conjunction
	1.42	process nouns
	1.41	abstract nouns
	1.39	predictive modals
	1.35	non-finite *-ed* clause
	1.31	suasive verbs
	1.14	*by* passive
	1.10	clausal coordinating conjunction
	1.04	short passives
	1.03	definite articles
	1.01	human nouns
	0.92	conditional subordinating conjunction
	0.91	*wh*- relative clause, object position with prepositional fronting (‘pied piping’)
Medium	0.80	place nouns
	0.78	verbs of causation
	0.63	quantity nouns
	0.54	technical nouns
Small	0.49	premodifying nouns
	0.49	topical adjectives
	0.45	modals of possibility
	0.45	non-finite *-ing* clause
	0.37	*be* verb
	0.35	group noun
	0.28	attributive adjective
	0.17	indefinite article
	0.13	*wh*- relative clause, object position
	0.13	*to* complement clauses controlled by verbs of modality, causation, and effort
	0.07	*to* complement clauses controlled by stance noun
	0.06	stranded preposition

Table 4: Key features for statuary law

The negative effect sizes indicate markedly less frequent use of verb-associated features in the statutory law corpus in comparison to the PWL corpus. Features with medium to large negative effect sizes include: progressive aspect (d = -2.05), past tense (d = -1.81), perfect aspect (d = -1.32), time and place adverbs (d = -1.29; d = -1.04), present tense (d = -0.85), split auxiliaries (d = -0.79), the infinitive (d = -0.77), and a variety of semantic domains of verbs (mental verbs, communication verbs, activity verbs, aspectual verbs, verbs of likelihood, verbs of existence). Typically, narrative features with large effect sizes also included a variety of pronouns (first person, third person, pronoun *it*) and contractions. Proper nouns had the largest effect size in the PWL corpus, which is unsurprising based on the selection of registers in the present study, which can concern an unlimited number of different people and places.[3] Various clausal features also had a large effect size in the PWL corpus, including *that* complement clause controlled by verbs, *that* relative clauses, and the *wh-* relative clauses with the pronoun in the subject position (though the latter three features had small effect sizes).

Notably, features that appeared in less than 50 per cent of the texts in the PWL corpus (and over 90 % in the statutory law corpus) and were key in statutory law with medium to large effect sizes included *wh-* relative clauses with the pronoun in the object position, *wh-* 'pied-piping' relative clauses, and suasive verbs (e.g., *ask*, *command*, *insist*). The low dispersion of these features across the PWL coupled with the high keyness in statutory law makes these highly important features to consider in the discussion of readability.

[3] Also recall that, in Biber and Egbert's (2018) key feature study, proper nouns had a very high effect size in both news reports and sports reports.

Effect size	*d*	Feature
Large	-2.21	proper nouns
	-2.05	progressive aspect
	-2.00	contractions
	-1.81	past tense
	-1.76	3rd person pronoun
	-1.56	1st person pronoun
	-1.36	mental verbs
	-1.32	perfect aspect
	-1.29	time adverb
	-1.26	pronoun *it*
	-1.21	communication verb
	-1.17	activity verb
	-1.04	adverb of place
	-0.91	*have* verb
	-0.88	indefinite pronoun
	-0.88	downtowner
	-0.86	aspectual verbs
	-0.85	present tense
Medium	-0.79	split auxiliary
	-0.77	infinitive
	-0.76	verbs of occurrence
	-0.67	subordinating adverbial other
	-0.61	*that* complement clause controlled by verb
	-0.59	verb of likelihood in other contexts
Small	-0.39	verbs of existence
	-0.34	*that* relative clauses
	-0.32	predicative adjective
	-0.22	linking adverbial
	-0.20	cognitive noun
	-0.17	*wh-* relative clause, subject position

Table 5: Key features for popular written language

5. DISCUSSION

The literate nature of statutory language is seen in part in the variety of semantic domains of nouns that are key in the statutory law corpus, corroborating the findings by Williams (2013), who noted the nominal nature of legislative texts. While there seems to be a large number of key semantic domains of nouns for a register with content that is far more restricted than that of popular language (which has great freedom in topic), the semantic domains represented are clearly oriented towards topics typically discussed in law. These domains include process nouns (e.g., *system*, *meeting*; *d*=1.42), abstract nouns (e.g., *agreement*; *d*=1.41), human nouns (e.g., *person*, *governor*; *d*=1.01), place nouns (e.g., *town, city*; *d*=0.80), and technical nouns (e.g., *jurisdiction*; *d*=0.54*).* Statutes typically contain descriptions of the people, settings, and contexts in which a law takes effect, meaning that the semantic domains named above complement one

another well. Excerpt 1, below, demonstrates the use of a variety of nouns from different semantic domains (in bold) working together to describe people, setting, and subject matter in a highly specific context. In particular, this excerpt uses a large number of abstract nouns, such as *discretion* and *compliance*.

(1) **Excerpt 1**: The **director** shall have **discretion** to assess an administrative **penalty** of not more than two hundred fifty dollars ($250) per **offense** against any **insurance company** that fails to notify the **director** as required in this section. The **director**, in his or her **discretion**, may bring a **civil action** to collect all assessed civil **penalties**. The workers' **compensation court** shall have **jurisdiction** to enforce **compliance** with any order of the **director** made pursuant to this **section**. (R.I. § 28-36-12).

In contrast, popular written language makes a more frequent use of various verb-associated features, which is characteristic of oral and narrative language. Note that while the topical content of the historical article below concerns matters of law (see excerpt 2), the narrative, story-telling aspect of the text is reflected in the use of past tense, perfect aspect, and proper nouns (underlined), in particular, when compared to excerpt 1. While excerpt 2 narrates a historical event, excerpt 3 appears to narrate an individuals' personal thoughts, making use of both present tense and past tense, and the perfect and progressive aspects. The variety of semantic domains of verbs which are key in written popular language is also notable, including mental verbs (*think*) and activity verbs (*spend*, *move*).

(2) **Excerpt 2**: Historical Article. The hearings **had run** for eleven days. The hearing three years earlier to confirm Fortas as associate justice **had run** for three hours. At the beginning of October, Fortas's nomination **went** to the full Senate for a vote. For four days straight, senators **defended** or **lambasted** Fortas until a cloture petition to end the debate **was introduced**. (https://www.neh.gov/humanities/2009/septemberoctober/feature/supremely-contentious)

(3) **Excerpt 3**: Sports Report. The NHL should **step in** and **cough up** a few $$$. **I think** the NFL **helps** out with new stadiums. Bettman **has spent** millions **to keep** a money losing franchise in PHX. He **could spend** a few more **to keep** a money making one in EDM. Even if the league **approved** relocation for the Oilers, there **would be** 8 teams in line to **move** to Edmonton. This **has** nothing to do with the city and everything to do with Katz not **wanting to spend** any of his $200 Billion. (http://ca.sports.yahoo.com/blogs/nhl-puck-daddy/oilers-talking-relocation-seattle-playing-arena-deal-hardball-012114557--nhl.html)

While the preference for verbs is associated with oral and narrative varieties and the dense use of nouns is associated with statutory language, the key feature analysis

showed exceptions to this based on semantic domain. There are two semantic domains of verbs that are key in statutory language with relatively large effect sizes: suasive verbs (e.g., *ask*, *command*, *insist*; *d*=1.31) and verbs of causation (e.g., *let*, *permit*; *d*=0.78). These two domains of verbs serve highly specific purposes in statutory language: suasive verbs mandating or giving direction (or excusing from responsibility) (excerpt 4) and verbs of causation giving permission to act (excerpt 5).

(4) **Excerpt 4**: Zoo animals loaned pursuant to this section are not deemed to be surplus property, and **no** motion **is required** to enter into an agreement for the loaning of zoo animals. (S.D. § 6-13-16).

(5) **Excerpt 5**: If the tax collector fails to respond at any step in the process under this section within the prescribed period of time, then the governing body **shall be permitted** to remove the tax collector from office as provided in paragraph V. (N.H. § 41:40).

Excerpt 5 also demonstrates the use of modal *shall* (common in legislation) and the passive voice, the latter of which is another characteristic feature of literate varieties such as formal academic writing. The passive voice has historically been given lots of attention in the conversation surrounding readability of texts and is frequently targeted in text simplification. Two forms of the passive voice are key with large effect sizes in statutory law (*by* passives and short passives), corroborating past findings by Williams (2004) and Bulatović (2013). While short passives are typically favored when the agent is unknown, as is common in academic writing (Biber *et al.* 2002: 168), they seem instead to be favored in statutory law for the purpose of inclusiveness. In many cases, leaving out the agent necessarily implies ‘everybody’ or ‘anybody’ who commits an act, which, importantly, makes it clear that all citizens of that state are subject to that particular law, and the legal consequences should they not act in accordance with it. On the other hand, the passive + *by* construction is used in statutory language for the opposite purpose: to indicate exactly who has the authority or power to act. Note the passive constructions in excerpts 6 and 7, which are used in two different ways: 1) to indicate that an action applies to everyone, and 2) to give a person or entity authority.

(6) **Excerpt 6**: The official flag of the state shall **be displayed** with the flag of the United States only from sunrise to sunset, or between the hours **designated by** proper authority. However, the flag may **be displayed** after sunset upon special occasions when it is desired to produce a patriotic effect. (A.K. § 44.09.030).

(7) **Excerpt 7**: If the date of the special election **conducted** pursuant to § 12-11-1.1 requires that absentee ballots **cast by** absent uniformed services voters or overseas voters arriving after election day be counted as **required by** 2 USC Chapter 1 § 8 as of January 1, 2008, these absentee ballots shall **be processed and counted by** the provisional ballot counting board. (S.D. § 12-11-2.1).

Results of the key feature analysis show that statutory language exhibits the use of both clausal and phrasal features, confirming the findings by Goźdź-Roszkowski (2011). Notable phrasal features that are key in statutory law include prepositions and phrasal conjunctions, which serve the purpose of providing as much detail as possible in the description of the person, context, or situation in which a law applies. Prepositions, which signal embedded prepositional phrases and phrasal verbs, have the largest effect size of any feature in the corpus of statutory law (d=1.84). In statutory language, they function predominantly to provide qualifying details in order to narrow the identity or scope of the noun that they modify. The use of this contributes to the dense packaging of referential information, as they are more compact than clausal postmodifiers (Biber *et al.* 1999: 607). Excerpt 8, below, comprises a single sentence with ten prepositions (in bold), which together function to provide an operational definition of a term. One of these prepositions belongs to a single complex prepositional phrase (*with respect to*), one is a part of a prepositional verb (*deal with*), and six prepositions head a prepositional phrase. These prepositional phrases come in a variety of forms, including genitive/postmodifying (e.g., *law [of this state]*), and adverbial (e.g., *property [within this state]*).

(8) **Excerpt 8**: (a) "Charitable trust" means any fiduciary relationship **with** respect **to** property arising under the law [**of** this state] or [**of** another jurisdiction] as a result [**of** a manifestation] [**of** intention] to create it and subjecting the person by whom the property is held to fiduciary duties to deal **with** the property [**within** this state] for any charitable, nonprofit, educational, or community purpose. (N.H. § 7:21).

Phrasal embedding also appears in the form of phrasal coordinating conjunctions, which, along with the use of causal coordinating conjunctions, contribute to the long, drawn-out sentences that are packed with information and tend to make sentences hard to follow (Tiersma 1999). Phrasal and clausal coordinating conjunctions are both key in the corpus of statutory law with large effect sizes over 1.0 (d=1.64; d=1.10). Phrasal coordinators in statutory language are most often used to directly identify a highly specific list of individuals, entities, or objects that the statute applies to. See, for instance, excerpt 9, which lists a set of qualifying items (*labor*, *material*, or *rental*

equipment), that must be furnished by the person in order for the statute to apply to them. The length of excerpt 9, which contains a total of ten phrasal and clausal conjunctions, is attributed to the thorough description of the circumstances under which an individual has the right to sue. This results in the subject (a person) being separated from the corresponding verb phrase *has the right to*, by a string of embedded clauses and phrases, including seven phrasal and clausal conjunctions. This format is not uncommon, as describing the characteristics of the subject that a statute pertains to, or the context in which the statute takes effect, is an important characteristic of statutory language.

> (9) **Excerpt 9**: (c) **A person** *who has furnished labor,* material, or rental equipment to a bonded contractor or his subcontractors for the work specified in the contract, and *who has not been paid in full* for it before the expiration of a period of ninety days after the day on which the last of the labor was done or performed by the person or material or rental equipment was furnished or supplied by the person for which the claim is made, **has the right to sue** on the payment bond for the amount, or the balance of it, unpaid at the time of institution of the suit and to prosecute the action for the sum or sums justly due the person. (S.C. § 11-35-3030).

The use of multiple phrasal and clausal coordinators in quick succession to one another can result in confusion, as it can be easy to mistake one type of conjunction for the other. For example, in excerpt 9, if one reads: *the last of the labor was done or performed by the person or material or rental equipment (...)*, it is easy to mistake the second clausal conjunction (*person or material…*) for a phrasal conjunction. This is resolved semantically in the clause, but is still challenging to process in real time as the reader looks for a conclusion to the long sentence in the form of a phrasal conjunction, and is instead introduced to yet another clause.

Conditional subordinating conjunctions (e.g., *if, unless*) are frequently associated with statutory language for the same reasons mentioned above: they contribute to the specification of the conditions under which authorizations, mandates, or prohibitions take effect, or do not take effect. Excerpt 10, below, includes five conditional statements in a list format, each moving further from the initial clause that the conditional statement is dependent upon for meaning. Because of this, and because conditional subordinating clauses have flexibility in their syntactic position (i.e., beginning, medial, final), the conditional statement can start to read as though it occupies the beginning syntactic position. This is particularly problematic for readability when the sentence

potentially reads more smoothly with the conditional statement in a different syntactic position from which it appears.

(10) **Excerpt 10**: (j) Upon conviction by a court of a person of an offense described in (a)(7) of this section, the department shall disqualify that person from driving a commercial motor vehicle for the following periods:

1. **if the person has not been previously convicted of violating** an out-of-service order, not less than 180 days;
2. **if the person has been previously convicted once** of violating an out-of-service order, not less than two years;
3. **if the person has been previously convicted more than once** of violating an out-of-service order, not less than three years;
4. **if the person operates a commercial motor vehicle transporting hazardous materials** or a vehicle designed to transport 16 or more passengers, including the driver, in violation of an out-of-service order, not less than 180 days;
5. **if the person has been previously convicted of operating a commercial motor vehicle transporting hazardous materials** or a vehicle designed to transport 16 or more passengers, including the driver, in violation of an out-of-service order two or more times in separate incidents within a 10-year period, not less than three years. (A.K. § 28.33.140).

While highly clausal language is frequently associated with decreased readability of statutes, the key feature analysis in the present study revealed that the preference for the type of clause may be what distinguishes statutory language from other forms of popular language. In particular, there was a difference in the distribution of finite and non-finite causal constructions: statutory language uses markedly more non-finite clauses relative to popular written language, and markedly fewer finite clauses. All key clausal constructions in the PWL were finite, including *that* relative clauses, *that* complement clauses and *wh-* subject position relative clauses, and nearly all key clausal constructions in the statutory law corpus were non-finite, including post-nominal *-ing* and *-ed* clauses, and two types of *to-* complement clauses. The exception to this pattern was the *wh*-relative clause with the pronoun in the object position (both with prepositional fronting and without) which appeared alongside the non-finite constructions in statutory law. This distribution should be interpreted with caution, however, as only five of the nine clausal constructions meet the Cohen's *d* threshold for 'key' (>-/+0.20). The distribution of non-finite and finite clauses is represented in Figure 2.

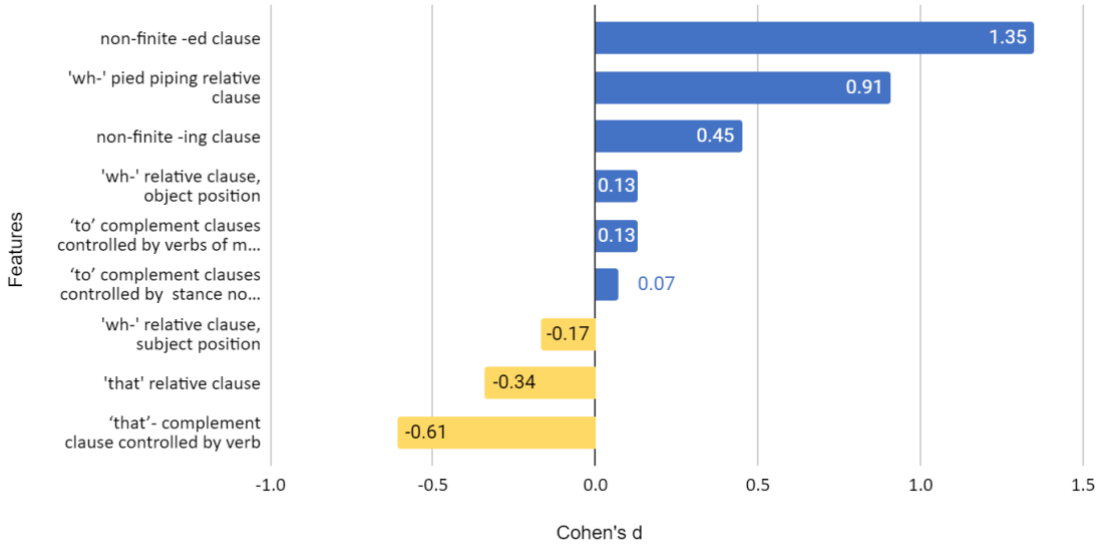


Figure 2: Distribution of finite and non-finite clause constructions

Non-finite clauses, which are favored in statutory law, are more compact and less explicit than finite clauses and often lack an explicit subject or subordinator (Biber *et al.* 1999: 198). This lack of explicitness is directly related to the condensed nature of statutes, as the drafters attempt to pack as much information as possible into a small space. The condensed nature of the non-finite *-ing* and *-ed* clauses can be seen in excerpts 11 and 12.

(11) **Excerpt 11**: A mutual bank may, with the approval of the department, establish and operate branches inside the state. Before **approving the establishmen**t and operation of a branch xoffice, the department shall make the findings required before the granting of a charter to a mutual bank with respect to the **branch proposed**. (A.K. § 06.15.290).

(12) **Excerpt 12**: The governing body, within sixty days after the filing of any such delinquent list, shall examine such list and, on **being** satisfied that any of the taxes **so listed** are not collectible, it shall, by resolution, release the collector from the collection thereof and order **the same canceled.** (N.J. § 54:4-91.2).

In contrast, finite clauses such as *that* complement clauses controlled by verbs and *that* relative clauses were markedly less common in the statutory law corpus. *That* complement clauses controlled by verbs, which have a medium effect size of d = -0.61 in the PWL corpus, are often used in reported or quoted speech, which is highly common in texts narrating past events or recalling conversations, as shown in excerpt 13.

(13) **Excerpt 13**: Encyclopedia article. Speaking about Niall Horan, who we made his boyfriend when we created Nithan Syran, Nat Han said: "I love Niall...Can I just **say that** Niall is one of the nicest lads you'll ever. (http://www.sugarscape.com/main-topics/celebrities/784664/exclusive-nathan-sykes-wanted-and-one-direction-being-lovers)

That relative clauses (key in the PWL corpus with a small effect size of d = -0.34) are used in the post-modification of a noun phrase in either the restrictive form (to establish a reference) or non-restrictive form (to provide additional information about the antecedent, not required for identification). These two forms can be seen in the encyclopedia excerpt (14) below.

(14) **Excerpt 14**: Encyclopedia article. While in common parlance anything **that** attempts to provide an explanation for a cause can be dubbed a "theory", a scientific theory has a much more specific meaning. Scientific theory is far more than just a casual conjecture or some Joe's guesswork. A theory in this context is a well-substantiated explanation for a series of facts and observations **that** is testable and can be used to predict future observations. (*Scientific Theory*: https://rationalwiki.org)

While these functions are also important in statutory language, statutes appear to favor *wh-* relativizers. This is consistent with the findings of Biber *et al.* (1999: 611), who found a preference for *wh-* relativizers in formal academic prose over *that* relativizers (twice the frequency of occurrence).

The notable exception to the finite/non-finite split between the two registers is the pied-piping *wh-* relative clause construction (d=0.91). In this construction, there is a preposition located at the beginning of the clause preceding the relative object pronoun, resulting in classic formal phrases such as *to whom*, *for which*, and *in which*. The prepositional fronting does not serve any immediate, unique function in statutory law, but is instead considered stylistic and highly characteristic of statutory language. In excerpt 15, below, a single sentence holds four instances of this construction. Excerpt 16 also makes use of four pied-piping constructions embedded within an even longer sentence, which has been truncated in order to conserve space. It should also be noted that the pied-piping construction is frequently passive, a characteristic again shared with academic prose as found by Biber *et al.* (1999), who noted that object position relative clauses in particular are frequently found alongside the passive voice.

(15) **Excerpt 15**: A remote claimant has a right of action on the payment bond only upon giving written notice to the contractor within ninety days from the date **on which the person did or performed the last of the labor** or furnished or supplied the last of the material or rental equipment **upon which the claim is made**, stating with substantial accuracy the amount claimed as unpaid and the name of the party **to whom the material or rental equipment was furnished or supplied** or **for whom the labor was done or performed**. (S.C. § 11-35-3030).

(16) **Excerpt 16**: When any owner, tenant or subtenant of a lot or lots or tract of land shall file in any court of competent jurisdiction within the county **in which said lot or lots or tract of land may be situated**, his or her affidavit, or the affidavit of any other creditable person for them, stating that from knowledge, information or belief the party or parties owning, controlling or working the adjoining lot or lots or tract of land, and **upon which said party or parties are sinking shafts**, mining, excavating and running drifts, and that said drifts, **in which said parties are digging, mining and excavating** any mineral ore or veins of coal, extend beyond the lines and boundaries of said lot or lots or tract of land owned, controlled or worked by them, and have entered into and upon the premises of the party or parties making said affidavit, or **for whom said affidavit is made**, the judge of such court shall issue his or her written order [...]. (K.S. § 49-109).

An important finding of the present study is that clausal embedding as a whole cannot necessarily be considered highly characteristic of statutory language relative to other forms of written language. This is based on two findings: 1) different types of clausal constructions appeared key in both corpora, and 2) several constructions had effect sizes approaching zero, indicating similar use in statutory language and popular written language. The latter finding is demonstrated in the two excerpts below, which demonstrate similar use of a variety of clausal features in statutes and written popular language.

(17) **Excerpt 17**: *Wh-* relative clause, subject position (*who*); *Wh-* relative clause, object position (*which*)
(c) Any member **who** is aggrieved by a denial of benefits to be provided under this section may appeal the denial in accordance with regulations of the department of health, **which** have been promulgated pursuant to chapter 17.12 of title 23. (R.I. § 27-30-1).

(18) **Excerpt 18**: Encyclopedia article. *Wh-* relative clause, subject position.
The remainder of your companions in the following order of priority, minus whoever is already included in your active party and those **who** have sided against you before this point [...]. (https://dragonage.fandom.com/wiki/The_Last_Straw)

This suggests that the *type* of clausal construction may matter quite a bit for readability. For this reason, it seems that the discussion surrounding clausal constructions that are particularly problematic should focus more narrowly on constructions that are markedly less common in popular written language, and particularly characteristic of statutory language, such as the *wh-* pied-piping construction and the condensed non-finite *-ed* and *-ing* clauses.

6. Conclusions and future research

This study has provided a large-scale, detailed description of what the register of statutory language looks like and, in particular, how it differs from language that the lay person is exposed to on an everyday basis. It is important that we continue to make these comparative analyses when we attempt to describe statutory language so that we understand not just how frequently a feature appears in register, but how *characteristic* it is of that register. For example, Hiltunen (2012) reported that around a quarter of subordinating clauses in legislation are adverbial, and while this seems like a large proportion, it may not paint the full picture of the use of this feature. In the present study, the feature 'other adverbial subordinating clauses' actually had a medium effect size in the PWL corpus (see Table 5), meaning that it is markedly *less* frequent in statutes compared to other forms of popular written language.

This study has also showed us that we need to be looking at clausal embedding at a more fine-grained level as opposed to making blanket statements about the challenges that it poses for readability. The present study has demonstrated that several types of finite clauses, for example, are in fact key in the PWL corpus, or not key at all (exceedingly small effect sizes, under +/- .20).

Future research of this kind would benefit from a more detailed analysis of clausal embedding, with a specific focus on adverbial clauses and centrally-embedded clauses, which both Charrow and Charrow (1979) and Bhatia (1993) argue are highly characteristic of legislative language and problematic for readability. Future research may also expand on this information to examine readability from the reader's perspective. It is hoped that this study may provide a constructive path forward in addressing lack of readability in legislative texts, both by demonstrating the use of

empirical methods to identify differences between statutory and popular language, and identifying features that may be less familiar to the lay person.

REFERENCES

Alasmary, Abdullah. 2019. Lexical bundles in contract law texts: A corpus-based exploration and implications for legal education. *International Journal of English Linguistics* 9/2: 244–257.

Andersson, Dan. 2007. *Deontic Modal Verbs in EU Legislation: A Comparative Study of Documents in Four Germanic Languages*. Stockholm: University of Stockholm dissertation.

Bednarek, Monika and Helen Caple. 2014. Why do news values matter? Towards a new methodological framework for analysing news discourse in Critical Discourse Analysis and beyond. *Discourse & Society* 25/2: 135–158.

Bhatia, Vijay Kumar. 1983. Simplification v. easification: The case of legal texts. *Applied Linguistics* 4/1: 42–54.

Bhatia, Vijay Kumar. 1993. *Analysing Genre: Language Use in Professional Settings*. London: Longman.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Language in Contrast* 14/1: 7–34.

Biber, Douglas and Susan Conrad. 2009. *Register, Genre, and Style.* Cambridge: Cambridge University Press.

Biber, Douglas and Jesse Egbert. 2016. Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics* 44/2: 95–137.

Biber, Douglas and Jesse Egbert. 2018. *Register Variation Online*. Cambridge: Cambridge University Press.

Biber, Douglas and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing.* Cambridge: Cambridge University Press.

Biber, Douglas and Bethany Gray. 2019. Are law reports an 'agile' or an 'uptight' register? Tracking patterns of historical change in the use of colloquial and complexity features. In Teresa Fanego and Paula Rodríguez-Puente eds. *Corpus-based Research on Variation in English Legal Discourse*. Amsterdam: John Benjamins, 147–170.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd and Marie Helt. 2002. Speaking and writing in the university: A multidimensional comparison. *Tesol Quarterly* 36/1: 9–48.

Biber, Douglas, Stig Johansson, Geofrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman Grammar of Spoken and Written English*. London: Longman.

Biel, Lucja. 2009. Corpus-based studies of legal language for translation purposes: Methodological and practical potential. In Carmen Heine and Jan Engberg eds. *Reconceptualizing LSP. Proceedings of the XVII European LSP Symposium* Aarhus: Aarhus University, 1–15.

Biel, Lucja. 2014. Phraseology in legal translation: A corpus-based analysis of textual mapping in EU law. In Le Cheng, King Kui Sin and Anne Wagner eds. *The Ashgate Handbook of Legal Translation*. London: Routledge, 178–192.

Biel, Lucja. 2017. Lexical bundles in EU law: The impact of translation process on the patterning of legal language. In Stanislwaw Goźdź-Roszkowski and Gianluca Pontrandolfo eds. *Phraseology in Legal and Institutional Settings: A Corpus-Based Interdisciplinary Perspective.* London: Routledge, 10–26.

Breeze, Ruth. 2013. Lexical bundles across four legal genres. *International Journal of Corpus Linguistics* 18/2: 229–253.

Bulatović, Vesna. 2013. Legal language: The passive voice myth. *ESP Today* 1/1: 93–112.

Caliendo, Giuditta, Gabriella Di Martino and Marco Venuti. 2005. Language and discourse features of EU secondary legislation. In Anna Duszak and Guiseppina Cortese eds. *Identity, Community, Discourse: English in Intercultural Settings*. Bern: Peter Lang, 381–404.

Charrow, Robert P. and Veda Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review* 79: 1306–1374.

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Davies, Mark. 2012. *Oppositions and Ideology in News Discourse*. London: Bloomsbury Academic Press.

Davies, Mark. 2013. *Corpus of Global Web-Based English*. https://corpus.byu.edu/glowbe/

Egbert, Jesse and Margaret Wood. (Under review). *Constructing and Designing a Specialized Corpus of Statutory Law* (CorUSSS).

Foley, Roger. 2002. Legislative language in the EU: The Crucible. *International Journal for the Semiotics of Law* 15/4: 361–374.

Fowler, Roger. 2013. *Language in the News: Discourse and Ideology in the Press*. London: Routledge.

Gibová, Klaudia. 2011. On modality in EU institutional-legal documents. In Alena Kačmárová eds. *English Matters II: A Collection of Papers by the Institute of British and American Studies Faculty*. Prešov: University of Prešov, 6–12.

Goźdź-Roszkowski, Stanislaw. 2011. *Patterns of Linguistic Variation in American Legal English: A Corpus-based Study*. Bern: Peter Lang.

Hiltunen, Risto. 2012. The grammar and structure of legal texts. In Lawrence M. Solan and Peter M. Tiersma eds. *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press, 39–51.

Jablonkai, Réka. 2010. English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes* 29/4: 253–267.

Özyildirim, Işil. 2011. A comparative register perspective on Turkish legislative language. *Law Review* 1/1: 79–94.

Pontrandolfo, Gianluca. 2015. Investigating judicial phraseology with COSPE: A contrastive corpus-based study. In Claudio Fantinuoli and Federico Zanettin eds. *New Directions in Corpus-based Translation Studies*. Berlin: Language Sciences Press, 137–160.

Rodríguez-Puente, Paula. 2019. Interpersonality in legal written discourse: A diachronic analysis of personal pronouns in law reports, 1535-present. In Teresa Fanego and Paula Rodríguez-Puente eds. *Corpus-based Research on Variation in English Legal Discourse*. Amsterdam: John Benjamins, 171–199.

Scollon, Ron. 2014. *Mediated Discourse as Social Interaction: A Study of News Discourse*. London: Routledge.

Seracini, Francesca. 2020. *The Translation of European Union Legislation: A Corpus-based Study of Norms and Modality*. Milan: LED Edizioni Universitarie.

Staples, Shelley, Jesse Egbert, Douglas Biber and Bethany Gray. 2016. Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication* 33: 149–183.

Sun, Yuxiu and Le Cheng. 2017. Linguistic variation and legal representation in legislative discourse: A corpus-based multi-dimensional study. *International Journal of Legal Discourse* 2/2: 315–339.

Tapia, Ana M. Gates and Douglas Biber. 2014. Lexico-grammatical stance in Spanish news reportage: Socio-political influences on *que*-complement clauses and adverbials in Ecuadorian broadsheets. *Spanish Journal of Applied Linguistics* 27/1: 208–237.

Tiersma, Peter M. 1999. *Legal Language*. Chicago: The University of Chicago Press.

Trebits, Anna. 2009. The most frequent phrasal verbs in English language EU documents: A corpus-based analysis and its implications. *System* 37/3: 470–481.

Williams, Christopher. 2004. Legal English and plain language: An introduction. *ESP across Cultures* 1/1: 111–124.

Williams, Christopher. 2007. *Tradition and Change in Legal English: Verbal Constructions in Prescriptive Texts*. Bern: Peter Lang.

Williams, Christopher. 2013. Changes in the verb phrase in legislative language in English. In Bas Aarts, Jo Close and Sean Wallis eds. *The Verb phrase in English: Investigating Recent Language Change with Corpora*, 353–371.

Xie, Qin. 2018. Critical discourse analysis of news discourse. *Theory and Practice in Language Studies* 8/4: 399–403.

Yana, Dewi. 2015. The lexico grammatical features of the political register analysis in the editorial of the Jakarta Post Newspaper. *ANGLO-SAXON* 6/2: 15–23.

Corresponding author
Margaret Wood
Northern Arizona University
Department of English
705 S Beaver St
Flagstaff, AZ 86001
United States
E-mail: mkw57@nau.edu

received: December 2021
accepted: June 2022

APPENDIX 1: Statutory law and popular written language varieties situational characteristics

Register	Participants	Relationship among participants	Production circumstances	Setting	Purposes	Topic
State Codes	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Unequal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit	Varied (divorce, wills, personal injury, welfare, crime, real estate)
News Reports	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied (politics, economy, entertainment, business, health)
Sports Reports	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Sports
Encyclopedia Articles	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied
Historical Articles	*Addressor:* Individual or group knowledgeable in area *Addressee:* General public	Non-interactive, Impersonal, Equal power relationship	Planned Revised Edited	Contemporary Public Not face-to-face	Inform Exposit Narrate	Varied

RiCL Research in Corpus Linguistics

A corpus-based study of reporting verbs in academic Portuguese

Marine Laísa Matte – Elisa Marchioro Stumpf
Federal University of Río Grande do Sul / Brazil

Abstract – Referring to other sources is a cornerstone in academic writing and one way of framing someone else's ideas is through reporting verbs. There is little research on this phenomenon in academic Portuguese. Most of these studies analyze reporting practices without focusing on linguistic aspects (Bessa 2011; Hoffnagel 2010), with few studies on reporting verbs (Souza and Mendes 2012). The aim of this paper is to analyze how reporting verbs are used in the *Corpus of Portuguese for Academic Purposes* (CoPEP; Kuhn and Ferreira 2020), a corpus of research articles in Brazilian and European Portuguese. CoPEP was divided into two subcorpora: one with texts related to Hard Science (engineering, exact-earth science, and health science), and another with texts related to Soft Science (applied social science and humanities). *Sketch Engine* (Kilgarriff *et al.* 2014) was used to extract the verbs that are used before and after the lemma *autor* 'author'. Results indicate that texts in Hard Science have a slightly higher frequency of reporting verbs than texts in Soft Science, but both rely on similar reporting verbs to cite the voice of others. There is preference for the present tense in comparison with past and future, for the active voice in detriment of the passive voice, and for the order 'author + verb'.

Keywords – reporting verbs; academic Portuguese; citation practices; disciplinary variation

1. INTRODUCTION

In the past two decades, institutions of higher education in Brazil have witnessed a growth in the number of new campuses, courses, and students. This is largely due to public policies such as the *Program for the Restructuring and Expansion of Federal Universities* (REUNI) and affirmative action programs, started by *Law of Social Quota* in 2012. Portugal, in turn, has also registered an increase in the number of international student enrollment in the last 15 years (Oliveira *et al.* 2015). Moreover, there has been an influx of students who may not be used to features of academic discourse: non-traditional entrants in higher education in Brazil and an increasing number of foreign students in Portugal. This calls attention to the need for research into Portuguese for academic purposes, so as to help students face the demands of coursework at university.

In spite of recent severe budget cuts in science and technology,[1] Brazil is one of the top producers of scientific knowledge in Latin America (Kowaltowski *et al.* 2021) with twelve percent of its researchers publishing articles in Portuguese,[2] together with 3 percent from Portugal (Hernández Bonilla 2021).

The language of publication seems not to be an either-or matter, meaning that scholars may choose to publish in more than one language, considering different purposes, genres, and audiences (Pérez-Llantada 2021). Besides, there is a strong link between languages of publication and disciplinary areas, with scholars from harder and health science speakers of Portuguese as L1 tending to publish their work mostly in English. In a study on language choice in scholarly publication, Solovova *et al.* (2018) analyzed the choice between English and Portuguese in articles from three disciplines (linguistics, information and library sciences, and pharmacology and pharmacy) written between 1998 and 2017. The authors state that

> a comparison between Portuguese-written and English-written articles during a 20-year period divided in two decades (1998–2007 and 2008–2017) shows a rise in *both* languages within the Social Sciences and Humanities. Overall figures are substantially higher in English, but *relative* figures indicate the comparatively higher rise in Portuguese articles in the second decade (Solovova *et al.* 2018: 12, authors' italics).

Despite the budget cuts, there is a body of research being published in Portuguese, this meaning that both students and researchers need support in their publishing endeavors. However, research on academic Portuguese is still scarce, corroborating Kuhn's (2017) argument that not only Portuguese is less researched when compared to English and other languages, but also that most of the research tends to focus on text and discourse features, with few lexico-grammatical descriptions.

Furthermore, the few studies on academic Portuguese tend to focus on teaching, although didactic materials and teaching resources are still scarce (Stumpf 2021). This indicates the need for more research on academic Portuguese and shows that reported speech is a relevant feature in this discipline. Among the conventions of academic discourse, successfully integrating quotations, that is, using sources and citing the work

[1] More details about the situation can be found in Kowaltowski *et al.* (2021) and Quintans-Júnior *et al.* (2021).

[2] Although we acknowledge that Portuguese is a pluricentric language spoken in nine different countries spread over four continents, we bring data related to education and research from Portugal and Brazil, which are countries with a larger number of higher education institutions and journals.

of others is paramount (Coffin *et al.* 2005). It seems clear that academic texts, whether written or oral, rely on external sources to build arguments and link them to certain fields of knowledge whose citation practices can differ substantially. Hence, when incorporating other sources into their own writing, authors reveal their identity, and work towards belonging to specific discourse academic communities (Hoffnagel 2010).

This study presents the initial findings of a larger research on reporting practices in academic Portuguese, more precisely, on the use of reporting verbs in the *Corpus of Portuguese from Academic Journals* (CoPEP; Kuhn and Ferreira 2020). The paper addresses two research questions: 1) What are the reporting verbs that are mostly used to cite the work of others in Hard Science and Soft Science? and 2) To what extent are there differences and similarities in relation to how both registers use reporting verbs?

It is worth mentioning that our motivation to carry out this research was mostly based on the perceived needs of our students, who used a somewhat limited number of structures to report the work of other authors. Our intention was to find different patterns so that our students could expand their repertoire and improve their writing skills by mastering this particularly important feature of academic texts. Charles (2006: 327), in her study of phraseological patterns of citations, highlights the pedagogical applications of such a research and states that bringing the patterns to the classroom is "beneficial in raising student awareness of contextual factors and in enhancing their understanding of what lies behind the language choices evident on the page."

The corpus, containing 9,900 texts from academic journals in both Brazilian and European Portuguese, was divided into two subcorpora: one accounting for texts related to Hard Science and another for texts related to Soft Science. It should be borne in mind, however, that there is a fine line between what is considered Hard or Soft Science, even more so in the age of interdisciplinary research. Moreover, Soft Science has been considered inferior to Hard Science historically (Smith *et al.* 2000). In addition, aspects such as verifiability, replicability and more methodological rigor have been associated to Hard Science, making Soft Science seem less robust and scientific. For our purposes, however, we consider a traditional classification of the different disciplines as belonging to those areas, similar to the way Kuhn and Ferreira (2020) organized CoPEP. Kuhn and Ferreira (2020) follow the division proposed by the *Coordination for the Improvement of Higher Education Personnel* (CAPES) in Brazil and classify the texts into three main disciplines: College of Life Sciences (Biology,

Agrarian and Health Sciences), College of Humanities (Humanities, Applied Social Sciences and Linguistics, Literature and Arts) and College of Exact Sciences, Technology and Multidisciplinary (Earth and Exact Sciences, Engineering and Multidisciplinary).

The paper is structured as follows. First, we briefly discuss some aspects related to reporting practices and, more specifically, reporting verbs in Portuguese (Section 2). Then we present our methodology (Section 3) and discuss our results, comparing them to other studies of reporting verbs in academic written language (Section 4). We conclude the paper with a summary and some final remarks highlighting limitations and suggestions for follow-up investigations (Section 5).

2. LITERATURE REVIEW

2.1. Citation practices

Research on academic language is more commonly carried out in English. Nevertheless, works such as those of Hyland (1999, 2002) can be useful for other languages, such as Portuguese. Considering these aspects, some conventions of academic practices are spread along different languages, as the language itself is a means of communication. Thus, in academic settings, language choices are shaped, among other factors, according to the specificities of particular academic communities that follow certain conventions.

Citing is a common academic practice that helps the writer be part of a research community by creating a rhetorical space (Hoffnagel 2010). According to Hyland (1999: 341), "one of the most important realizations of the research writer's concern for audience is that of reporting, or reference to prior research," which, in practice, happens with the use of citations. Swales (1990) argues that it is a way of indicating to which field of knowledge writers belong, as they contribute to the production of knowledge by exploring and explaining specific topics of their area and thus bringing the voice of other authors. Swales' (1990) taxonomy includes two types of citations: integral citations and non-integral citations. The present paper focuses on integral citations which, according to Thompson (2005: 312) are "placed within the sentence and play an explicit role within the syntax of the sentence." Hyland and Jiang (2017) show how preference for non-integral forms of citation has increased since the 1960s in four disciplines (applied linguistics, biology, engineering, and sociology), which points to a

phenomenon where importance is given to the facts and contributions from previous work without the focus on the authors.

Another way of classifying citations is by focusing on the reporting verbs. Hyland (1999, 2002) offers a typology that divides them into 'research (real-world) acts', 'cognition acts' and 'discourse acts'. Verbs indicating research acts refer to activities and processes that take place in the real world, such as *observe*, *discover*, *analyze*, and *calculate*. Verbs representing cognition acts are those related to mental actions of the researcher, such as *believe*, *assume*, and *view*. Discourse acts are related to the verbal expression of either cognitive or research acts, such as *discuss*, *report*, and *state*. In some cases, however, these categories are not clear-cut and may overlap.

In an analysis of academic texts produced in Portuguese and published in anthropology and psychology Brazilian journals, Hoffnagel (2010) indicates that in integral citations the writer introduces the discourse being cited with the use of reporting verbs, which are one of the various aspects that make up the text. It is worth noting that verb choice is also rhetorical, suggesting that certain verbs are linked to disciplinary practices. Thus, the selection of specific reporting verbs in detriment of others is not random.

In English, this can be clearly seen in Hyland's (1999) results where writers use more reporting verbs in philosophy than in physics. There is also a prominence of verbs such as *say*, *argue*, *think*, and *suggest* in the humanities, while harder sciences favor *use*, *report*, *describe*, and *show*. Accordingly, Soft Science tends to use more verbs expressing discourse acts, while texts related to engineering and science adopt verbs related to research acts.

2.2. Reported speech in Portuguese

A fair amount of research on reported speech and on reporting verbs in Portuguese focuses on journalistic (Corbari and Ramos 2018) or literary registers (Saburi Costa and Freitas 2017), and there has been little research on this phenomenon in academic language. Most studies in academic Portuguese analyze reporting practices more broadly (Hoffnagel 2010; Bessa 2011), with few studies focusing on reporting verbs other than some isolated hints here and there (Souza and Mendes 2012).

Bessa (2011) discusses the use of reporting verbs as a mandatory practice in academia, as an academic piece of writing is only valid when including arguments and theories discussed by other authors. From a dialogical perspective, Bessa (2011) puts forward the idea that following writing manuals on how to cite in academic articles is not enough to master this aspect of academic writing. Using someone else's voice is much more complex than simply reporting their ideas mechanically. As Bessa (2011: 426) argues, there are eight main reasons why writers cite:

> (i) introducing a point of view, (ii) signaling belonging to a framework, a school of thought, (iii) referring to previous works, to trace the state of a problem, (iv) supporting a definition; (v) substantiating an assertion; (vi) discussing an assertion, moving away from a position; (vii) justifying a behavior; and (viii) introducing a new idea.[3]

As Bessa (2011) aptly notices, understanding citation in academic texts should not be restricted to technical features; it should also encompass an enunciative dimension, for the author's positioning comes into play. As highlighted earlier, this positioning is key to the development of an authorial voice, since authors can tell apart what has been studied by others from what they are doing. Likewise, being able to properly quote the work of others helps frame the author as an insider in the field, whereby they demonstrate their knowledge of references.

In a study dealing with a theoretical and pedagogical reflection about text production in academic settings, Motta-Roth and Hendges (2010) use four academic genres as the basis of their discussion, namely, reviews, research projects, academic articles, and abstracts. Based on Swales's (1990) socio-rhetorical framework, they translate the verbs into Portuguese and analyze these academic genres in terms of organization, structure, and linguistic features in relation to academic practices accepted in academia. It is worth mentioning that although one chapter of the book is centered on different types of citation and verb classification, it is not clear whether the reporting verbs that came up as the result of their analysis are used in integral or non-integral citations. Nevertheless, Motta-Roth and Hendges' (2010) results can serve as a possible framework to meet the goals of our own study.

[3] Our translation. Original version: "(i) introduzir um ponto de vista; (ii) marcar o pertencimento a uma corrente, a uma escola; (iii) referir-se a trabalhos anteriores, para traçar o estado de uma problemática, (iv) sustentar uma definição; (v) fundamentar uma afirmação; (vi) discutir uma afirmação, se afastar de uma posição; (vii) justificar um comportamento; e (viii) introduzir uma ideia nova." (Bessa 2011: 426).

Table 1 presents the verbs that are frequently used in the subjects that we consider part of Hard Science and Soft Science, respectively.

	Hard Science		**Soft Science**
	Biology, physics, electrical engineering, mechanical engineering, epidemiology, nursing and medicine		**Marketing, applied linguistics, psychology, sociology, education, philosophy**
1	*Descrever* 'describe'	1	*Sugerir* 'suggest'
2	*Desenvolver* 'develop'	2	*Descobrir* 'discover/to find out'
3	*Propor* 'propose'	3	*Argumentar* 'argue'
4	*Descobrir* 'discover/find out'	4	*Dizer* 'say'
5	*Mostrar* 'show'	5	*Mostrar* 'show'
6	*Reportar* 'report'	6	*Descrever* 'describe'
7	*Usar* 'use'	7	*Notar* 'notice'
8	*Sugerir* 'suggest'	8	*Explicar* 'explain'
9	*Estudar* 'study'	9	*Reportar* 'report'
10	*Demonstrar* 'demonstrate'	10	*Alegar* 'claim'
11	*Discutir* 'discuss'	11	*Propor* 'propose'
12	*Identificar* 'identify'	12	*Demonstrar* 'demonstrate'
13	*Observar* 'observe'	13	*Analisar* 'analyze'
14	*Expandir* 'expand'	14	*Destacar* 'highlight'
15	*Publicar* 'publish'	15	*Enfocar* 'focus'
16	*Dar* 'give'	16	*Discutir* 'discuss'
17	*Examinar* 'examine'	17	*Fornecer* 'provide'
18	*Indicar* 'indicate/point out'	18	*Pensar* 'think'

Table 1: Verbs used in Hard and Soft Science (adapted from Motta-Roth and Hendges 2010: 99)

Hoffnagel (2010) analyzed citations in 16 articles dealing with psychology, with 1,292 citations, and 16 articles dealing with anthropology, with 1,025 citations. According to Hoffnagel (2010), there is an enormous variety of reporting verbs in both genres: 135 verbs in anthropology and 90 verbs in psychology. It is worth mentioning, however, that around 50 percent of these verbs were used only once in the corpus. The top five verbs used in texts dealing with anthropology are: *dizer* 'say', *afirmar* 'claim', *citar* 'quote', *apontar* 'point out', and *mostrar* 'show', while in psychology they are: *realizar* 'make/do', *observar* 'observe', *propor* 'propose', *sugerir* 'suggest', and *apontar* 'point out'.

While this literature highlights the importance of the rhetorical function of citations and the choice of reporting verbs, it must be said that our work focuses on a single type of citation in order to find patterns distributed across the two large areas of Hard Science and Soft Science. Thus, due to the number of excerpts, we decided to explore the forms that were found in the corpora and relate them to what has been already published in the field by focusing on the aspects that match the purposes of our study.

3. DATA AND METHODS

3.1. The corpus

The corpus used in this investigation is CoPEP[4] (Kuhn and Ferreira 2020), which contains 9,900 texts from academic journals balanced in both Brazilian and European Portuguese. These academic journals are all indexed in the *Scientific Electronic Library On-line* (SciELO).[5] In order to meet the goals of the study, CoPEP was divided into two subcorpora, one subcorpus (Hard Science) containing texts from engineering, exact-earth science, and health science, and another subcorpus (Soft Science) with texts from applied social science and humanities. Table 2 provides information on the number of tokens and texts in both subcorpora.

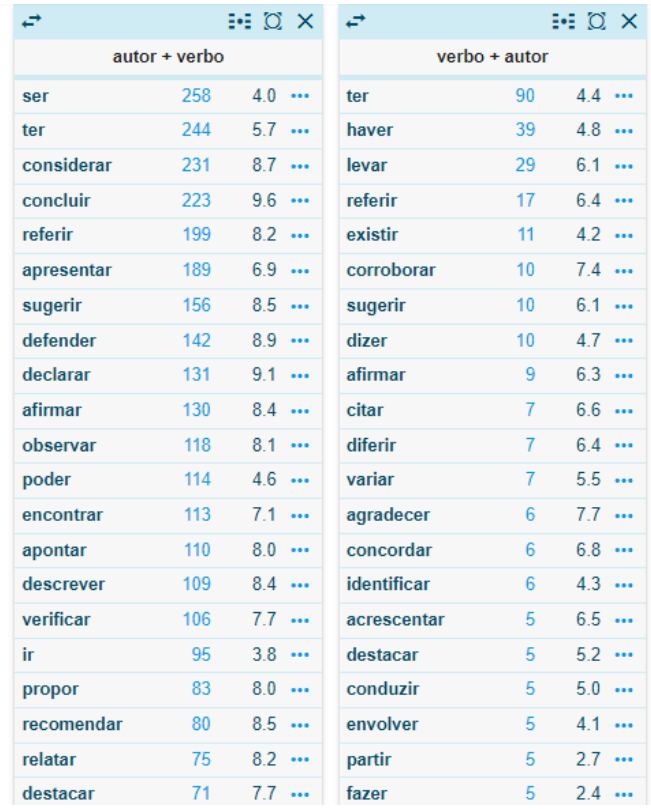
	Words	**Number of texts**	**Average number of words per text**
Soft Science	25,744,456	4,636	5,553.2
Hard Science	14,678,555	5,264	2,788.48
Total	40,423,011	9,900	8,3411.68

Table 2: Structure of CoPEP

3.2. Methodological procedures

In order to answer our research questions, four main steps were undertaken in both the Soft Science and Hard Science subcorpora. First, we determined that our analysis would be based on the verbs that go together with the lemma *autor* 'author', as the focus of this investigation is on reporting verbs and how external author's ideas are framed. Thus, as illustrated in Figures 1 and 2, we have used the *Word Sketch* tool in *Sketch Engine* (Kilgarriff *et al.* 2014) to generate the list of verbs that collocate before and after *autor* 'author' in both subcorpora.

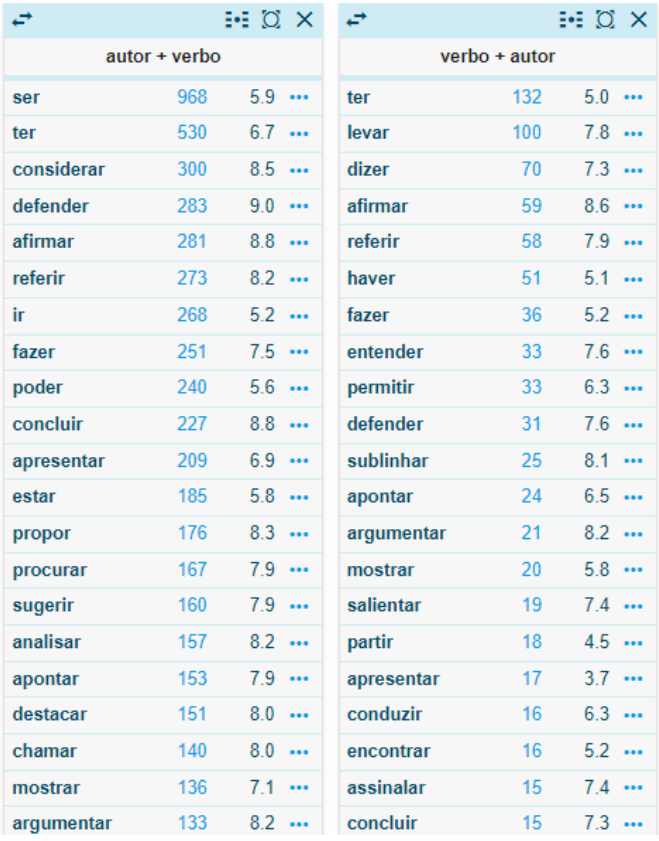
[4] CoPEP is available on *Sketch Engine* and is balanced in terms of fields of knowledge and language variety, since it includes texts published in Brazilian and European Portuguese. For more information regarding the corpus metadata and compilation, please, refer to Kuhn and Ferreira (2020).
[5] https://scielo.org/es/



autor + verbo		
ser	258	4.0
ter	244	5.7
considerar	231	8.7
concluir	223	9.6
referir	199	8.2
apresentar	189	6.9
sugerir	156	8.5
defender	142	8.9
declarar	131	9.1
afirmar	130	8.4
observar	118	8.1
poder	114	4.6
encontrar	113	7.1
apontar	110	8.0
descrever	109	8.4
verificar	106	7.7
ir	95	3.8
propor	83	8.0
recomendar	80	8.5
relatar	75	8.2
destacar	71	7.7

verbo + autor		
ter	90	4.4
haver	39	4.8
levar	29	6.1
referir	17	6.4
existir	11	4.2
corroborar	10	7.4
sugerir	10	6.1
dizer	10	4.7
afirmar	9	6.3
citar	7	6.6
diferir	7	6.4
variar	7	5.5
agradecer	6	7.7
concordar	6	6.8
identificar	6	4.3
acrescentar	5	6.5
destacar	5	5.2
conduzir	5	5.0
envolver	5	4.1
partir	5	2.7
fazer	5	2.4

Figure 1: *Word Sketch* results for the most frequent verbs in Hard Science



autor + verbo		
ser	968	5.9
ter	530	6.7
considerar	300	8.5
defender	283	9.0
afirmar	281	8.8
referir	273	8.2
ir	268	5.2
fazer	251	7.5
poder	240	5.6
concluir	227	8.8
apresentar	209	6.9
estar	185	5.8
propor	176	8.3
procurar	167	7.9
sugerir	160	7.9
analisar	157	8.2
apontar	153	7.9
destacar	151	8.0
chamar	140	8.0
mostrar	136	7.1
argumentar	133	8.2

verbo + autor		
ter	132	5.0
levar	100	7.8
dizer	70	7.3
afirmar	59	8.6
referir	58	7.9
haver	51	5.1
fazer	36	5.2
entender	33	7.6
permitir	33	6.3
defender	31	7.6
sublinhar	25	8.1
apontar	24	6.5
argumentar	21	8.2
mostrar	20	5.8
salientar	19	7.4
partir	18	4.5
apresentar	17	3.7
conduzir	16	6.3
encontrar	16	5.2
assinalar	15	7.4
concluir	15	7.3

Figure 2: *Word Sketch* results for the most frequent verbs in Soft Science

Next, we searched for the top 15 verbs that collocate with *autor* ‘author’ by using *Corpus Query Language* (CQL) in order to have access to the concordance lines of these verbs combined with *autor* (‘author’) in a 5-word window. The following CQL queries were used (cf. Table 3). [6]

Hard Science	**Soft Science**
Query 1: [lemma="autor"] [] {0,5} [lemma="considerar|concluir|referir|apresentar|sugerir|defender|afirmar|observar|encontrar|apontar|descrever|verificar|propor|recomendar|relatar"]	Query 1: [lemma="autor"] [] {0,5} [lemma="considerar|defender|afirmar|referir|fazer|concluir|apresentar|propor|procurar|sugerir|analisar|apontar|destacar|chamar|mostrar"]
Query 2: [lemma="referir|corroborar|sugerir|dizer|afirmar|citar|diferir|concordar|identificar|acrescentar|destacar|conduzir|fazer"] [] {0,5} [lemma="autor"]	Query 2: [lemma="ter|dizer|afirmar|referir|fazer|entender|permitir|defender|sublinhar|apontar|argumentar|mostrar|salientar|apresentar|conduzir"] [] {0,5} [lemma="autor"]

Table 3: *Corpus Query Language*

Based on these two steps, we realized that the verbs *ter* ‘have’, *ser* ‘be’, *haver* ‘there is/there are’, and *fazer* ‘do/make’ presented interesting behaviors. Hence, we decided to run a new CQL search and analyze them separately, as they can be used as auxiliary verbs for compound tenses and on verb phrases. We decided to consider valid cases where *ter*, *ser*, and *haver* were used as auxiliary verbs (and not as the main verb) and *fazer* was used as the main verb to indicate something that was done by the author(s), excluding idioms and cases like those illustrated in examples (1) to (3).

(1) Ao refletir sobre a dinâmica regional da economia brasileira, diferentes autores **fazem** uso de importantes ressalvas para pensar o processo de desconcentração produtiva verificado a partir da Região. ‘Reflecting on the regional dynamic of the Brazilian economy, different authors **make** use of important caveats to think about the process of deconcentration verified from the region’.

(2) Este autor não **fazia** parte do seleto grupo dos intelectuais vinculados à academia. ‘This author did not **make** part of a select group of intellectuals linked to academia / This author was not part of a select group of intellectuals linked to academia’.

[6] Translation of the verbs to English:
Hard Science query 1: ‘consider’, ‘conclude’, ‘refer’, ‘present’, ‘suggest’, ‘defend’, ‘state’, ‘observe’, ‘find’, ‘point out’, ‘describe’, ‘verify’, ‘propose’, ‘recommend’, ‘report’.
Hard Science query 2: ‘refer’, ‘corroborate’, ‘suggest’, ‘say’, ‘state’, ‘cite’, ‘differ’, ‘agree’, ‘identify’, ‘add’, ‘highlight’, ‘conduct’, ‘do/make’.
Soft Science query 1: ‘consider’, ‘defend’, ‘state’, ‘refer’, ‘do/make’, ‘conclude’, ‘present’, ‘propose’, ‘intend’, ‘suggest’, ‘analyze’, ‘point out’, ‘highlight’, ‘call, to show’.
Soft Science query 2: ‘have’, ‘say’, ‘state’, ‘refer’, ‘do/make’, ‘understand’, ‘allow’, ‘defend’, ‘underline’, ‘point out’, ‘argue’, ‘show’, ‘stress’, ‘present’, ‘conduct’.

(3) Este autor **tinha** como objetivo desenvolver um site onde os próprios usuários poderiam gerar conteúdo. 'This author **had** as objective developing a site where the users could generate content'.

Other verbs that were initially excluded are *declarar* 'declare' and *agradecer* 'thank', since all of them occurred in formulaic expressions as in, for instance, *the author(s) declare(s) that there is no conflict of interest* and *we thank the editor and two anonymous reviewers*. The verb *levar* 'lead, take' was also excluded since it was mainly used in sentences such as *this led authors to state/defend*, and we focused on verbs coming afterwards. Finally, *existir* 'exist' and *partir* 'leave' were also excluded since they do not function as reporting verbs.

Valid occurrences were then classified according to: 1) the discipline in which they occurred, 2) voice (passive or active), 3) number (singular or plural of 'author'), 4) order ('verb + author' or 'author + verb'), 5) tense, aspect and mood or non-finite verb forms (converb, past participle, or infinitive). Since we aimed at finding patterns of use, we also classified the reporting verb according to Hyland's (1999, 2002) typology, which considers the type of activity that the verbs refer to. Besides, the classification of the verbs was partly based on Shaw (1992) and Hyland and Jiang (2017), whose studies account for the tense and the aspect of verbs. In cases where two valid verbs were used, the sentence was classified twice, once for each verb, as in (4), below. Likewise, verb phrases denoting time or modality were classified according to the first verb, as in (5)–(6).

(4) A autora **citando** Bourdieu (1983) **afirma** (*citando* 'citing'): converb, *afirma* 'states': 3rd person singular simple present)

'The author, **citing** Bourdieu (1983), **states** (…)'

(5) A autora **continua referindo** (*continua* 'continues'): 3rd person singular simple present)

'The author **continues referring** (…)'

(6) Os autores **deverão referir-se** (*deverão* 'should': 3rd person plural future simple)

'The authors **should refer themselves** (…)'

4. RESULTS AND DISCUSSION

In this section, we first present the overall frequency of reporting verbs distributed in the corpus. Normed counts per 10,000 words are presented between brackets. As stated earlier, the aim is to find the most frequent reporting verbs in the registers related to Hard and Soft Science, together with their patterns of use, and to discuss possible differences and similarities between the registers. Thus, the quantitative results shown in the tables are followed by the discussion of the data. When possible, we refer to other studies conducted in English to try to support and provide the motivation for our findings.

As Table 4 shows, 6,103 occurrences of reporting verbs used before or after the word *autor* 'author' in a five-word window were valid. Of these, 3,716 (normed frequency 1.44) are attested in the texts related to Soft Science and 2,387 (normed frequency 1.62) in the texts related to Hard Science. These frequencies indicate a higher frequency of reporting verbs with *autor* 'author' in Hard Science. Although "softer disciplines tend to employ more citations" (Hyland 1999: 346), the differences in the normed counts could be partially explained by the preference in Soft Science to use proper names to refer to authors, instead of using the lemma 'author' in a more general way.

	Excerpts with reporting verbs
Soft Science	3,716 (1.44)
Hard Science	2,387 (1.62)
Total	6,103

Table 4: Frequency of excerpts with reporting verbs in each register

Table 5, below, provides information on voice, order, and number of reporting verbs in each subcorpus. The data include raw frequencies and the percentages between brackets. Overall, there is a clear preference for the active voice and the order *autor* + *verbo* 'author + verb' in both registers, with 87 percent of the occurrences in the active voice in Soft Science and 90 percent in Hard Science. As regards percentages, the order *autor* + *verbo* 'author + verb' is more frequent in both registers (80% of the cases in Soft Science and 90% of the cases in Hard Science) when compared to the order *verbo* + *autor* 'verb + author'. Soft Science exhibits a slightly higher variation in terms of order, showing more excerpts with *verbo* + *autor* 'verb + author' order, as shown in Table 5. Finally, when it comes to number, in Soft Science, the choice for the singular

or the plural is balanced: 53 percent of the cases are attested in the singular form and 47 percent in the plural form. By contrast, in Hard Science, the preference is for the plural form (84% of instances) to the detriment of the singular form (16% of instances). The preference for the plural form in Hard Science might be related to the fact that in this register publications with multiple of authors are common.

	Excerpts with reporting verbs	**Active voice**[7]	**Passive voice**	***Autor* + verb order**	**Verb + *autor* order**	**Singular form**	**Plural form**
Soft Science	3,716	3,261 (87%)	92 (2.4%)	2,992 (80%)	725 (20%)	2,003 (53%)	1,714 (47%)
Hard Science	2,387	2,156 (90%)	48 (2%)	2,162 (90%)	226 (10%)	392 (16%)	1,995 (84%)
Total	6,103	5,417	140	5,154	951	2,395	3,709

Table 5: Number of excerpts with reporting verbs by voice, order, and number

The results on tense, aspect, and mood are shown in Table 6.

		Hard Science	**Soft Science**
***Pretérito* 'past'**	*Pretérito perfeito* (preterite perfect)	699 (0.48)	370 (0.14)
	Pretérito imperfeito (preterite imperfect)	10 (0.01)	40 (0.02)
	Pretérito imperfeito contínuo (preterite imperfect continuous)	0 (0)	2 (0.001)
	Pretérito mais-que-perfeito (past perfect)	3 (0.002)	5 (0.002)
	Pretérito mais-que- perfeito composto (compound past perfect)	0 (0)	6 (0.002)
	Pretérito perfeito composto (compound preterite perfect)	54 (0.04)	93 (0.04)
		766 (0.52)	**516 (0.2)**
***Presente* 'present'**	*Presente simples* (simple present)	1,428 (0.97)	2,817 (1.09)
	Presente contínuo (present continuous)	2 (0.0013)	2 (0,001)
		1,430 (0.97)	**2,819 (1.09)**
***Futuro* 'future'**	*Futuro perifrástico* (compound future)	0 (0)	6 (0.002)
	Futuro perifrástico contínuo (compound future continuous)	0 (0)	1 (0.0004)
	Futuro do presente (simple future)	2 (0.001)	4 (0.002)
	Futuro do pretérito (conditional tense - would)	2 (0.001)	2 (0.001)
		4 (0.003)	**13 (0.005)**

Table 6: List of tenses, moods, and aspects with raw and normed frequencies

Notably, the present is the preferred tense. Hard Science makes a more frequent use of the past tense (0.52), whereas Soft Science makes a more frequent use of the present (1.09), and the future is rarely used in both registers, but slightly more frequent in Soft Science (0.005) than in Hard Science (0.003). Some examples are provided in (7)–(12).

(7) (...) a grande maioria dos autores **não observou** (*observou* 'did not observe': 3rd person singular preterite perfect). Hard Science.

'(…) the vast majority of authors **did not observe** (…)'

[7] The sum of the occurrences in the active and the passive voice does not match the number of excerpts with reporting verbs in each subcorpus because occurrences of non-finite verb forms cannot be considered either active or passive voice.

(8) O autor **afirmava** (*afirmava* ‘stated’: 3rd person singular preterite imperfect). Hard Science.

‘The author **stated** (…)’

(9) O autor do texto **estava fazendo** (*estava fazendo* ‘was doing’: 3rd person singular preterite imperfect continuous). Soft Science.

‘The author of the text **was doing** (…)’

(10) O *autor* ***apontara*** (*apontara* ‘had pointed out’: 3rd person singular past perfect). Soft Science.

‘The author **had pointed out** (…)’

(11) O autor **havia afirmado** (*havia afirmado* ‘had stated’: 3rd person singular compound past perfect). Soft Science.

‘The author **had stated** (…)’

(12) (...) autoras/es feministas **têm feito** (*têm feito* ‘have done’: 3rd person plural compound preterite perfect)

‘(…) the feminist authors **have done** (…)’

When further looking at the distinct forms these three tenses may have, there is a stronger preference for the *pretérito perfeito* ‘preterite perfect’ in Hard Science (0.48) than in Soft Science (0.14). Despite these differences, it is possible to observe that, although *pretérito imperfeito contínuo* (‘preterite imperfect continuous’) and *pretérito mais-que-perfeito composto* (‘compound past perfect’) are rare tenses of reporting verbs in Soft Science (0.001 and 0.002, respectively), they are not used in Hard Science. As for the preterit perfect compound, indicating an action that started in the past but is still ongoing (akin to the present perfect tense in English), there are 54 occurrences in Hard Science and 93 in Soft Science, making it the second most frequent past tense in the corpus. According to Hyland and Jiang (2017), there has been an increasing trend in using reporting verbs in the present tense in sociology and engineering alike (which would belong to our Soft and Hard Science subcorpora, respectively), followed by the past tense. However, since Hyland and Jiang (2017) analyze four disciplines, each of them belonging to different registers, their data is difficult to compare with ours, mainly because the languages under analysis (English and Portuguese) have their own peculiarities.

In the present tense, the simple (cf. (13)) and continuous (cf. (14)) aspects are used, the latter with an extremely low frequency. *Presente simples* 'simple present', on the other hand, is a very common verb tense when citing the voice of others and incorporating sources. In the excerpts with reporting verbs, this tense is higher in Soft Science (normed frequency 1.09) than in Hard Science (normed frequency 0.97).

(13) (...) autor australiano **destaca** (*destaca* 'highlights') 3rd person singular simple present)

'(…) the Australian author **highlights** (…)'

(14) (...) autor **está defendendo** (*está defendendo* 'is defending' 3rd person singular present continuous)

'(…) the author **is defending** (…)'

Finally, there are few occurrences of future tense in both subcorpora, besides the fact that the comparison between the areas is balanced. Some of these occurrences are shown in (15)–(18) below.

(15) (...) autor **vai chamar** (*vai chamar* 'will call': 3rd person singular compound future). Soft Science

'(…) the author **will call** (…)'

(16) (...) autor **vai defendendo** (*vai defendendo* 'keeps defending': 3rd person singular compound future continuous). Soft Science

'(…) the author **keeps defending** (…)'

(17) (...) autores que **apresentarão** (*apresentarão* 'will present': 3rd person plural simple future). Hard Science

'(…) the authors that **will present** (…)'

(18) (...) autores **defenderiam** (*defenderiam* 'would defend': 3rd person plural conditional tense). Soft Science.

'(…) the authors **would defend** (…)'

The examples provided above are all in the indicative mood. Occurrences in the subjunctive mood, as in (19) and (20) are rarely attested in the corpus (65 overall), even if the subjunctive is required after some subordinating conjunctions in Portuguese, such as *embora* 'although'. Most cases are in the present tense, followed by instances in the compound preterite perfect.

(19) (...) autores **defemdam** (*defendam* ‘defend’: 3rd person plural simple sujunctive). Hard Science

‘(…) authors **defend** (…)’

(20) (...) autores **tenham apresentado** (*tenham apresentado* ‘had presented defend’: 3rd person singular compound preterit prefect subjunctive conditional tense). Soft Science.’

‘(…) authors **had presented** (…)’

Another interesting finding in the analysis is the use of modalization. Modalization is mainly used when the writer makes a stand towards the voice of the author being cited, instead of bringing the voice of the other in a more impartial way, as in (21)–(22).

(21) (...) autor parece **defender**. Soft Science.

‘(…) the author **seems to defend** (…)’

(22) (...) autor **precisa considerer**. Soft Science.

‘(…) the author **needs to consider** (…)’

We also coded non-finite verb forms that appeared as dependent clauses in complex sentences with reporting verbs, if the verb was a verb in our list (cf. Table 1). As shown in Table 7, among these verb forms, the past participle (cf. (23)) is the one with the highest frequency in the corpus, with 333 occurrences of which 236 were attested in Soft Science and 97 in Hard Science. There are 135 occurrences of the infinitive (cf. (24)), evenly distributed in both subcorpora (0.03). As for converb forms, which are non-finite verb forms used in adverbial subordination (cf. (25)), these are also evenly distributed (0.0.1). The compound converb (cf. (26)), however, is slightly more frequently attested in Hard Science (0.005) than in Soft Science (0.002).

	Past participle	**Infinitive**	**Converb**	**Compound converb**
Soft Science (364)	236 (0.09)	85 (0.03)	36 (0.01)	7 (0.002)
Hard Science (184)	97 (0.06)	50 (0.03)	29 (0.01)	8 (0.005)
Total	**333**	**135**	**65**	**15**

Table 7: Number of non-finite verb forms in Soft and Hard Science

(23) (...) no caso **apresentado** pelos autores (*apresentado* ‘presented’: past participle). Hard Science.

‘(…) in the case **presented** by the authors (…)’

(24) (...) palavras ou expressões utilizadas pelos autores para **descrever** (*descrever* 'describe': infinitive). Hard Science.

'(…) words or expressions utilized by the authors to **describe** (…)'

(25) (...) Alguns autores, **observando** a formação do enfermeiro (*observando* 'observing': converb). Hard Science.

'(…) Some authors, **observing** the education of the nurse (…)'

(26) (...) não **tendo** os autores **encontrado** (*tendo encontrado* 'having found': compound converb). Hard Science.

'(…) not **having** authors **found** (…)'

Tables 8 and 9 provide information regarding the most frequent verbs with the lemma *autor* 'author'. In total, 27 different reporting verbs make up 2,387 occurrences in Hard Science. Out of these, 2,196 in the *autor* + *verbo* 'author + verb' order, and 191 in the *verbo* + *autor* 'verb + author' order. Concerning Soft Science, 26 different reporting verbs make up 3,716 occurrences. Out of these, 2,686 are used in the *autor* + *verbo* 'author + verb' order, and 1,030 are used in the *verbo* + *autor* 'verb + author' order.

		Hard Science ***Autor + verbo***	**Hard Science** ***Verbo + autor***	**TOTAL**
1	*Referir* 'refer'	185	83	268
2	*Considerar* 'consider'	248	2	250
3	*Concluir* 'conclude'	235	2	237
4	*Sugerir* 'suggest'	164	40	204
5	*Apresentar* 'present'	169	1	170
6	*Defender* 'defend'	158	0	158
7	*Afirmar* 'state'	139	11	150
8	*Observar* 'observe'	127	1	128
9	*Encontrar* 'find'	119	2	121
10	*Verificar* 'verify'	110	3	113
11	*Apontar* 'point out'	110	2	112
12	*Descrever* 'describe'	104	1	105
13	*Recomendar* 'recommend'	93	1	94
14	*Propor* 'propose'	83	0	83
15	*Relatar* 'report'	79	0	79
16	*Referir-se* 'refer oneself/themselves'	26	0	26
17	*Citar* 'quote'	22	1	23
18	*Fazer* 'do/to make'	14	0	14
19	*Destacar* 'highlight'	1	9	10
20	*Acrescentar* 'add'	0	9	9
21	*Corroborar* 'corroborate'	0	8	8
22	*Dizer* 'say'	0	8	8
23	*Identificar* 'identify'	0	7	7
24	*Propor-se* 'propose oneself/themselves'	7	0	7
25	*Contribuir* 'contribute'	1	0	1
26	*Apresentar-se* 'introduce oneself/themselves'	1	0	1
27	*Reportar* 'report'	1	0	1
	Total	**2,196**	**191**	**2,387**

Table 8: Frequencies according to order in Hard Science

		Soft Science *Autor + verbo*	Hard Science *Verbo + autor*	TOTAL
1	*Defender* 'defend'	313	105	418
2	*Afirmar* 'state'	285	75	360
3	*Fazer* 'do/make'	254	76	330
4	*Referir* 'refer'	220	96	316
5	*Considerar* 'consider'	0	305	305
6	*Apresentar* 'present'	174	88	262
7	*Concluir* 'conclude'	245	0	245
8	*Apontar* 'point out'	165	71	236
9	*Analisar* 'analyze'	181	0	181
10	*Sugerir* 'suggest'	161	0	161
11	*Mostrar* 'show'	136	24	160
12	*Propor* 'propose'	150	0	150
13	*Destacar* 'highlight'	146	0	146
14	*Chamar* 'call'	139	0	139
15	*Dizer* 'say'	0	75	75
16	*Referir-se* 'refer oneself/themselves'	62	0	62
17	*Sublinhar* 'underline'	0	40	40
18	*Salientar* 'stress'	0	33	33
19	*Propor-se* 'propose oneself/themselves'	31	0	31
20	*Argumentar* 'argue'	0	29	29
21	*Entender* 'understand'	0	13	13
22	*Destacar-se* 'stand out'	8	0	8
23	*Mostrar-se* 'show oneself/themselves'	8	0	8
24	*Apresentar-se* 'introduce oneself/themselves'	5	0	5
25	*Encontrar* 'find'	2	0	2
26	*Defender-se* 'defend oneself/themselves'	1	0	1
	Total	**2,686**	**1,030**	**3,716**

Table 9: Frequencies according to order in Soft Science

When it comes to the order, some verbs tend to occur in the *verbo + autor* 'verb + author' order and are hardly attested in examples of *autor + verbo* 'author + verb'. In Hard Science, this is the case for *destacar* 'highlight', *acrescentar* 'add', *corroborar* 'corroborate', *dizer* 'say', and *identificar* 'identify', most of them being verbs used to express discourse acts. In Soft Science, apart from *considerar* 'consider' and *entender* 'understand' (classified as verbs indicating cognition acts), the other verbs that are only used in the 'verb + author' order are *dizer* 'say', *sublinhar* 'underline', *salientar* 'stress', and *argumentar* 'argue' are also all expressing discourse acts.

The five most frequent reporting verbs are different in Hard and Soft Science. *Referir* 'refer', *considerar* 'consider', *concluir* 'conclude', *sugerir* 'suggest', and *apresentar* 'present' are the five most frequent reporting verbs in Hard Science (Table 8), whereas *defender* 'defend', *afirmar* 'claim', *fazer* 'do/make', *referir* 'refer', and *considerar* 'consider' are the five most frequent reporting verbs in Soft Science, as shown in Table 9.

Considering the five most frequent verbs in Soft Science, it is not surprising that *defender* ‘defend’ is the most frequent verb, since it is a verb that clearly shows a strong stance from the author’s perspective (by author, we mean the author(s) of the original text, the one being reported). According to Hyland’s framework of evaluative meaning of reported verbs, *defender* could be considered a neutral verb when related to the writer’s opinion (writer refers to the author of the text reporting other works), which, in turn, can indicate a positive view from the author’s perspective (Hyland and Jiang 2017). One could also argue that texts in Soft Science value clear positioning of authors and writers alike more than text in Hard Science do. Among the five most frequent verbs in the Hard Science corpus, the verb *sugerir* ‘suggest’ stands out, as it shows a more tentative stance from the author.

Furthermore, there are verbs that are used both in Soft and Hard Science together with *autor*, whether before or after the lemma. However, most appear exclusively in either Hard or Soft Science. The reporting verbs that are used in both Soft and Hard Science are *afirmar* ‘state’, *apontar* ‘point out’, *apresentar* ‘present’, *concluir* ‘conclude’, *considerar* ‘consider’, *defender* ‘defend’, *propor* ‘propose’, *referir* ‘refer’, and *sugerir* ‘suggest’. The reporting verbs used exclusively in Hard Science are *observar* ‘observe’, *encontrar* ‘find’, *verificar* ‘verify/check’, *descrever* ‘describe’, *recomendar* ‘recommend’, and *relatar* ‘report’. The verbs that are only used in Soft Science are *fazer* ‘make/do’, *analisar* ‘analyze’, *mostrar* ‘show’, *chamar* ‘call’, *destacar* ‘highlight’, and *procurar* ‘intend’.

These results are in line with Motta-Roth and Hendges (2010) who determined a list of the 18 reporting verbs most frequently used in subjects that correspond to what we understand as Hard Science. Out of these, only four match our results, namely *descrever* ‘describe’, *propor* ‘propose’, *sugerir* ‘suggest’, and *observar* ‘observe’. When it comes to the Soft Science subcorpus, the reporting verbs we retrieved that coincide with the Motta-Roth and Hendges’ (2010) list are five, namely, *sugerir* ‘suggest’, *mostrar* ‘show’, *propor* ‘propose’, *analisar* ‘analyze’, and *destacar* ‘highlight’. These discrepancies can be explained because we have narrowed down our analysis to the sequences *autor + verbo* ‘author + verb’ and *verbo + autor* ‘verb + author’. This was not the case in Motta-Roth and Hendges (2010), who included other types of citations in their study.

There is also partial agreement between our results and those of Hoffnagel (2010). Although our Soft Science subcorpus comprises more disciplinary registers than that of Hoffnagel, which is restricted to texts dealing with psychology and anthropology, the ten most frequent reporting verbs in both areas also appear within the first 15 positions in Soft Science, with the exception of *citar* 'cite' and *observar* 'observe', which are not attested in this subcorpus. For this analysis, we considered *realizar* and *fazer* as synonyms, since both may mean 'do/make'.

The reporting verbs were also classified according to Hyland's (1999, 2002) typology. The only verb that was excluded was *fazer* 'do/make', as it is a delexical verb whose meaning is extremely light since it is attached to the noun linked to it, as in *fazer referência* 'make reference' or *fazer uma discussão* 'make a discussion'. The results are shown in Table 10.

	Hard Science	**Soft Science**
Research (real-world)	*Concluir* 'conclude' *Apresentar* 'present' *Observar* 'observe' *Encontrar* 'find' *Verificar* 'verify' *Apresentar-se* 'introduce oneself'	*Apresentar* 'present' *Concluir* 'conclude' *Analizar* 'analyze' Mostrar-se 'show oneself' *Apresentar*-se 'introduce oneself' *Encontrar* 'find'
Cognition	*Referir* 'refer' *Considerar* 'consider' *Defender* 'defend' *Recomendar* 'recommend' *Corroborar* 'corroborate'	*Defender* 'defend' *Referir* 'refer' *Considerar* 'consider' *Referir*-se 'refer oneself' *Entender* 'understand' *Defender-se* 'defend oneself'
Discourse	*Sugerir* 'suggest' *Afirmar* 'state' *Apontar* 'point out' *Descrever* 'describe' *Propor* 'propose' *Relatar* 'report' *Referir-se* 'refer oneself' *Citar* 'quote' *Destacar* 'highlight' *Acrescentar* 'add' *Dizer* 'say' *Propor-se* 'propose oneself' *Contribuir* 'contribute' *Reportar* 'report'	*Afirmar* 'state' *Apontar* 'point out' *Sugerir* 'suggest' *Propor* 'propose' *Destacar* 'highlight' *Chamar* 'call' *Dizer* 'say' *Sublinhar* 'underline' *Salientar* 'stress' *Propor-se* 'propose oneself' *Argumentar* 'argue' *Destacar-se* 'stand out'

Table 10: Classification and occurrences of reporting verbs based on Hyland's (1999, 2002) typology

As shown in Table 10, the most frequent reporting verbs in the data belong to the group of discourse acts. This might be because the scope of our research on reporting verbs was restricted to the sequences 'author + verb' and 'verb + author'. Thus, verbs such as

afirmar ‘state’, *apontar* ‘point out’, *descrever* ‘describe’, and *relatar* ‘report’ are relatively neutral verbs that do not convey a negative or a positive tone. According to Hyland and Jiang (2017), this phenomenon is becoming more frequent, as there is a tendency for authors to use neutral forms to refer to verbal activities.

5. SUMMARY AND FINAL REMARKS

Our research has addressed an existing gap in the literature regarding the analysis of reporting verbs in academic Portuguese, for verbs either following or preceding the lemma *autor* ‘author’. For this purpose, we used CoPEP and analyzed two subcorpora, one representing softer sciences and the other harder sciences. From the 6,103 valid occurrences of reporting verbs extracted from CoPEP, 3,716 (1.44) are used in the Soft Science corpus, while 2,387 (1.62) are used in Hard Science corpus, which shows a higher frequency of reporting verbs with *autor* ‘author’. The results showed that, from the 15 most used verbs in both Soft and Hard Science, there are nine verbs used in both corpora: *afirmar* ‘state’, *apontar* ‘point out’, *apresentar* ‘present’, *concluir* ‘conclude’, *considerar* ‘consider’, *defender* ‘defend’, *propor* ‘propose’, *referir* ‘refer’, and *sugerir* ‘suggest’, even though there is no agreement in the order in which they appear (judging by the number of occurrences).

Our list partially matches the verbs mentioned in previous studies (Motta-Roth and Hendges 2010), although it is difficult to compare and contrast the data since this study differs in the way citations were collected. In our data, the reporting verbs are used mainly in the simple present and preterit perfect tenses, with a preference for the use of active voice and the order *autor* + *verbo* ‘author + verb’. These patterns of use might be related to the genre under analysis. Nevertheless, further research which includes genre as a variable could help support this argument. Some patterns regarding different uses of reporting verbs in English according to disciplinary areas have not been found in our corpus, making it difficult to draw comparisons between English and Portuguese. Searching for other citation patterns can help shed further light into this. For now, one could also speculate that disciplinary differences are not so marked in Portuguese, which points to more established patterns in English.

As stated earlier, the pedagogical concerns leading to this study explain why it is more related to the form of the occurrences and not so much to their rhetorical function

in the texts. The number of excerpts with reporting verbs made it difficult to code them manually with relation to this pragmatic aspect. Although it is a large number, it needs to be acknowledged that different types of citation are not included, for example, cases with the use of proper nouns or where nouns such as *researchers* or *scholars* were used, among others. This means that there is still room for more studies that encompass other forms of citation, which could then account for the disciplinary variation that can be seen in these structures. While it is understood that form, meaning, and function are intertwined, the focus on form here also helps address the lack of studies dealing with lexico-grammatical features of academic Portuguese, as pointed out by Kuhn (2017).

Other aspects that were not controlled for in our study include the section of the paper from which the reporting verb came from, which can influence the verb tense, and whether they came from the same article. Therefore, one way of continuing this study would be to broaden the search so as to encompass other forms, while gathering and coding for more information about the occurrences. Another possibility is looking into the differences regarding the language varieties represented in CoPEP. Once accounting for a representative sample with different types of reporting occurrences, more patterns can be brought to light and can then be compared and contrasted to other more widely researched languages, such as English.

Despite the limitations of this study, our findings can be useful and represent, to the best of our knowledge, a first step into the large-scale study of reporting verbs in different disciplinary areas in Portuguese. They can aid the development of much-needed pedagogical materials aimed at novice researchers and learners of academic Portuguese, whether as a first or a second language. They can also be used for studies with learner corpora, since CoPEP is a representative sample of academic language both in the Brazilian and European varieties of Portuguese.

REFERENCES

Bessa, José Cezinaldo Rocha. 2011. (Re)pensando a citação em textos acadêmico-científicos. *Signum: Estudos Da Linguagem* 14/2*:* 421–439.

Charles, Maggie. 2006. Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes* 25/3: 310–331.

Coffin, Caroline, Mary Jane Curry, Sharon Goodman, Ann Hewings, Theresa Lillis and Joan Swann. 2005. *Teaching Academic Writing: A Toolkit for Higher Education*. New York: Routledge.

Corbari, Alcione Tereza and Quézia Cavalheiro M. Ramos. 2018. Verbos dicendi na notícia: Pontos de um continuum argumentativo na construção da intertextualidade. *Fórum Linguístico* 15/1: 2903–2923.

Hernández Bonilla, Juan Miguel. 2021. How to end the hegemony of English in scientific research. *El País* English Edition. https://english.elpais.com/usa/2021-07-30/how-to-end-the-hegemony-of-english-in-scientific-research.html (3 Jul, 2022.)

Hoffnagel, Judith C. 2010. A prática de citação em trabalhos acadêmicos. *Cadernos de Linguagem e Sociedade* 10/1*:* 71–88.

Hyland, Ken. 1999. Academic attribution: Citation and the construction of disciplinary knowledge. *Applied Linguistics* 20/3: 341–367.

Hyland, Ken 2002. Activity and evaluation: Reporting practices in academic writing. In John Flowerdew ed. *Academic Discourse*. London: Longman, 115–130.

Hyland, Ken and Feng Jiang. 2017. Points of reference: Changing patterns of academic citation. *Applied Linguistics* 40/1: 64–85.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

Kowaltowski, Alicia, Marcus Oliveira and A. Ariel Silver Hernan Chaimovich. 2021. The push for open access is making science less inclusive. Times Higher Education (THE). https://www.timeshighereducation.com/opinion/push-open-access-making-science-less-inclusive (3 Jul, 2022.)

Kuhn, Tanara Zingano. 2017. *A Design Proposal of an Online Corpus-driven Dictionary of Portuguese for University Students*. Lisboa: Universidade de Lisboa dissertation.

Kuhn, Tanara Zingano and José Pedro Ferreira. 2020. O Corpus de Português Escrito em Periódicos - CoPEP. *Documentação de Estudos em Lingüística Teórica e Aplicada* 36/2: 2020360209. https://doi.org/10.1590/1678-460x2020360209

Motta-Roth, Désiréé and Graciela Rabuske Hendges. 2010. *Produção Textual na Universidade*. São Paulo: Parábola Editorial.

Oliveira, Isabel Tiago de, Madalena Ramos, Ana Ferreira and Sofia Gaspar. 2015. Estudantes estrangeiros em Portugal: Evolução e dinâmicas recentes (2005/6 a 2012/13). *Revista de Estudos Demográficos* 54: 39–56.

Pérez Llantada, Carmen. 2021. Genres and languages in science communication: The multiple dimensions of the science policy interface. *Language & Communication* 78: 65–76.

Quintans-Júnior, Lucindo José, George Rego Albuquerque, Sérgio Campello Oliveira and Robério Rodrigues Silva. 2020. Brazil's research budget: Endless setbacks. *EXCLI Journal* 19: 1322–1324.

Saburi Costa, Bianca Freitas and Cláudia Freitas. 2017. Verbos de elocução em português: Um estudo descritivo com base em grandes corpora e motivado pela linguística computacional. *Fórum Linguístico* 14/3: 2266–2285.

Smith, Laurence Daniel, Lisa A. Best, Donald Allan Stubbs, John Johnston and Andrea Bastiani Archibald. 2000. Scientific graphs and the hierarchy of the sciences: A latourian survey of inscription practices. *Social Studies of Science* 30/1: 73–94.

Shaw, Philip. 1992. Reasons for the correlation of voice, tense, and sentence function in reporting verbs. *Applied Linguistics* 13/3: 302–319.

Solovova, Olga Santos, Joana Vieira and Joaquim Veríssimo. 2018. Publish in English or perish in Portuguese: Struggles and constraints on the semiperiphery. *Publications* 6/2: 1–14. https://doi.org/10.3390/publications6020025

Souza, Medianeira and Wellington Vieira Mendes. 2012. Uma análise sistêmico-funcional do dizer em artigos científicos de graduandos. *Documentação de Estudos em Lingüística Teórica e Aplicada* 28: 537–560.

Stumpf, Elisa Marchioro. 2021. Portuguese as an additional language for academic purposes: Contributions from the academic literacies model to course design. *Revista Linguagem & Ensino* 24/2: 317–331.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Thompson, Paul. 2005. Points of focus and position: Intertextual reference in PhD theses. *Journal of English for Academic Purposes* 4/4: 307–323.

Corresponding author
Marine Laísa Matte
Federal University of Rio Grande do Sul
Department of Modern Languages
Bento Gonçalves Avenue 9500
91540-000 Porto Alegre
Brazil
E-mail: marine.matte@ufrgs.br

received: October 2021
accepted: June 2022

RiCL Research in Corpus Linguistics

Cross-linguistic transfer in academic journal writing: Preliminary evidence from lexical bundle analysis in Russian and English

Maria Kostromitina
Northern Arizona University / USA

Abstract – Lexical bundles are considered a fundamental feature of academic writing and have been extensively studied by corpus linguists. However, while learner corpus-based studies have noted the differences between first (L1) and second languages (L2) in the production of lexical bundles, few of them have assessed the underlying causes of such differences, particularly regarding cross-linguistic transfer. The present study investigates the use of lexical bundles in professional writing in the field of Educational Psychology produced by L1 English and L1 Russian authors in order to evaluate the evidence of cross-linguistic transfer in the writing of L2 English learners with L1 Russian background and examine the patterns of L2 English lexical bundle use that mirror L1 English production. This exploratory study compares the frequency and discourse functions of lexical bundles produced by native speakers of English to those used by Russian speakers in their L2 English professional writing, as well as professional writing in their L1. The results of the study indicate that L2 English writing produced by Russian speakers displays overlap in the composition and use of lexical bundles in L1 Russian writing pointing at possible L1 transfer.

Keywords – lexical bundles; professional writing; L1 transfer; cross-linguistic analysis; Russian; English

1. INTRODUCTION

Linguistic features of academic writing registers have been scrutinized by corpus researchers over the last few decades (see Hyland 2015). One of those features are lexical bundles (LBs), or recurrent lexical sequences identified through corpus analysis (Pan *et al*. 2016). As described by Paquot (2013), lexical bundles may be grammatically complete or incomplete phrasal (e.g., *at the same time*, *the results of the*) or clausal (e.g., *I think that*, *is used as the*) segments that fulfil certain discourse functions. As such, LBs have been found to generally act as referential markers (e.g., *at the end of*), text organizers (e.g., *as shown in figure*), and stance markers (e.g., *it is possible that*) in written registers (Biber *et al*. 2003). Corpus linguistic studies have often compared native speaker (L1) and second language (L2) learner production of LBs in academic writing in a target

language (e.g., Chen and Baker 2010; O'Donnell *et al*. 2013). One major limitation of such studies, however, is that they do not consider the native language of the writers and its possible influences on the way lexical bundles are patterned in academic texts. In other words, the common approach contrasting L1 and L2 LBs in a target language without examining the third component, academic writing in learners' L1, does not provide us with insights into the possible reasons behind the observed LB patterns. While studies have recognized the active role that cross-linguistic transfer may play in L2 writing (e.g., Bybee 2008; Paquot 2014), evidence of transfer has been limited. Moreover, the existing evidence of cross-linguistic transfer has been largely inconsistent in motivating the approaches to the assessment of transfer in previous studies and thus potentially weakening the validity of the results.

The goal of the present study is to explore the potential L1 influence in L2 English professional writing produced by Russian authors through the analysis and comparison of high-frequency LBs in three corpora of academic articles in the field of Educational Psychology published in L1 English, L2 English, and L1 Russian. The purpose is to contribute to our understanding of the use of LBs in L2 published academic writing and provide insights into the possible causes of discrepancies in the use of LBs in L1 and L2 writing. First, the study compares the patterns of LB use in L1 and L2 English writing produced by Russian writers to provide further evidence regarding the development of L2 academic writing. Second, the role of L1 influence in the use of LBs by L2 English learners is examined. More specifically, the study compares the use of LBs in two varieties of writing within one discipline: L2 English written by Russian native speakers and L1 Russian. Applying Jarvis's (2000) intra-L1-group congruity criterion of the L1 influence identification framework, the study aims to determine the extent to which the use of LBs in L2 English academic writing made by Russian native speakers differs from the L1 Russian norms. To that end, a functional analysis of LBs in the three language varieties (L1 English, L2 English, L1 Russian) is also performed to collect additional evidence of L1 influence in published L2 English writing.

To the author's knowledge, this is the first study that examines evidence of L1 English-likeness and possible L1 transfer in the use of LBs by Russian speakers of English. Importantly, the study examines expert writing from the discipline of Educational Psychology to avoid confounding "register/discipline differences with the

difference between groups of writers" (Pan *et al*. 2016: 62). More precisely, the study addresses the following research questions:

1. To what extent does L1 Russian writers' use of high-frequency LBs in L2 English writing in the field of Educational Psychology can be attributed to writers' L1 influence and/or proximity to L1 English production?
   a. What are the bundles that are shared between the three language varieties?
   b. What are the bundles that are shared between L1 and L2 English writing?
   c. What are the bundles that are shared between L2 English and L1 Russian writing? (Jarvis's (2000) Intra-L1-group congruity criterion).
2. What are the differences in discourse functions of the identified LBs used in L1 English, L2 English, and L1 Russian expert writing in Educational Psychology?

The paper starts with a brief overview of current literature on LBs in academic writing as well as the role of writers' L1 in LB use (Section 2). Section 3 follows with a description of the corpus used in the study as well as the methodology used. Finally, the results are presented and discussed in Section 4 with regard to Jarvis's (2000) framework of L1 transfer as well as within the domain of academic writing.

2. Overview of the literature

2.1. LBs in academic writing

Since the introduction of the concept of LBs, or "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber *et al*. 1999: 990), studies in corpus linguistics have examined their role in L2 writing (e.g., Simpson-Vlach and Ellis 2010; Salazar 2014). Particularly in the domain of English for Academic Purposes (EAP), researchers have agreed that learners' control of formulaic sequences, such as LBs, is essential for successful academic writing as this register exhibits "a distinct set of lexical bundles, associated with [its] typical communicative purposes" (Biber and Barbieri 2007: 265). Further investigating this argument, Hyland (2008) explored the forms, structure, and functions of LBs in a large corpus of academic writing within four disciplines. He found that bundles were not only important for academic discourse, but also for differentiating texts by discipline (Hyland 2008: 57). Increasingly, in the field of EAP, studies have used this framework to compare and analyze the use of LBs by native

speakers and L2 learners of English in academic writing. So far, it has been shown that the use of formulaic language largely depends on the language level of L2 writers. For example, Staples *et al*. (2013) examined learners' use of bundles in prompted Test of English as a Foreign Language (TOEFL) writing tasks. The study showed that high proficiency learners used fewer bundles compared to low proficiency learners, thus lending support to the hypothesis that learners move towards self-constructed rather than formulaic language with an increase in their target language proficiency (Ellis 2002).

Although LBs are likely to be observed in advanced academic writing, it is still unclear whether highly proficient L2 English learners use them effectively. Research seems to agree that learners misuse L1 English bundles and fail to understand their pragmatic functions in agreement with L1 conventions (Granger 1998; Nekrasova 2009). For example, Chen and Baker (2010) compared LBs retrieved from a corpus of published academic texts with LBs in two corpora of student academic writing (L1 and L2). The study demonstrated that L2 learners employed a smaller range of LBs in their writing; furthermore, they overused certain expressions which were rarely used by native speakers (Chen and Baker 2010: 43). Adapting Chen and Baker's (2010) methodology, Ädel and Erman (2012) investigated the use of English-language LBs in advanced learner writing in comparison with native-speaker writing. For their analysis, the researchers focused on writing by undergraduate university students in the discipline of linguistics. The study found that native speakers included a larger and more varied number of LB types in their writings, including negations, unattended *this*-bundles, existential *there*-bundles, and hedging bundles (Ädel and Erman 2012: 86).

Regarding the discourse functions of LBs in L2 writing, English learners' language production has been found to exhibit lack of register awareness, as well as phraseological and semantic misuse (Gilquin *et al*. 2007; Paquot 2014). Pan *et al*. (2016) conducted a corpus-driven analysis of LBs used by L1 English and L2 English (L1 Chinese) academic professionals writing for telecommunications research journals. The study found major structural and functional differences in LBs between L1 and L2 writing. More specifically, L1 and L2 professionals employed structurally different bundles serving similar functions (Pan *et al.* 2016: 69). On the other hand, a few studies have argued that the use of LBs in L1 and L2 academic writing is largely similar (Swales and Feak 2004; Wulff and Römer 2009). Claims have been made that even though L2 English writers

overuse high frequency LBs, they use the same amount of bundles as L1s overall (Durrant and Schmitt 2009).

In sum, many learner corpus-based studies have noted the differences in L1 and L2 production and use of LBs in discourse. Emphasizing the frequency information of L1 and L2 bundles, studies have explained patterns of overuse or underuse of LBs in learner texts (e.g., Gilquin 2008; Chen and Baker 2010). However, research has largely overlooked the possible underlying explanations for learner deviations in LB use as well as approaches to the investigation of these explanations. In other words, although the findings of the studies mentioned above are valuable in that they provide insights into the differences in the use of LBs in L1 and L2 writing, they do not necessarily investigate the possible causes behind the observed discrepancies. The following section provides an overview of current research of one of such causes, namely, L1 influence.

2.2. L1 influence in the use of LBs

It has been hypothesized that misuse of LBs in an L2 is in part related to L1 influence or transfer, defined as a statistically significant process "occurring from the native language to the foreign language" (Jarvis 2000; see also Selinker 1966: 103; Odlin 2003). One way of investigating such an influence in L2 writing has been Contrastive Interlanguage Analysis. The aim of such analysis is to identify the over- and under-use of chosen features (i.e., LBs) in L2 learners' production in order to detect L1 interference (Granger 2002; Rica Peromingo 2012). For instance, Lu and Deng (2019) compared the use of LBs in dissertation abstracts written by doctoral students who were L1 English speakers and L2 English learners from China. The four-word bundles identified in the study were categorized structurally and functionally revealing substantial differences in the frequencies of use across categories. More specifically, Chinese students demonstrated an underuse of bundles containing indefinite articles that the authors linked to the lack of the article system in Chinese. In a similar study, Esfandiari and Barbary (2017) contrasted four-, five-, and six-word LBs in psychology research articles written by L1 English and L2 English speakers from Iran. The study found that Persian writers used fewer LBs overall and in structurally and functionally different ways when compared to L1 English writers. As such, Persian writers utilized significantly more dependent clauses and significantly fewer research-oriented bundles. Additionally, the study found a substantial amount of LBs (between 20% and 25%) that were shared between the two corpora.

Finally, Pérez-Llantada (2014) compared LBs across three language varieties of expert academic writing (L1 English, L2 English written by Spanish speakers, and L1 Spanish). After analyzing the structures and functions of bundles specific to one or two language variables, she argued that the use of LBs by L2 writers deviated from L1 norms and concluded that L2 expert writers' formulaicity was 'hybrid' ―largely, but not completely, native-like (Pérez-Llantada 2014: 93).

Additional studies on the L1 influence in L2 academic writing offered further insights into the processes behind the phenomenon. Rica Peromingo (2012) investigated L1 transfer in argumentative essays by Spanish learners of English. In particular, the study looked at linking adverbial LBs that create textual cohesion (e.g., *in other words*). The learners in the study demonstrated overuse of L2 English adverbials that had a similar meaning to those used in Spanish (e.g., *in conclusion = en conclusión*). Rica Peromingo hypothesized that the structural and semantic similarity of the LBs could explain the observed transfer. L1 transfer in learners' production of LBs that are semantically and structurally similar in learners' L1 and target L2 was also supported by Allen (2011). The study suggested that the overuse of certain LBs (e.g., *it can be said (that)*) in final course research papers written by Japanese learners of English might occur due to the proximity of these bundles to similar L1 Japanese bundles. Allen (2011: 119) attributed this transfer pattern to lexical priming in one's L1 that may facilitate writing in an L2.

While the studies above have provided some evidence for possible L1 transfer in the use of LBs, this evidence is based solely on the finding that a certain construction found in L2 writing exists in learners' L1. Paquot (2013) argued that such an approach may be problematic as it involves post-hoc guessing on the side of the researcher. In order to address this issue, she examined the effects of transfer on French EFL learners' use of LBs applying Jarvis's (2000) framework for the study of L1 transfer that consists of three potential sources of transfer evidence (see Section 2.3 below). Conducting a LB analysis on the French part of the *International Corpus of Learner English* (ICLE),[1] Paquot (2013) detected that learners' application of three-word LBs in writing was associated with lexico-grammatical as well as functional frequency patterns in French. Based on these results, Paquot argued that the first language of learners may prompt them to use LBs in a way that is not typical for English. In a follow-up study, Paquot (2017) investigated the

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html

preferred use of LBs expanding the analysis in the writing of French and Spanish learners of English. Using the frequency data, Paquot found strong positive correlations between the frequency of discourse organizational and stance-oriented LBs in learners' written production and its equivalent form in the learners' L1. Making use of the same framework, Güngör and Uysal (2020) recently investigated the cross-linguistic influence of L1 Turkish on L2 English on the learners' production of four-word LBs. The study revealed that 45 percent of bundles in L2 English writing were distinctive to Turkish authors.

Taken together, previous studies pointed out deviations in learners' use of LBs. Some have compared L1 and L2 LBs and argued that learners, irrespective of their L2 proficiency levels, misuse the formulaic sequences in L2 English academic writing (e.g., Chen and Baker 2010; Salazar 2011; Ädel and Erman 2012; Esfandiari and Barbary 2017). Although these studies claimed that the misuse of LBs in L2 texts might be due to the L1 transfer, they oftentimes assumed L1 interference just based on the analysis of the L2 texts without analyzing the data in L1 (Gilquin and Paquot 2008). At the same time, those studies that included learners' L1 as another point of comparison (e.g., Pérez-Llantada 2014) have disregarded the importance of evidence that is rooted in established frameworks. Lastly, the studies that made use of such frameworks are limited to certain L1s and need to be expanded to learners from other L1 backgrounds.

2.3. L1 influence identification framework

As argued in the previous section, few studies that examined L1 transfer evidence in L2 learners' production of LBs in academic writing grounded their investigations in transfer frameworks. To this end, Paquot (2013) adapts Jarvis's (2000) framework for assessing L1 transfer. According to Paquot (2013: 393–394), the framework requires three types of comparisons to be considered by studies in order for transfer to be supported by sufficient evidence: (1) intra-L1-group homogeneity in learners' L2 performance where learners that share an L1 display similar patterns of use of a specific L2 feature; (2) inter-L1-group heterogeneity in learners' L2 performance where learners from different L1s do not share the same patterns; and (3) intra-L1-group congruity between learners' L1 and L2 performance where the comparison of learners' use of a feature in their L1 and L2 reveals similarities. In her later study, Paquot (2017), referring to Jarvis (2000: 258), emphasized that intra-L1-group congruity is the strongest type of evidence for L1 influence, as the comparison of learners'

L1 and L2 production can demonstrate L1 features that motivate patterns of use of similar features in learners' L2. Additionally, intra-L1-group congruity lends itself to a statistical approach to L1 transfer examination, which is crucial in Jarvis's framework.

3. CORPUS AND METHODOLOGY

The corpora examined in this study were comprised of research articles in the field of Educational Psychology. These articles were written by L1 English (PSY-ENG1), L2 English (PSY-ENG2), and L1 Russian (PSY-RUS1) expert writers. It is important to remember that for the sake of comparability, all of the L2 English articles were written by Russian native speakers (see below). The articles came from three major peer-reviewed journals in the field of psychology: *American Psychologist* (L1 English), *Psychology in Russia* (L2 English), and *Национальный Психологический Журнал (Nacionalniy Psihologicheskiy Zhurnal)* (L1 Russian). *American Psychologist* was chosen on the basis of its high impact factor (4.856) and the fact that it is the official journal of the *American Psychological Association*. Since impact factor is not calculated for Russian psychological journals, the other two periodicals were selected because they are published by the leading research universities in Russia. Overall, the corpora in this study were designed for contrastive descriptive research of LBs in written discourse of L1 English, L2 English, and L1 Russian academic professionals and, therefore, were made comparable with regard to register, discipline, communicative purposes, and authors' level of expertise.

One concern that emerged during the first stages of data collection was determining the first language of a writer. Following Pan *et al*. (2016: 63), L1 Russian (and thus L2 English) writers were defined as authors affiliated to an institution located in a country where Russian was spoken as the first language. Additionally, the author's first and last names had to be considered native to these countries. Articles by writers with arguable names were excluded from the corpora. The same procedure was implemented to identify L1 English writers. The final corpus structure is shown in Table 1.

	PSY-ENG1 (L1 English)	PSY-ENG2 (L2 English)	PSY-RUS1 (L1 Russian)
Number of texts	61	85	91
Average number of words per text	6,730.50	4,842.80	4,525.10
Total number of words	410,558	411,637	411,787
Total number of types	19,025	17,149	53,399
TTR	5.26	4.77	12.97
Standardized TTR	5.09	4.74	12.53

Table 1: Summary of built corpora

The process of corpus building for this study consisted of two steps. During the first step of data collection, articles published between 2017 and 2019 were downloaded for each corpus. Importantly, only research articles, descriptions or research methodology, and literature reviews were included in the corpora; that is, other types of texts published in the journals (e.g., editor's notes, reviews, opinions) were excluded from the analyses. After the extraction, all articles were cleaned of meta-data and references as well as text in languages other than the target ones. For example, if an L1 Russian article contained text in a language other than Russian, this text was removed from the article before its inclusion in the corpus. In order to match the corpora on the number of words, additional articles from 2015 and 2016 were downloaded from the journals in the PSY-ENG2 and PSY-RUS1 corpora. This resulted in three corpora with the same number of words, although slightly different text counts (see Table 1).

3.1. Identification of lexical bundles

In order to retrieve the frequency lists of bundles and compute tokens and types of LBs from the collected corpora, the study used the *Natural Language Toolkit* (NLTK) library of *Python* (Bird *et al*. 2009) and followed the LB extraction steps outlined in Ren (2021). Log-likelihood values were calculated in *R*, a free statistical environment (R Core Team 2019) and compared to establish whether the frequency of the bundles used only by L1–L2 English and the frequency of the bundles used only by L2 English–L1 Russian writers differed significantly. Significant differences in the use of similar LBs between L1 and L2 English corpora would indicate that Russian learners of English demonstrate professional writing that is different from L1 English writing. Conversely, lack of significance in the use of similar LBs between L2 English and L1 Russian corpora would suggest L1 transfer in the writing of Russian authors in English.

Three criteria were considered in the identification of LBs: bundle length, frequency, and dispersion. The study focused on four-word bundles for PSY-ENG1 and PSY-ENG2 to make the analysis more manageable and comparable to those of other studies (e.g., Chen and Baker 2010; Pérez-Llantada 2014; Pan *et al.* 2016). Moreover, this length seems to display a wider variety of structures and functions for analysis than three- and five-word bundles (Cortes 2004; Hyland 2008). Cortes (2004: 401) also noted that three-word bundles are often embedded in four-word bundles (e.g., *at the end* and *at the end of*). However, it was deemed necessary to also include three-word LBs in the process of retrieval and analysis of LBs in the PSY-RUS1 corpus. The Russian language has a rich and highly inflectional morphological system. Importantly, inflectional morphemes embedded in a word can indicate tense, voice, and number (cf. *on the other hand* vs. *с другой стороны* [*s drugoy storony*]). Moreover, some functional words, for example, definite and indefinite articles do not exist in Russian. Therefore, it is often the case that a four-word bundle in English has a three-word equivalent in Russian (*the table shows that* vs. *таблица показывает что* [*tablitsa pokazyvaet chto*]). Thus, both three-word and four-word LBs were analyzed from the PSY-RUS1 corpus.

As for the criterion of LB frequency, recent studies made use of varied thresholds ranging between 20–40 times per million words (e.g., Biber *et al*. 2004; Hyland 2008; Chen and Baker 2010). For this study, a high cut-off of 40 per million was set. This threshold is helpful in filtering out content bundles as well as bundles containing discipline-specific nouns (Ädel and Erman 2012; Pérez-Llantada 2014). The dispersion criterion for this study was set at 10 percent. This means that a lexical bundle had to appear in at least 10 percent of the texts in a corpus to be considered for inclusion in the analysis. Previously, researchers have chosen different dispersion criteria for their studies varying between three to five texts in a corpus (Biber and Conrad 1999; Chen and Baker 2010; Ädel and Erman 2012). Pan *et al*. (2016), for example, established a LB dispersion threshold of five texts for an 87-text corpus (5.7%) and ten texts for a 179-text corpus (5.6%). Although this approach is effective for comparing corpora with the same number of texts, it can present a methodological problem if the corpora are not matched for this number (Hyland 2008). Setting a percentage dispersion threshold was especially important for the second step of lexical bundle extraction in this study since the three corpora differed in the number of texts (see Table 1). The established dispersion threshold was also considered adequate given the previous practices.

LBs for the analysis were identified on the basis of their word forms and not lemmas. In other words, inflected variants of the same lemma were treated independently. This decision was especially important in the case of the PSY-RUS1 corpus since, as mentioned above, Russian has a highly inflectional morphology, and the identification of LBs based on lemmas might have caused loss of important comparison points between the corpora. The retrieved bundles were checked manually for the remaining area-specific content bundles. Content bundles involving proper nouns (*American Psychological Association*) were excluded and the bundles related to conducting research in general (e.g., *majority of the informants*) were kept. Following Chen and Baker (2010:33), the overlapping bundles in the PSY-RUS1 list were merged; thus, three-word bundles that were parts of four-word bundles in the list and occurred with the exact same dispersion and frequency were merged. For example, the three-word bundle *то же время* (*to zhe vremya*) 'the same time' appeared in 23 texts and had a frequency of 150 words per million. A similar four-word bundle *в то же время* (*v to zhe vremya*) 'at the same time' has the same dispersion and frequency. Therefore, the two overlapping bundles are combined into *(в) то же время*+ (*(v) to zhe vremya*+) / '(at) the same time+' in the final list. The merged bundles are indicated with a plus (+) sign in the compete lists provided in Appendix 1.

3.2. Application of Jarvis's (2000) framework for additional L1 transfer evidence

As mentioned above, to provide further statistical evidence of L1 influence on Russian L2 English writers' production of LBs, the study used the L1 transfer assessment framework proposed by Jarvis (2000). Following Paquot's (2017: 6) claim that the intra-L1-group congruity between learners' L1 and L2 performance presents the strongest type of evidence for L1 influence (also Jarvis 2000: 258) and for the sake of feasibility, the present study made use of this effect to further examine L1 transfer in Russian writers' LB use in L2 English writing. As Paquot (2013: 400) notes, the simplest way to test the intra-L1-group congruity criterion is to check whether there are bundles that are shared between learners' L1 and L2 writing. Thus, frequent LBs in PSY-ENG2 and PSY-RUS1 were compared for the presence of overlapping bundles.

3.3. Translation and analysis of bundles

To single out the bundles shared in the three language varieties as well as bundles shared by only PSY-ENG2 and PSY-RUS1 (Jarvis's (2000) intra-L1-group congruity), the L1 Russian LBs were translated into English by two researchers (the author and another applied linguistics scholar) whose native language was Russian and who had done similar translation work before. The translations from Russian to English were done with the help of the *Collins Russian-English Dictionary*.[2] Importantly, the translations were maintained as close as possible to the original. In other words, the researchers aimed at word-for-word translations; however, in cases where it was not possible, a lexical bundle with the most similar meaning was used. The translations provided by both researchers were compared in order to ensure the validity of the English equivalents for the Russian LBs. All discrepancies were discussed and resolved reaching 100 percent agreement between the two translators. The three LB lists were then compared manually. Log-likelihood analyses were performed with the bundles shared between PSY-ENG1 and PSY-ENG2, as well as between PSY-ENG2 and PSY-RUS1, to find out the significant differences in bundle frequencies in these language varieties. LBs unique to only one corpus were also identified.

After the quantitative analysis regarding the LBs extracted from the three corpora, the 50 most frequent bundles in each list were classified. Biber *et al*.'s (2004) framework (modified by Hyland 2008 and Pan *et al*. 2016) was used to compare the LBs based on their discourse functions. LBs were classified into three major categories: research-oriented (parallel to 'referential' bundles in Biber's *et al*. (2004) framework), text-oriented (parallel to 'discourse-organizing'), and stance-oriented bundles. Bundles identified as research-oriented were those that explained the procedures in a study as well as its structure (e.g., *at the same time*). Text-oriented bundles (e.g., *in addition to*) were those involved in organization of the text of an article and its argumentative elements. Finally, stance-oriented bundles (e.g., *it is possible that*) had the function of conveying an author's evaluation and attitude towards the reported information. The bundles were first classified by two raters trained in the field of corpus linguistics and familiar with the framework. The initial agreement rate between the raters was 82 percent. After an inter-rater norming session was held, disagreements in functional identification of LBs were resolved resulting in 100 percent agreement. As the final step of the functional analysis,

[2] https://dictionary.reverso.net/russian-english/

Chi-square tests were also performed to check for significant differences in the functional distribution of bundle types in the three corpora.

4. Results and discussion

The established frequency and dispersion cut-offs resulted in 82 bundles identified in PSY-ENG1, 223 bundles in PSY-ENG2, and 264 bundles in PSY-RUS1. Appendix 1 provides a complete list of the extracted LBs with their frequencies normalized per million words (pmw). Overall, the amount of LBs retrieved from the three corpora supports the view that the academic written register can be clearly characterized by formulaicity and fixedness of expressions (Pérez-Llantada 2014). If compared to previous research in the area of LBs in academic writing, Cortes (2004) reported 54 frequent bundles in her corpus of writing in history and 109 bundles in biology writing. Pérez-Llantada (2014) was able to retrieve a total of 56 bundles in L1 English, 77 in L2 English, and 114 in L1 Spanish writing. With regard to the total number of LBs in the three corpora, L1 English writing displayed the lowest amount of frequent LBs (83), especially since both L2 English and L1 Russian writing contained more than twice the amount of bundles (227 and 264). A similar trend was displayed in Hyland (2008) and Römer (2009) with L2 writers producing a larger number of bundles than L1 English writers. This finding offers support to Ellis (2002) who suggested that L2 production is oftentimes more formulaic than L1 production. Additionally, the finding also seems to support the hypothesis expressed by Pérez-Llantada (2014), who suggests that an observed wider range of bundles can be interpreted in terms of lexical variety of a given language. Thus, the fact that PSY-RUS1 showed the highest total number of word types (53,399), Type-Token Ratio (TTR) (12.97), and Standardized Type-token Ratio (STTR) (12.53) compared to PSY-ENG1 and PSY-ENG2, as indicated in Table 1, could be viewed as indirect evidence for the lexical richness of the Russian language and, consequently, the higher number of the extracted LBs. However, this hypothesis does not explain the large number of bundles in PSY-ENG2 with the word types, TTR, and STTR being close to PSY-ENG1. Another explanation for the differing numbers of frequent LBs can be the possibility of L1 influence in writing (Paquot 2014). Russian learners of English might be adapting some of the LBs from their native language into L2 English writing. Finally, it is also possible that because the PSY-ENG1 corpus included a smaller number of texts,

it yielded fewer bundles despite the same dispersion cut-off (see Chen and Baker 2010: 43).

4.1. Core bundles

To identify the bundles that were shared between all three corpora, the extracted LBs were compared manually. A total of six bundles were shared between three corpora, representing 7.3 percent of the bundles in L1 English writing, and three percent in L2 English writing as well as in L1 Russian writing (see Table 2). It can be assumed that these core bundles are extremely useful in both English and Russian for various discourse purposes. Supporting Pan *et al*. (2016: 68), the majority of these core bundles serve the text-organizing function (*at the same time*, *as well as the*, *in the case of*), with two bundles functioning as research-organizers (*at the end of*, *is one of the*) and one bundle having a stance function (*it is important to*).

Interestingly, some of these bundles had differing normalized frequencies; for instance, *at the same time* was the most frequently occurring LB in the PSY-ENG2 corpus (303 pmw), but barely met the threshold in the PSY-ENG1 corpus (20 pmw). In contrast, *it is important to* appeared 118 times per million words in PSY-ENG1 and only 32 times in PSY-RUS1.

The use of the core LBs in L2 professional writing might extend on more than just the two languages under analysis. After comparing the core LBs in our study to those in Chen and Baker (2010) and Ädel and Erman (2012), five out of six bundles overlapped in the two studies. The only exception was *at the end of*, which was identified as a shared lexical bundle only in Ädel and Erman (2012). Recall that both studies compared L1 English academic writing to learner writing in by native speakers of other languages (Swedish and Chinese). It seems, therefore, that these core bundles are acquired by L2 English writers with different L1 backgrounds and are not indicative of L1 transfer.

Lexical bundle	Frequency, pmw		
	PSY-ENG1	PSY-ENG2	PSY-RUS1
it is important to	118	75	32
as well as the	60	170	85
at the end of	38	63	23
in the case of	25	168	32
is one of the	24	113	49
at the same time	20	303	150

Table 2: Bundles shared by all three corpora

4.2. Bundles shared in L1 English and L2 English

A total of 17 bundles were found to overlap in L1 and L2 English writing, representing 20.3 percent of the L1 English writing and 7.5 percent of the L2 English writing. If we add these bundles to the core bundles shown in Table 3, PSY-ENG1 and PSY-ENG2 share a total of 23 LBs (28% and 10.1% of L1 and L2 writing respectively). It appears that this amount of overlap in bundles is quite large, especially in comparison to the results by Chen and Baker (2010), who found 16 percent of LBs overlapping between L1 and L2 English writing.

Lexical bundle	Frequency, pmw			
	PSY-ENG1		PSY-ENG2	
in the context of	135		142	
it is important to	**118***		75	
as well as the	60	**170***		
one of the most	48	**97** *		
in the form of	41		72	
at the end of	38	**63***		
in the development of	38		35	
it is possible that	38		38	
with respect to the	38		35	
in addition to the	33		28	
the nature of the	33		28	
as a result of	31	**72***		
the context of the	31		38	
in terms of the	25		28	
in the case of	25	**168***		
it is possible to	24	**69***		
is one of the	24	**113***		
and the development of	21		35	
it should be noted (that)+	21	**91***		
the importance of the	21		22	
at the time of	20		28	
for the development of	20	**85***		
at the same time	20	**303***		

* = significant at $p<0.05$

Table 3: Bundles shared only between PSY-ENG1 and PSY-ENG2

Following Simpson-Vlach and Ellis (2010), log-likelihood values were calculated for the overlapping bundles. The list of overlapping bundles is presented in Table 3 together with the results of the log-likelihood analysis with the core bundles underlined and the numbers in bold indicating overuse. The log-likelihood statistics indicate that L2 English writing displays an overuse of some of the shared bundles (e.g., *as well as the*, *in the case of*, *at the same time)* including all of the core bundles. Similar findings were reported by Ädel and Erman (2012) who found that shared LBs were overused in L2 writing. It may be the case that L2 writers are more familiar with these bundles and feel confident using them in writing (Granger and Rayson 1998; Pérez-Llantada 2014). Ellis (2008) also suggests that L2 writers might have memorized these LBs and routinized them in their writing. Only one bundle (*it is important to*) was underused in the PSY-ENG2 corpus. This underuse may be due to the fact that Russian academic writers tend to use fewer stance bundles, as illustrated in the functional analysis in Section 4.5 below, pointing at possible L1 transfer.

Compared to the complete PSY-ENG1 and PSY-ENG2 lexical bundle lists, the data seem to support Swales's (2005: 10) and Ädel and Erman's (2012) observation that attended *this*-bundles with the meta-discursive head nouns (*of this study is*, *this point of view*, *the results of this study*) are more common in non-native writing.

4.3. Bundles shared in L2 English and L1 Russian (intra-L1-group congruity)

To further investigate the L1 transfer evidence within Jarvis's (2000) framework, the extracted lexical bundle lists were compared to find bundles that overlapped in PSY-ENG2 and PSY-RUS1 (see Table 4). A total of 22 bundles were shared between PSY-ENG2 and PSY-RUS1 corpora, which comprise 9.7 percent of the frequent bundles in L2 English writing and 8.3 percent in L1 Russian writing. If merged with the core bundles, there is a total of 28 bundles shared between the two corpora (12.3% and 10.6% in PSY-ENG2 and PSY-RUS1, respectively). The overlapping LBs between L2 English writing produced by Russians and L1 Russian writing further suggest the possibility of L1 influence. Yet, the significant log-likelihood values of the overlapping bundles presented in Table 3 indicate, in accordance with Pérez-Llantada's (2014) findings, that very few bundles in L2 English writing are used in a Russian native-like manner. This suggests that L2 writers' usage of bundles is not fully native-like and represents a combination of both L1 English and L1 Russian academic writing. At the same time, bundles like *in the*

present study / в данном исследовании (*v dannom issledovanii*), *we can say that / мы можем сказать что* (*my mozhem skazat chto*), or *an important role in / важную роль в* (*vazhnuyu rol v*) do not significantly differ in their use in L2 English and L1 Russian pointing at the possible L1 transfer, especially since these bundles do not occur in the corpus of L1 English writing (see Appendix 1). Even those LBs that are used in L1 Russian writing significantly more often than in L2 English writing (e.g., *on the one hand / с одной стороны* [*s odnoy storony*], *in this case the / в этом случае* [*v etom sluchaye*], *and on the other / а с другой* [*a s drugoy*]) are potentially indicative of cross-linguistic transfer as they do not appear in L1 English writing at all.

Some other important observations about L1 Russian LBs emerged after examining the lists more closely. Similarly to English, a lot of four-word Russian bundles had three-word bundles embedded in them (e.g., *свидетельствуют о том (что)+* [*svidetelstvuyut o tom (chto)+*] */ indicate (that), вывод о том (что)+* [*vyvod o tom (chto)+*] */ conclusion about that (that)+*, *несмотря на то (что)+* [*nesmotrya na to (chto)+*] */ despite the fact (that)+*, *так же как (и)+* [*tak zhe kak (i)+*] */ same as (and), (с) нашей точки зрения+* [*(s) nashey tochki zreniya+*] */ (from) our point of view*). It seems that this embedding is dictated by the syntactic structure of the language: the words in brackets in the examples are prepositions and conjunctions that are, in most cases, required by the words they follow or precede.

Bundles like *вывод о том (что)+* (*vyvod o tom (chto)+*) */ conclusion about that (that)+* deserve special attention in this study. In this bundle, the demonstrative pronoun *том* (*tom*) ('that' in prepositional case) acts as the head noun in the noun phrase of the prepositional phrase *о том* (*o tom*; 'about that'). This prepositional phrase can be roughly translated as 'about the fact' (*о том факте* [*o tom factye*]) with the Russian version being an acceptable and widely used phrase. However, the noun *fact* is often omitted in Russian because it is contextually predictable and, therefore, redundant (Jaeger and Tily 2011: 328). In PSY-RUS1, 22 out of 231 bundles (9.5%) had a similar structure with the pronoun *то/том* (*to/tom*; 'that'/'that' in prepositional case) taking the place of the noun. Interestingly, there were seven bundles in PSY-ENG2 that contained the word *fact* (*the fact that the, to the fact that, by the fact that, due to the fact that, in the fact that, of the fact that, explained by the fact*) and none in PSY-ENG1. This finding serves as another indicator that L1 transfer may be happening in L2 writing. It is noteworthy that apart from the 22 shared bundles there were cases when the bundles closely resembled each other in

PSY-ENG2 and PSY-RUS1 LB lists. For example, the PSY-ENG2 bundle *in the present study* did not occur in PSY-RUS1, but it is very similar to an L1 Russian bundle *в данной статье* (*v dannoy statye* 'in the present article'). Corresponding cases include L2 English bundles *the study showed that, we assume that, in his opinion* that have close equivalents in L1 Russian writing.

The comparison of PSY-ENG2 and PSY-RUS1 bundles also provided further methodological considerations with regard to the length of *n*-grams in English and Russian. It has been argued above that a four-gram is the most commonly studied length of *n*-grams in most studies on lexical bundles. However, the present analysis revealed that the length of similar *n*-grams in Russian and English often does not match. When the retrieved three- and four-grams in PSY-RUS1 were translated into English, their length changed; many three-grams in Russian became one-grams in English (*в том числе* [*v tom chisle*] *including*, *в настоящее время* [*v nastoyashee vremya*] */ currently*, *тем не менее* [*tem ne menee*] *nevertheless*, *включает в себя* [*vklyuchayut v sebya*] *includes*, *на сегодняшний день* [*na segodnyashniy den*] / *nowadays*, *по всей видимости* [*po vsey vidimosti*] */ evidently*, *в последнее время* [*v poslednee vremya*] *recently*). A few bundles also became longer after being translated from Russian into English (*важно отметить что* [*vazhno otmetit chto*] */ it is important to note that*, *следует подчеркнуть что* [*sleduet podcherknut chto*] */ it should be emphasized that*, *это связано с* [*eto svyazano s*] */ it is connected to*). Therefore, it is evident that a larger range of LB lengths needs to be included in cross-linguistic bundle studies, especially when one of the compared languages is so morphologically rich, as is the case with Russian.

Lexical bundle	Frequency, pmw			
	PSY-ENG2		PSY-RUS1	
at the same time / в то же время	**303***			150
a high level of / высокий уровень того	**224***			39
as well as the / так же как и	**170***			85
in the case of / в том случае	**168***			32
on the other hand / с другой стороны	**145***			105
in the process of / в процессе того	**142***			39
is one of the / является одним из	**113***			49
with a high level / с высоким уровнем	**110***			39
with the help of / с помощью того	**101***			29
on the one hand / с одной стороны		94	**153***	
it is important to / является важным то	**75***			32
at the end of / в конце того	**63***			23
are presented in table / представлены в таблице	**56***			35
in this case the / в этом случае		50	**91***	
as well as to / так же как и чтобы		38		39
(it) can be assumed that +/ можно предположить что		38	**91***	
in the present study / в данном исследовании		38		39
(at) the same time they + / в то же время они		38		23
as well as in / так же как и в		35	**85***	
in the fact that / в том что		32		37
we can say that / мы можем сказать что		32		26
an important role in / важную роль в		28		26
as well as a / так же как и		28		23
and on the other / а с другой		25	**42***	
as well as their / так же как и их		25		20
not only in the / не только в		20	**55***	
to a lesser extent / в меньшей степени		20	**35***	

* = significant at $p<0.05$

Table 4: Bundles shared only between PSY-ENG2 and PSY-RUS1

4.4. Functional classification

To answer the second question in the study, the first 50 bundles in the three lists were classified according to their discourse function in the articles. The complete analysis of the first 50 bundles can be found in Appendix 2. As seen in Figure 1, PSY-ENG1 and PSY-ENG2 display similar proportions of the three main functional categories. Research-oriented bundles constitute the largest category in both corpora, with 42 percent and 60 percent respectively, whereas stance bundles comprise 22 percent and 12 percent of the 50 most frequent LBs in the two corpora. Similarly, Pérez-Llantada (2014) found that the bundles shared by L1 and L2 English most commonly perform a referential function. Turning to PSY-RUS1 bundles, text-oriented LBs clearly rank as the largest category with 72 percent, followed by research-oriented bundles (18%) and stance bundles (10%). This LB distribution partially supports Pan *et al*. (2016), who also found stance to be the smallest functional category in L1 Chinese writing; however, the text-oriented category

was the largest one in L1 and L2 English writing in contrast to what happens in the current study, which shows the dominance of research-oriented bundles in L1 and L2 English.

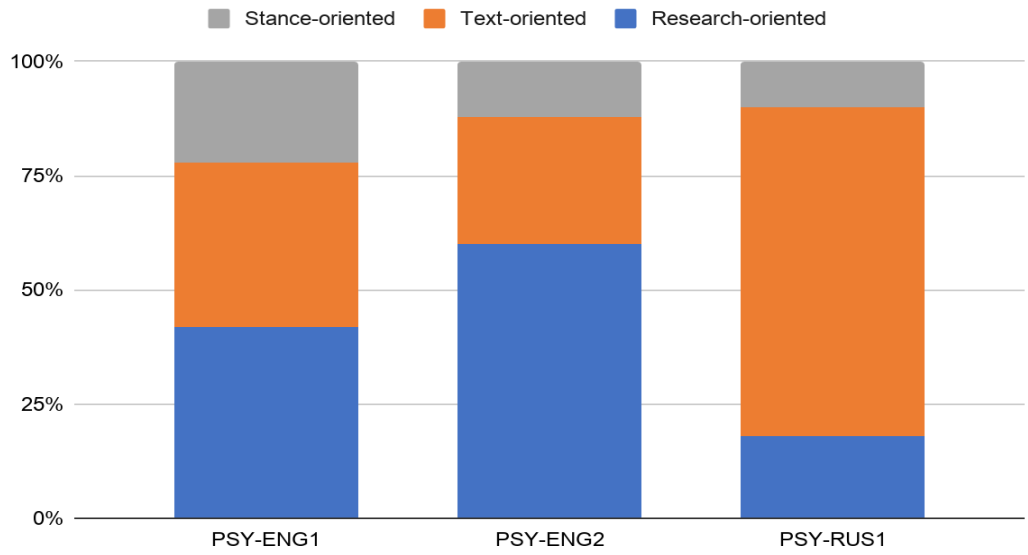


Figure 1: Functional distribution of the 50 most frequent bundles in PSY-ENG1, PSY-ENG2, and PSY-RUS1

Table 5 presents the results of a Chi-square test showing a significant medium-sized difference in the functional distribution of the bundles between the three corpora ($\chi 2$= 22.71, 4 df, Cramer's V = 0.272, p<0.05).

			Function		**Total**
		Research-Oriented	**Text-oriented**	**Stance-oriented**	
PSY-ENG1	Count	21	18	11	50
	Expected Count	20	22.7	7.3	50
	Adjusted Residual	0.4	-1.6	1.8	
	Probability value	0.4839	0.1615	0.3681	
PSY-ENG2	Count	30	14	6	50
	Expected Count	20	22.7	7.3	50
	Adjusted Residual	3.5	-3	-0.7	
	Probability value	0.0005	0.0027	0.4839	
PSY-RUS1	Count	9	37	4	50
	Expected Count	20	22.7	7.3	50
	Adjusted Residual	-3.9	4.6	-1.1	
	Probability value	0.00009	0.000004	0.2713	

Table 5: Results of the Chi-square test of the functional distribution of the 50 most frequent bundles in PSY-ENG1, PSY-ENG2, and PSY-RUS1

To find out where exactly the significance lies, post-hoc tests were conducted, and adjusted residuals and probability values were calculated and compared to the

Bonferroni-adjusted $p<0.005$. As revealed in Table 5, the distribution of functions was not significantly different from the expected counts in PSY-ENG1 bundles; however, research-oriented and text-oriented bundles were found significantly more frequent in L1 Russian writing ($p<0.005$). It is noteworthy that the results of the Chi-square test for these two functions were also significant with regard to L2 English bundles, although the distribution was the opposite. There were significantly fewer text-oriented bundles and more research-oriented bundles than expected.

Looking closely at the functional subcategories of the bundles (available in Appendix 2), we can notice that, in line with Hyland (2008), research-oriented bundles in the three corpora are represented by the following subcategories: description, location, quantification, procedure, and topic. It seems that L1 English and L2 English professional writers make use of description bundles more often, focusing on providing identification of new information for the readers (Biber 2009). Research-oriented bundles in L1 Russian writing are remarkably less common. However, within the subcategories of text-oriented bundles, there is a prevalence of transition signals in L1 Russian writing with 72.4 percent of all the bundles in this category. This subcategory is also the largest one in L2 English writing (42.6%). The main function of text-oriented bundles is to establish textual cohesion through signaling transition or discussion of results, framing the discussion, and guiding the reader through the overall structure of the article. In other words, these bundles can be described as meta-discourse (Ädel and Erman 2012). It has been previously reported (Ädel 2006) that L2 learners tend to overuse meta-discourse in academic writing; however, this is not the case in my data, perhaps due to a higher L2 proficiency of expert writers.

In contrast, framing signals are the most prominent subcategory in L1 English writing. Framing is the only subcategory in text-oriented bundles where L2 writers use bundle tokens significantly less frequently than L1 writers do (Pan *et al*. 2016). With regard to this subcategory, it is interesting to note that two out of five LBs used by L2 English writers overlap with L1 English writing (*in the context of* and *in the case of*). Römer (2009) and Chen and Baker (2010) noticed that the bundle *in the context of* was rarely used by novice L2 learners; however, it is highly frequent in the PSY-ENG2 corpus (146 times pmw). This, again, may indicate that the use of LBs becomes more native-like with the growing proficiency of L2 professional writers. Additionally, the high frequency

of the bundle might have occurred due to writers' L1 influence, although an equivalent bundle was not detected in the PSY-RUS1 corpus.

Stance features and engagement features were used to convey the author's interpretation in professional writing, but the proportions were somewhat small, as noted in previous research (Biber *et al*. 2004; Hyland 2008; Chen and Baker 2010). Although the distribution of stance-oriented bundles was not found significantly different from the expected counts, it is still noteworthy that the PSY-ENG1 list contained almost twice as many stance bundles as PSY-ENG2 and PSY-RUS1. Similar observations about the lack of control of formulaic language expressing stance in L2 professional writing were made by previous studies (Ellis 2008; Granger and Meunier 2008; Pérez-Llantada 2014). This lack of control may be attributed to L1 syntactic and lexical transfer. If we compare the stance-oriented bundles in L2 English and L1 Russian writing, several similarities emerge, the most outstanding being the use of quite a direct noun *fact* (*the fact that the, to the fact that,* and *мот факт что* [*tot fakt chto*] */ the fact that*). It seems, therefore, that the stance feature bundles in L2 English and L1 Russian writing might not display enough hedging. Pérez-Llantada (2014) hypothesized that the paucity of stance meanings that builds a potentially face-threatening discourse can be attributed to the mismatch of L1 pragmatic norms. Pragmatic mismatches have also been reported in Philippine scholars (Salazar 2011: 193) and in Finnish undergraduates who show less variation in stance bundles than their L1 English counterparts (Ädel and Erman 2012). As explained in Granger (1998) and Chen and Baker (2010), the L2 English writers use fewer hedges because they have not acquired full pragmatic competence yet. At the same time, the presence of overlapping stance bundles in PSY-ENG1 and PSY-ENG2 (*it is important to*, *it is possible that*) points at a developing proficiency in L2 English writing and suggests that the use of stance bundles in L2 English writing is influenced by both L1 English and L1 Russian distribution of LBs.

5. Conclusion and future directions

The present study explored the use of LBs in L1 and L2 professional writing in the field of Educational Psychology. In particular, the study investigated the nature and functions of LBs in L1 and L2 English, as well as L1 Russian articles, in an effort to examine the similarities between L1 English and L2 English writing and detect possible evidence for

cross-linguistic transfer between L1 Russian and L2 English writing produced by Russian speakers.

Regarding the first research question that centered around the frequency evidence of cross-linguistic transfer, the results indicated that Russian authors display some evidence of L1 transfer in their L2 English writing (Bybee 2008; Paquot 2014). Specifically, the intra-L1-group congruity evidence collected within the framework proposed by Jarvis (2000) showed that a number of identified bundles were shared between L2 English and L1 Russian writing and did not occur in L1 English writing. A similar trend was uncovered in Güngör and Uysal (2020), where bundles specific to Turkish learners of English constituted almost 50 percent of the LB list. Further evidence of transfer was found in the functional analysis of LBs. That is, the high-frequency bundles in L2 English and L1 Russian writing included fewer stance-oriented LBs than in L1 English. Additionally, within text-oriented bundles, transition signals were the largest subgroup proportionally compared to L1 English bundles, where framing was the most common function of text-oriented LBs. On the other hand, L1 and L2 English writing demonstrated similar distribution of functions overall with research-oriented bundles being the largest category, while text-oriented bundles were the most common in Russian. Finally, my analysis revealed a list of core bundles that were shared among L1 and L2 English speakers. Thus, the study offered some evidence for cross-linguistic transfer in English writing produced by Russian authors, although it was not pervasive in the analysis of the extracted LBs and their functions.

The corpus-driven approach of the study supported the current research in LBs, showing that formulaic sequences are a fundamental feature of the academic register across language variables. However, the number of frequent LBs was found higher in PSY-ENG2 and PSY-RUS1 writing in comparison to PSY-ENG1. This result disconfirms previous research (e.g., Chen and Baker 2010; Ädel and Erman 2012), which found that non-native speakers possess a more restricted inventory of bundles than native speakers. Thus, the present study contributes to a unique strand of research (cf. Pérez-Llantada 2014) that uses corpus evidence to demonstrate that the L2 English writing reflects a ‘hybrid’ nature of formulaic language. In this study, L2 English displays a small number of register-determined bundles also shared by L1 English and L1 Russian. At the same time, it also includes a considerable percentage of formulaic sequences used by the L1 English writers as well as bundles transferred from L1 Russian. Furthermore, through

the functional analysis of the most frequent LBs it was found that both L2 English and L1 Russian employ fewer stance-oriented bundles, and the number of text-oriented bundles is closer between L1 and L2 English writing. Finally, one cannot forget about the case of *fact* in PSY-RUS1 writing that seems to influence the composition of LBs in PSY-ENG2. In brief, L2 English professional writing is partly, but not fully, native-like, possibly due to cross-linguistic influences from the writers' L1.

The present exploratory study poses several directions for future research. To control for content-specific bundles, the study only focused on one discipline. However, research with monolingual corpora has empirically confirmed the existence of 'discipline-sensitive' bundles in the context of research article writing (e.g., Cortes 2004; Hyland 2008). It would be worth conducting interlinguistic comparison of bundles across the disciplines to determine what bundles are specific to those disciplines and what discourse functions these bundles perform in L1 and L2 writing. It would also be of theoretical interest to further investigate the hybrid formulaic nature of L2 English research articles in languages other than Russian and Spanish (Pérez-Llantada 2014). With regard to methodology, another limitation of the current study has been the absence of a L2 English corpus that was produced by learners with the L1 background different from Russian. While the study was able to make use of previous comparable research that identified LBs in order to meet one of the criteria in Jarvis's (2000) framework (intra-L1-group congruity), a corpus built specifically for the study would facilitate a more fine-grained search of the bundles that could be shared between L2 English learners from different backgrounds and thus contribute to our understanding of L1 transfer in Russian by providing the other two types of evidence from Jarvis (2000). It also needs to be stressed that only four-word bundles were considered in PSY-ENG1 and PSY-ENG2. The process of translation of Russian bundles into English showed that some of the translated LBs did not match in length to the original. It was often the case that a three-word Russian bundles could be translated as one word in English. Thus, a fuller picture of the use of formulaic language across the corpora could have been given if more bundle lengths had been included. Finally, as pointed out by one of the reviewers, comparative analyses of translations are inherently problematic as not all LBs have exact equivalents between languages.

REFERENCES

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Ädel, Annelie and Britt Erman. 2012. Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31/2: 81–92.

Allen, David. 2011. Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education* 1: 105–127.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics* 14/3: 275–311.

Biber, Douglas and Frederica Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26/3: 263–286.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. *Language and Computers* 26: 181–190.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2003. Lexical bundles in speech and writing: An initial taxonomy. In Andrew Wilson, Paul Rayson and Tony McEnery eds. *Corpus Linguistics by the Lune.* Bern: Peter Lang, 71–93.

Biber, Douglas, Susan Conrad and Viviana Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25/3: 371–405.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Bird, Steven, Ewan Klein and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit.* Sebastopol, California: O'Reilly Media Inc.

Bybee, Joan. 2008. Usage-based grammar and second language acquisition. In Peter Robinson and Nick C. Ellis eds. *Handbook of Cognitive Linguistics and Second Language Acquisition*. London: Routledge, 226–246.

Chen, Yu-Hua and Paul Baker. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14/2: 30–49.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23/4: 397–423.

Durrant, Philip and Norbert Schmitt. 2009. To what extent do native and non-native writers make use of collocations? *IRAL-International Review of Applied Linguistics in Language Teaching* 47/2: 157–177.

Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in Second Language Acquisition* 24/2: 143–148.

Ellis, Nick C. 2008. The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. *The Modern Language Journal* 92/2: 232–249.

Esfandiari, Rajab, and Fatima Barbary. 2017. A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes* 29: 21–42.

Gilquin, Gaëtanelle. 2008. Combining contrastive and interlanguage analysis to apprehend transfer: Detection, explanation, evaluation. In Gaëtanelle Gilquin, Szilvia Papp and María Belén Díez-Bedmar eds. *Linking up Contrastive and Learner Corpus Research.* Leiden: Brill, 1–33.

Gilquin, Gaëtanelle and Magali Paquot. 2008. Too chatty: Learner academic writing and register variation. *English Text Construction* 1/1: 41–61.

Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6/4: 319–335.

Granger, Sylviane. 1998. Prefabricated patterns in advanced EFL writing: Collocations and formulae. In Anthony P. Cowie ed. *Phraseology: Theory, Analysis, and Applications*. Oxford: Clarendon Press, 145–160.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson eds. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching.* Amsterdam: John Benjamins, 3–33.

Granger, Sylviane and Fanny Meunier. 2008. Phraseology in foreign language learning and teaching. In Fanny Meunier and Sylviane Granger eds. *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins, 15–20.

Granger, Sylviane and Paul Rayson. 1998. Automatic profiling of learner texts. In Sylviane Granger ed. *Learner English on Computer*. London: Routledge, 119–131.

Güngör, Fatih and Hacer Hande Uysal. 2020. Lexical bundle use and crosslinguistic influence in academic texts. *Lingua* 242: 102859. https://doi.org/10.1016/j.lingua.2020.102859

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27/1: 4–21.

Hyland, Ken. 2015. *Teaching and Researching Writing*. London: Routledge.

Jaeger, T. Florian and Harry Tily. 2011. Language processing complexity and communicative efficiency. *WIRE: Cognitive Science* 2/3: 323–335.

Jarvis, Scott. 2000. Methodological rigor in the study of transfer: Identifying L1 influence in them interlanguage lexicon. *Language Learning* 50/2: 245–309.

Lu, Xiaofei and Jinlei Deng. 2019. With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. *Journal of English for Academic Purposes* 39: 21–36.

Nekrasova, Tatiana M. 2009. English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59/3: 647–686.

Odlin, Terrence. 2003. Cross-linguistic influence. In Catherine J. Doughty and Michael H. Long eds. *The Handbook of Second Language Acquisition*. Oxford: Blackwell Publishing, 436–486.

O'Donnell, Matthew Brook, Ute Römer and Nick C. Ellis. 2013. The development of formulaic sequences in first and second language writing. *International Journal of Corpus Linguistics* 18/1: 83–108.

Pan, Fan, Randi Reppen and Douglas Biber. 2016. Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes* 21: 60–71.

Paquot, Magali. 2013. Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics* 18/3: 391–417.

Paquot, Magali. 2014. Cross-linguistic influence and formulaic language: Recurrent word sequences in French learner writing. *EUROSLA Yearbook* 14/1: 240–261.

Paquot, Magali. 2017. L1 frequency in foreign language acquisition: Recurrent word combinations in French and Spanish EFL learner writing. *Second Language Research* 33/1: 13–32.

Pérez-Llantada, Carmen. 2014. Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes* 14: 84–94.

R Core Team. 2019. *R*: *A Language and Environment for Statistical Computing. R* Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Ren, Junqiang. 2021. Variability and functions of lexical bundles in research articles of applied linguistics and pharmaceutical sciences. *Journal of English for Academic Purposes* 50, 100968. https://doi.org/10.1016/j.jeap.2021.100968

Rica Peromingo, Juan Pedro. 2012. Corpus analysis and phraseology: Transfer of multi-word units. *Linguistics and the Human Sciences* 6/1–3: 321–343.

Römer, Ute. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7/1: 140–162.

Salazar, Danica. 2011. *Lexical Bundles in Scientific English: A Corpus-based Study of Native and Non-native Writing*. Barcelona: University of Barcelona dissertation.

Salazar, Danica. 2014. *Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching*. Amsterdam: John Benjamins.

Selinker, Larry. 1966. *A Psycholinguistic Study of Language Transfer*. Washington: Georgetown University dissertation.

Simpson-Vlach, Rita and Nick C. Ellis. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31/4: 487–512.

Staples, Shelley, Jesse Egbert, Douglas Biber and Alyson McClair. 2013. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes* 12/3: 214–225.

Swales, John M. 2005. Attended and unattended "this" in academic writing: A long and unfinished story. *ESP Malaysia* 11/1: 1–15.

Swales, John M. and Christine B. Feak. 2004. *Academic Writing for Graduate Students: Essential Tasks and Skills, Vol. 1*. Michigan: University of Michigan Press.

Wulff, Stefanie and Ute Römer. 2009. Becoming a proficient academic writer: Shifting lexical preferences in the use of the progressive. *Corpora* 4/2: 115–133.

Corresponding author
Maria Kostromitina
Department of English
Northern Arizona University
705 S Beaver St.,
Flagstaff, AZ, 86011, USA
e-mail: mk782@nau.edu

received: October 2021
accepted: June 2022

APPENDICES

Appendix 1: Complete lexical bundle list.

PSY-ENG1

Rank	Frequency	Freq. per mil.	Range	Bundle
1	122	294	9	*the online supplemental materials*
2	79	190	7	*in the online supplemental*
3	56	135	26	*in the context of*
4	49	118	26	*it is important to*
5	33	79	20	*are more likely to*
6	28	67	17	*the extent to which*
7	25	60	20	*as well as the*
8	25	60	13	*more likely to be*
9	24	58	8	*online supplemental materials for*
10	23	55	15	*a wide range of*
11	21	48	10	*one of the most*
12	21	50	6	*see the online supplemental*
13	18	43	8	*the degree to which*
14	17	41	10	*a meta analysis of*
15	17	41	9	*in the form of*
16	16	38	9	*as a function of*
17	16	38	9	*as part of a*
18	16	38	8	*at the end of*
19	16	38	8	*in the development of*
20	16	38	9	*it is possible that*
21	16	38	6	*with respect to the*
22	15	36	8	*has been shown to*
23	15	36	11	*have been shown to*
24	15	36	8	*health and well being*
25	15	36	9	*(is) important to note that +*
26	15	36	9	*were more likely to*
27	14	33	7	*can be used to*
28	14	33	11	*in addition to the*
29	14	33	6	*(the) science and practice of +*
30	14	33	7	*the nature of the*
31	14	33	8	*the ways in which*
32	13	31	11	*as a result of*
33	13	31	8	*national institutes of health*
34	13	31	11	*of this article is*
35	13	31	6	*the context of the*
36	13	31	11	*to the extent that*
37	12	29	10	*in a sample of*
38	12	29	10	*in a way that*
39	12	29	8	*in the general population*
40	12	29	9	*research is needed to*
41	12	29	7	*we were able to*
42	11	25	7	*across the life span*

43	11	25	9	*been shown to be*
44	11	25	9	*has the potential to*
45	11	25	9	*in terms of the*
46	11	25	8	*in the case of*
47	11	25	9	*it may also be*
48	11	25	8	*over the past years*
49	10	24	9	*a wide variety of*
50	10	24	7	*at the university of*
51	10	24	8	*in the absence of*
52	10	24	9	*is one of the*
53	10	24	8	*it is possible to*
54	10	24	8	*physical and mental health*
55	10	24	8	*(the) purpose of this article (is to) +*
56	9	21	7	*a risk factor for*
57	9	21	6	*and physical well being*
58	9	21	8	*and the development of*
59	9	21	6	*in the face of*
60	9	21	7	*in this article we*
61	9	21	7	*is likely to be*
62	9	21	8	*it is clear that*
63	9	21	6	*it should be noted (that) +*
64	9	21	6	*the importance of the*
65	9	21	6	*the magnitude of the*
66	8	20	7	*a wide array of*
67	8	20	6	*as part of the*
68	8	20	7	*at the same time*
69	8	20	6	*at the time of*
70	8	20	8	*for the development of*
71	8	20	6	*has focused on the*
72	8	20	6	*in light of the*
73	8	20	6	*it may be that*
74	8	20	6	*of health and human*
75	8	20	7	*over a year period*
76	8	20	7	*research has shown that*
77	8	20	6	*the full range of*
78	8	20	6	*the national institutes of*
79	8	20	8	*to be associated with*
80	8	20	8	*was associated with a*
81	8	20	7	*with a focus on*
82	8	20	7	*within the context of*

PSY-ENG2

Rank	Frequency	Freq. per mil.	Range	Bundle
1	96	303	41	*at the same time*
2	72	224	18	*a high level of*
3	57	180	34	*the results of the*
4	55	170	27	*as well as the*
5	54	168	20	*in the case of*
6	48	151	26	*on the basis of*
7	46	145	30	*on the other hand*
8	45	142	26	*in the context of*
9	45	142	21	*in the process of*
10	36	113	22	*is one of the*
11	35	110	20	*it is necessary to*
12	35	110	6	*with a high level*
13	34	107	6	*russian version of the*
14	32	101	16	*the relationship between the*
15	32	101	18	*with the help of*
16	31	97	20	*one of the most*
17	30	94	25	*on the one hand*
18	29	91	18	*(it) should be noted that +*
19	28	88	9	*a higher level of*
20	27	85	12	*for the development of*
21	26	82	22	*is based on the*
22	26	82	13	*the end of the*
23	25	78	17	*the fact that the*
24	24	75	15	*in the course of*
25	24	75	16	*it is important to*
26	23	72	14	*as a result of*
27	23	72	18	*in the form of*
28	23	72	15	*the analysis of the*
29	22	69	11	*in the field of*
30	22	69	15	*it is possible to*
31	22	69	12	*on the level of*
32	21	66	13	*it was found that*
33	20	63	10	*a low level of*
34	20	63	14	*at the end of*
35	20	63	9	*in the structure of (the) +*
36	20	63	12	*that there is a*
37	19	60	13	*turned out to be*
38	18	56	11	*are presented in table*
39	18	56	11	*the basis of the*
40	18	56	13	*the level of the*
41	18	56	14	*the same time the*
42	18	56	14	*to the fact that*
43	17	53	12	*a number of studies*
44	17	53	14	*in accordance with the*

45	17	53	8	*in the group of*
46	17	53	8	*level of development of*
47	17	53	11	*the first stage of*
48	16	50	12	*be explained by the*
49	16	50	12	*in this case the*
50	16	50	7	*the case of the*
51	16	50	12	*the development of the*
52	16	50	11	*to the study of*
53	15	47	9	*of the relationship between*
54	15	47	9	*studies have shown that*
55	15	47	6	*the content of the*
56	15	47	10	*the study of the*
57	15	47	10	*the total number of*
58	15	47	7	*with different levels of*
59	14	44	11	*can be explained by*
60	14	44	11	*in the study of*
61	14	44	7	*of the level of*
62	14	44	11	*of this study is*
63	14	44	11	*the beginning of the*
64	14	44	7	*the dynamics of the*
65	14	44	9	*to the development of*
66	13	41	8	*and the level of*
67	13	41	8	*as one of the*
68	13	41	9	*at the level of*
69	13	41	12	*by the fact that*
70	13	41	11	*is considered to be*
71	13	41	8	*of the development of*
72	13	41	9	*the development of a*
73	13	41	9	*the purpose of this*
74	13	41	9	*to the conclusion that*
75	12	38	9	*as well as to*
76	12	38	8	*(it) can be assumed that +*
77	12	38	10	*in contrast to the*
78	12	38	8	*in other words the*
79	12	38	7	*in the learning process*
80	12	38	9	*in the number of*
81	12	38	6	*in the present study*
82	12	38	10	*it is possible that*
83	12	38	9	*the context of the*
84	12	38	7	*the same time they*
85	12	38	9	*the value of the*
86	11	35	7	*an increase in the*
87	11	35	7	*and the development of*
88	11	35	10	*as well as in*
89	11	35	9	*for each of the*
90	11	35	11	*in a number of*

91	11	35	8	*in the development of*
92	11	35	10	*makes it possible to*
93	11	35	9	*of this study was*
94	11	35	7	*one of the first*
95	11	35	10	*results of the study*
96	11	35	10	*the rest of the*
97	11	35	11	*the role of the*
98	11	35	9	*took part in the*
99	11	35	7	*with respect to the*
100	10	32	6	*as in the case*
101	10	32	9	*due to the fact*
102	10	32	8	*in the fact that*
103	10	32	8	*is consistent with the*
104	10	32	9	*is determined by the*
105	10	32	9	*is related to the*
106	10	32	7	*it was shown that*
107	10	32	6	*of the dynamics of*
108	10	32	9	*of the most important*
109	10	32	9	*point of view of*
110	10	32	6	*the concept of the*
111	10	32	8	*the formation of the*
112	10	32	7	*the influence of the*
113	10	32	6	*the meaning of the*
114	10	32	8	*was found that the*
115	10	32	7	*we can assume that*
116	10	32	8	*we can conclude that*
117	10	32	8	*we can say that*
118	10	32	7	*with the results of*
119	10	32	7	*with the use of*
120	9	28	7	*an important role in*
121	9	28	6	*and the degree of*
122	9	28	8	*as a basis for*
123	9	28	8	*as well as a*
124	9	28	9	*at the beginning of*
125	9	28	6	*at the time of*
126	9	28	7	*can be found in*
127	9	28	7	*from the perspective of*
128	9	28	8	*in addition to the*
129	9	28	7	*in terms of the*
130	9	28	8	*in the works of*
131	9	28	7	*of the fact that*
132	9	28	8	*take into account the*
133	9	28	8	*the characteristics of the*
134	9	28	7	*the differences between the*
135	9	28	7	*the nature of the*
136	9	28	9	*the result of the*

137	9	28	9	*the results of our*
138	9	28	6	*the same time it*
139	9	28	8	*the use of the*
140	9	28	8	*the validity of the*
141	9	28	7	*this study was to*
142	9	28	6	*to be the most*
143	9	28	8	*was based on the*
144	9	28	7	*which is based on*
145	9	28	6	*within the framework of*
146	8	25	6	*a negative impact on*
147	8	25	7	*a result of the*
148	8	25	8	*and at the same*
149	8	25	8	*and on the other*
150	8	25	8	*and the ability to*
151	8	25	8	*as well as their*
152	8	25	7	*be noted that the*
153	8	25	7	*can serve as a*
154	8	25	6	*did not differ from*
155	8	25	7	*explained by the fact*
156	8	25	8	*in front of the*
157	8	25	7	*in our case the*
158	8	25	6	*in our opinion the*
159	8	25	6	*in our research we*
160	8	25	6	*in the current study*
161	8	25	6	*in the educational process*
162	8	25	7	*in which a person*
163	8	25	6	*of the ability to*
164	8	25	6	*of the study we*
165	8	25	6	*on the development of*
166	8	25	6	*the aim of the*
167	8	25	6	*the conclusion that the*
168	8	25	8	*the one hand the*
169	8	25	8	*the quality of the*
170	8	25	8	*the results of this*
171	8	25	7	*the second stage of*
172	8	25	8	*there were no significant*
173	8	25	7	*this point of view*
174	8	25	8	*to take into account*
175	8	25	8	*under the influence of*
176	8	25	7	*us to conclude that*
177	7	22	6	*a great number of*
178	7	22	6	*an analysis of the*
179	7	22	6	*and as a result*
180	7	22	6	*and the number of*
181	7	22	6	*as the result of*
182	7	22	7	*can conclude that the*

183	7	22	6	*considered to be an*
184	7	22	6	*correlation analysis of the*
185	7	22	6	*for the first time*
186	7	22	6	*for the study of*
187	7	22	7	*in order to achieve*
188	7	22	6	*in our study we*
189	7	22	6	*in the same way*
190	7	22	6	*in this study the*
191	7	22	6	*is due to the*
192	7	22	6	*is understood as a*
193	7	22	7	*it turned out that*
194	7	22	7	*of the results of*
195	7	22	7	*of the study was*
196	7	22	7	*on the results of*
197	7	22	7	*one of the main*
198	7	22	6	*significant differences in the*
199	7	22	6	*the differences in the*
200	7	22	6	*the idea of the*
201	7	22	6	*the importance of the*
202	7	22	6	*the other hand the*
203	7	22	6	*the point of view*
204	7	22	6	*the research was conducted*
205	7	22	6	*to the analysis of*
206	6	20	6	*a high degree of*
207	6	20	6	*and the results of*
208	6	20	6	*at the age of*
209	6	20	6	*considered to be the*
210	6	20	6	*during the process of*
211	6	20	6	*for a long time*
212	6	20	6	*in line with the*
213	6	20	6	*in the formation of*
214	6	20	6	*is associated with the*
215	6	20	6	*make it possible to*
216	6	20	6	*not only in the*
217	6	20	6	*of the study is*
218	6	20	6	*one of the key*
219	6	20	6	*results of this study*
220	6	20	6	*that the role of*
221	6	20	6	*the form of a*
222	6	20	6	*the one hand and*
223	6	20	6	*the reliability of the*
224	6	20	6	*the study was to*
225	6	20	6	*to a lesser extent*
226	6	20	6	*to the theory of*
227	6	20	6	*turned out that the*

PSY-RUS1

Rank	Freq	Freq. per mil.	Range	Bundle	Translation
1	232	758	58	*о том что*	*about this [the fact] that*
2	103	336	30	*по сравнению с*	*in comparison with*
3	86	281	37	*в том что*	*in this [the fact] that*
4	81	248	34	*и т п*	*and similar*
5	79	242	39	*в том числе*	*including*
6	74	238	33	*в связи с*	*in connection to*
7	74	238	30	*и т д*	*and so on*
8	66	215	29	*в зависимости от*	*in dependence with/ depending on*
9	66	215	36	*в отличие от*	*in contrast with*
10	60	196	29	*в соответствии с*	*in agreement with/in contrast with*
11	60	196	36	*на то что*	*to this [the fact] that*
12	58	189	29	*вместе с тем*	*at the same time/together with this*
13	55	179	28	*с точки зрения*	*from the point of view*
14	51	166	27	*по отношению к*	*in relation to*
15	51	166	21	*тех или иных*	*these or others*
16	48	156	22	*в первую очередь*	*in first turn/ firstly*
17	48	156	23	*в то же*	*at the same*
18	47	153	25	*с одной стороны*	*from the one side/ on the one hand*
19	46	150	22	*на наш взгляд*	*in our view*
20	46	150	23	*то же время*	*at the same time*
21	46	150	23	*(в) то же время +*	*at the same time*
22	44	143	9	*в социальных сетях*	*in social networks*
23	44	143	26	*тот факт что*	*the fact that*
24	40	130	24	*следует отметить что*	*needed to point out that*
25	39	127	27	*в то время*	*at the time/ while*
26	38	124	20	*тем не менее*	*nevertheless*
27	37	120	15	*было показано что*	*was shown that*
28	37	120	28	*в настоящее время*	*at present time/ currently*
29	36	117	17	*в большей степени*	*to a greater extent/degree*
30	36	117	17	*в свою очередь*	*in its turn*
31	35	114	20	*связи с этим*	*connection with this*
32	34	111	23	*то время как*	*at the time when*
33	35	111	20	*(в) связи с этим+*	*concerning that/ in connection to*
34	33	108	20	*в этом случае*	*in this case*
35	34	107	23	*в то время как*	*while/ at the same time as*
36	32	105	21	*с другой стороны*	*from the other side/ on the other hand*
37	32	105	9	*там же с*	*also there with*
38	32	104	16	*вывод о том что +*	*conclusion that / conclusion about the fact that*

39	28	91	23	*в данном случае*	*in this case*
40	28	91	19	*можно предположить что*	*can assume that [it can be assumed that]*
41	30	90	30	*описание хода исследования*	*study process description*
42	27	88	18	*речь идет о*	*talk is about/ this is about*
43	27	88	18	*с тем что*	*with this [idea] that*
44	27	88	14	*сделать вывод о*	*make a conclusion about*
45	25	85	10	*как видно из*	*as seen from*
46	25	85	18	*так и в*	*also in/ as well as in*
47	24	78	16	*в возрасте от*	*in the age from*
48	24	78	12	*в ответ на*	*in response to*
49	24	78	6	*в реальной жизни*	*in real life*
50	24	78	17	*той или иной*	*this or that*
51	23	75	6	*на уровне тенденции*	*on the tendency level*
52	22	75	16	*а так же*	*as well as*
53	22	75	11	*в подростковом возрасте*	*in adolescent age*
54	22	75	15	*до сих пор*	*until now*
55	22	75	18	*не только*	*not only*
56	21	68	16	*включает в себя*	*includes*
57	21	68	8	*детей и подростков*	*children and adolescents*
58	21	68	17	*для того чтобы*	*in order to*
59	21	68	10	*на самом деле*	*in reality*
60	21	68	13	*при этом в*	*at the same time in/ with this in*
61	21	68	11	*состоит в том*	*consists of*
62	20	65	14	*на вопрос о*	*to the question of*
63	20	65	13	*так же как*	*as well as*
64	20	65	16	*таким образом в*	*thus/therefore*
65	19	62	17	*вопрос о том*	*quiestion about that [the fact that]*
66	19	62	12	*по нашему мнению*	*in our opinion*
67	18	59	11	*в данной работе*	*in this work*
68	18	59	12	*на этом этапе*	*at this stage*
69	18	59	11	*не может быть*	*cannot be*
70	18	59	16	*том что в*	*this [this fact] that in*
71	18	58	9	*сделать вывод о том*	*conclude that/ make a conclusion that*
72	17	55	10	*вне зависимости от*	*independent of*
73	17	55	14	*не только в*	*not only in*
74	17	55	13	*но при этом*	*but at the same time*
75	17	55	10	*те или иные*	*these or others*
76	17	55	15	*том числе и*	*including and*
77	17	55	15	*в том числе и*	*also including*
78	17	55	10	*состоит в том что*	*consists of this [the fact] that*
79	16	52	10	*в нашем исследовании*	*in our study*
80	16	52	7	*друг от друга*	*from each other*

81	16	52	12	*и др в*	*and others in*
82	16	52	15	*несмотря на то (что) +*	*despite the fact that*
83	16	52	13	*о том как*	*about how*
84	16	52	15	*несмотря на то*	*despite the fact that*
85	15	49	11	*исследовании приняли участие*	*took part in the study*
86	15	49	10	*можно рассматривать как*	*can be viewed as*
87	15	49	12	*отметить что в*	*note that in*
88	15	49	11	*является одним из*	*is one of the*
89	14	46	10	*так же как и +*	*as well as*
90	14	45	8	*более высокий уровень*	*a higher level*
91	14	45	12	*друг с другом*	*with each other*
92	14	45	9	*зависимости от того*	*depending on*
93	14	45	11	*к тому что*	*to this [the fact] that*
94	14	45	10	*как и в*	*as in/like in*
95	14	45	12	*на первый план*	*in the foreground*
96	14	45	10	*на сегодняшний день*	*nowadays/ to date*
97	14	45	8	*с другими людьми*	*with other people*
98	14	45	10	*так и на*	*as well as on*
99	14	45	11	*того или иного*	*that or the other [gen]*
100	14	45	11	*является одной из*	*is one of*
101	13	43	9	*в исследовании приняли участие*	*took part in the study*
102	13	43	10	*заключается в том что*	*consists in this [the fact] that*
103	13	42	9	*а с другой*	*and on the other*
104	13	42	9	*в исследовании приняли*	*in the study took*
105	13	42	10	*заключается в том*	*can be summarized in*
106	13	42	9	*и тем самым*	*and with that*
107	13	42	10	*по всей видимости*	*evidently/apparently*
108	13	42	11	*после того как*	*after this [the fact] that*
109	12	40	9	*в той или иной*	*in one or another*
110	12	40	6	*говорить о том что*	*talk about*
111	12	40	8	*с нашей точки зрения +*	*from our point of view*
112	12	40	8	*свидетельствует о том что*	*indicates that*
113	12	40	9	*свидетельствуют о том что +*	*indicate that*
114	12	39	6	*было установлено что*	*was established that*
115	12	39	8	*в данном исследовании*	*in the present study*
116	12	39	9	*в той или*	*in this or*
117	12	39	6	*говорить о том*	*talk about*
118	12	39	10	*можно сделать вывод*	*can be concluded*
119	12	39	6	*особый интерес представляет*	*presents special interest*
120	12	39	10	*с высоким уровнем*	*with a high level of*

121	12	39	8	*свидетельствует о том*	*provides evidence to [sing]/ indicates that/ indicates the fact that*
122	12	39	10	*так и для*	*as well as for/to*
123	12	39	10	*таким образом*	*therefore*
124	12	39	12	*хода исследования в*	*study process in*
125	12	39	6	*юношей и девушек*	*young men and women*
126	10	37	9	*в том что в*	*in that/ in the [fact] that*
127	10	37	8	*можно сделать вывод о*	*we can conclude/ can be concluded ...about*
128	10	37	7	*обращает на себя внимание*	*noteworthy/draws attention upon itself*
129	11	35	7	*а также на*	*as well as on*
130	11	35	8	*а также с*	*as well as with*
131	11	35	8	*более или менее*	*more or less*
132	11	35	7	*в меньшей степени*	*to a lesser degree/ extent*
133	11	35	11	*в нашей работе*	*in our work*
134	11	35	9	*в работе с*	*in the work with*
135	11	35	9	*исследования показали что*	*of the study showed that*
136	11	35	8	*можно говорить о*	*can be talked about*
137	11	35	7	*на себя внимание*	*attention on itself*
138	11	35	8	*необходимо отметить что*	*it is necessary to note that*
139	11	35	6	*по всей выборке*	*throughout the sample*
140	11	35	10	*по крайней мере*	*at least*
141	11	35	7	*представлены в таблице*	*are presented in table*
142	11	35	10	*то что в*	*this [the fact] that in*
143	9	33	6	*в том числе в*	*including in*
144	9	33	6	*вне зависимости от того*	*no matter/ independent of [the fact] that*
145	9	33	7	*как видно из таблицы*	*as can be seen from table*
146	10	32	8	*а также в*	*as well as in*
147	10	32	9	*в данной статье*	*in the present article*
148	10	32	6	*в обеих группах*	*in both groups*
149	10	32	6	*в отечественной психологии*	*in the fatherland psychology*
150	10	32	9	*в последнее время*	*recently*
151	10	32	6	*в сочетании с*	*together with / in conjunction with*
152	10	32	7	*в том случае*	*in the case of*
153	10	32	9	*в частности в*	*in particular in*
154	10	32	9	*важно отметить что*	*it is important to note that*
155	10	32	8	*видно из таблицы*	*seen from table*
156	10	32	8	*и в целом*	*and in general*
157	10	32	6	*на этой стадии*	*at this stage*
158	10	32	7	*обращает на себя*	*draws upon itself*
159	10	32	8	*по их мнению*	*in their opinion*
160	9	29	7	*в конечном счете*	*in the end [as a result]*

161	9	29	7	*в некоторых случаях*	*in some cases*
162	9	29	7	*в полной мере*	*fully/ in full capacity*
163	9	29	6	*в этой области*	*in this area*
164	9	29	6	*и уверенность в*	*and confidence in*
165	9	29	7	*из того что*	*from [the fact] that*
166	9	29	8	*кроме того в*	*Besides, in*
167	9	29	6	*могут быть связаны*	*may be related*
168	9	29	8	*может привести к*	*can lead to*
169	9	29	6	*при этом не*	*while not*
170	9	29	8	*при этом они*	*while they*
171	9	29	9	*с опорой на*	*based on*
172	9	29	9	*с помощью*	*via/with the help of*
173	9	29	8	*с таким образом*	*with that way*
174	9	29	7	*так или иначе*	*anyway/ this or that way*
175	9	29	9	*таким образом можно*	*this way you can/we can/ it is possible*
176	9	29	6	*том числе в*	*including in*
177	9	29	8	*том что они*	*that they*
178	9	29	6	*человека и его*	*man and his*
179	9	29	6	*что по мере*	*that as*
180	9	29	9	*это может быть*	*it could be*
181	9	29	7	*это означает что*	*it means that*
182	8	29	7	*и т д в*	*and so on in*
183	8	29	7	*одни и те же +*	*same*
184	8	29	7	*указывает на то что*	*indicates that / [the fact] that/ points to the fact that*
185	8	29	7	*что в свою очередь*	*which in turn*
186	8	26	7	*было выявлено что*	*it was revealed that*
187	8	26	6	*в какой то*	*at some*
188	8	26	6	*в рамках данного*	*as part of this/in the frame o this*
189	8	26	6	*в ряде случаев*	*in some cases*
190	8	26	7	*в том чтобы*	*in that, to*
191	8	26	6	*в целом по*	*on the whole*
192	8	26	8	*важную роль в*	*important role in*
193	8	26	7	*друг к другу*	*to each other*
194	8	26	8	*и при этом*	*and wherein*
195	8	26	6	*и таким образом*	*and thus*
196	8	26	7	*и те же*	*the same*
197	8	26	8	*и то что*	*and [the fact] that*
198	8	26	8	*может быть связано*	*may be related*
199	8	26	6	*можно сказать что*	*we can say that*
200	8	26	7	*на то чтобы*	*in order to*
201	8	26	6	*не могут быть*	*can not be*
202	8	26	7	*не только на*	*not only on*
203	8	26	8	*но и в*	*but also in*

204	8	26	6	*по мере увеличения*	*as you increase/as we increase*
205	8	26	7	*с ним в*	*with him in*
206	8	26	7	*становится все более*	*getting more/becoming increasingly*
207	8	26	7	*так и не*	*as well as not*
208	8	26	7	*так и с*	*as well as with*
209	8	26	7	*указывает на то*	*indicates that*
210	8	26	7	*что в свою*	*which in its*
211	8	26	7	*что может быть*	*what could be*
212	8	26	7	*что он не*	*that he not*
213	7	25	6	*в том случае если*	*in case if*
214	7	25	6	*вместе с тем в*	*at the same time in*
215	7	25	6	*внимание на то (что)+*	*attention to [the fact] that*
216	7	25	6	*о том что в*	*about that in/ about [the fact] that*
217	7	23	7	*а также о*	*as well as about*
218	7	23	6	*в значительной степени*	*to a large extent*
219	7	23	6	*в конце x*	*at the end of x*
220	7	23	6	*в одном из*	*in one of*
221	7	23	7	*в последние годы*	*in recent years*
222	7	23	7	*в том числе*	*including*
223	7	23	7	*в целом и*	*in general and*
224	7	23	6	*внимание на то*	*attention to [the fact]*
225	7	23	7	*время от времени*	*occasionally/ from time to time*
226	7	23	6	*исследование показало что*	*the study showed that*
227	7	23	7	*их связи с*	*their relationship with*
228	7	23	6	*мы предположили что*	*we assumed that*
229	7	23	6	*на этот вопрос*	*to this question*
230	7	23	6	*предположить что в*	*suggest that in*
231	7	23	6	*при этом у*	*at the same time in*
232	7	23	7	*равно как и*	*as well as*
233	7	23	6	*с тем в*	*with that in*
234	7	23	7	*связано с тем*	*due to the/ connected to the*
235	7	23	6	*сделать следующие выводы*	*draw the following conclusions*
236	7	23	6	*том случае если*	*in case if*
237	7	23	7	*это связано с*	*it's connected with*
238	6	20	6	*p при этом*	*p in this case*
239	6	20	6	*а не на*	*and not on*
240	6	20	6	*а также их*	*as well as their*
241	6	20	6	*в исследовании были*	*in the study were*
242	6	20	6	*в которой он*	*in which he*
243	6	20	6	*в последние десятилетия*	*in recent decades*
244	6	20	6	*в целом в*	*generally in*
245	6	20	6	*до того как*	*before*

246	6	20	6	*и в отношении*	*and regarding/and in relation to*
247	6	20	6	*исследования в исследовании*	*research in research*
248	6	20	6	*исследования в статье*	*research in the article*
249	6	20	6	*к тому же*	*in addition*
250	6	20	6	*как раз и*	*just and*
251	6	20	6	*можно выделить несколько*	*there are several/it is possible to single out (highlight) several*
252	6	20	6	*мы предполагаем что*	*we assume that*
253	6	20	6	*мы считаем что*	*we believe that*
254	6	20	6	*на первом этапе*	*at the first stage*
255	6	20	6	*на этой основе*	*on this basis*
256	6	20	6	*не только для*	*not only for*
257	6	20	6	*но и на*	*but also on*
258	6	20	6	*отметить что на*	*note that on*
259	6	20	6	*по его мнению*	*in his opinion*
260	6	20	6	*тем или иным*	*one way or another*
261	6	20	6	*том что он*	*that he [the fact] that he*
262	6	20	6	*человек в возрасте*	*elderly person*
263	6	20	6	*это проявляется в*	*it manifests itself in*
264	6	20	6	*связано с тем что*	*due to [the fact] that/connected to [the fact] that*

Appendix 2: Functional classification of the first 50 frequent bundles

Research-oriented	
Description	PSY-ENG1 *the online supplemental materials, online supplemental materials for, in the development of, the nature of the, in a way that, the ways in which* PSY-ENG2 *the relationship between the, is based on the, for the development of* *in the form of, in the field of, on the level of, turned out to be, the basis of the, the level of the, level of development of, be explained by the* PSY-RUS1 *в возрасте от / in the age from, той или иной / that or other*
Location	PSY-ENG1 *in the online supplemental, at the end of, over the past years, at the university of* PSY-ENG2 *the end of the, in the course of, at the end of, in the structure of (the)+, that there is a, in the group of* PSY-RUS1 *в настоящее время / at present time, там же с / also there with*
Quantification	PSY-ENG1 *the extent to which, a wide range of, one of the most, the degree to which, to the extent that, a wide variety of* PSY-ENG2 *a high level of, is one of the, with a high level, one of the most, a higher level of, a low level of, a number of studies* PSY-RUS1 *в большей степени / to a greater extent*
Procedure	PSY-ENG1 *a meta analysis of, as part of a, in the general population, we were able to* PSY-ENG2 *the results of the, in the process of, with the help of, the analysis of the, the first stage of* PSY-RUS1 *в ответ на / in response to, в то время / at the time*
Topic	PSY-ENG1 *health and wellbeing of, (the) science and practice of +, national institutes of health, across the lifespan of* PSY-ENG2 *russian version of the* PSY-RUS1 *в социальных сетях / in social networks, в реальной жизни / in real life*
Text-oriented	
Framing signals	PSY-ENG1 *in the context of, in the form of, as a function of, with respect to the, the context of the, in the case of*

	PSY-ENG2 *in the case of, on the basis of, in the context of, in this case the, the case of the* PSY-RUS1 *в том числе / including, в зависимости от / depending on, по отношению к / in relation to, речь идет о / the talk is about*
Transition signals	PSY-ENG1: *as well as the, in addition to the, in terms of the* PSY-ENG2: *at the same time, as well as the, on the other hand, on the one hand, the same time the, in accordance with the* PSY-RUS1: *в то время как / however, о том что / about this [the fact] that, по сравнению с / in, comparison to, в том что / in [the fact] that, и т п / and so on, в связи с / in connection to, и т д / and so forth, в отличие от / in contrast with, в соответствии с / in agreement with, на то что / to [the fact] that, вместе с тем /together with this, с точки зрения / from the point of view, в первую очередь / firstly, с одной стороны / on the one hand, в то же время / at the same time, тем не менее / nevertheless, в свою очередь / in its turn, (в) связи с этим + / in connection to this, с другой стороны / on the other hand, с тем что (more specifically), так и в / so in*
Structuring signals	PSY-ENG1: *see the online supplemental, of this article is, in a sample of* PSY-ENG2: *are presented in table* PSY-RUS1: *описание хода исследования / study process description, как видно из / as seen from*
Resultative signals	PSY-ENG1: *has been shown to, have been shown to, as a result of* PSY-ENG2: *as a result of, it was found that* PSY-RUS1: *вывод о том что + / conclusion about [the fact] that,* *сделать вывод о / make a conclusion about*
Stance-oriented	
Engagement features	PSY-ENG1: *it is important to, (is) important to note that +* PSY-ENG2: *it is necessary to, (it) should be noted that +, it is important to* PSY-RUS1: *можно предположить что / it can be assumed that* *следует отметить что / should be pointed out that*
Stance features	PSY-ENG1: *are more likely to, more likely to be, it is possible that, were more likely to, can be used to, research is needed to, been shown to be, has the potential to, it may also be* PSY-ENG2: *the fact that the, it is possible to, to the fact that* PSY-RUS1: *тот факт что / the fact that, на наш взгляд / in our view*

RiCL Research in Corpus Linguistics

The metadiscourse of Arabic academic abstracts: A corpus-based study

Mai Zaki
American University of Sharjah / United Arab Emirates

Abstract – Research on metadiscourse and rhetorical features in modern Arabic academic writing is scarce both in quantity and in scope. Abstracts, in particular, are a severely understudied academic register. This study aims to fill a gap in the study of academic abstracts in Arabic by providing a more comprehensive analysis of metadiscourse in Arabic academic abstracts. The data for the study includes a corpus of 400 Arabic abstracts, which have been labeled according to two variables: (a) abstract type (journal or dissertation); and (b) author gender (male, female, mixed gender). The analysis follows the theoretical framework proposed by Hyland (2019), as the data has been annotated for both textual metadiscourse (transition markers, frame markers, evidentials, endophorics and code glosses) and interpersonal metadiscourse (hedges, boosters, attitude markers, engagement markers and self-mentions). Results show that Arabic academic abstracts are rich in both types of metadiscourse features. Transition and frame markers have the highest frequency in the textual domain, while boosters and self-mentions are highly frequent in the interpersonal domain. Endophoric markers and hedges are the least used types of metadiscourse in the data, but engagement markers are surprisingly more frequent than previously thought.

Keywords – metadiscourse; abstracts; Arabic; academic writing; corpus

1. INTRODUCTION

Academic writing is a discourse domain that reflects reader-writer relationship in rather specific ways. Many elements which linguistically embody this relationship have been discussed in the literature as 'metadiscourse'. According to Hyland (2019: 3), this term was coined by the linguist Zellig Harris in 1959 in an attempt to highlight the aspects of perception, reception and interaction in/of a text. Later research built on this concept, most notably in the works of Kopple (1985) and Crismore (1989), who discussed a wide range of discoursal features acting as metadiscourse devices and setting the tone of the rhetorical structure of a text. In pursuit of persuading readers of an academic argument, writers are seen to employ a heterogenous group of "cohesive and interpersonal features" (Hyland and Tse 2004: 157) in order to guide and engage readers along a certain interpretive path.

While there is now a considerable body of research on metadiscourse in general and in academic texts in particular, most of this research is restricted to texts in English. There is a visible need to fill a big gap in the research on metadiscourse in other languages. This study intends to partly fill this gap by exploring the use of metadiscourse markers in abstracts of Arabic journals and dissertations in the field of humanities and social sciences through these two research questions: (1) What is the overall distribution of metadiscourse markers in Arabic academic abstracts? (2) Are there distributional variations for metadiscourse elements in Arabic academic abstracts across the two variables: abstract type and author gender? The study adopts a corpus-based approach and employs both quantitative and qualitative analyses to the data, making a contribution to the field in two ways. First, it provides a detailed study of metadiscourse in abstracts of academic Arabic which, in contrast to existing studies, is based on a sizable corpus. Second, it presents useful insights on metadiscourse devices in relation to two variables: type and gender. By making a distinction between abstracts of journal papers and those of academic dissertations on the one hand, and between authors' genders on the other, this study provides a broader view of the use of metadiscourse in academic Arabic and thus facilitates our understanding of this discipline while at the same time laying the foundation for further studies.

This paper is organized as follows. Section 2 reviews some of the relevant literature on metadiscourse and academic writing. Section 3 introduces the corpus data collected for this study and the methodology followed. Section 4 presents the results of the analysis in relation to the two variables: abstract type and author gender. Section 5 discusses the results in light of the theoretical model and in comparison to other studies. Finally, Section 6 ends with some concluding remarks and recommendations for future research.

2. LITERATURE REVIEW

2.1. Metadiscourse and academic abstracts

The study of metadiscourse as a functional category is concerned with the way in which personalities, attitudes and assumptions play a role in the writing and receiving of a text. According to Hyland (2019: 44),

> metadiscourse is the cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community.

As such, the concept of metadiscourse which can be seen as "intuitively attractive" (Hyland and Tse 2004: 156) has been defined in various ways over time (*inter alia*, Halliday 1977; Kopple 1985; Crismore 1989; Thompson and Thetela 1995). Yet, one of the well-known classifications today is based on the work of Hyland (1998, 1999, 2004a, 2019), who developed over the years an operational model for metadiscoursal devices in academic writing. The basic distinction within this model is between two major categories of metadiscourse: the textual (or interactive) and the interpersonal (or interactional). Despite the change in the terms, both dimensions of interaction retain the same underlying conceptual distinction.[1] The textual/interactive dimension has to do with the writer-reader relationship and all the linguistic devices which the writer employs in the organization of discourse to reflect his/her awareness of the readers' rhetorical expectations and processing abilities. This includes the use of transition markers, frame markers, endophoric markers, evidentials and code glosses. Each one of these sub-types manipulates the way propositional information is organized in the discourse and act as an interpretive guide to the readers. On the other hand, the interpersonal/interactional dimension has to do with the way writers project themselves onto the propositional content of the discourse as a form of establishing an authorial 'voice' in the text. This is done through various linguistic means which serve to comment, evaluate, express solidarity, allow engagement or manipulate the readers' attention throughout the text. The sub-types which fulfil this role include boosters, hedges, attitude markers, self-mention and engagement markers.

While studies of metadiscourse have applied its principles to a wide array of academic texts, ever since Ventola's (1994) seminal work, special attention has been given to the abstract as a genre. Young (2006: 64) describes abstracts as "an exercise in precise, accurate language;" therefore, by nature of their function, certain cognitive and linguistic skills are involved in writing abstracts regardless of the language. Functionally, an abstract is a form of academic writing, but it is also an independent piece of writing in

[1] The terminology used in Hyland's classification of metadiscourse has evolved over time. In Hyland (1998, 1999), he used the terms 'textual-interpersonal' then Hyland and Tse (2004) and Hyland (2019) shifted to the terms 'interactive-interactional' in distinguishing the two main domains of metadiscourse. For simplicity, in this paper the original terms 'textual' and 'interpersonal' are used throughout with no theoretical implications.

itself, that is, a type of a "stand-alone *mini-text*" as Huckin (2006: 93) calls it. It performs the important role of presenting the gist of the article or dissertation in an appealing and informative way for the readers. As such, abstracts set the relationship between the authors and the readers early on, and abstract writers have to take into consideration certain elements such as the readers' needs, shared knowledge within the discipline, expectations of objectivity and degree of commitment to the communicated message. According to Huckin (2006), abstracts are an important medium of communication between writers and readers in more than one way. Abstracts can be 'screening devices' used by readers as a shortcut to the decision of reading the whole paper/dissertation. Crucially, abstracts are also "previews, creating an interpretive frame that can guide reading" (Huckin 2006: 93). Metadiscourse markers contribute a great deal to creating such an interpretive frame.

With a firm position in standard policies on academic publication in many languages, interest has grown steadily in studying the language, structure and metadiscourse of abstracts across various disciplines (*inter alia*, Swales 1990; Stotesbury 2003; Fischer and Zigmond 2004; Lorés 2004; Dahl 2004a; Swales and Feak 2004; Miech *et al.* 2005; Ayers 2008; Gillaerts and van de Velde 2010). There has also been particular interest in the study of metadiscourse in abstracts by language learners (e.g., Ozdemir and Longo 2014; Jin and Shang 2016; Nugroho 2019). On the other hand, the relationship between gender, metadiscourse and academic writing in general has not received the same attention. With hardly any studies particularly focused on gender and academic abstracts, there has been some work on gender-based differences in different types of academic writing (e.g., Robson *et al.* 2002; D'Angelo 2008; Serholt 2012; Pasaribu 2017). Tse and Hyland (2008), who admit that gender has been far less studied in academic writing compared to other factors, examine gender differences of metadiscourse in a corpus of academic book reviews. Their results show that males use more metadiscourse overall, but also highlight that gender alone is not the decisive factor in academic writing style, as variation in discipline plays a major role. Similarly, Parasibu (2017), who studied metadiscourse in academic essays, finds that male students use more interpersonal markers but that field-specific differences are more significant than gender-based ones.

2.2. Metadiscourse in Arabic academic writing

Whereas academic writing in English is well-documented and well-established in the areas of publication and teaching, the story is somewhat different in modern standard academic Arabic. There is, in fact, very little research which examines in detail the linguistic and/or rhetorical features of modern Arabic academic writing, in general, let alone abstracts in particular.

General reference works on academic writing in Arabic are fairly similar in their accounts. Hassan (1996: 68) argues that an academic writer should always aim to "highlight the facts with honesty and objectivity"[2] while avoiding influencing the reader. He also acknowledges the importance of an abstract, in that it is the first part to be read in an academic document after the title, although he focusses on the editorial formalities rather than the language of abstracts. Even though Hassan (1996) does not discuss markers of metadiscourse directly, he highlights a few aspects of academic writing which would contribute to the metadiscourse structure, such as the use of personal pronouns and expressions of emphasis. Al-Shahrani (2010), on the other hand, lists aspects of academic writing which he deems important, including objectivity, explicitness, precision, formality and hedging, albeit with no specific examples. Even when he discusses some linguistic issues related to academic language, the discussion is general at best and seems to reinforce the importance of academic writing being informative, impersonal and minimally rhetorically interactive. Al-Sharif (1996: 153) agrees with this depiction of academic style, noting that repetition, exaggeration, and the use of the first-person are examples of poor academic style. It is worth noting, however, that he does not mention abstracts as a part of an academic research paper.

Amidst this uncertain place for abstracts in Arabic academic writing, it can be said that abstract writing conventions in Arabic are not as standardized as they are in other languages. To start with, as seen in our data, there are at least three different equivalents to the term 'abstract' in Modern Standard Arabic, the variety used for formal writing by all Arabic speakers. An abstract in Arabic can be *mulaxxaṣ*, *mustaxlaṣ* or *xulāṣa*, three derivations of one root in Arabic meaning 'summarize' or 'outline'. Furthermore, previous studies analyzing Arabic abstracts are scarce and most of them simply compare Arabic and English abstracts and are based on very small data sets. Alharbi and Swales (2011), for example, described some similarities and differences between the two

[2] All translations of Arabic quotes and examples are mine.

languages, based on 28 paired abstracts, as they explored the degree of interactivity in the texts. They looked at linguistic features such as the use of first person pronouns, evaluative language and rhetorical moves and indicated a "broad degree of correspondence between the abstracts in the two languages" (Alharbi and Swales 2011: 83). Alotaibi (2015), on the other hand, worked with 44 paired Arabic-English abstracts and noted an overuse of textual markers in both sets. He also reported no variation in the rhetorical organization of the abstracts between the two languages. Alzarieni *et al.* (2019) focused on interpersonal metadiscourse markers in their study of 60 patent abstracts written by native Arabic speakers. Their results show that boosters, hedges and attitude markers are the most frequently used types of interactional markers. The extensive use of boosters is seen as consistent with the field of patents, as they are mainly used to assert the importance of the inventions discussed in the abstracts.

Among the scarce studies on Arabic academic abstracts, some adopted a genre analysis of data rather than focusing on metadiscourse *per se*. Alhuqbani (2013), for example, carried out a genre-based analysis of Arabic abstracts in different fields, with a focus on move structure, based on a corpus of 40 abstracts. He notes that the move structure in abstracts of various disciplines differs greatly, attributing this to the fact that "the Arabic journals' publication policy […] leaves the writing of abstracts at the researchers' disposal" (Alhuqbani 2013: 379). Similarly, Fallatah (2016) compared the move structure of abstracts written by native and non-native speakers by analyzing a total of 93 abstracts divided into those written by Saudi authors in English, Saudi authors in Arabic and international authors. Adopting the genre analysis framework of Swales and Feak (2004), the author concludes that abstracts written by native English and Saudi Arabic speakers reflect a more consistent move structure than those written by Saudi non-native speakers of English. Finally, Bouziane and Metkal (2020) compared the move structures of 112 abstracts in the areas of applied linguistics in three languages: Arabic, French and English. As far as Arabic is concerned, the authors note a difference in conformity with conventions of abstract writing between abstracts written by Middle Eastern writers and those written by writers from North Africa. Despite its small data set for each language, Bouziane and Metkal's (2020) study sheds some light on the differences in abstract writing among the various Arabic-speaking countries, another area which warrants further research.

It is also worth mentioning that none of these studies on metadiscourse in Arabic academic discourse takes the gender factor into consideration. Only Alqahtani and Abdelhalim (2020) explored gender-based differences in textual metadiscourse of Arabic EFL learners. Therefore, there is a big gap in the literature on metadiscourse in Arabic academic writing, and this study aims to fill this gap by providing some insights on gender-based patterns in the use of textual and interpersonal metadiscourse. Finally, there are other studies which are not directly relevant to the research at hand, although they do shed light on different aspects of metadiscourse and academic Arabic writing or academic English as a second language (L2) by Arabic speakers (*inter alia*, El-Seidi 2000; Sultan 2011; Alhumidi and Uba 2016; Briones 2018; Al-Ghoweri and Al Kayed 2019). Sultan (2011), for example, compared the discussion sections of Arabic and English linguistics research papers written by native speakers of Arabic and English. Sultan finds that Arab writers use significantly more metadiscourse markers than English writers. Briones (2018), on the other hand, analyzed 29 abstracts written in English by native Arabic speakers extracted from three academic journals. Besides presenting insightful observations on the move structure of the data, the paper also raises some questions as to the influence of cultural norms and/or ethnicity of the authors' writing style.

3. Data and methodology

The data for this study was collected from online sources of Arabic academic texts, including websites of Arabic academic journals and research repositories of some universities in the Arab world.[3] The academic field of the collected data was restricted to the humanities and social sciences for two reasons. First, to control for possible variations among different disciplines; and secondly, to allow for a bigger corpus, since it is more difficult to find academic papers in the scientific/medical fields written in Arabic. The data has been divided according to type into two categories: (a) abstracts of journal papers; and (b) abstracts of masters/doctoral dissertations. The data has been further divided by gender into three categories: male, female and mixed gender (i.e. multiple authors of different genders, this only applying to journal papers). All abstracts are written by native Arabic speakers. The total number of abstracts in this corpus is 400, amounting to approximately 73,000 words. Table 1 below shows the breakdown of the corpus.

[3] The author wishes to acknowledge the contribution of Aya Alchayah, MA in Translation and Interpreting student and graduate research assistant, in collecting the necessary data for this study.

	No. of abstracts / No. of words (male)	No. of abstracts / No. of words (female)	No. of abstracts / No. of words (mixed gender)	Total
Journals	136 / 22,342	84 / 14,430	50 / 8,366	270 / 45,138 (67.5%)
Dissertations	65 / 13,568	65 / 14,229	-	130 / 27,797 (32.5%)
Total	201 / 35,910 (50.25%)	149 / 28,659 (37.25%)	50 (12.5%)	400 / 72,935

Table 1: Corpus data for the study

The corpus was then uploaded to a text annotation program, maintaining the type and gender distinctions.[4] Each sub-corpus was manually annotated for linguistic features of metadiscourse following Hyland's (2019) framework. All data was analyzed word by word, and no pre-existing list of metadiscourse elements was used; however, automatic bulk annotation of certain lexical items was carefully applied and manually checked. A second annotator was asked to review the annotations and the initial percentage of agreement was 85 percent. The quantitative results of the annotations were then downloaded for further processing and visualization purposes. The quantitative aspect in this study is deemed of high importance, given the reasonable size of the corpus on which it is based and the scarcity of other reliable data. It remains a limitation, however, that these results are restricted to the field of humanities and social sciences. Table 2 below summarizes the sub-types of metadiscourse with examples from Arabic.

Category	Function	Examples
Textual/Interactive		
1. Transitional markers	express semantic relations between clauses	بالإضافة إلي ('in addition to') فضلاً عن ('as well as') في المقابل ('in contrast') لكن ('but') بينما ('whereas') بالرغم من ('in spite of') بالتالي ('hence') مِن ثَم ('therefore')

Table 2: Sub-types of metadiscourse according to Hyland and Tse (2004: 169) with added Arabic examples

[4] The text annotation tool used was *Recogito* (Rainer *et al*. 2016), which allows for manual identification and tagging of any word in the corpus with multiple tags (e.g., textual, frame marker). The tool also allows for automatic bulk annotation. For example, if the conjunction لكن (*lākin* 'but') was tagged as a transitional marker, the tool can automatically apply the same annotation to all instances of this conjunction in the corpus. However, even bulk annotations were manually checked for accuracy.

Category	Function	Examples
2. Frame markers	explicitly refer to discourse shifts, sequences or text stages	أولاً (‘firstly’) أخيراً (‘lastly’) أما (‘as for’) تهدف الدراسة إلى (‘the study aims’) خَلُص البحث إلى (‘the research concludes’) ثُم (‘then’) وفي النهاية (‘in the end’)
3. Endophoric markers	refer to information in other parts of the text	الآتي (‘the following’) التالي (‘the following’) مما سبق (‘from the previous’) ما تقدم (‘what precedes’) ما يأتي (‘what follows’)
4. Evidentials	refer to sources of information from other texts	بالنسبة لـ + اسم (‘according to ..’) قال + اسم (‘…says’) يقول النحاة/ الفلاسفة ... (‘grammarians/philosophers say’)
5. Code glosses	help readers grasp meanings of ideational material	أي (‘that is’) يعني (‘meaning’) بمعنى آخر (‘in other words’) مثل/ كـ (‘such as/like’) يُسمى (‘is called’)
Interpersonal/Interactional		
1. Boosters	emphasize force or writer’s certainty in message	قد + فعل ماضي (*qad* + past tense verb) وضح / تبيّن (‘is clear’) يدل بوضوح (‘clearly shows’) أكدت النتائج (‘results confirm’) خاصةً (‘especially’) لا سيّما (‘particularly’)
2. Attitude markers	express writer's attitude to propositional content	للأسف (‘unfortunately’) تميَّز بـ (‘is distinguished’) يتّفق (‘agrees with’) مخالفاً (‘in disagreement’) هام (‘important’) بارز (‘significant’) رائع (‘excellent’) نادراً (‘rarely’)
3. Hedges	withhold writers’ full commitment to statements	قد + فعل مضارع (*qad* + present tense verb) ربما (‘maybe’) يُمكن، من الممكن (‘is possible’) لعل (‘perhaps’)
4. Self-mention	explicit reference to the author(s)	الباحث/ة (‘the researcher’) دراستنا (‘our study’) قُمنا بـ (‘we did’)
5. Engagement markers	explicitly refer to or build relationship with reader	القارئ (‘the reader’) يجب أن/علينا (‘we have to’) من المعلوم أن (‘it is known’) ينبغي (‘should’)

Table 2 (continuation)

4. RESULTS

4.1. Metadiscourse elements by abstract type

The first round of analysis takes abstract type as the main variable. The total annotations are shown in Figure 1 below. The analysis shows that textual metadiscourse markers are the most frequent (2,575 or 53%) of all metadiscourse elements in the corpus. If we compare abstract types, the results show a narrow majority of textual features (1,485 or 51%) of all journal abstracts, while dissertation abstracts recorded a slightly higher majority with 1,090 (or 55%) of total metadiscourse elements being in the textual domain.

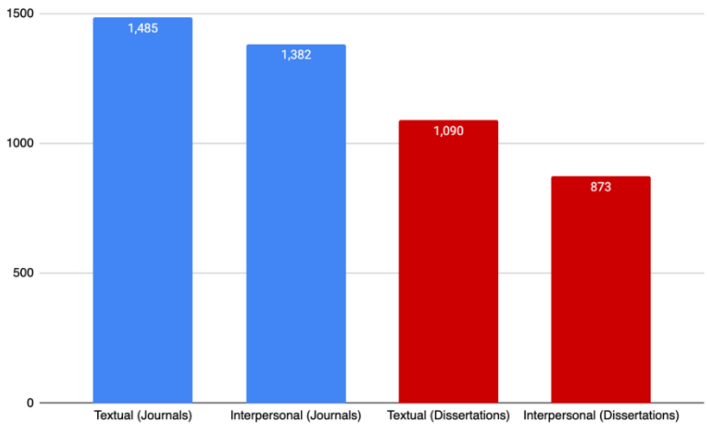


Figure 1: Total metadiscourse elements by type

4.1.1. Textual metadiscourse (TM)

Table 3 below shows a clearer picture of the distribution of textual elements, illustrates the breakdown of its various types.[5] The quantitative results reveal that there is a small difference between journal and dissertation abstracts in the top two sub-types —frame markers and transition markers— which, when combined, constitute between 80 and 84 percent of all textual markers.

[5] See also Appendix 1 for the quantitative data on metadiscourse elements by abstract type with normalized frequency per 1,000 words, and a comparison graph by percentage.

Textual metadiscourse	Journal abstracts	% of total TM	Dissertation abstracts	% of total TM
Transition markers	549	37	416	38.2
Frame markers	634	42.7	506	46.4
Endophoric markers	75	5	28	2.5
Evidentials	138	9.3	68	6.3
Code glosses	89	6	72	6.6

Table 3: Distribution of textual metadiscourse by abstract type

As far as frame markers are concerned, it is noted that the majority of those in both abstract types largely correspond to the main move structures for abstracts (for a chronological summary of abstract moves as identified in the literature, see Briones 2018). In particular, frame markers in the data seem to be following a five-move structure, as a combination of Hyland (2000) and Belcher (2009), favoring a verbal sentence structure as move openers as illustrated in Table 4.

Move type	Corresponding frame markers
Introduction	يتناول البحث ('the research discusses')
	يتعرض هذا البحث لـ ('this research tackles')
	تقدّم هذه الدراسة ('this study presents')
	يسلّط هذا البحث الضوء على... ('this research sheds light on')
Aim/purpose	تهدف الدراسة إلى ('the study aims to')
	استهدف البحث ('the research aims to')
	يسعى البحث إلى ('the research attempts to')
	أهداف الدراسة... ('the objectives of the study')
Method/process	يتركز البحث على ('the research focuses on')
	اعتمد البحث على ('the research depends on')
	منهجية البحث ('the research methodology')
	عينة البحث ('the research sample')
Results/conclusions	توصل البحث إلى ('the research arrived at')
	خلصت / انتهت الدراسة إلى ('the study concluded that')
	أسفرت الدراسة عن ('the study resulted in')
	نتج عن هذه الدراسة ('it resulted from this study')
	من نتائج الدراسة ('among the results of the study')
Recommendations	أوصت الدراسة بـ ('the study recommended that')
	قدّمت الدراسة توصيات بـ ('the study presented recommendations to')

Table 4: Move types and corresponding frame markers in the data

Apart from signaling text stages in this way, many frame markers in the data use a variety of nouns to refer to different sections of the paper/dissertation, often combined with ordinal specifications such as the first, second, etc., including البند (*al-band* 'the-item'); الفصل (*al-faṣl* 'the-chapter'); المحور (*al-miḥwar* 'the-axis'), in addition to the nouns القسم (*al-qism* 'the-section') and الباب (*al-bāb* 'the-part/chapter') particularly in dissertation abstracts. Finally, the other two items which feature frequently as frame markers in both types of abstracts are the particles أمّا (*'ammā* 'as for'), usually used to signal shifting from one topic to another, and ثُمَّ (*ṯumma* 'then').

Transition markers, which include a variety of conjunctions and prepositional phrases in the data, are also frequently used with the aim to "help readers interpret pragmatic connections between steps in an argument" (Hyland 2019: 59). He further argues for three sub-types of transition markers which are reflected in the data, as illustrated in Table 5.

Type of transition marker relation	Examples from the data
Addition	('also') كما، ('and also') وأيضاً ('in addition to') بالإضافة إلى، فضلاً عن، علاوة على
Comparison	('similarly') كذلك ('but')لكن ('in spite of') على الرغم من ('on the other hand') ومن ناحية أخرى ('whereas') غير أن، في حين، بينما ('except that') إلا أن ('in contrast') بالمقابل
Consequence	('so') فـ.. ('therefore') ولهذا، ولذلك، ولذا ('and with that') وبذلك ('consequently') ومن هنا، وبالتالي، وعليه، ومن ثَم ('as a result of') ونتيجة لذلك، يرجع ذلك إلى ('and in light of') وفي ضوء ('and based on that') وبناء على ذلك، ومن هذا المنطلق

Table 5: Types of transition marker relations according to Hyland (2019) with Arabic examples

Even though the analysis shows a low majority of the addition relation, it is the transition markers used for the consequence relation which show the broadest lexical variety in the data, ranging from one-letter prefix conjunctions (فـ *fa* 'so') to the use of both proximal (هذا *hāḏā* 'this') and distal (ذلك *ḏālika* 'that') demonstrative pronouns with or without a preposition and to a whole prepositional phrase (e.g., ومن هذا المنطلق *wa min hāḏal muntalaq* 'and from this perspective'). It is also worth noting that, as a stylistic feature of the Arabic language, it is common to combine conjunctions and/or start a clause with the conjunction و (*wa* 'and').

Finally, for journal abstracts, evidentials are ranked among the most used sub-types of textual markers, followed by code glosses and endophoric markers. Journal abstracts use more evidential markers, that is, references to opinions/ideas external to the author; in turn, there is no difference in the frequency of using code glosses, that is, restatement or explanation of ideas in the text, between the two abstract types. Moreover, it is noted that dissertation abstracts show the lowest use of endophoric markers (2.5%) while journal abstracts use double that figure.

4.1.2. Interpersonal metadiscourse (IM)

As for interpersonal metadiscourse, Table 6 below presents the detailed distribution of all its elements in the data. As can be noticed, boosters are the most frequent interpersonal devices in both types of abstracts, especially in journal abstracts which show a higher percentage of boosters than dissertation abstracts. In terms of form, almost 47 percent of all boosters in the data are divided between two particles only: قد (*qad*) and إنَّ (*'inna*) which are defined as particles for emphasis in Arabic grammar. Examples (1) and (2) illustrate their use.

(1) اختيرت عينة عشوائية من الطالبات.. **و(قد)** (Journals_F)
wa qad 'uxtīrat ʔayyina ʔašwā'iyya min aṭ-ṭālibāt
'And a random sample of female students have (*in fact*) been chosen'

(2) **فـ(إنَّ)** هذا الأمر يستلزم إلقاء الضوء علي كل المتغيرات التي تؤثر على تحقيق هذا الهدف (Dissertations_M)
fa-'inna hāḏa l-'amr yastalzim 'ilqā' aḍ-ḍaw' ʔalā kul al-mutaġayyirāt al-latī tu'aṯṯir ʔalā taḥqīq hāḏal-hadaf
'So, this matter (*does*) necessitate shedding light on all the variables which influence achieving this goal'

Note that in these examples both particles are also combined with the frame markers in the form of the conjunctions *and* and *so*, respectively. The essential characteristic of these two examples is that the emphatic particles are grammatically unnecessary, that is, they are optional, but are chosen by the writer to emphasize the message. Other boosters in the data include the verbs أظهرت (*'ađharat* 'demonstrated'), أثبتت (*'aṯbatat* 'proved'), أوضحت/بيّنت (*'awḍaḥat*/*bayyanat* 'clarified'), كشفت (*kašafat* 'revealed') and the adverbs خاصةً/لا سيّما (*xāṣatan*/*lāsiyyamā* 'especially').

Interpersonal Metadiscourse	Journal abstracts	% of total IM	Dissertation abstracts	% of total IM
Boosters	565	40.9	300	34.4
Attitude markers	220	16	211	24.1
Hedges	53	3.8	39	4.5
Self-mention	443	32	173	19.8
Engagement markers	101	7.3	150	17.2

Table 6: Distribution of interpersonal metadiscourse by abstract type

One outstanding difference between the two types of abstracts lies in the use of self-mention metadiscourse markers. In journal abstracts, the use of self-mention metadiscursive devices represents 32 percent of all interpersonal markers, being the second most frequent device attested. In dissertation abstracts, self-mention

metadiscursive devices only represent 19.8 percent of all interpersonal markers and are the third most frequent devices found in the corpus. It is worth noting here that in the analysis, the category of self-mention includes the following:

- Instances of using the noun الباحث / الباحثة (*al-bāḥiṯ* / *al-bāḥiṯa* 'the researcher-masculine/the researcher-feminine') or its dual/plural forms.
- Instances of the possessive pronoun suffixes ـي- / ـنا- (*nā* / *iy* 'my/our')
- Instances of verbs conjugated with the first person أنا (*'anā* 'singular') or نحن (*naḥnu* 'plural'), taking into consideration that the use of first-person plural could be an engagement marker especially in dissertation abstracts where the author is singular.
- Instances of a passive verb construction which refer to actions performed by the author.

In fact, as Table 7 below shows, the last sub-type, that is, passive verb constructions referring to the author, constitute 51% of the overall self-mentions in the corpus. Therefore, even though self-mention as a whole is used more commonly in journal abstracts compared to dissertation abstracts, authors of both types of texts equally use passive verb constructions referring to the author in almost half the occurrences.

Self-mention markers in the data			
Type	Journal abstracts	Dissertation abstracts	Total
Nouns, pronouns, first person conjugated verbs	212 (47.8%)	90 (52%)	302 (49%)
Passive verb constructions	231 (52.2%)	83 (48%)	314 (51%)

Table 7: Self-mention markers in the data by abstract type

The overall results of interpersonal metadiscourse also show that there are two types of interpersonal elements that are more frequently used in dissertation abstracts compared to journal abstracts. First, attitude markers are the second most frequent type of metadiscursive marker in dissertations (24%). In journal abstracts, by contrast, they only represent 16 percent of the metadiscursive markers found. Second, engagement markers are 10 percent more frequent in dissertation abstracts. According to Hyland (2019: 63), in practice, it is often difficult to distinguish between attitude markers and engagement markers, since both can be considered affective devices. This has been noted in the analysis presented here. Attitude markers, that is, expressions of the writer's appraisal of propositional content in terms of various emotions, such as surprise, agreement, obligation or importance, are mostly manifest in a variety of adjectives in the data.

However, other syntactic structures are also used to that effect. Some examples of attitude markers are illustrated in Table 8 below.

Type of attitude marker	Example
Adjectives	هام/مهم (‘important’), رائع (‘magnificent’), جاد (‘serious’) بليغ (‘eloquent’), مؤثّر (‘effective’)
Prepositional phrases	في غاية الأهمية (‘extremely important’) من الصعب (‘it is difficult’) من الضروري (‘it is necessary’) من المفضّل (‘it is preferrable’)
Verbs	يستدعي (‘requires’), يعاب عليه (‘he/she is shamed’)

Table 8: Syntactic types of attitude markers

Engagement markers, on the other hand, may include a number of syntactic features which have relational implications on the discourse and help to bring the reader as a participant in the process of reading. According to Hyland (2019: 58), this could be in the form of an explicit address to the reader with the use of second person pronouns or inclusive *we*. In the data, only the latter was easily identifiable in dissertation abstracts which clearly have a single author. In these cases, all references to the author using a plural pronoun (e.g., أننا *’annanā* ‘that we’; دراستنا *dirāsatunā* ‘our study’) or verb conjugation (e.g., نحاول *nuḥāwil* ‘we try’; لاحظنا *lāḥaḏnā* ‘we noticed’) were counted as an engagement marker. No instances of second person pronouns were attested in the data, although there were a few instances of the term القارئ (*al-qāri’* ‘the reader’) as in (3).

(3) يقف **القارئ** مذهولاً أمام الدافع الحقيقي وراء كتابة هذا الكتاب ذائع الصيت.. (Journals_M)
yaqifu al-qāri’ maḏhūlan ’amām ad-dāfiʔ al-ḥaqīqī warā’ kitābat hāḏal kitāb ḏā’iʔ al-ṣīt
‘*The reader* stands in amazement regarding the real motive behind writing this famous book’

In this example, the writer is trying to involve the reader in the topic of the paper not only by mentioning the noun explicitly but also by using affective language to signal a shared perspective. Other engagement markers which can serve as devices “rhetorically positioning the audience” (Hyland 2019: 63) include questions, obligation modals (e.g., يجب أن (*yajib ’an* ‘ought to’); لابد (*lābud* ‘must’); علينا (*ʔalainā* ‘we have to’) and references to shared knowledge (e.g., من المعلوم أن *minal maʔlūm ’an* ‘it is known that’; كما نعلم *kamā naʔlam* ‘as we know’).

Hedges are the least frequent interpersonal element in both types of abstracts, constituting only 3.8 percent and 4.5 percent respectively of all interpersonal metadiscoursive markers. Contrary to boosters, hedges indicate the writer’s reluctance to commit to the propositional content and therefore allow for information to be presented

as a personal opinion rather than as a fact. In English, these are typically expressed via adverbs such as *possibly*, *perhaps*, *rather*, etc. In the data, hedges were represented via a variety of lexical items including the verb يمكن (*yumkin* 'could'), the structure قد (*qad* 'may') plus a present tense verb to indicate possibility, the adverbs لعل (*laʔalla* 'perhaps') and ربما (*rubbamā* 'maybe'), and the prepositional phrase إلى حد ما (*'lā ḥaddin mā* 'to an extent').

4.2. Metadiscourse elements by gender

The second round of analysis takes gender as the main variable. There are three categories under gender: male, female and mixed gender (for journal papers only). It is worth noting here that the author gender distinction does not correspond to author number; in other words, only the mixed gender category necessarily implies multiple authors, while the other two categories include both single and multiple authors of the same gender. The total number of annotations are shown in Figure 2. As in the previous section, the analysis shows a majority of textual metadiscourse across gender groups, making up 55 percent of all metadiscourse elements used by males (1,326 instances), while this percentage decreases slightly to 51 percent (955 and 294 instances, respectively) for females and mixed gender.

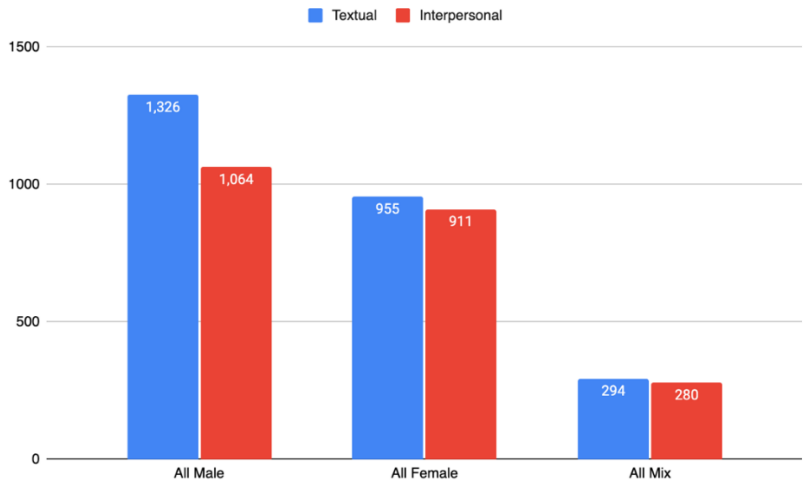


Figure 2: Total metadiscourse elements by author gender

4.2.1. Textual metadiscourse (TM)

In order to obtain a more detailed picture of the distribution of textual elements by author gender, Table 9 below summarizes the breakdown of the various sub-types.[6]

Textual metadiscoure	Male	% of total TM	Female	% of total TM	Mixed gender	% of total TM
Transition markers	508	38.3	361	37.8	96	32.7
Frame markers	595	45	419	43.9	126	42.8
Endophoric markers	50	3.8	32	3.3	21	7.2
Evidentials	86	6.4	89	9.3	31	10.5
Code glosses	87	6.5	54	5.7	20	6.8

Table 9: Distribution of textual metadiscourse by author gender

On the basis of the data, frame markers are the most frequent type of textual metadiscourse used by all gender groups, with male authors using 45 percent of the total metadiscourse markers of this type. This is not a surprising result, given the close correspondence between the use of frame markers and the overall discourse structure of abstracts, as discussed in the previous section. Similarly, transition markers are in second place, with male authors using 38.3 percent of all textual metadiscourse markers of this type.

It is interesting to note that the mixed gender category shows more frequent use of evidentials and endophoric markers in comparison to the other two genders. The female authors are also noted to use more evidentials than male authors, in whose writings only 6.4 percent of the total textual markers are evidentials.

Evidentials are defined as "metalinguistic representations of an idea from another source" (Thomas and Hawes 1994: 129), a crucial feature in academic texts, as evidentials provide important support for the author's arguments and adds a certain validity to their academic worth. In English academic texts, these markers are typically expressed by an in-text citation from a source, often associated with structures such as *according to x*, *x argues that*, etc. In the Arabic data, two types of evidentials were detected: (a) specified references and (b) unspecified references. The former typically mention the name/year of the reference, sometimes preceded by the expression وفقاً لـ/حسب (*wifqan li / ḥasb* 'according to'). The latter do not mention a specific name, but the author rather makes a general reference to opinions from external sources. These can be expressed by a variety

[6] See also Appendix 2 for the quantitative data on metadiscourse elements by author gender with normalized frequency per 1000 words, and a comparison graph by percentage.

of nouns, including النقاد العرب (*an-nuqqād al-ʔarab* ‘Arab critics’), جمهور الفقهاء (*jumhūr al-fuqahā’* ‘the majority of jurists’) or just the vague البعض (*al-baʔḍ* ‘some’).

Endophoric markers, on the other hand, are scarcely used by the female group (3.3%) in comparison to the other two groups and to all other types of textual markers. The main function of endophoric markers is to establish linking relations between different parts of the text. Therefore, expressions referring to preceding or following parts of the text would act as endophoric markers. In our data, most of the endophoric markers were used to refer to parts of the text yet to come, that is, making cataphoric (rather than anaphoric) reference. Example (4) illustrates the two types of endophoric references in the data.

(4) وإنطلاقاً **مما سبق**، فإن الدراسة عكفت بالاستناد إلى مقولات النقد النسوي على تحقيق **ما سبق** من خلال ثلاثة محاور هي **على النحو الآتي:** (Journals_Mix)
wa-nṭilāqan mimmā sabaq fa’inna d-dirāsa ʔakafat bil-istināد ’ilā maqūlāt al-naqd al-nasawī ʔalā taḥqīq mā sabaq min xilāl ṯalāṯat maḥāwir hiya ʔala n-aḥw at-tālī
‘And based on *what has been mentioned*, the study, relying on works of feminist criticism, set out to achieve *what has been mentioned* through three axes which are *the following*’

In addition to the example above, other endophoric markers in the data include the expressions التالي (*at-tālī* ‘the following’), ما يأتي (*mā ya’tī* ‘what follows’), فيما يلي (*fīmā yalī* ‘in what follows’) for cataphoric reference (usually followed by a colon), as well as ما تقدّم (*mā taqaddam* ‘what was mentioned before’) for anaphoric reference.

The figures of code glosses are rather similar across all gender groups with an average of approximately 6 percent, albeit the female author category is noted to have the lowest percentage (see Table 9). This type of metadiscourse helps writers to explain or elaborate on their propositional content with the aim of facilitating comprehension for the reader. This can be done by providing examples, giving additional explanation, rephrasing part of the text, etc. In the data, it was found that providing examples is the most common purpose of code glosses, where examples are introduced by various expressions from the standard كـ/مثل (*ka / miṯl* ‘for example’) to the more elaborate على سبيل المثال (*ʔalā sabīl al-miṯāl* ‘as way of an example’). In addition, different forms of code glosses were used for introducing explanations or rephrasing including أي (*’ayy* ‘that is’), بمعنى آخر (*bimaʔnā ’āxar* ‘in other words’), ما يسمّى (*mā yusammā* ‘what is called’), هذا يعني أن (*hāḏā yaʔnī ’an* ‘this means that’), etc.

4.2.2. Interpersonal metadiscourse (IM)

As far as interpersonal metadiscourse is concerned, Table 10 below presents the breakdown of the distribution of all its elements in the data by author gender. As expected, boosters retain their position as the most widely used feature of interpersonal metadiscourse. The analysis also reveals that the category of mixed gender has the highest percentage of boosters (42.5%), while the female author category has the lowest percentage (35.8%). Example (5) below illustrates a case of double boosters from the mixed gender category.

Interpersonal metadiscourse	Male	% of total IM	Female	% of total IM	Mixed gender	% of total IM
Boosters	420	39.5	326	35.8	119	42.5
Attitude markers	197	18.5	184	20.2	50	17.8
Hedges	43	4	39	4.3	10	3.8
Self-mention	283	26.6	244	26.8	89	31.7
Engagement markers	121	11.4	118	12.9	12	4.2

Table 10: Distribution of interpersonal metadiscourse by author gender

(5) ... التي **الهامة** على المكانة أيضاً **ويدلّ** على [...] **يدلّ بوضوح** الأمر الذي (Journals_Mix)
al-’amr al-laḏī yadullu biwuḍūḥ ʔalā … wa yadullu ‘ayḍan ʔalā l-makānat al-hāma al-latī
‘The fact that *clearly shows* that […] and also *indicates* the *important* status which…’

Attitude markers, on the other hand, rank third place across all gender groups, with females showing the highest percentage (20.2%). As, examples (6) and (7) illustrate, female authors tend to accumulate several attitude markers in the same phrase.

(6) ويعتبر استخدام الواقع الافتراضي في العلوم المختلفة **ضرورة حتمية لا مفر منها** (Journals_F)
wa-yuʔtabar ’istixdām al-wāqiʔ al-iftirāḍiyy fī l-ʔulūm al-muxtalifa ḍarūra ḥatmiyya lā mafarr minha
‘The use of virtual reality in different disciplines is considered an *absolute inevitable necessity*’

(7) تشير نتائج هذه الدراسة إلى **أهمية** تقديم خدمة إذاعية **هادفة ومفيدة ومثيرة للإهتمام** (Dissertations_F)
tušīru natā’ij hāḏihi l-dirāsa ’ilā ’ahamiyyat taqdīm xidmah ’iḏāʔiyya hādifa wa mufīda wa muṯira lil’ihtimām
‘The results of this study indicate the *importance* of providing a broadcasting service which is *purposeful, beneficial and interesting*’

Results also show that self-mention is the second most frequent type of interpersonal metadiscourse across all gender groups, with the category of mixed gender exhibiting the highest percentage (31.7%). Looking at the detailed numbers for types of self-mention in Table 11, it is observed that this category shows the highest percentage of passive verb

constructions referring to the author (54%). As mentioned before, passive verb constructions referring to actions performed by the author(s) seem to contribute heavily to the representation of the author(s) in the texts. Passive grammatical constructions in general are impersonal structures in language, and as such they are frequently associated with academic writing. In Arabic specifically, there are two ways to formulate the passive (Ryding 2005: 657): (a) the inflectional passive, which is constructed by altering the vowel pattern of the verb; and (b) the periphrastic passive, which is constructed with the help of a dummy verb meaning completed/finished such as تمّ (*tamma* 'done'). Examples (8) and (9) illustrate the two types respectively.

Type	Male	Female	Mixed gender
Nouns, pronouns, first person conjugated verbs	135 (47.7%)	126 (51.6%)	41 (46%)
Passive verb constructions	148 (52.3%)	118 (48.4%)	48 (54%)

Table 11: Self-mention markers in the data by author gender

(8) ...وقد **اختيرت** عينة عشوائية من معلمي رياض الأطفال (Journals_F)
wa qad 'uxtīrat ʔayyina ʔašwā'iyya min muʔallimī riyāḍ al-aṭfāl
'And a random sample of kindergarten teachers *were chosen*…'

(9) ...**وتم تطبيق** الأداتين على عينة بلغت خمس وعشرون طالب وطالبة **تم اختيارهم** عشوائياً من كليتين (Journals_M)
wa tamma taṭbīq al-'adātatain ʔalā ʔayyina balaġat xams wa ʔušrūn ṭālib wa ṭāliba tamma 'ixtiyārahum ʔašwā'iyyan min kuliyyatain
'And the two tools *were applied* to a sample of 25 male and female students who *were* randomly *chosen* from two colleges…'

In fact, an in-depth analysis of types of passive verb constructions shows that 72.3 percent of all passive constructions in the data are realizations of the periphrastic passive. This percentage is even higher in the male and female groups individually with 80 percent in each being periphrastic passives. There is very little research on the stylistic differences between the two structures; however, Larcher and Girod (1990, quoted in Mansouri 2016: 234) maintain that one of the reasons for the dominance of periphrastic passive in modern standard Arabic is to avoid confusion between active and passive readings of the verb in unvowelized texts. All of our data (with very few exceptions) and most of Arabic academic writing being unvowelized, this explanation seems plausible.

Attitude and engagement markers are most frequently used by the female gender category, while the mixed gender category has the lowest frequency of hedges and engagement markers. In addition to the use of inclusive *we* and modal verbs, expressions of shared knowledge, which can be linguistically manifest in various ways, were one of

the trickiest types of engagement markers to detect. Examples (10) and (11) illustrate some cases from the data:

(10) ومن **الأفكار الذائعة** في القانون فكرة ارتباط التأمين بالخطر (Disserations_M)
wa min al-’afkār al-ḏā’iʔa fil-qānūn fikrat ’irtibāṭ at-ta’mīn bil-xaṭar
‘And *among the common ideas* in law is the idea which associates insurance with danger’

(11) **وبما أن** العقل هو المحرك الأساسي للإنسان وللعقل عاداته التي نتصرف بها.. (Journals_F)
wa bimā ’anna l-ʔaql huwa l-muḥarrik al-’asāsī lil-’insān wa lil-ʔaql ʔādātuh al-latī nataṣarraf bihā
‘And *given that* the mind is the main drive for the human being, and the mind has its habits to which we behave accordingly’

Finally, hedges are the least frequent interpersonal metadiscourse markers across all gender groups, with similar range of percentages compared to the results by abstract type.

5. Discussion

In general, writing Arabic academic abstracts does not seem to be governed by explicit rhetorical rules. As part of academic writing in modern standard Arabic, it can be said that abstracts assume the status of a ‘borrowed genre’ (Najjar 1990; see also Al-Qahtani 2006) that has been influenced by academic practices in English (and in French in some parts of the Arab world). As Hyland (2014: 13) explains “academic writers do not simply produce texts that plausibly represent an external reality, but use language to acknowledge, construct and negotiate social relations;” hence, the various features of metadiscourse examined in this study have shown how writers of academic abstracts in Arabic establish such social relations with their audience.

This study attempted to address two research questions: (a) what the overall distribution of metadiscourse elements in Arabic academic abstracts is; (b) what the distributional variations of metadiscourse markers across both abstract type and author gender are. As far as the first research question is concerned, the quantitative analysis in this study is the first to offer a comprehensive overview of the overall distribution of metadiscourse markers based on a substantial corpus of 400 Arabic abstracts. The results have shown that native Arabic speakers writing in the academic genre are aware of the writer-reader relationship nuances and how they can be manipulated. The rhetorical dynamics of Arabic academic abstracts is rich and reflects an intricate mix of writer-oriented and reader-oriented metadiscoursal features. This is manifest in the abundance of both textual and interpersonal markers, where the textual ones only slightly edge over

the interpersonal across all types and groups. This is consistent with previous studies on metadiscourse in English academic texts (e.g., Hyland 1998 on research articles; Hyland 2019 on academic dissertations) and in Arabic ones (e.g., Alotaibi 2015). Therefore, writers of Arabic academic abstracts actively use metadiscourse elements to influence both the readers' understanding of the text and the writer's attitude to its propositional content.

As regards the textual domain, whose function is "to help form a convincing and coherent text" (Hyland 1998: 442) by means of relating parts of the text to each other and to other texts, the analysis shows that transition markers and frame markers are by far the most frequent metadiscourse markers in the data. In the textual domain, such a high frequency of those two features has not been contested in other studies. Dahl (2004b), for instance, who conducted a comparative study of textual metadiscourse in academic papers written in English, French and Norwegian, firmly situates the abundance of textual markers within the Anglo-Saxon tradition of emphasizing the importance of communicating with the reader and making this "an explicit feature of the writing process" (Dahl 2004b: 1821). Similarly, in Arabic academic reference books, the position of the reader is omnipresent, and many justifications for prescribing certain rules in academic writing explicitly mention that it is for the benefit of the reader. Al-Shahrani (2010) even argues that academic writing has a 'special audience' who judge the quality of that writing on the basis of both scientific and linguistic standards. Therefore, it is no surprise that transition and frame markers abound in the data to provide explicit links between different parts of the text. Alotaibi (2015) also concluded that Arabic texts rely heavily on transition markers, especially of the addition relation, and noted the extensive use of frame markers as move openings in Arabic abstracts, which is consistent with the findings in the present study. On the other hand, the results for interpersonal metadiscourse markers reflected some patterns of usage that do not necessarily match those of other studies. Interpersonal metadiscourse, in particular, has been heavily researched as the most personal type of metadiscourse, where the "author's perspective towards both propositional information and readers themselves" is expressed (Hyland 2019: 61). In the present analysis, the results show a strong use of interpersonal elements in Arabic academic abstracts, despite the emphasis on objectivity in Arabic academic writing.

The second research question is concerned with the more specific distributional variations in relation to the two variables: text type and author gender. The analysis yielded some interesting results. Starting with textual markers, evidentials have shown the highest frequencies in journal abstracts and by mixed gender authors. However, in this study the category of evidentials was not restricted to explicit references in the form of an in-text citation. Since Hyland (2019: 61) explains that this type of metadiscourse serves to "distinguish who is responsible for a position" with the aim of strengthening an argument, it was deemed appropriate also to include here what has been labeled as 'unspecified reference', such as *some say, Arab critics argue*, etc. It is unclear what role such expressions play in academic writing of other languages, but since they do assign certain opinions to sources external to the author, they were included in the data. It is interesting to note, though, that evidentials generally play a more prominent role in English academic writing (e.g., Hyland 1998, 2019), and are especially more frequent in soft disciplines (e.g., Hyland 2004b; Khedri 2018). Also, Ozdemir and Longo (2014) found that native speakers of English used evidentials more frequently in their thesis abstracts compared to non-native speakers. Therefore, it seems that academic expectations regarding references to others' work can be culture-specific. Patterns of evidentials use in academic texts by gender are scarce, but both Pasaribu (2017) and Yeganeh and Ghoreyshi (2015) found that females tend to use more evidentials to support their arguments. Our results from Arabic are consistent with this tendency, which even seems to influence the mixed gender group to have the highest percentage of evidentials (10.5%) in comparison to the male group (6.4%).

Code glosses and endophoric markers, on the other hand, play a minor role in Arabic academic abstracts. The highest frequency of endophoric markers was 7.2 percent out of the total textual metadiscourse elements (by mixed gender authors), while the lowest was 2.5 percent (in dissertation abstracts). The high frequency in the mixed gender group could be attributed to the fact that those abstracts are written by multiple authors, although this justification needs to be verified by further examination of all multiple author abstracts in the other gender groups. Tse and Hyland (2008) found no major gender-based differences in the use of code glosses and endophoric markers, but in the current study it was noted that female authors have the lowest frequencies of both.

As for interpersonal markers, it is noted that boosters, which emphasize the force of propositions and "imply certainty" (Hyland and Tse 2004: 168), were the most frequent

element in the data with an average of 37.6 percent by abstract type and 39.2 percent by author gender of the total interpersonal metadiscourse markers. It is also noted that journal abstracts had 6 percent more boosters than dissertation abstracts, and mixed gender abstracts had almost 7 percent more boosters than female authors. This contrasts with the use of hedges, which function to downplay the writer's commitment to any certainty. In fact, hedges presented the lowest proportion of interpersonal metadiscourse markers in the data with an average of just 4 percent across abstract types and author genders. The high percentage of boosters in Arabic abstracts is comparable to what Alzarieni *et al.* (2019) found in their data of Arabic patent abstracts, where boosters formed 53 percent of the total interpersonal metadiscourse elements. However, the specificity of patents as a field might be more influential than expected in the way writers balance their commitment to their statements. The same study found that hedges form 42 percent of all interpersonal metadiscourse features. This combination of boosters and hedges shows that hedges can sometimes be used to mitigate boosters. Gillaerts and van de Velde (2010) note the same phenomenon of hedged boosters in their data of English journal abstracts. However, in our study, hedges do not seem to play any significant role, and writers, both male and female, do not shy away from boosting their arguments. In fact, Al-Gublan (2013), who compares the use of hedges in English and Arabic scientific research papers, highlights that English writers tend to use hedges more frequently, while Arab writers avoid them in order to be "more precise and accurate" (Al-Gublan 2013: 205). The high frequency of hedges in English academic writing in general is attested in the literature (e.g., Hyland 1998, 2014, 2019), whereas there is little research on hedging in Arabic academic writing. However, it seems that the high frequency of boosters and low frequency of hedges in Arabic abstracts reflect a tendency for being more straightforward and not open for interpretation.

Self-mention is another interpersonal marker which provides interesting results. According to Hyland and Tse (2004: 170) this type of metadiscourse "reflect[s] the degree of author presence," typically through the use of personal and possessive pronouns. While they acknowledge that English academic writing preaches the avoidance of using the first person, they emphasize the importance of self-mention in creating a "scholarly identity" (2004: 172) for the author. However, care must be taken when comparing frequencies of self-mention for two reasons: (a) studies vary in their delineation of what counts as self-mention; and (b) the use of self-mention in abstracts only cannot be compared to its use in academic writing in general due to the specific nature of abstracts and academic

expectations regarding author presence in them. In our study, self-mentions had the highest frequency in journal abstracts and by mixed gender authors at an average of 31.8 percent of all metadiscourse elements. There was no variation in the frequency of self-mention markers between male and female authors.[7] This relatively high frequency seems to run counter to results in other studies. Alotaibi (2015: 8), for example, argues that the low frequency of self-mentions in his results suggests that "Arabic-speaking writers tend to avoid self-mentions whether they are writing in their first language or in English." Similarly, Alzarieni *et al.* (2019) report only two percent frequency for self-mentions in their Arabic data. Alharbi and Swales (2011), who only analyze first person pronouns, report a low frequency of self-mention in the Arabic abstracts and attribute this to "cultural perceptions that the written description of 'research' properly requires a more formal style employing the passive and/or self-referring expressions such as 'this paper/study/research'" (Alharbi and Swales 2011:75). While these results are not completely in line with ours, some observations tie in with our findings regarding the use of passive constructions referring to the author. In fact, the high frequency of passive constructions in the data encourages further investigation into the function of these constructions, and whether they are mainly used to maintain textual cohesion or also contribute to the creation of writer stance (Baratta 2009).

To further confound the issue of self-mention in academic texts, in their study of 72 research article abstracts in English, Gillaerts and van de Velde (2010) ignore two types of the interpersonal metadiscourse elements in their analysis: self-mention and engagement markers. Their reason for excluding self-mentions is that "there is no agreement on their interpersonal effect" (Gillaerts and van de Velde 2010: 131), and that the use of first person pronouns can make a text even less subjective than when they are implicit. Due to limitations of space, this paper does not discuss the different types of self-mention and their role in creating writer identity in discourse in detail. It is worth mentioning, though, that the results of the present study go along the lines of Ivanič's (1998: 26) assertion that "writers differ considerably in how far they claim authority as the source of the content of the text, and in how far they establish an authorial presence

[7] One reviewer pointed out that the high percentage of self-mentions in the mixed gender group could probably relate to the fact that it is the multiple author group, and that there is a cultural tendency to give credit to the authors in a multi-authored article but not in single authored ones. While there is no direct research which supports this tendency in Arabic academic writing, it might be worth investigating this aspect in future research (but see Al-Shujairi 2020 who studies self-mention in single-authored research articles only because it guarantees the use of self-mention in a way that multi-authored work does not).

in their writing." In this study, and as far as Arabic academic abstracts are concerned, a broad definition of self-mentions is warranted as it gives a clearer picture of how far Arab authors establish their authorial presence, whether using explicit references (e.g., personal or possessive pronouns, the noun phrase *the researcher*) or implicit references (e.g., passive constructions).

Finally, another result that does not seem to be consistent with previous studies has to do with engagement markers. Both Alotaibi (2015) and Alzarieni *et al.* (2019) report on an absence of engagement markers in Arabic abstracts. Alotaibi (2015: 8) considers this a surprising result but explains that it indicates that

> Arab writers perceive the genre of abstract, whether the English or the Arabic one, to be free from any engagement with the reader as this may project a conversational and an informal tone.

Even in English, Gillaerts and van de Velde (2010: 131) exclude engagement markers from their study on the basis that they are "virtually absent" in abstracts and "because the few elements that may qualify as engagement markers are hardly distinguishable from attitude markers." Likewise, both Ozdemir and Longo (2014) and Nugroho (2019) report zero occurrences of engagement markers in both native and non-native abstracts. The latter justifies this by arguing that, due to their length and function, abstracts are not an ideal ground to establish a direct relationship with the reader, a feature that is more appropriate for an introduction or a discussion section in academic writing. Indeed, Hyland and Tse (2004) found that engagement markers are a fifth of all interpersonal metadiscourse in dissertations, while in Hyland (1998) engagement markers were the least frequent metadiscourse element with only 3.5 percent of the total. However, in Hyland (2019), engagement markers were strongly present in doctoral dissertations, with the highest frequency in the humanities (applied linguistics).

In the present study, engagement markers were definitely present although not with a high frequency. The highest occurrences were in dissertation abstracts, which formed 17 percent of total interpersonal metadiscourse features. According to Hyland (2014), there are two main purposes for writers to use engagement strategies: (a) to establish a relationship with the readers and include them in the argument; and (b) to position the audience in a particular path and guide their thinking. For the former, in our data, this was exclusively done through the use of inclusive *we*, which appeals to solidarity with the reader, despite warnings in some Arabic academic writing books against using this

structure (see Hassan 1996: 70). For the latter purpose, our data show a variety of elements including questions, obligation modals and representations of shared knowledge. It is believed that a combination of the corpus size and the specification of discipline in this study has led to the detection of engagement markers in Arabic abstracts. After all, Arabic academic writing acknowledges the importance of establishing a relationship with the reader "so that the reader understands the text in the way intended by the researcher" (Al-Shahrani 2010: 15). Engagement markers, though small in number, help to achieve this goal. It was also noted that dissertation abstracts had higher frequencies of both attitude and engagement markers compared to journal abstracts, which may be attributed to the more 'personal relationship' of the writer with their dissertation, as a result of a much longer time commitment and dedication. Finally, the results show that both the male and female groups use engagement markers more frequently with little variation (an average of 12%), whereas this percentage drops to just 4 percent when authors belong to a mixed gender group. More research is needed to ascertain the effects of mixed gender on academic writing, but the results here suggest that it leads to the authors focusing more on referring to themselves and less on engaging with the reader.

6. Concluding remarks

Since its early days when the concept of metadiscourse was vaguely defined as "discourse about discourse" (Kopple 1985: 83), the term has proven to be usefully operationalized for a better representation of the tools at the writers' disposal to engage with their audience. Hyland's (2019: 4) theoretical framework employed in this study has helped in understanding abstracts as a "social engagement," by exploring the ways writers project themselves onto the discourse. This study has shown that writers of Arabic academic abstracts use a wide variety of textual and interpersonal metadiscourse features, and that they use them quite homogeneously. The results indicate high frequencies of transition and frame markers on the one hand, and boosters and self-mentions on the other. The lowest frequencies were for endophoric markers and hedges. The results for self-mentions and engagement markers, in particular, raise some questions regarding the scope of these elements in academic Arabic. In terms of text types, dissertation abstracts have more attitude and engagement markers than journal abstracts, while authors use self-mention

markers more in journal abstracts. In terms of author gender, male authors use frame markers more frequently, while female authors use engagement markers the most.

One important finding in the analysis is that when it comes to the study of metadiscourse, it is worthwhile analyzing data word by word instead of relying on a pre-set data of lexical items for each type. Time-consuming as this task may be, it is a more accurate and reliable method in order to capture instances of metadiscourse. As the analysis here has shown for a language such as Arabic, many categories of metadiscourse can be expressed through a wide variety of lexical items and syntactic structures.

This study intends to provide a solid foundation for the study of metadiscourse in Arabic academic abstracts, whether for stylistic or pedagogical purposes. Yet, one of the limitations of the study is the scope of the discipline. The data analysed here belongs to the fields of humanities and social sciences only. As previous studies have shown, other academic disciplines may have their own rhetorical particularities. Therefore, it is hoped that this study will pave the way for more detailed studies of metadiscourse in Arabic across disciplines, as well as for contrastive studies of Arabic/English academic abstracts based on larger corpora and paying special attention to variations in the nature of metadiscoursal expressions in both languages, as well as to gender-based variations in Arabic academic discourse.

REFERENCES

Al-Ghoweri, Helen A. and Murad M. Al Kayed. 2019. A comparative study of hedges and boosters in English and Jordanian Arabic: Economic newspaper articles as a case study. *Theory and Practice in Language Studies* 9/1: 52–59.

Al-Gublan, Badriah. 2013. Hedging strategies in English and Arabic scientific written discourse. *Arab Journal for the Humanities* 31/124: 183–208.

Alharbi, Lafi M. and John Swales. 2011. Arabic and English abstracts in bilingual language science journals: Same or different? *Languages in Contrast* 11/1: 70–86.

Alhumidi, Hamed and Sani Uba. 2016. College students' use of metadiscourse across two languages: A case of college students at the College of Basic Education, Kuwait. *International Journal of English Language Teaching* 4/9: 14–29.

Alhuqbani, Mohammed Nasser. 2013. Genre-based analysis of Arabic research article abstracts across four disciplines. *Journal of Educational and Social Research* 3/3: 371–382.

Alotaibi, Hmoud. 2015. Metadiscourse in Arabic and English research article abstracts. *World Journal of English Language* 5/2: 1–8.

Al-Qahtani, Abdulkhaleq. 2006. *A Contrastive Rhetoric Study of Arabic and English Research Article Introductions*. Oklahoma: Oklahoma State University dissertation.

Alqahtani, Sahar and Safaa Abdelhalim. 2020. Gender-based study of interactive metadiscourse markers in EFL academic writing. *Theory and Practice in Language Studies* 10/10: 1315–1325.

Al-Shahrani, Saad. 2010. Al-kitāba Al-akādimiyya: Xaṣā'iṣuha wa mutaṭallabātuha al-luġawiyya (Academic writing: Its characteristics and linguistic requirements). https://educationalresearchers.files.wordpress.com/2018/12/d8a7d984d983d8aad8a7d8a8d8a9-d8a7d984d8a3d983d8a7d8afd98ad985d98ad8a9-d8aed8b5d8a7d8a6d8b5d987d8a7-d988d985d8aad8b7d984d8a8d8a7d8aa.pdf (5 August, 2021.)

Al-Sharif, Abdullah. 1996. *Manāhij Al-baḥṯ Al-ʔilmiyy: Dalīl aṭ-ṭālib fī kitābat al-abḥāṯ al-ʔilmiyya* (*Methodologies of Scientific Research: The Student's Guide to Writing Scientific Research*). Alexandria: Maktabat Al-'išʔāʔ Publisher.

Al-Shujairi, Yasir. 2020. What, which and where: Examining self-mention markers in ISI and Iraqi local research articles in applied linguistics. *Asian Englishes* 22/1: 20–34.

Alzarieni, Mahmoud, Manal Intan Safinaz Zainudin, Norsimah Mat Awal and Mohamed Zain Sulaiman. 2019. Interactional metadiscourse markers in the abstract Sections of Arabic patents. *Arab World English Journal* 10/2: 379–393.

Ayers, Gael. 2008. The evolutionary nature of genre: An investigation of the short texts accompanying research articles in the scientific journal *Nature*. *English for Specific Purposes* 27/1: 22–41.

Baratta, Alexander M. 2009. Revealing stance through passive voice. *Journal of Pragmatics* 41/7: 1406–1421.

Belcher, Wendy Laura. 2009. *Writing your Journal Article in 12 Weeks: A Guide to Academic Publishing Success*. Thousand Oaks, California: Sage Publications

Bouziane, Abdelmajid and Fatima Ezzahra Metkal. 2020. Differences in research abstracts written in Arabic, French, and English. *English Studies at NBU* 6/2: 233–248.

Briones, Roy Randy. 2018. Move analysis of abstracts in applied linguistics research: The Middle East and North Africa (MENA) perspective. *Asian Journal of English Language Studies* 6: 25–55.

Crismore, Avon. 1989. *Talking with Readers: Metadiscourse as Rhetorical Act*. Bern: Peter Lang.

Dahl, Trine. 2004a. Some characteristics of argumentative abstracts. *Akademisk Prosa* 2: 49–69.

Dahl, Trine. 2004b. Textual metadiscourse in research articles: A marker of national culture or of academic discipline? *Journal of Pragmatics* 36/10: 1807–1825.

D'Angelo, Larissa. 2008. Gender identity and authority in academic book reviews: An analysis of metadiscourse across disciplines. *Linguistica e Filologia* 27: 205–221.

El-Seidi, Mohamed. 2000. Metadiscourse in English and Arabic argumentative writing: A cross-linguistic study of texts written. In Zaynab Ibrahim, Sabiha Aydelott and Nagwa Kassabgy eds. *Diversity in Language: Contrastive Studies in Arabic and English Theoretical and Applied Linguistics*. Cairo: AUC Press, 111–126.

Fallatah, Wafaa. 2016. Features of Saudi English research articles abstracts. *Arab World English Journal* 7/2: 368–379.

Fischer, Beth A. and Michael J. Zigmond. 2004. Components of a research article. http://sites.psu.edu/bee11/wp-content/uploads/sites/16010/2011/03/Components-of-a-Research-Article.pdf (5 August, 2021.)

Gillaerts, Paul and Freek van de Velde. 2010. Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes* 9/2: 128–139.

Halliday, Michael A. K. 1977. *Explorations in the Functions of Language*. London: Edwards Arnold.

Hassan, Ahmad. 1996. *Usūl al-Baḥṯ al-ʔilmiyy*. (*Principles of Scientific Research*) *Parts I and II*. Cairo: Al-Maktaba Al-Akādīmiyya.

Huckin, Thomas. 2006. Abstracting from abstracts. In Martin Hewings ed. *Academic Writing in Context: Implications and Applications.* London: Bloomsbury Publishing, 93–103.

Hyland, Ken. 1998. Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics* 30/4: 437–455.

Hyland, Ken. 1999. Talking to students: Metadiscourse in introductory coursebooks. *English for Specific Purposes* 18/1: 3–26.

Hyland, Ken. 2000. *Disciplinary Discourses: Social Interactions in Academic Writing.* London: Longman.

Hyland, Ken. 2004a. *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor: University of Michigan Press.

Hyland, Ken. 2004b. Disciplinary interactions: Metadiscourse in L2 postgraduate writing. *Journal of Second Language Writing* 13/2: 133–151.

Hyland, Ken. 2014. Engagement and disciplinarity: The other side of evaluation. In Gabriella Camiciotti Del Lungo and Elena Tognini Bonelli eds. *Academic Discourse: New Insights into Evaluation*. Bern: Peter Lang, 13–30.

Hyland, Ken. 2019. *Metadiscourse: Exploring Interaction in Writing*. London: Bloomsbury.

Hyland, Ken and Polly Tse. 2004. Metadiscourse in academic writing: A reappraisal. *Applied Linguistics* 25/2: 156–177.

Ivanič, Roz. 1998. *Writing and Identity: The Discoursal Construction of Identity in Academic Writing*. Amsterdam: John Benjamins.

Jin, Xin and Yan Shang. 2016. Analyzing metadiscourse in the English abstracts of BA theses. *Journal of Language Teaching and Research* 7/1: 210–215.

Khedri, Mohsen. 2018. Evidentials in research articles: A marker of discipline. *Journal of Social Sciences and Humanities* 26: 145–158.

Kopple, William J. Vande. 1985. Some exploratory discourse on metadiscourse. *College Composition and Communication* 36/1: 82–93.

Larcher, Pierre and Alain Girod. 1990. Passif grammatical, passif periphrastique et categorie d'auxiliaire en Arabe classique moderne. *Arabica* 37/2: 137–150.

Lorés, Rosa. 2004. On RA Abstracts: From rhetorical structure to thematic organisation. *English for Specific Purposes* 23/3: 280–302.

Mansouri, Aous. 2016. *Stative and Stativizing Constructions in Arabic News Reports: A Corpus-based Study*. Boulder: University of Colorado dissertation.

Miech, Edward J., Bill Nave and Frederick Mosteller. 2005. The 20,000 article problem: How a structured abstract can help practitioners sort out educational research. *Phi Delta Kappan* 86/5: 396–400.

Najjar, Hazem. 1990. *Arabic as a Research Language: The Case of the Agricultural Sciences*. Michigan: University of Michigan dissertation.

Nugroho, Ardi. 2019. Exploring metadiscourse use in thesis abstracts: A cross-cultural study. *Journal of English Language and Culture* 9/2: 113–127.

Ozdemir, Neslihan Onder and Bernadette Longo. 2014. Metadiscourse use in thesis abstracts: A cross-cultural study. *Procedia-Social and Behavioral Sciences* 141: 59–63.

Pasaribu, Truly. 2017. Gender differences and the use of metadiscourse markers in writing essays. *International Journal of Humanity Studies* 1/1: 93–102.

Rainer, Simon, Elton Barker, Leif Isaksen, Pau de Soto, Valeria Vitale and Rebecca Kahn. 2016. *Recogito*. Austrian Institute of Technology GmbH. https://recogito.pelagios.org/

Robson, Jocelyn, Becky Francis and Barbara Read. 2002. Writes of passage: Stylistic features of male and female undergraduate history essays. *Journal of Further and Higher Education* 26/4: 351–362.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Serholt, Sofia. 2012. *Hedges and Boosters in Academic Writing – A Study of Gender Differences in Essays Written by Swedish Advanced Learners of English*. Gothenburg: University of Gothenburg BA dissertation. http://hdl.handle.net/2077/29526

Stotesbury, Hilkka. 2003. Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes* 2/4: 327–341.

Sultan, Abbas. 2011. A contrastive study of metadiscourse in English and Arabic linguistics research articles. *Acta Linguistica* 5/1: 28–41.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, John and Christine B. Feak. 2004. *Commentary for Academic Writing for Graduate Students: Essential Tasks and Skills*. Ann Arbor: University of Michigan Press.

Thomas, Sarah and Thomas P. Hawes. 1994. Reporting verbs in medical journal articles. *English for Specific Purposes* 13/2: 129–148.

Thompson, Geoff and Puleng Thetela. 1995. The sound of one hand clapping: The management of interaction in written discourse. *Text – Interdisciplinary Journal for the Study of Discourse* 15/1: 103–127.

Tse, Polly and Ken Hyland. 2008. Robot Kung fu: Gender and professional identity in biology and philosophy reviews. *Journal of Pragmatics* 40: 1232–1248.

Ventola, Eija. 1994. Abstracts as an object of linguistic study. In Světla Čmejrková, František Daneš and Eva Havlová eds. *Writing vs. Speaking: Language, Text, Discourse, Communication.* Tübingen: Gunter Narr Verlag, 333–335.

Yeganeh, Maryam Tafaroji and Seyedeh Marzieh Ghoreyshi. 2015. Exploring gender differences in the use of discourse markers in Iranian academic research articles. *Procedia – Social and Behavioral Science* 192: 684–689.

Young, Petey. 2006. *Writing and Presenting in English: The Rosetta Stone of Science*. Amsterdam: Elsevier.

Corresponding author
Mai Zaki
Department of Arabic and Translation Studies
American University of Sharjah
PO Box 26666, Sharjah
United Arab Emirates
e-mail: mzaki@aus.edu

received: October 2021
accepted: June 2022

APPENDICES

Appendix 1: Textual and interpersonal metadiscourse elements in journal abstracts and dissertation abstracts

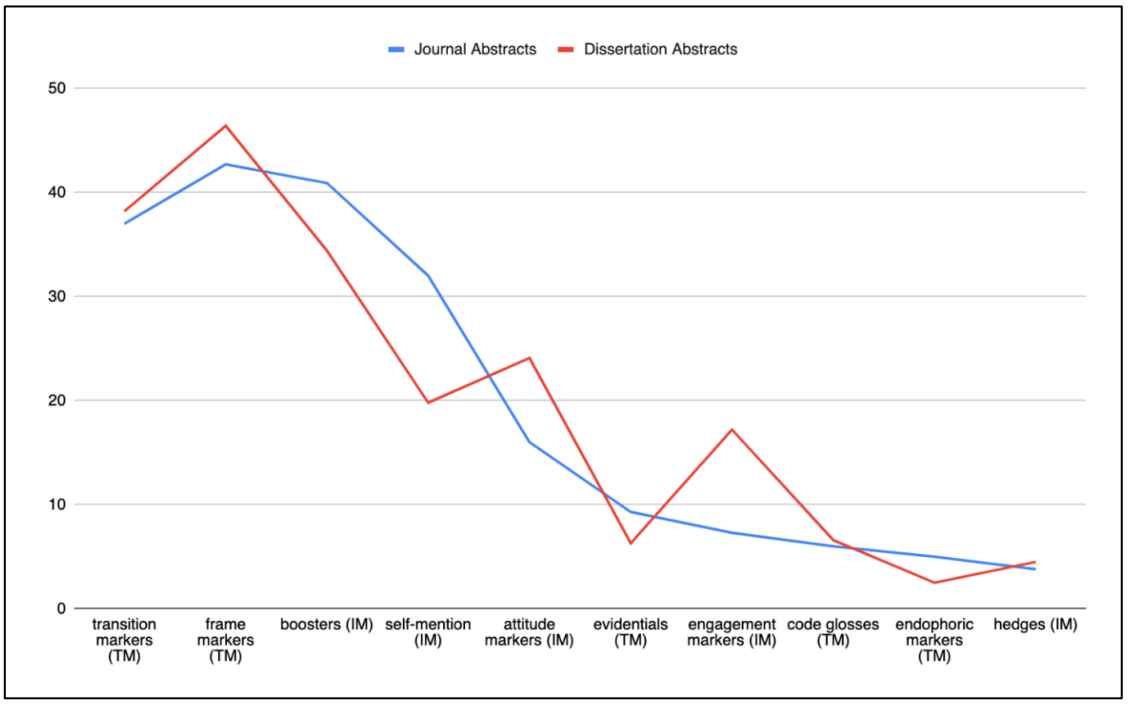
Textual metadiscourse elements in journal abstracts per 1,000 words:

Textual metadiscourse per 1,000 words	Journals	Dissertations	All
transition markers	12.16	14.97	13.23
frame markers	14.05	18.20	15.63
endophoric markers	1.66	1.01	1.41
evidentials	3.06	2.45	2.82
code glosses	1.97	2.59	2.21

Interpersonal metadiscourse elements in dissertation abstracts per 1,000 words:

Interpersonal metadiscourse per 1,000 words	Journals	Dissertations	All
boosters	12.52	10.79	11.86
attitude markers	4.87	7.59	5.91
hedges	1.17	1.40	1.26
self-mention	9.81	6.22	8.45
engagement markers	2.24	5.40	3.44

Comparing percentage of total textual metadiscourse (TM) and of total interpersonal metadiscourse (IM) by abstract type:



Appendix 2: Textual and interpersonal metadiscourse elements in abstracts by gender

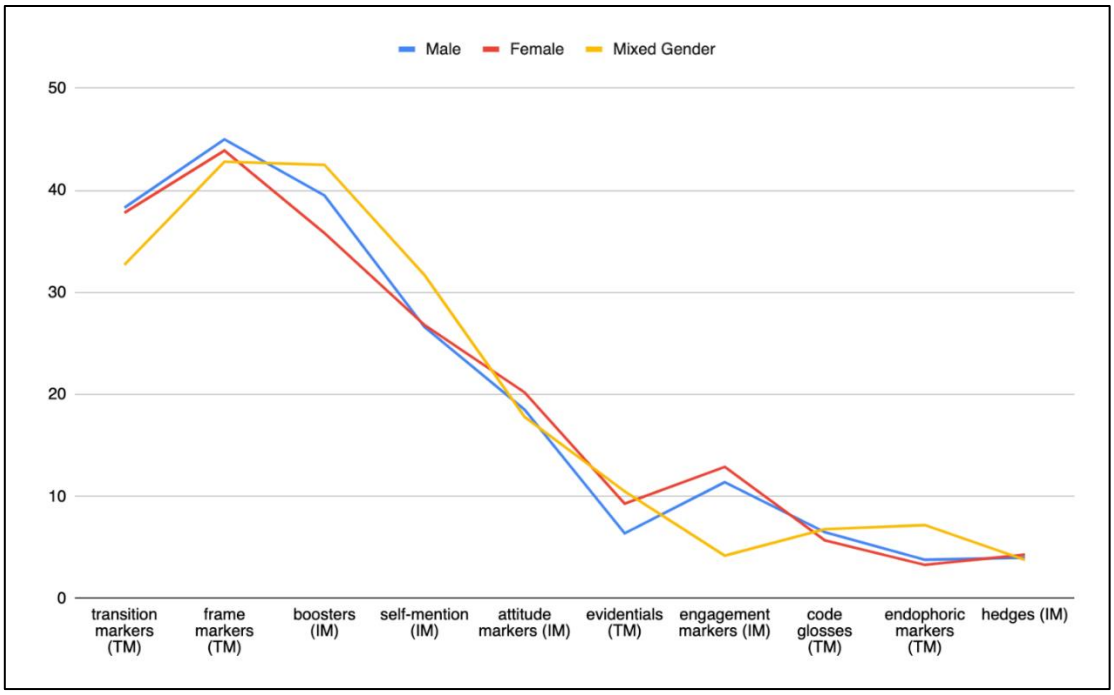
Textual metadiscourse elements in abstracts by gender (per 1,000 words):

Textual metadiscourse per 1,000 words	Male	Female	Mixed Gender
transition markers	14.15	12.60	11.48
frame markers	16.57	14.62	15.06
endophoric markers	1.39	1.12	2.51
evidentials	2.39	3.11	3.71
code glosses	2.42	1.88	2.39

Interpersonal metadiscourse elements in abstracts by gender (per 1,000 words):

Interpersonal Metadiscourse per 1,000 words	Male	Female	Mixed Gender
boosters	11.70	11.38	14.22
attitude markers	5.49	6.42	5.98
hedges	1.20	1.36	1.20
self-mention	7.88	8.51	10.64
engagement markers	3.37	4.12	1.43

Comparing percentage of total textual metadiscourse (TM) and of total interpersonal metadiscourse (IM) by author gender:



Appendix 3: Transliteration system for Arabic

Consonants:

ص	ش	س	ز	ر	ذ	د	خ	ح	ج	ث	ت	ب	ا / أ
ṣ	š	s	z	r	ḏ	d	x	ḥ	j	ṯ	t	b	’
ي	و	هـ	ن	م	ل	ك	ق	ف	غ	ع	ظ	ط	ض
y	w	h	n	m	l	k	q	f	ġ	ʔ	đ	ṭ	ḍ

Vowels:

Short vowels			Long vowels		
ُ	ِ	َ	و	ي	ا
u	i	a	ū	ī	ā

RiCL Research in Corpus Linguistics

Review of Laporte, Samantha. 2021. *Corpora, Constructions, New Englishes. A Constructional and Variationist Approach to Verb Patterning.* Amsterdam: John Benjamins. ISBN: 978-9-027-20850-7. https://doi.org/10.1075/scl.100

Martin Hilpert
University of Neuchâtel / Switzerland

The book under review is a contribution to a growing literature that approaches the study of New Englishes on the basis of corpus data (Hundt 2020). The book attempts a theoretical and methodological synthesis that draws in equal measures on Edgar Schneider's Dynamic Model of postcolonial Englishes (Schneider 2003), Patrick Hanks's Theory of Norms and Exploitations (Hanks 2013), and Adele Goldberg's work on Construction Grammar (Goldberg 1995, 2006). As the subtitle indicates, a main empirical focus of the book is on verb patterning, more specifically on the lexico-grammatical behavior of the high-frequency verb *make*. English *make* is a highly multifunctional verb. It encodes the idea of creation (*Let's make some dinner!*), but it also functions as a causative marker (*It made me smile*), it has resultative uses (*This makes things more difficult for us*), and it features in a broad range of lexicalized patterns, as in *make it* 'arrive', *make sure* 'verify', *make up* 'invent', etc. The overall goal of the book is to compare the lexico-grammatical profile of *make* across four different varieties of English. The inner-circle variety of British English serves as the basis for a comparison with the outer-circle varieties (Kachru 1985) of Hong Kong English, Indian English, and Singapore English, as represented by their respective ICE corpora.[1]

The overall aim of the study is to test how different corpus-based measures can inform the analysis of developmental stages in postcolonial Englishes. The analyses in the book thus seek to identify aspects of language use that map onto the degree to which a variety of English is institutionalized. In Schneider's dynamic model, Hong Kong

[1] http://ice-corpora.net/ice/index.html

English, Indian English, and Singapore English can be arranged on a cline of increasing institutionalization. In other words, Schneider's model yields theoretical predictions that can be tested on the basis of corpus data. Does that mean that the three varieties exhibit predictable variation with regard to the lexico-grammatical profile of *make*? As the book makes clear, and as will be discussed in the paragraphs below, the answer to that question is not a straight-forward *yes* or *no*, but instead, it requires a bit of nuance.

Before the ideas and results of the book are fleshed out in more detail, a few comments on its general structure are in order. The book is divided into eight chapters. A short introduction presents the main objectives of the study, which is guided by four research questions. The first of these aims to determine the lexico-grammatical profile of *make* in British English. The second research question asks how abstract argument structure constructions and more concrete patterns with *make*, for which the book uses the term 'lexically-bound constructions', are mutually related in that lexico-grammatical profile. The third research question turns to postcolonial Englishes, asking how outcomes of structural nativization can be observed in corpus data. Research question number four addresses how different linguistic phenomena reflect the degree of institutionalization of a variety of English. The following four chapters flesh out the theoretical and methodological background. Chapter 2 is dedicated to the World Englishes paradigm; Chapter 3 addresses the topic of structural nativization in postcolonial Englishes; Chapter 4 discusses Construction Grammar and how it can be combined with the Theory of Norms and Exploitations; and Chapter 5 fleshes out methodological aspects of data handling and analysis. The main empirical contributions of the book are found in Chapters 6 and 7. Chapter 6 focuses on British English and offers a detailed bottom-up analysis of how the verb *make* is used in abstract argument structure constructions as well as in more concrete constructions that are partially or fully lexically filled. Based on the findings of that analysis, a constructional network is proposed that represents the full lexico-grammatical profile of *make* in British English. Chapter 7 takes that network as the reference point for a comparison with Hong Kong English, Indian English, and Singapore English. The comparison takes different levels of abstraction into account. First, the four varieties of English are analyzed with regard to their respective uses of argument structure constructions that involve *make*. The chapter also highlights variation in light verb constructions with *make*. The findings are discussed in the light of proposals by Hoffmann (2014) that link the developmental stages of postcolonial Englishes in Schneider's

Dynamic Model with the cognitive representation of linguistic knowledge. The book closes with Chapter 8, which summarizes the empirical insights and theoretical contributions of the book.

On the basis of this general overview, the following paragraphs highlight some of the findings that merit particular attention. First, the analysis of British English that is offered in Chapter 6 is innovative in its approach, which aims to account near-exhaustively for the constructions that are used with *make*. The corpus data reveal that *make* is used with a range of argument structure constructions that exhibit a Zipfian distribution with regard to their frequency. Transitive *make* accounts for the majority of examples, resultative and causative uses are already considerably less frequent, and they are followed by a number of other patterns that barely register in the corpus data. With regard to lexically specified constructions, a similar picture emerges. The analysis reveals a small number of highly conventionalized patterns that account for a large share of the data and a wide variety of expressions that occur only once or twice in the data. The empirical findings are used to sketch out a constructional network that captures how *make* is used across its different valency frames and lexically specified patterns. The chapter offers a useful discussion of the different links that connect the constructions in that network.

While the insights offered in Chapter 6 are highly interesting in themselves, they are of course mainly intended as a basis for the comparisons across varieties that are undertaken in Chapter 7. Here, a number of striking parallels are observed. Not only are the overall frequencies of *make* very similar across the four ICE corpora, but it also emerges that the four varieties are indistinguishable with regard to the relative frequencies of the argument structure constructions and the most frequent lexically-bound constructions in which *make* appears. However, there are also interesting contrasts, specifically with regard to light verb constructions (e.g. *make a decision*, *make a mistake*). For example, postcolonial Englishes differ from British English in the use of zero articles, specifically with regard to singular nouns. Expressions such as *make choice*, *make distinction*, or *make correction*, which are attested in Hong Kong English and Indian English, are not used in the same way in British English. An analysis of the collocational behavior of light verb constructions further indicates that expressions such as *make use (of something)* are overrepresented in postcolonial Englishes.

The central issue that underlies the analysis in Chapter 7 is the question of whether the similarities and differences that are observed can be aligned with the developmental stages of Schneider's Dynamic Model. In the existing literature, Hong Kong English is viewed as being institutionalized to a lesser degree than Indian English, which in turn is not quite as strongly institutionalized as Singapore English. In the terminology of the Dynamic Model, Hong Kong English has gone through the phase of exonormative stabilization and is now undergoing nativization. Indian English has completed the phase of nativization and has entered the stage of endonormative stabilization. Singapore English is currently in that stage, but has started to undergo differentiation, which is the last stage of the Dynamic Model. Importantly, existing research has yielded inconclusive results with regard to the alignment of developmental stages and corpus-based measures. For example, Mukherjee and Gries (2009) study the collocational profile of the ditransitive construction across British English, Hong Kong English, Indian English, and Singapore English, finding that postcolonial Englishes gradually emancipate themselves with increasing institutionalization, so that the differences become more and more substantial over time. In a follow-up study, Gries and Mukherjee (2010) investigate whether this observation generalizes to n-grams of different lengths. The results do not converge with the earlier findings on the ditransitive construction. In other words, the degree of institutionalization is not easily mapped onto association strength in n-grams. Mixed findings are not only obtained by Gries and Mukherjee (2010), but also by Edwards and Laporte (2015), Werner (2016), and Deshors (2017). Hoffmann (2014) further argues that a development along the stages of Schneider's Dynamic Model may also lead to greater convergence between postcolonial Englishes and inner-circle varieties, specifically when it comes to highly abstract and productive constructions. Coming back to the analyses that are presented in the book, some predictions are clearly borne out while others are disconfirmed. With regard to the former, it is found that highly abstract constructions vary to greater extent in later stages of Schneider's Dynamic Model, which is in line with findings presented by Hoffmann (2014). Also, at the level of collocational preferences, later developmental stages correspond to greater lexical variation. With regard to intermediate levels of structure however, the results are in conflict with the predicted clines, so that the three postcolonial Englishes do not line up according to their respective developmental stages.

The fact that not all of the results can be neatly accounted for in terms of Schneider's Dynamic Model or Hoffmann's re-interpretation of that model does not take anything away from the important contribution that the book makes. As a proof of concept study, it illustrates the potential of an original, highly promising approach to the study of postcolonial Englishes. It is shown that mapping out the constructional network of a multifunctional verb in a reference variety and comparing that network against data from other varieties can yield stimulating insights that usefully inform theoretical questions, not only concerning World Englishes but also with regard to Construction Grammar and corpus-linguistic methodology. It is further written in an accessible style that makes it easy to follow the arguments that are made. The many strengths of the book notwithstanding, there are a few minor weaknesses. First of all, the undisputed star of the book, the verb *make*, would have deserved a place on the title page, so that researchers who are interested in that verb would be able to find this research. Second, the theoretical and methodological chapters in the first half of the book are more extensive than they would need to be. A thorough introduction to the general background is of course beneficial, but most readers will pick up this volume for its empirical results and its conclusions, which are of great interest. Third, while the combination of Construction Grammar and the Theory of Norms and Exploitations makes perfect sense, similar bottom-up corpus-based approaches to the study of constructions have been in practice in a variety of projects, notably in efforts to build up constructional networks in different languages (Lyngfelt *et al.* 2018). The research presented in this book, notably with regard to levels of abstraction and links between constructions, would connect beautifully to existing work in that area. That said, the book already succeeds in creating important links between different research traditions. In summary, this is a book that deserves attention, and that will leave the reader with many stimulating ideas.

References

Deshors, Sandra C. 2017. Structuring subjectivity in Asian Englishes: Multivariate approaches to mental predicates across genres and functional uses. *English Text Construction* 10/1: 132–163.

Edwards, Alison and Samantha Laporte. 2015. Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels. *English World-Wide* 36/2: 135–169.

Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.

Gries, Stefan Th. and Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics* 15: 520–548.

Hanks, Patrick. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge: MIT Press.

Hoffmann, Thomas. 2014. The cognitive evolution of Englishes: The role of constructions in the Dynamic Model. In Sarah Buschfeld, Thomas Hoffmann, Magnus Huber and Alexander Kautzsch eds. *The Evolution of Englishes: The Dynamic Model and Beyond, Varieties of English Around the World*. Amsterdam: John Benjamins, 160–180.

Hundt, Marianne. 2020. Corpus-based approaches to World Englishes. In Daniel Schreier, Marianne Hundt and Edgar W. Schneider eds. *The Cambridge Handbook of World Englishes*. Cambridge: Cambridge University Press, 506–533.

Kachru, Braj B. 1985. Standards, codification and sociolinguistic realism: The English language in the outer circle. In Randolph Quirk and Henry George Widdowson eds. *English in the World: Teaching and Learning the Language and Literatures*. Cambridge: Cambridge University Press, 11–30.

Lyngfelt, Benjamin, Kyoko Ohara, Tiago Timponi Torrent and Lars Borin. 2018. *Constructicography: Constructicon Development across Languages*. Amsterdam: John Benjamins.

Mukherjee, Joybrato and Stefan Th. Gries. 2009. Collostructional nativisation in New Englishes. Verb-construction associations in the International Corpus of English. *English World-Wide* 30/1: 27–51.

Schneider, Edgar. 2003. The Dynamics of New Englishes. From identity construction to dialect birth. *Language* 79/2: 233–281.

Werner, Valentin. 2016. Overlap and divergence – Apects of the present perfect in World Englishes. In Elena Seoane and Cristina Suárez-Gómez eds. *World Englishes: New Theoretical and Methodological Considerations*. Amsterdam: John Benjamins, 113–142.

Reviewed by
Martin Hilpert
University of Neuchâtel
Institut de langue et littérature anglaises
Espace Tilo Frey 1
CH-2000 Neuchâtel
Switzerland
E-mail: martin.hilpert@unine.ch

RiCL Research in Corpus Linguistics

Review of Julia Lavid-López, Carmen Maíz-Arévalo and Juan Rafael Zamorano-Mansilla. 2021. *Corpora in Translation and Contrastive Research in the Digital Era*. Amsterdam: John Benjamins. ISBN: 978-9-027-20918-4. DOI: https://doi.org/10.1075/btl.158

Mikhail Mikhailov
Tampere University / Finland

1. INTRODUCTION

The book is a collection of articles written by authors from Spain, Great Britain, Germany, Switzerland and other countries. The papers are based on the presentations held at the *International Symposium PaCor* 2018 (*Parallel Corpora: Creation and Applications*, Madrid, November 2018) hosted by the research group FUNCAP[1] in collaboration with the Institute of Modern Languages and Translation and members of the Department of English Studies at the Complutense University of Madrid (UCM).

The papers present research in the field of corpus-based translation and contrastive studies. The authors work with different pairs of languages: English and Spanish, English and German, English and Chinese, English and Portuguese, English and Turkish, and even Old English and Modern English. All the papers deal with parallel or comparable corpora.

The usefulness of multilingual corpora in contrastive and translation studies has been promoted by many researchers starting in the 1990s (see, e.g., Baker 1995; Johansson 2007; McEnery and Xiao 2008). However, multilingual corpora have traditionally got less attention than monolingual corpora (see, e.g., Kenning 2010;

[1] https://www.ucm.es/funcap/el-grupo

Mikhailov and Cooper 2016: 1–2). The first conference devoted to parallel corpora was organized in Uppsala, Sweden, in 1999 (Borin 2002). Then, after a fifteen-year-long break, the topic was resumed at the *PaCor* 2016 symposium at the University of Santiago de Compostela (Doval and Sánchez-Nieto 2019). More and more publications have appeared on this subject (see, e.g., Bernardini 2011; Frankenberg-Garcia 2009; Tiedemann 2012, among others), yet one cannot claim that no stone has been left unturned.

The book is divided into two parts. In the first part ("Corpus resources and tools") the issues of collecting and querying corpora are discussed. The second part ("Corpus-based studies and explorations") consists of case studies based on findings from parallel and comparable corpora.

2. Summary

The introductory chapter by Julia Lavid-López is not only a guide of the volume (as often happens), but also explains the idea of the book, which is to show the scope of available data and to introduce the tools that can be used for its querying.

The chapter begins with a solid overview of the corpus resources with an emphasis on parallel corpora. The author does a brief historical tour which is very helpful for the readers with little background in the field. The most prominent and important projects are mentioned: *English Norwegian Parallel Corpus* (ENPC),[2] ACTRES,[3] *The European Parliament Proceedings Parallel Corpus* (Europarl),[4] *Multilingual Text Tools and Corpora* (Multext),[5] *The Open Parallel Corpus* (OPUS),[6] and CLARIN ERIC.[7] The chapter also introduces the main challenges of compiling parallel corpora: limited availability of parallel texts from certain domains, genres, time spans, and for certain pairs of languages, as well as imbalance in the direction of the translations.

The next section is devoted to corpus-related tools: Translation Memory (TM) systems and corpus management tools. Personally, I would not have assigned TM systems

[2] https://www.hf.uio.no/ilos/english/services/knowledge-resources/omc/enpc/
[3] https://actres.unileon.es/wp/
[4] https://www.statmt.org/europarl/
[5] https://cordis.europa.eu/project/id/LRE62050
[6] https://opus.nlpl.eu/
[7] https://www.clarin.eu/

to corpus software. They are used for entirely different purposes: facilitating and automating translation process (Computer Aided Translation tools). Corpus-related functions, like parallel concordancing, are add-ons, and *Trados*[8] or *WordFast*[9] concordancing is much less flexible than in real corpus management systems like *Sketch Engine* (Kilgarriff *et al*. 2014) or the *IMS Open Corpus Workbench* (CWB).[10] However, there is some logic in introducing TMs together with corpus tools, because TM technology is little by little catching up with corpus technologies, and, as we will see, one of the chapters (see Ranasinghe *et al*. below) deals with making TM more intelligent. The overview of corpus management tools demonstrates that they are still very much oriented on monolingual corpora. Most of the tools mentioned in the chapter were initially developed for monolingual corpora and have additional functionality for querying parallel corpora as well. The current developments include the constantly growing role of web-based software and extensive use of Corpus Query Language (CQL) querying.[11] The most popular tool is *Sketch Engine*, and it is not only a research tool but is quite suitable for practical tasks, like copyediting or translating.

The chapter shows that the development of parallel corpora and corpus tools should serve both contrastive studies and translation studies and, at the same time, can be available for translation practitioners.

2.1. Part I: Corpus resources and tools

The first chapter in Part I ("A fresh look at language technologies and resources for translators and interpreters") by Gloria Corpas Pastor and Fernando Sánchez Rodas provides a brief outline of IT-resources for translators and interpreters. The authors point out that cardinal changes have took place in translation process. These are expansion of Computer Aided Translation (CAT) tools and Neural Machine Translation (NMT), cloud technologies, and crowdsourcing. Post-editing machine-translated texts becomes a routine, not an occasional task. The 'traditional translation' is being rapidly displaced. Although the field of interpreting is more conservative towards technologies and, at the moment, is still falling behind translation, it is also experiencing significant changes.

---

[8] https://www.trados.com/products/trados-studio/
[9] https://www.wordfast.com/
[10] https://cwb.sourceforge.io/
[11] https://www.sketchengine.eu/documentation/corpus-querying/

Different kinds of remote interpreting have become part of everyday life, and on-site events with interpreter online (telephone-mediated interpreting, video-mediated interpreting) are being replaced by cloud events with all participants communicating online via teleconferencing. As a result, the cloud-based interpreting is experiencing a fast growth.

The authors point out that, in spite of many advantages they give to translators, CAT tools still have many weak points and do not provide optimal solutions in many cases. Text corpora can in many cases complement CAT and Machine Translation (MT) and assist translators in many tasks. The main advantage of the corpora is the availability of huge amounts of data. Using corpora in interpreting is less obvious, yet there is some development here as well. Corpora are used by interpreters mainly in the preparation phase, and there also exist interpretation corpora that are used in interpreting studies and for interpreter training.

Another type of tool mentioned in the chapter is computer-assisted interpreting (CAI). However, these are multi-purpose tools, among them digital pens, note-making tools, and terminology management tools. These instruments are designed for a large group of users including interpreters.

Currently, translators have significantly more tools at their disposal than interpreters. Among translators, the most active users of CAT tools and MT are working in the field of specialized translation. Literary translators usually reject these tools but refer favorably on corpora that help them in looking up better equivalents, translation solutions, or check usage of a certain word or phrase.

The authors claim that the four stages of machine translation acceptance defined by Sgourou (2019; (1) nescience, (2) contempt, (3) reluctant adoption and shame, and (4) acceptance) are applicable to acceptance of all kinds of technological innovations in translating and interpretation. Translators are now in stage (4), while interpreters are somewhere in between stages (3) and (4).

Actually, the authors of the chapter present two different kinds of technologies: those supporting technical processes of translation and interpretation (scheduling, data sharing, teamwork, transmission) and those supporting language services (checking lexical units and terminology, grammar check, looking up translation equivalents, etc.). It would have been better to deal with them separately and to point out the differences

between the utilities designed especially for translators (e.g., CAT tools), for a wide range of language service providers (e.g., corpora), and for all users (Optical Character Recognition, MT).

Chapter 2 by Yi Gu and Ana Frankenberg-Garcia ("ZHEN: A directional parallel corpus of Chinese source texts and English translations") is devoted to parallel corpora with the language pair of English and Chinese. The authors point out that most of existing parallel corpora of this language pair are collections of translations from Chinese into English, a large number of them being translated by native speakers of Chinese with post-editing by English native speakers. Although the existing corpora contain a certain amount of Chinese-English translations, it is difficult to detect those because the source text is not specified for official documents, as is the case with United Nations (UN) texts. The English-Chinese parallel texts are usually more difficult to obtain, and many of them are old texts from the nineteenth century.

In the chapter, a new English-Chinese corpus is introduced. The authors outline the criteria for selecting texts for the corpus, existing difficulties in searching and collecting parallel texts, and the technique for looking up source texts and translations. The resulting corpus represents various text genres, such as government documents, white papers, UN documents, fiction, political speeches, movie subtitles, academic abstracts, etc. The source texts are written in Mainland Mandarine Chinese and are published after 1990 (with few exceptions). The corpus was compiled with *Sketch Engine* and can be shared with other researchers. The authors show the advantages of the resource compared to other English-Chinese datasets and outline its possible uses.

In Chapter 3, "Word alignment in a parallel corpus of Old English (OE) prose. From asymmetry to inter-syntactic annotation," Javier Martín-Arista presents a parallel corpus with Old English texts and their translations into the Present-day English (PDE), with multiple examples that demonstrate morphological and syntactic differences between OE and PDE. This type of corpus is not very common, and it has certain technical issues that need to be solved. The *Open Access Annotated Parallel Corpus Old English* (ParCorOE)[12] targets 300,000 running words, and consists of OE texts of various genres and their translations into PDE. The collection is fairly large for this type of corpora. The texts are aligned at sentence and word levels, lemmatized, and include morpho-syntactic

[12] https://www.nerthusproject.com/search-parcoroe

annotation. The syntactic structures can be visualized as graphs. The corpus is freely accessible online.

Chapter 4 by Tharindu Ranasinghe, Ruslan Mitkov, Constantin Orăsan and Rocío Caro Quintana is entitled "Semantic textual similarity based on deep learning: Can it improve matching and retrieval for Translation Memory tools?" Current TM tools are based on string matching techniques (Levenshtein's distance, Dice-Sørensen index; see Levenshtein 1966 and Sørensen 1948, respectively). These methods work well on pairs of sentences which are lexically and syntactically close. Using semantic similarity helps to find sentences with other lexemes and/or other grammatical constructions used to express the same meaning. Semantic similarity measures are distance measures of semantic vectors of sentences which are the result of pairwise comparing sentences from large datasets. In this chapter, the authors try to find out whether using semantic textual similarity has perspectives. They test various semantic sentence encoders (InferSent,[13] Universal Sentence Encoder,[14] and SBERT[15]) and compare the results with Okapi[16] which uses Dice-Sørensen index. The testing is done on the English-Spanish Directorate-General for Translation of the European Commission (DGT) TM. The results show that the semantic encoders are fast enough to be used in industry and that they are more efficient with the sentences with low string similarity. They also produce less bad matches resulting of partial coincidence of the sentences.

In Chapter 5, "TAligner 3.0: A tool to create parallel and multilingual corpora," Zuriñe Sánz-Villar and Olaia Andaluz-Pinedo introduce a tool for working with parallel corpora. Unlike many existing corpus tools, *TAligner 3.0*[17] is an open source and cross-platform tool that can align multiple translations of the same text and even retranslations. The software has also special features for aligning dramatic texts. The search routines provided are frequency lists and parallel concordancing.

The tool belongs to the third generation software, that is, it works on workstations, and not on servers. The authors are aware that this creates certain problems with

[13] https://github.com/facebookresearch/InferSent
[14] https://www.tensorflow.org/hub/tutorials/semantic_similarity_with_tf_hub_universal_encoder
[15] https://www.sbert.net/
[16] https://okapiframework.org/wiki/index.php/Main_Page
[17] https://addi.ehu.es/handle/10810/42445?locale-attribute=en

installations on local computers and sharing corpora (especially in the case of large corpora).

Chapter 6, "Developing a corpus-informed tool for Spanish professionals writing specialised texts in English," by María Pérez Blanco and Marlén Izquierdo demonstrates direct practical applications of multilingual corpora. *Promociona-TÉ*, a tool for generating product descriptions for the tea industry, was a result of cooperation between the ACTRES research group and a tea manufacturer Pharmadus Botanicals, S.L.[18] The tool is based on the data from a comparable English-Spanish corpus. The instruments of this kind are very important for small enterprises which cannot afford commissioning translators, and for which machine translation of specialized texts does not yield sufficient quality because of the limited availability of parallel texts.

2.2. Part II: Corpus-based studies and explorations

In Chapter 7 ("English and Spanish discourse markers in translation: Corpus analysis and annotation"), Julia Lavid-López presents an analysis of the English discourse markers (DM) *in fact*, *actually*, and *really* and their Spanish equivalents. The author uses large parallel corpora from the OPUS corpus collection available at *Sketch Engine*. The Spanish translation correspondences of the three DMs are first collected from English-Spanish parallel concordances. The corpus provided the most typical Spanish translation correspondence for the DMs in question. The analysis of data also makes it possible to define the meanings of both English and Spanish DMs. The chapter contains many usage examples and interesting findings on usage, meanings, frequencies, and interrelation of the markers.

The OPUS datasets provide large amounts of data which helps to find the most typical pairs of equivalents and perform quantitative analysis. However, the information on direction of translation is not available and it is therefore not possible to define subcorpora with original English texts and their Spanish translations, and with original Spanish texts and their English translations. Another problem is that source texts can be written by non-native speakers, and also some translations can be performed by non-native speakers of the target language. Finally, a translator may misunderstand the text

[18] https://www.pharmadus.com/

and use a wrong DM as equivalent. These issues can influence the use of DMs in both languages (see also Mikhailov 2021). However, large, clean and reliable parallel corpora are still to be acquired, and results obtained from noisy data are also valuable.

In Chapter 8, "The discourse markers *well* and *so* and their equivalents in the Portugese and Turkish subparts of the TED-MDB corpus," Amália Mendes and Deniz Zeyrek continue the discussion on DMs in original texts and their translations. The researchers study cross-lingual correspondences of the English DMs *well* and *so*. The data used comes from TED Talks transcripts. The TED Talks presentations are transcribed by volunteers and translated by other volunteers into other languages. The data is freely available, thus providing a valuable multilingual dataset. The TED MDB corpus (see Zeyrek *et al.* 2020) is compiled of such parallel texts with discourse relations annotated. The information on discourse relations is still not available from large corpora, and most research is therefore still carried out on small data. The research demonstrates that discourse markers of the source text are often omitted in translations, but the tendencies are different for different language pairs. The English marker *well* is sometimes kept by Portuguese translators, while Turkish translators leave it out. At the same time, the marker *so* is usually left out in Portuguese translations and often preserved in Turkish talks.

Although the methodology is interesting and promising, more data would be needed. The size of the corpus is less than 20,000 running words with about 7,000 tokens per language. The case studies are carried out on 12 examples of *well* and 30 examples of *so*.

In Chapter 9, "Variation of evidential values in discourse domains: A contrastive corpus-based study (English and Spanish)," Juana I. Marín-Arrese studies evidentiality, that is, marking the source of information, in oral and written communication in English and Spanish. The research is carried out on comparable corpora. Two types of evidentiality markers, indirect-inferential evidentiality (IIE) and indirect-reportative evidentiality (IRE), are compared. The study demonstrates that IIE markers are more extensively used both in English and in Spanish, and that the use of evidentiality markers is different in oral and written language. The data from both languages demonstrate the same tendencies with some minor differences.

Chapter 10 ("Translation for dubbing of Westerns in Spain: An exploratory corpus-based analysis") by John D. Sanderson, presents an analysis of the lexis in the American westerns dubbed into Spanish. The study is based on a parallel corpus which, at the

moment of publication, included transcripts of 20 American westerns from 1939 to 2012 aligned with the transcripts of the films dubbed into Spanish. The author studies the impact of the censorship of Franco's dictatorship on the choice of equivalents. For example, the culture-bound word *marshal* tends to be domesticated in earlier films (*comisario, alguacil*) and foreignized in later films (*sheriff*); rude expressions like *son of a bitch* are avoided: although there exists an exact match in Spanish (*hijo de puta* 'son of a whore'), an artificial *hijo de perra* ('son of a she-dog') is used (although the word *puta* 'whore' is nevertheless used as a separate lexeme). The study points out that a special sociolect for translating American westerns has been developed and some equivalents are still being used even now, many years after Franco's decease. In the chapter the practical use of parallel corpora of film transcripts for translating is also mentioned.

In Chapter 11, "Generic analysis of mobile application reviews in English and Spanish: A contrastive corpus-based study," Natalia Mora López explores the composition of texts in the genre of online review. She compares English and Spanish reviews from *Google Play Store*. The data is a small corpus of 200 texts (100 English, 100 Spanish) drawn from a larger text collection. The study is based on the Appraisal Theory, which aims at detecting positive and negative attitudes expressed in texts. A number of patterns are found and their features studied. In many cases the attitudes can be detected on the lexical level, although some texts, especially spam, can be misallocated. No significant differences between English and Spanish reviews were found.

In Chapter 12 ("Exploring variation in translation with probabilistic language models"), Alina Karakanta, Heike Przybyl, and Elke Teich compare the language of translations and interpretations in relation to the language of original written texts and speech. The data used are obtained from the Europarl-UdS corpus, with written texts and translations originated in the European Parliament, and several interpreting corpora. The target languages are English-German and English-Spanish translations and interpretations. The metadata of the corpora make it possible to select the data produced by native speakers both for source texts/speeches and for translations/interpretations.

The method used is Kullback-Leibler Divergence (KLD), which allows to measure probability disruptions in the data being compared. The findings are visualized as word clouds. The word probabilities are compared pairwise for translations vs. originals, interpretations vs. originals, and translations vs. interpretations. The method allows to

detect the words typical of a certain type of data, such as for German original texts as opposed to texts translated from English into German, etc. The results demonstrate that the language of translations and interpretations differs from that of the texts/speeches originally produced in the same language. Possible reasons include differences in the process of creating original text/speech and translation/interpretation, as well as the 'shining through' of the source language in translations/interpretations. However, some similar effects are detected in German and Spanish data, which demonstrates that not all features can be interpreted in terms of 'shining-through'.

In Chapter 13 ("Binomial adverbs in Germanic and Romance languages: A corpus-based study"), Johannes Graën and Martin Volk present a method of extracting binomial adverbs (*more or less*, *here and now*, etc.) from large multilingual corpora. The study is performed on the *large-Scale PARallel Corpora to study LINGuistic variation* (SPARCLING)[19] and includes six languages: English, French, German, Italian, Spanish, Swedish. Detecting multiword expressions, and binomial expressions among them, is very important both for linguistic research and for automated language processing (parsing, MT). Direct queries like Adv + Conj + Adv do not have enough recall because of the parsing errors; therefore, lists of adverbs based on morphological annotation were extracted from the corpus and the searches performed on these lists after their cleaning up. The candidates were filtered out using MI-scores and the boundaries of the multiword expressions were checked with the help of entropy values. The interlingual correspondences in parallel corpora worked as additional criteria for detecting binomial adverbs.

3. Discussion

The book addresses researchers working in the fields of translation studies, contrastive studies, and corpus linguistics. Some of the papers in the first part deal with the issues of language technologies, and many papers from the second part are connected with discourse analysis. The book shows well enough the state of the art in the field: the studies presented use different methods and approaches and are performed on data of very different nature. The volume also reveals the main tendencies in modern corpus research:

[19] https://www.cl.uzh.ch/en/texttechnologies/research/corpus-linguistics/sparcling.html

the size of datasets is growing, new languages and language pairs are being studied with the help of corpora, new kinds of data (e.g., interpretation corpora) are being collected, descriptive statistics is being replaced by sophisticated quantitative methods, etc. It also becomes clear that the data available is not sufficient for all kinds of research and that automated annotation has many weak points.

All chapters present original research and fit well into the composition of the book. Although the studies are devoted to different language pairs, they are focused on methodological issues rather than on findings in particular languages and therefore all of them are of interest for researchers working with other languages.

As it often happens with conference volumes, the chapters are written by different authors and present different topics, which makes the book rather heterogeneous. Some papers are very easy to read and are more practically oriented, others make use of complicated methods and need more effort. Still, the volume does not require special background. All in all, the book is a suitable reading for someone interested in multilingual corpora and their use in contrastive and translation studies. It will hopefully inspire more research in the field.

References

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7/2: 223–243.

Bernardini, Silvia. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS – A Journal of Professional Communication* 26: 2–13.

Borin, Lars ed. 2002. *Parallel Corpora, Parallel Worlds. Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Amsterdam: Brill.

Doval, Irene and María Teresa Sánchez Nieto. 2019. Parallel corpora in focus: An account of current achievements and challenges. In Irene Doval and María Teresa Sánchez Nieto eds. *Parallel Corpora: Creation and Applications*. Amsterdam: John Benjamins, 1–15.

Frankenberg-Garcia, Ana. 2009. Compiling and using a parallel corpus for research in translation. *International Journal of Translation* XXI/1: 57–71.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora: On the Use of Corpora in Contrastive Studies.* Amsterdam: John Benjamins.

Kenning, Marie Madeleine. 2010. What are parallel and comparable corpora and how can we use them? In Michael McCarthy and Anne O'Keeffe eds. *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 487–500.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

Levenshtein, Vladimir I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10/8: 707–710.

McEnery, Tony and Richard Xiao. 2008. Parallel and comparable corpora: What is happening? In Gunilla Anderman and Margaret Rogers eds. *Incorporating Corpora: The Linguist and the Translator*. Clevedon: Multilingual Matters, 18–31.

Mikhailov, Mikhail and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research.* London: Routledge.

Mikhailov, Mikhail. 2021. Mind the source data! Translation equivalents and translation stimuli from parallel corpora. In Vincent X. Wang, Lily Lim and Defeng Li eds. *New Perspectives on Corpus Translation Studies*. Singapore: Springer, 259–279.

Sgourou, Maria. 2019. The four stages of machine translation acceptance in a freelancer's life. In *Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019), Varna, Bulgaria*. Shoumen, Bulgaria: Incoma Ltd., 134–135. https://aclanthology.org/W19-8700/

Sørensen, Thorvald. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Kongelige Danske Videnskabernes Selskab* 5/4: 1–34.

Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC' 2012). Istanbul, Turkey: European Language Resources Association, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

Zeirek, Deniz, Amália Mendes, Yulia Grishina, Murathan Kurfali, Samuel Gibbon and Maciej Ogrodniczuk. 2020. TED Multilingual Discourse Bank (TED-MDB): A parallel corpus annotated in the PDTB style. *Language Resources & Evaluation* 54: 587–613.

Reviewed by
Mikhail Mikhailov
Languages Unit, ITC
Tampere University
FI-33014 Tampere University, Finland
e-mail: mikhail.mikhailov@tuni.fi

RiCL Research in Corpus Linguistics

Review of Carrió-Pastor, María Luisa ed. 2020. *Corpus Analysis in Different Genres: Academic Discourse and Learner Corpora*. London: Routledge. ISBN: 978-0-367-49993-8. https://doi.org/10.4324/9780367815905

Nanxi Bian – Ge Lan
City University of Hong Kong / Hong Kong

A plethora of research located at the intersection of discourse analysis and linguistic studies has adopted a corpus approach in the past decade. Corpora provide empirical evidence for observed linguistic patterns, showing that research findings are traceable, objective, and scientific. This book is a collection of studies with two foci that are reflected by two sections: Section 1 ("Corpus studies on academic discourse") with eight chapters and Section 2 ("Studies on learner corpora") with ten chapters. This review consists of 1) a summary of the major contents of each chapter and 2) a review of the book content based on the two sections mentioned above and the genres and the linguistic features analyzed in the different chapters.

In the first chapter, the author conducts a corpus analysis to examine the use of the metadiscourse device self-mention in research papers. The research purpose is to identify the patterns in which writers show an authorial persona and figure out the variations of the use of self-mentions in three different academic disciplines. The corpus comprises 150 research papers written by English native speakers distributed in three types of papers: engineering, linguistic, and medicine papers. The results demonstrate that writers show an authorial persona in all three disciplines, while the use of self-mentions varies in frequency across disciplinary genres. Researchers are encouraged to explore self-citations for future research so as to have a complete picture of how writers construct their authorial persona.

Chapter 2 focuses on expressions of gratitude in the prefaces of linguistic books. The author conducts a corpus analysis to examine the forms and functions of thanking expressions. The corpus comprises 72 prefaces extracted from books written in English. After a searching process with *CasualConc*,[1] the retrieved thanking expressions are classified. Results show that the thanking expressions include both routinized thanking formulae and creative ones. Furthermore, the results show that the main function of the thanking expressions is to show appreciation to people and institutions for their help and support, indicating that thanking expressions are related to showing academic modesty and honesty.

Chapter 3 investigates the expressions of criticism in two time periods of the 1980s (USSR) and 2010s (contemporary Russia) and how the changes and evolution in criticism expressions took place in these periods. The data includes the reviews of the 1980s and the 2010s published in *Issues in Linguistics*,[2] but only the ones that have both authors and reviewers from the Soviet era or Russia are included. The author manually tags and calculates the negative critical acts in the corpus and compares the critical acts of the soviet with modern periods. The results reveal that the reviews in the 1980s are less critical than those in the 2010s, which demonstrates a tendency to use a more critical attitude.

An investigation on the attitudinal qualifications conveyed by the use of modal verbs within the genre of medical abstracts is conducted in the fourth chapter. The corpus consists of 48 abstracts of medical research papers. The results reveal a massive use of the dynamic modality and epistemic modality, which shows potentiality and possibility respectively. Epistemic *may* is the most frequently attested modal verb used in background sections, introductions, and sections stating the results in the abstracts. The dynamic meaning is mainly found in modals *can* and *could*. The authors conclude that the use of dynamic and epistemic modals allows writers to present their ideas and external facts without imposing their views.

Chapter 5 studies the pragmatic functions of the adverb *fairly* as a metadiscourse device in scientific writing. Specifically, the disciplinary differences are explored. The authorial stance of mitigating effect expressed with *fairly* is also examined. The corpora used are the *Corpus of History English Texts* and the *Corpus of English Texts on*

[1] https://sites.google.com/site/casualconc/
[2] https://www.linguisticsociety.org/issues-linguistics

Astronomy (Moskowich *et al.* 2019), both are included in the *Coruña Corpus of English Writing (1700-1900)*. The results indicate that the adverb *fairly* tends to function as a mitigating device. At the same time, differences in syntactic patterns and pragmatic functions are observed among scientific registers.

Chapter 6 focuses on lexical bundles in academic journal descriptions (JD). The study investigates the frequency of occurrences and the functions of the lexical bundles in a multidisciplinary corpus. The corpus comprises 80 JDs divided into four disciplines: linguistics, sociology, biology, and mechanical engineering. The author categorizes 24 lexical bundles into referential, discourse organizing, and stance bundles, and conducts an N-gram analysis and a manual observation of occurrences. The referential type is the most frequent bundle attested. The results show a high frequency of lexical bundles with inconspicuous disciplinary differences, which suggests that JDs are highly formulaic and standardized texts. For future research, the author encourages comparative studies about JDs in less prestigious periodicals, as well as comparative studies about other book sections.

Chapter 7 focuses on the collaborative work in corpus compilation within the genre of medical research articles. The aim is to clarify the rationality of adopting an ethnographic approach in the corpus compilation process. Another goal is to raise linguists' and ESP teachers' awareness about turning to authentic texts and professional's expertise in field-specific genre corpus compilation in order to get access to representative data. A detailed description which includes the criteria for corpus compilation is presented. The proposed ethnographic methodology for corpus compilation goes from context to text allowing more effective and consistent linguistic research outcomes.

Chapter 8 focuses on conducting qualitative research on language use in academic discourse with the help of *Computer-Aided Qualitative Data Analysis* (CAQDAS). Screenshots of the CAQDAS are presented and make the demonstration clear to readers. The data includes research articles in the top-tier journals such as *English for Specific Purposes*[3] or *Journal of Second Language Writing*.[4] The findings reveal that CAQDAS efficiently supports the qualitative analysis of academic discourse. The author claims

[3] https://www.journals.elsevier.com/english-for-specific-purposes
[4] https://www.sciencedirect.com/journal/journal-of-second-language-writing

that the access to and the specific training on computational tools for researchers are highly expected.

Chapter 9 investigates non-native learners' knowledge of cohesion and coherence. The authors investigate contrastive discourse markers in academic argumentative essays written by learners of English and German. The corpus consists of two sub-corpora (a sub-corpus of English and a sub-corpus of German) each containing 40 argumentative essays in humanities and social sciences. *Sketch Engine* (Kilgarriff *et al.* 2014) is used for lexical search. The error analysis indicates that non-native learners of both English and German tend to overuse or misuse certain connectors and that an imprecise use of discourse markers can disrupt coherence or mislead readers. The data also indicates a low variability in discourse markers used by non-native learners. These problems are attributed to an intensive exposure of learners to explicit teaching. The findings suggest that the explicit teaching of cohesive devices use should avoid oversimplification. Data-driven learning is recommended in the learning of cohesive devices.

Chapter 10 explores what kind of personal metadiscourse markers (PMM) are used in Final Degree Dissertations (FFD) and investigates the functions these markers perform. The analysis is based on Ädel's (2006) reflexive modal approach to personal metadiscourse. The self-compiled corpus for this study, the *Trabajos de Fin de Estudio del Grado de Educación Primaria* (TFE-Prim), includes 130 FFDs and is divided into three sub-corpora: TFE-Did (pedagogic proposals), TFE-Inv (research), and TFE-Rev (literary review). The results reveal that PMMs are more frequently attested in TFE-Inv. The main function of PMMs in the observed data is to address the receiver during the reading process. It is also observed that the typology of FDD has an influence on the use of PMM. The qualitative results demonstrate a strong preference for discursive functions such as saying and reminding. The author points out that further work about raising the author's awareness in FDD in education sciences is required.

Chapter 11 examines the use and distribution of metadiscourse interactional features in 55 explanatory essays written by Spanish native speakers with a C1 CEFER level of English.[5] The author searches manually for the interactional metadiscourse features listed in Hyland (2005) and analyzes their frequencies of occurrence. The quantitative results show that engagement markers are most frequently used, while self-mentions and boosters are less frequently attested. The qualitative results indicate that

[5] http://cvc.cervantes.es/obref/marco/cvc_mer.pdf

Spanish native speakers with a C1 CEFER level of English know a very small amount of the interactional devices listed in Hyland (2005). Interactional features not included in Hyland (2005) list are marked in the corpus. These new interactional devices can be considered as specific interactional metadiscourse devices used by these Spanish native speakers who are learning English.

Chapter 12 compares the rhetorical functions of citations which Spanish and American students use in their native language in the writing of their Master Theses. The corpus consists of 24 Masters Theses in applied linguistics: 12 by Spanish native postgraduate writers and 12 by American native postgraduate writers. The writing by students is compared with that of expert writers. Based on Petrić's (2007) typology, citations are manually coded in terms of their rhetorical functions. It is shown that authors who write in English use many citations with complex rhetorical functions. The expert-novice comparison reveals that postgraduate students tend to adopt an expository style, while expert writing makes use of a more conventional dialogic style.

Chapter 13 assesses linguistic complexity in native and non-native academic English writing through an inventory of 24 numeric measures provided by automatic analyzers. The aim is to test the hypothesis that linguistic complexity and academic language proficiency are correlated. The corpus consists of academic essays written by both native and non-native writers. The native data is retrieved from the *Louvain Corpus of Native English Essays* (LOCNESS; Granger 1998), and the non-native data is retrieved from the *Written Corpus of Learner English* (WriCLE; Rollison and Mendikoetxea 2010). Software tools L2SCA[6] and Coh-Metrix[7] are used for pre-processing the texts, analyzing the syntactic structures, and identifying significant indexes, revealing linguistic complexity, and validating the results. Principal Component Analysis and Logistic Regression Analysis are used to figure out the most significant groups of features. The hypothesis that a higher level of academic language proficiency indicates a higher level of linguistic complexity is revealed to be only partial. The trends per proficiency level suggested by the statistical model are considerably irregular.

Chapter 14 presents the new corpus *Corpus for the Learning of Catalan for Specific Purposes* (CALEC), which is an important aid for the teaching and learning of

[6] https://aihaiyang.com/software/
[7] http://cohmetrix.com/

languages for specific purposes within the university framework in Catalan. CALEC was compiled by collecting descriptive texts produced by university students doing degrees in computer engineering and industrial engineering. Error analysis is conducted to pinpoint the areas of learning difficulties and the level of students' communicative competence. Observing that students have insufficient command of spelling in Catalan and that English has a high level of interference in the terminology of the subject-matter, the study systematizes students' errors and figures out their needs, which supports the design of teaching materials pedagogically.

Chapter 15 aims to identify word sequences in written academic tasks of Spanish undergraduate students. The authors conduct a Contrastive Interlanguage Analysis by comparing native and non-native learners' writings. The native students' writings are further compared with native experts' writings. The following corpora are analyzed: 1) the *Academic Corpus of the University of Valencia* (ACUV), which contains research articles by expert native writers; 2) the *British Academic Written English* (BAWE; Nesi *et al.* 2008), which contains novice writing by native English writers; and 3) the *Corpus of Learners of English as a Foreign Language* (CASTLE),[8] which contains non-native English writings by students. The results observe a sizeable number of overused four-word bundles, indicating leaners' incomplete command of the pragmatic complexity of long sequences. Additionally, a large number of overused lexical bundles reflect personal stance features, indicating non-native characteristics. The authors believe that students should get more exposure to the lexical bundle inventories and more intense contact with academic registers.

Chapter 16 takes cognitive linguistics to explore the use of three verbs of vision *regard, see,* and *view* in academic English corpora of native expert, native non-expert, and non-native non-expert writers, with the focus on the non-literal meaning and metaphorical senses of the verbs, and the patterns of use of the non-literal meanings. Based on the *Professional English Research Consortium Corpus* (PERC)[9] and two sub-corpora of BAWE (native non-expert corpus and non-native non-expert corpus), the author studies the correlation among the use of the non-literal vision verbs, the native and non-native use of English, and the level of expertise in academic writing. It is concluded that non-native non-expert writers most frequently use *regard* and *view*,

[8] http://corefl.learnercorpora.com/
[9] https://scnweb.japanknowledge.com/register/PERC/index.html

while overusing the non-literal meaning of *regard* and underusing the non-literal meaning of *view*, when compared with native expert and non-expert writings. Non-native non-expert writers also tend to overuse the non-literal *see* in comparison to native expert writers, but tend to underuse the non-literal *see* if compared to native non-expert writers.

Chapter 17 studies the expression of emotion in master's theses by native English speakers (NE) and non-native English speakers (NNE). The corpora used in the study consist of master's theses by NE and NNE in the disciplines related to engineering, natural sciences, health, and human sciences. The frequency analysis shows that most types of emotion expressions attested in both NE and NNE texts are boosters and modal verbs. There is a more frequent use of emotion expressions in NE texts, which implies that NE speakers are less concerned about showing their opinions or feelings. Moreover, NNE students follow more traditional patterns and avoid sharp and emphatic words. Thus, pragmatic awareness should be raised in the language classroom and in instructions regarding academic English writing. Students are recommended to get more exposure to authentic texts to obtain more explicit ideas about the disciplinary-specific expressions.

The volume ends with Chapter 18, which investigates the online production of university students who study English as a Foreign Language when English is used as the vehicular language in the classroom. The study analyzes the students' act when they realize that they have made a grammar or spelling mistake on an online forum. The analysis provides students with techniques to overcome incorrectness in online writing and help them get proper awareness of Foreign Language Anxiety (FLA). The investigation makes use of TICOR, a sub-corpus corpus of ENTERCOR (Torrado-Crespón 2018), which is divided into two sub-corpora: ICT (from pre-school education degree) and TIC (from primary education degree). Findings reveal a lack of proofreading by students before they submit their online production, and that they simply apologize for their mistakes when they realize the teacher is reading their productions. The author suggests to explicitly advise students to proofread and emphasize that the teacher will take spelling mistakes into account in the final mark. Additionally, auto-corrective software is recommended for online writing.

This edited volume covers two important and interrelated types of corpus studies according to the nature of the corpora, namely corpus studies on academic texts

produced by expert writers (e.g., authors of published journal articles) and corpus studies on academic texts produced by learner writers (e.g., university students). In corpus linguistics, scholars have been exploring the linguistic and/or discursive characteristics of authentic academic texts produced by expert writers to expand our understanding of academic genres. Likewise, to leverage language teaching in academic contexts, an increasing number of scholars have been investigating academic texts from student writers. These two groups of studies are not only important as two individual research areas but also are interrelated, since student writers are expected to learn and ultimately handle linguistic and/or discursive characteristics of academic texts from expert writers. It is not uncommon for expert texts to be integrated into language learning classroom as fitted examples for students to learn. Thus, this volume benefits a wide range of audience interested in researching and teaching academic discourse in different contexts.

The book includes corpus studies in diverse genres. In terms of expert writing, the genres include, but are not limited to, research papers in different fields of studies (e.g., engineering, medicine, astronomy), academic journal descriptions, book prefaces, and historical English texts. For learner texts, the genres cover theses/dissertations, explanatory essays or academic essays in general, descriptive texts, academic written tasks with specific prompts, and the sue of some existing corpora (e.g., PERC or BAWE). The learner texts not only include non-native texts that have received a lot of research attention in applied linguistic, but also native learner texts. Although the list of genres can never be exhaustive, meaning that there are always additional genres that can be studied (e.g., student writing from standardized language texts), the corpus studies with a fairly diverse group of academic genres in the book bring valuable insights to scholars who are interested in academic discourse from both experts and learners in general.

A broad range of linguistic or discursive features are studied in the volume, and their related discursive functions are also qualitatively analyzed. Numerous linguistic or discursive features can be studied from the perspective of discourse analysis with a corpus-based or a corpus-driven approach, ranging from individual words or phrases to types of lexical features (e.g., personal pronouns) and to grammatical complexity measures in general. All different linguistic and discursive features can be found in the studies in the book: a) particular words, such as the use of *fairly* as a metadiscourse

device; b) the use of a certain type of discourse markers, for example, markers for the expression of gratitude, markers of motion expressions based on boosters, and modal verbs; c) the use of linguistic patterns, such as N-grams and lexical bundles; and d) the overall linguistic patterns, such as syntactic structures and syntactic complexity. This broad range of linguistic or discursive features can meet the wide range of research interests from scholars in the interaction of corpus linguistics, discourse analysis, and text analysis in general.

The book encourages scholars to carry out empirical studies about academic discourse, with corpus linguistics as the research approach. Likewise, it can be an initial secondary resource for graduate students who are interested in reading recent literature on corpus studies dealing with academic discourse which is produced by expert or non-expert writers.

REFERENCES

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2*. Amsterdam: John Benjamins.

Granger, Sylviane. 1998. The computer learner corpus: A versatile new source of data for SLA research. In Sylviane Granger ed. *Learner English on Computer*. London: Routledge, 3–18.

Hyland, Ken. 2005. *Metadiscourse Exploring Interaction in Writing*. London: Continuum.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

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

Nesi, Hilary, Sheena Gardner, Paul Thompson and Paul Wickens. 2008. *British Academic Written English Corpus*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2539

Petrić, Bojana. 2007. Rhetorical functions of citations in high- and low-rated master's theses. *Journal of English for Academic Purposes* 6/3: 238–253.

Rollinson, Paul and Amaya Mendikoetxea.2010. Learner corpora and second language acquisition: Introducing WriCLE. In Jorge Luis Bueno Alonso, Dolores González Álvarez, Ursula Kirsten Torrado, Ana Elina Martínez Insua, Javier Pérez-Guerra, Esperanza Rama Martínez and Rosalía Rodríguez Vázquez eds. *Analizar datos>Describir Variación/Analysing Data>Describing Variation.* Vigo: Servizo de Publicacións Universidade de Vigo, 1–12.

Torrado-Crespón, Milagros. 2018. Interlanguage or technology when using English as a vehicular language: what influences students' productions online? https://m.riunet.upv.es/bitstream/handle/10251/120520/9924-46319-1-PB.html?sequence=2&isAllowed=y

Reviewed by
Nanxi Bian
City University of Hong Kong
Department of Linguistics and Translation
83, Tat Chee Avenue
Kowloon, Hong Kong SAR
E-mail: nbian7@cityu.edu.hk

Ge Lan
City University of Hong Kong
Department of English
83, Tat Chee Avenue
Kowloon, Hong Kong SAR
E-mail: gelan4@cityu.edu.hk

RiCL Research in Corpus Linguistics

Review of Fernández-Pena, Yolanda. 2020. *Reconciling Synchrony, Diachrony and Usage in Verb Number Agreement with Complex Collective Subjects*. New York: Routledge. ISBN: 978-0-367-41715-4 https://doi.org/10.4324/9780367815899

Lotte Sommerer
University of Freiburg / Germany

In her 2020 monograph, Yolanda Fernández-Pena analyzes subject-verb agreement patterns of English noun phrases (NPs) with collective nouns which also include an *of*-prepositional phrase (*of*-PP), as in *a group of students*, *a bunch of flowers*, *a couple of phone calls*, etc.

All together 23 singular collective nouns are investigated qualitatively and quantitatively: *band*, *batch*, *bunch*, *class*, *clump*, *couple*, *crowd*, *flock*, *gang*, *group*, *herd*, *host*, *majority*, *minority*, *number*, *pack*, *party*, *rash*, *series*, *set*, *shoal*, *swarm*, *troup*. These nouns are selected due to their relational nature. The (explicit or implicit) presence of the PP dependent specifying the members of the collective is obligatory, in contrast to more prototypical collective nouns, such as *committee* or *family*. In other words, it is much more likely that these words are followed by an *of*-complementation pattern.

It must be mentioned straight away that the presented analysis is limited in the sense that it investigates these collective nouns and their preferred subject-verb agreement exclusively in so-called 'complex collective subjects' (i.e., cases where an *of*-PP complement is present) but when doing so, also more complex examples than those listed above are analyzed, as in (1)–(3).

(1) [*A number of eminent scientists*] are active in promoting closer tics[sic] between scholarship and religion (COHA: 1985 MAG SatEvePost)

(2) [*The third set of case studies we discuss here*] was carried out by Barry Wilkinson (BNC:1985-1993, CAN 1101)

(3) [*A gang of bank robbers, masquerading as an unlikely string quartet*], engages in a battle of wills (BNC: 1985: 1933 HTT 46)

In these examples, the determiner is sometimes definite or one finds additional modification in the pre- or posthead. In other words, what is being investigated is the binominal structure of complex collective subjects following the constructional template [Det_1 (Mod) N_{coll} *of* (Det_2) (Mod) N_{pl} (Mod/Comp)].

In the literature, collective nouns have been researched extensively (Quirk *et al*. 1985: 757–759; Huddleston and Pullum 2002: 501–504; Corbett 2004; Keizer 2007). It is textbook knowledge that in English speakers have the option to either choose a verb that is singular or plural to follow these collective nouns, as in (4a) vs. (4b)

(4) a. [*The group*] has paid the entrance fee in advance.

b. [*The group*] have paid the entrance fee in advance.

The main explanation for the speaker's chosen verb agreement pattern, being either singular or plural, has long been the possibility of dual conceptualization of the collective noun. The variation is possible as the verb is either applicable to the collective as a whole or to the individuals that compose it (Biber *et al.* 1999: 188–189). If the collective noun *group* is interpreted as a conceptual unit, singular agreement is chosen; if it is seen as a homogeneous set of several visitors, where each member of the set has paid the entrance fee, then the plural verb form is preferred. Additionally, it has been suggested that the observable variation depends on register and region, with formal registers and American English showing a preference for singular agreement (Quirk *et al.* 1985: 19; Levin 2001: 60–70; Algeo 2006: 279–285; Hundt 2006, 2009). Diachronically, there also seems to be a growing overall preference for singular noun agreement. However, many examples do not reflect the postulated preferences and the observable variation is much more complex.

This is why, in recent years, several studies have been published which investigate additional factors that may be responsible for the chosen subject-verb agreement going beyond regional influence and dual conceptualization. For example, 'language-internal factors' like morpho-syntactic factors (e.g., type of determiner, distance between collective noun and verb), or semantic factors (e.g., animacy and type of the collective noun) have been shown to play of a role (Dekeyser 1975; Levin 1999, 2001; Depraetre 2003; Algeo 2006). Another crucial factor which might affect agreement, but which has

often been neglected in the literature so far, is whether the collective head noun takes an *of*-PP complement, as in (5).

(5) [*The group of visitors*] has/have paid the entrance fee in advance.

In these complemented NPs, the second so-called oblique noun (*visitors*) with its plural marking might affect subject-verb agreement choice as well, in the sense that in such cases it is more likely that speakers opt for plural agreement. This leads to the underlying main hypothesis of the monograph, namely that "the *of*-PPs and their constituent elements play a decisive factor in determining the pattern and present-day usage of the collective nouns that they accompany" (Fernández-Pena 2020: 4).

As a consequence, the monograph primarily investigates formal and lexico-semantic aspects of these prepositional constituents analyzing the potential interference and repercussions on the agreement relation. However, Fernández-Pena also looks at the nature of the chosen collective nouns, especially their quantifying potential, another aspect which has remained more or less unexplored so far. At the same time, the semantics of the verb is also investigated. Moreover, the existing research is also expanded by investigating the phenomenon diachronically.

In general, the following research questions are asked (RQs adapted from Fernández-Pena 2020: 4–5):

1. What determines verb number choice in the case of complex collective subjects: the collective noun, the PP or the structure as a whole?
2. To what extent (if at all) do the form and/or the semantics of the *of*-PP and/or the other elements in the subject affect the use of singular or plural verb number?
3. Are there any lexical biases? Is verb number agreement affected by the type of verb, type of collective noun or type of oblique noun?
4. Is there evidence of a diachronic evolution of those complex collective subjects, and in what way does it influence their current verbal agreement patterning and meaning?
5. What is the quantifying potential (if any) of complex collective subjects? To what extent does the interaction between the *of*-PP and verb agreement contribute to this use?

With regard to theoretical modeling, Fernández-Pena stresses that she uses a purely descriptive usage-based approach. Several theoretical frameworks are mentioned and

acknowledged, but the author does not openly subscribe to any particular theory. That being said, the monograph comes across as a functional-cognitive endeavor strongly inspired by the functionalist work of Keizer (2007) and Brems (2011), as well as by Langacker's (2008) Cognitive Grammar and by (Diachronic) Construction Grammar (e.g., Goldberg 2006; Traugott and Trousdale 2013).

The presented empirical studies are quantitative and corpus-based, using data that is extracted from three of the largest balanced corpora of English, namely the *Corpus of Contemporary American English* (COCA 1990–2012; Davies 2008–), the *Corpus of Historical American English* (COHA 1810–2009; Davis 2010–) and the *British National Corpus* (BNC 1960–1993; BNC Consortium).[1] Using these corpora, the author does not look at fine-grained dialectal or social variation, but what we get instead is an in depth, state-of-the-art multi-variate regression analysis with statistical testing of an extensive list of language-internal variables (for details see below).[2] Note, however, that at the end of the book, Fernández-Pena does investigate regional variation in more detail by incorporating data from the *Corpus of Global Web-Based English* (GloWbE; Davies 2013) analyzing some differences in six inner-circle varieties (American, Australian, British, Canadian, Irish, and New Zealand English).

With regard to length and chapter structure, the monograph is published in Routledge's *Studies in Linguistics* (volume 29) and is relatively concise (209 pages including references and indices) with only five main chapters including the introduction and conclusion. Chapter 2 summarizes the existing literature; the remaining two chapters are empirical and present first a diachronic corpus study (Chapter 3) and then a synchronic corpus study (Chapter 4). In the rest of this review, I will work through the individual chapters.

Chapter 2, "Complex collective subjects and verb number agreement in English: State of the art" (44 pages), is the main theoretical background chapter, which summarizes the current literature on the topic. It starts with a discussion of the internal differences of complex collective subjects (Section 2.1) showing that some of these binominal phrases have a partitive reading whereas others are pseudo-partitives. The collective noun can be

[1] Fernández-Pena uses the COCA and COHA versions provided by the online interface http://www.English-corpora.org, as well as the Lancaster Interface for the BNC at http://bncweb.lancs.ac.uk.
[2] The author also uses 'random forests' (Tagliamonte and Baayen 2012; Levshina 2015) and conducts some collexeme analysis (Stefanowitsch and Gries 2003). For the statistical analysis, the software *R* is used (R Core Team 2020).

interpreted referentially (*a bunch of flowers*, *a bunch of keys*), it can be given a partitive interpretation (*a bunch of the other guys*), or it can have a quantifier reading (*a bunch of guys*, in the sense of 'many guys'). Here *bunch* would be semantically bleached and refers to an indeterminate quantity. Fernández-Pena makes clear that, diachronically, the quantifying meaning can only develop from the constructional template [*a*/*an* N_{coll} *of* N_{pl}], with the indefinite article and a bare plural noun, as in *a number of guys* or *a group of people.* The chapter sheds light on the lexical-semantic differences of the various types (Section 2.1.1), but also discusses how these partitives and pseudo-partitives differ with regard to headedness, complexity, and compositionality (Section 2.1.2). This paves the way for the second part of the theoretical introduction, which is about verb-number agreement with subjects. Section 2.2 summarizes what the comprehensive grammars and syntactically oriented approaches have to say out about canonical and non-canonical agreement and its motivations: Corbett's canonical model (2004) and his 'agreement hierarchy' (Corbett 2006) are presented in Sections 2.2.1 and 2.2.2. Afterwards, Fernández-Pena continues to discuss alternative proposals, such as Langacker's (2008) Cognitive Grammar (Section 2.2.3). The last subsection (2.2.4) provides a discussion of the empirical studies which have been conducted so far. It also includes a short overview of the intra- and extralinguistic variables that have been identified in the existing literature which may affect the speakers' choice of agreement patterns. Fernández-Pena returns to these variables with a more detailed discussion in her empirical Chapter 4 (see below).

The theoretical background chapter is very well written and an easy read despite the complexity of the subject. It excels at summing up the current literature while pointing to many terminological inconsistencies in the current research. Especially useful for newcomers to the field is the introduction to measuring structural and syntactic complexity (the author's methodology is based on Rohdenburg 1996; Szmrecsányi 2004; Berlage 2014). Obviously, the chapter also prepares the ground for the two empirical chapters that follow.

Chapter 3, "Insights from diachrony: Reconciling form and meaning" (48 pages), is a diachronic investigation of only seven of the 23 collective nouns: *bunch*, *couple*,

group, *host*, *majority*, *minority*, and *number*. For the analysis, the author uses data from the COHA exclusively.[3] The following queries are run:

1. '(a/ the) (bunch/ couple/ group/ host/ majority/ minority/ number) of (*)(*.[NN2]/ people) *.[(VBZ/ VBDZ/ VDZ/ VHZ/ VVZ)]' for singular verbs;
2. '(a/ the) (bunch/ couple/ group/ host/ majority/ minority/ number) of (*)(*.[NN2]/ people) *.[(VBR/ VBDR/ VD0/ VH0/ VV0)]' for plural verbs.

After pruning the results, 4,776 examples are analyzed. Every collective noun is first discussed in a separate subsection which is then followed by a general discussion chapter. The main focus is on indefinite NPs with the indefinite article (e.g., *a group of people*), as this constructional template is the most frequent one and the only one susceptible to grammaticalization (i.e., development of a quantifier reading). However, the queries that are used also enable an investigation of examples with the definite article and/or modification (e.g., *the group of people I saw*, *the number of the people*). For all the collective NPs a potential increase or decrease in modification patterns is investigated as well as their (changing) verb agreement preferences over the years. When investigating verb agreement, the data set is reduced to those NP cases which are used in subject position and where the verb overtly marks singular/plural contrast. For the rest of the investigation (e.g., overall frequency increase), other argument positions are taken into consideration as well.

The main aim in Chapter 3 is to investigate the level of grammaticalization and the level of idiomaticity of the seven constructions. Signs of syntactic fixation and semantic opacity are explored. The question is to which extent the seven complex collective NPs have developed particular collocational and colligational preferences which indirectly could explain their verb agreement patterns in present-day English. Above all other things, Fernández-Pena investigates how often the collective noun combination, as in *a bunch of*, *a host of*, or *a number of*, has developed a quantifier function similar to *a lot of*, and if plural agreement increases for each type in time. The analysis reveals that the constructions do not form a homogeneous set and that the type of collective noun strongly conditions the binomial's structure and preferences (e.g., decrease in premodification). Although all seven types increase their syntactic fixation, show an increasing preference

[3] Although other diachronic historical corpora have been investigated as well, to a certain extent, pilot queries reveal that the COHA is the only diachronic corpus to provide enough data points for statistical investigation.

for the indefinite article, and an overall increase of plural agreement, one finds interesting individual differences.

The chapter represents an important contribution to the diachronic research on the topic, which so far has been rather scarce (Dekeyser 1975; Smitterberg 2006; Brems 2011; Shao *et al.* 2019). Especially, the presented classification schemes in Tables 3.4 and 3.5 (pp. 87–88) are an excellent attempt to determine the degree of grammaticalization and idiomatization. The constructions are positioned on a cline ranging from [*a number of* N_{pl}] as the most grammaticalized construction to [*a majority of* N_{pl}] as the least grammaticalized one. Additionally, the author includes a useful discussion of relative quantification as opposed to absolute quantification: out of the seven constructions, two of them show relative quantification (*minority* and *majority*) and five show absolute quantification (*bunch*, *couple*, *group*, *host*, and *number*). In general, it is shown that *minority of* and *majority of* behave slightly differently from the other collective nouns.

Chapter 4, "Modelling variation in verb number agreement with complex collective subjects in present-day English" (79 pages), is the main and longest chapter in the book. It reports the results of the synchronic corpus study. The corpora used are the BNC and the COCA in order to compare British with American English. Only five genres are investigated in the so-called 'original' version of the COCA (roughly 500 million words, before 2012 when the corpus was extended). Only the available written genres are used as a source because the spoken components of the BNC and COCA are not directly comparable. Regarding data retrieval, complex collective NPs in subject position are extracted. Again, various constructional templates are searched for. In contrast to the diachronic investigation, now the analysis is extended to the 23 collective nouns mentioned at the beginning. The data is again cleaned; for instance, examples with augmented subjects or with noun coordination are excluded.

After manual pruning of the data, the total number of valid instances is 5,406 tokens. The examples with the collective nouns *clump* and *couple* get excluded early on, as they do not show any variation in subject-verb agreement. In the end, 5,204 instances are analyzed. Those are coded for 25 variables, among them the dependent variable 'subject-verb agreement'. The other core variables are 'lexical type of collective noun' and 'lexical type of oblique noun'. Some of the coded morpho-syntactic variables are: 'type of Det_1' and 'type of Det_2', 'type of pre- and post-modification', and also

'morphological number of the oblique'. Here I would like to draw the attention to the author's classification scheme of morphological plural marking. POS (CLAWS7) tagged corpora[4] often show a lot of inconsistencies and errors when it comes to number distinction in noun tagging. In the chosen corpora, the used corpus tags NN0 (neutral for number), NN1 (singular) and NN2 (plural) are highly problematic for a number of reasons. On the one hand, NN0 is an extremely mixed bag and the NN1 tag subsumes singular and mass nouns. This grouping is seriously flawed as mass nouns are non-count nouns. In contrast to many researchers who simply ignore these issues, I applaud the author for her willingness to code the oblique nouns again using her own classification which is a useful scheme for future work in the field. Ultimately, the following bins are distinguished: 1) NN1 = singular nouns (*person*, *sample*); 2) NN2.s = plural marking by *– s*, (e.g., *bees*, *girls*, *computers*); 3) NN2.irregular = irregular plural marking by ablaut and other non-s strategies (e.g., *women*, *teeth*, *children*, *phenomena*); 4) NN0 = words which lack singular-plural contrast like mass nouns and others (e.g., *tuna*, *series*, *research*, *statistics*, *clothes*). The noun *people* constitutes its own category, due to its high frequency.

Fernández-Pena also investigates many variables related to structural complexity such as 'number of words preceding N_2', 'number of pre- and postmodifiers', 'syntactic configuration of the *of*-N_2 sequence', and also the 'distance between N_2 and the verb' counted by the number of intervening words. Additional lexico-semantic variables are 'lexical verb type', 'animacy of N_2', 'semantic number of N_2', and 'function of the NP/partition', deciding whether the binominal NP takes a partitive, a pseudo-partitive, or a referential reading. The extralinguistic variables are 'text' and 'variety'.

As an exploratory technique to determine the importance of the variables, a conditional random forest is run. After the random forest, a generalized linear mixed effects model with interactions is fitted. The random effects are variety, verb form, type of collective noun and type of oblique noun. Interactions are also fitted, namely between number of postmodifier of N_2 AND type of N_2 and between the number of intervening words between N_2 and verb AND type of N_2.

The results demonstrate that the patterns of agreement are mainly conditioned by the type of determiner, countability, animacy, semantic plurality, and morphological

[4] https://ucrel.lancs.ac.uk/claws7tags.html

number of the oblique, as well as by the syntactic complexity of the *of*-PP. For instance, non-human referents, which are less readily conceived of as aggregates of individuals, are significantly less likely to opt for plural verbal forms in comparison with human referents. At the same time, semantically singular and uncountable oblique nouns prefer singular agreement, while semantically plural and countable N_2s allow for greater variation. Regarding the morphological marking of the oblique noun, one can observe the so-called 'markedness effect', that is, singular oblique nouns are shown to be significantly less likely to occur with plural verbal forms in comparison with regular plural nouns in syntactically simple contexts. Importantly, the finding that irregular plural nouns favor more plural agreement than regular obliques contradicts previous studies. Morphologically unmarked nouns (NN0), including the semantically plural noun *people,* show a significant decrease in the likelihood of plural agreement with the increasing syntactic complexity of the noun phrase.

Syntactic complexity in terms of the number of postmodifiers is the only complexity measure to have a higher impact on agreement variation. In contrast, structural complexity counted in number of words is not a useful proxy of NP complexity. Most of the predictors that measure NP complexity in the study (number and length of the premodifiers, clause depth of the NP, and number of morphologically (un)marked nouns in the postmodifier) are discarded from the model for not improving its goodness-of- fit. The author also highlights strong lexical biases. Most of the variance in the data is accounted for by the collective noun, followed by the oblique noun and the verb.

> A series of collostructional analyses were used to examine these lexical factors more closely, producing further evidence of the collocational and colligational restrictions highlighted in Chapter 3. The most important findings concerned the interaction between the animacy of the referent and plural verb number, and the strong association of *bunch, couple*, *host*, *majority* and *minority* with the plural, as further evidence of their quantifying potential. (Fernández-Pena 2020: 177)

In general, it is shown that an intricate interplay of language internal (lexical, semantic, and formal) factors trumps extralinguistic factors like regional variety. Variety (British vs. American English) is still a significant predictor, but in NPs with *of*-PPs it plays less of a role.

To conclude this review, let us return to a more general evaluation of the monograph. As a reader I would have preferred to get the synchronic analysis first, with

the diachronic aspects being discussed only later. The interim presentation of the diachronic results in Chapter 3 somehow interrupts the theoretical discussion of the variables in Chapter 2 and their follow-up, hands-on coding and analysis in the regression model in Chapter 4. Moreover, it is a pity that no spoken data is investigated, a shortcoming that the author admits herself in the conclusion. Additionally, if one was desperately looking for criticism, what could be mentioned is that the book is a bit weak on the theoretical side, in the sense that it would have been nice to see a more elaborate discussion of some of the meta-theoretical concepts and what the empirical results 'mean' for usage-based, functional-cognitive or constructional models of language (change). In my opinion, the current length would have allowed for such an extension.

That being said, I would like to end with a clear recommendation: this book is an essential read for anyone interested in English binominals and agreement patterns. More generally, it will be of interest to students and researchers working in the field of language variation and change, corpus linguistics, and usage-based approaches to the study of language. The combined synchronic-diachronic analysis offers a much-needed, multi-faceted perspective and the large-scale quantitative analysis provides robust results. Moreover, from a didactic point of view, the book is truly a best practice example of how to explain and combine state-of-the-art quantitative methodology with meticulous qualitative analysis and substantial philological knowledge of the English NP. Especially the intelligent and motivated classification and categorization of the data as well as the thorough pruning of noise is something that no statistical (regression) model should do without. All this makes Fernández-Pena's monograph a highly valuable contribution to the field.

REFERENCES

Algeo, John. 2006. *British or American English? A Handbook of Word and Grammar Patterns*. Cambridge: Cambridge University Press.

Berlage, Eva. 2014. *Noun Phrase Complexity in English*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad and Geoffrey Leech. 2002. *The Longman Student Grammar of Spoken and Written English*. Harlow: Longman.

BNC Consortium. 2007. *The British National Corpus*. http://hdl.handle.net/20.500.12024/2554.

Brems, Lieselotte. 2011. *Layering of Size and Type Noun Constructions in English*. Berlin: Mouton de Gruyter.

Corbett, Greville G. 2004. *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2006. *Agreement.* Cambridge: Cambridge University Press.
Davies, Mark. 2008–. *The Corpus of Contemporary American English* (COCA): 520 million words, 1990–present. http://corpus.byu.edu/coca/.
Davies, Mark. 2010–. *The Corpus of Historical American English* (COHA): 400 million words, 1810–2009. http://corpus.byu.edu/coha/
Davies, Mark. 2013. *Corpus of Global Web-Based English* (GloWbE). https://corpus.byu.edu/glowbe/.
Dekeyser, Xavier. 1975. *Number and Case Relations in 19th Century British English: A Comparative Study of Grammar and Usage*. Antwerp: De Nederlandsche Boekhandel.
Depraetere, Ilse. 2003. On verbal concord with collective nouns in British English. *English Language and Linguistics* 7/1: 85–127.
Goldberg, Adele. 2006: *Constructions at Work. The Nature of Generalization in Language*. Oxford: Oxford University Press.
Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Hundt, Marianne. 2006. The committee has/ have decided…: On concord patterns with collective nouns in inner- and outer- circle varieties of English. *Journal of English Linguistics* 34/3: 206–232.
Hundt, Marianne. 2009. Concord with collective nouns in Australian and New Zealand English. In Pam Peters, Peter Collins and Adam Smith eds. *Comparative Studies in Australian and New Zealand English: Grammar and Beyond*. Amsterdam: John Benjamins, 207–224.
Keizer, Evelien. 2007. *The English Noun Phrase: The Nature of Linguistic Categorization*. Cambridge: Cambridge University Press.
Langacker, Ronald W. 2008. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press.
Levin, Magnus. 1999. Concord with collective nouns revisited. *ICAME Journal* 23: 21–33.
Levin, Magnus. 2001. *Agreement with Collective Nouns in English*. Lund: Lund Studies in English.
Levshina, Natalia. 2015. *How to Do Linguistics with R: Data Exploration and Statistical Analysis*. Amsterdam: John Benjamins.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.
R Core Team. 2020. *R version 4.2.0. A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. https://www.r-project.org/
Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7/2: 149–182.
Shao, Bin, Yingying Cai and Graeme Trousdale. 2019. A multivariate analysis of diachronic variation in a bunch of NOUN: A Construction Grammar account. *Journal of English Linguistics* 47/2: 150–174.
Smitterberg, Erik. 2006. Partitive constructions in nineteenth-century English. In Merja Kytö, Mats Rydén and Erik Smitterberg eds. *Nineteenth-century English: Stability and Change*. Cambridge: Cambridge University Press, 242–271.
Stefanowitsch, Anatol and Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8/2: 209–243.

Szmrecsányi, Benedikt M. 2004. On operationalizing syntactic complexity. In Gérald Purnelle, Cédrick Fairon and Anne Dister eds. *Le Poids des Mots: Actes des 7es Journées Internationales D'Analyse Statistique des Données Textuelles*. Louvain-la-Neuve: Presses Universitaires de Louvain, 1031–1038.

Tagliamonte, Sali and Harald R. Baayen. 2012. Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24/2: 135–178.

Traugott, Elizabeth Closs and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.

Reviewed by
Lotte Sommerer
University of Freiburg
Department of English
Rempartstr. 15
79085. Freiburg
Germany
E-mail: lotte.sommerer@anglistik.uni-freiburg.de

RiCL Research in Corpus Linguistics

Review of Elena Seoane and Douglas Biber eds. 2021. *Corpus-based Approaches to Register Variation.* Amsterdam: John Benjamins. ISBN: 978-9-027-21054-8. http://doi.org/10.1075/scl.103

Claudia Claridge
University of Augsburg / Germany

This book is a very rich collection, approaching register from various angles and with a wealth of methodologies and tools. Data types and contexts focused on in the contributions include modern (especially Chapters 5 and 6) and historical varieties of English (Chapters 10–12), as well as the analysis of English as a Foreign Language (EFL; Chapters 8–9). Both multiple- and single-register research (e.g., science, pop lyrics, or newspaper writing) is included, as well as studies approaching register from the point of view of one particular phenomenon (e.g., dative alternations, noun phrase modification). Needless to say, multi-dimensional analyses (MDA) are found here, both in the original (Chapter 9) and in the additive sense (cf. Berber Sardinha 2014, Chapter 8). A host of other methods are employed in these studies, such as mixed effects regression models (Chapter 3), random forests (Chapter 5), geometric multivariate analysis (Chapter 6), or generalised linear models (Chapter 9), just to name a few. Some tools receive a surprising but useful employment, such as the orthographic regulariser VARD (Baron and Rayson 2008) developed for historical data, being used here on modern pop lyrics (Chapter 8).

The majority of contributions are based on the understanding of the term 'register' by Biber (1988) and Biber and Conrad (2019), namely as "varieties associated with particular situational contexts that can be characterised for their audiences, medium […], interactivity, production circumstances, communicative purposes" (p. 2), and thus

with similar linguistic characteristics. While the latter have been investigated in quantitative ways, the characteristics of the situational contexts have neither been measured nor treated as continuous variables but have usually been described in a subjective and generalising manner. A critical discussion of such earlier approaches together with a comparison of the textlinguistic (Biber 1988 and following) and a systemic-functional perspective is presented by Biber, Egbert, Keller, and Wizner in Chapter 2, before they proceed with their undertaking of providing such measurements of the situational context. They report on two case studies from earlier papers (Biber *et al.* 2020 and Biber *et al.* 2021): one based on a range of web texts and one on conversational discourse types in the spoken component of the second *British National Corpus* (Spoken BNC2014; cf. Love *et al.* 2017). The web case study quantified 23 situational parameters on a six-point scale averaged across two independent raters and resulted in two dimensions of situational contexts. While Dimension 1 is characterised as opinionated discourse vs. technical information supported with evidence, Dimension 2 marks narrative, entertaining discourse versus other communicative purposes (explanatory, advice, and procedural discourse), from which dimension scores for individual texts can be derived. The scored texts form five clusters, which actually cut across registers. The second case study coded conversational discourse units for nine communicative purposes, leading to 16 clusters, characterised by labels such as ‘figuring-things-out’, ‘joking around’, or ‘conflict’. While an overall convincing start at the situational aspect, the second case study would have profited from more detail and illustration.

Two contributions (Chapters 3 and 5) deal with the dative alternation, which might have been placed in direct sequence. Chapter 3, by Engel, Grafmiller, Rosseel, Szmrecsanyi, and van de Velde, investigates and compares the effects of register and of various language-internal constraints on the choice of the dative realisation as either ditransitive or prepositional. Taking into account seven language-internal factors, in particular, recipient and theme definiteness as well as constituent length, and four registers as predictors (conversation, parliamentary debates, blogs, and newspaper articles) in a mixed effects regression analysis, they showed core grammar to be relatively stable across registers and register effects to be smaller than for other factors. Similarly, in Chapter 5 by Röthlisberger, register is marginally outranked in importance by variety of English (of which nine are investigated) and in some varieties (e.g., British

and Singapore English) there is little inter-register variation. Register differences across varieties are generally small, but with subtle distinctions regarding formal registers, which is hypothesised to be due to indigenisation effects. The internal constraints weight ratio and pronominal recipient turn out to be the most important factors overall. While the double object construction generally dominates, the prepositional variant is more likely in non-native Englishes.

Chapters 4 and 6 are two outliers in the volume in the sense that they both proceed from a Hallidayan systemic-functional perspective on register. In Chapter 4, Pérez-Guerra tests the hypothesis of theme choice being indicative of register, on the basis of themes having dual linguistic and situational/functional status just like registers. Secondly, the adequacy of two definitions of theme, those by Halliday (1985) and by Berry (1995) respectively, is tested as to their contribution to register characterisation. In an analysis encompassing 15 written registers of American English, the Hallidayan concept of theme (first ideational element) is claimed to be a plausible predictor of (dis)similarity between registers, while Berry's preverbal theme concept fares less well. The chapter would have gained in persuasiveness if the technical description had been somewhat more accessible and, in particular, if more linguistic illustration had been provided.

Neumann and Evert (Chapter 6) used 41 register-sensitive lexico-grammatical features for their analysis of 2,844 texts from the Hong Kong, Jamaica, and New Zealand components of the *International Corpus of English* (ICE).[1] They use a geometric multivariate analysis, inspired by multidimensional analysis (MDA), to explore and visualise the linguistic differences between texts. The resulting four dimensions are: 1) conceptual speaking/conceptual writing (e.g., ICE categories conversation, social letters, and news), 2) dialogic written/neutral (e.g., social letters, creative writing), 3) descriptive-narrative versus instructive-regulative (e.g., news, business letters, and administrative writing), and 4) neutral/online production (e.g., unscripted discourse). The three varieties differ in variance across the four dimensions, with the least variance in New Zealand and the most variance in Hong Kong texts, which may be due to less established conventions in the younger variety. Tenor-related (pragmatic) aspects seem to contribute more to variation than field-related aspects. The study is interesting not only for the visualisation aspect, but also for the overlap and

[1] http://ice-corpora.net/ice/index.html

differences it shows regarding Biber's dimensions, and for the potential problems inherent in the ICE text classification that it highlights. This very rich treatment also makes repeated reference to web materials for further illustration and corroboration, which, on the one hand, is laudable, but, on the other hand, detracts from the independence of the chapter.

Chapter 7, by Botha and van Zyl, focuses on the noun phrase, a feature that has previously been attested with interesting behaviour in variation and change. The novel approaches in this contribution concern using proportions of modifiers relative to the number of nouns used in a register and conducting as many as 45 pairwise comparisons for individual modifier forms in ten registers using effect size measures. The data combines the five registers of the *Corpus of Contemporary American English* (COCA; Davies 2008–) with five web-based registers from the *Corpus of Online Registers* (CORE; Davies 2016), which are matched for similar communicative purposes and intended audiences. Modification structures included in the study are premodifying (proper) nouns and adjectives (L1 position only) and postmodifying prepositional phrases, non-finite and relative clauses. High levels of premodification as well as of prepositional and non-finite postmodification were found to characterise written informational registers, while postmodification by *that*-relatives marked oral and involved registers. Some modification features showed more register-sensitivity than others, for instance, prenominal (proper) nouns and *which*-relatives as opposed to adjectives and *that*-relatives. The high number of comparisons allowed the observation of very fine-grained differences.

The next two chapters have a common focus in so far as both are concerned with EFL matters. Werner's contribution on pop lyrics points to them being a register in their own right, in contrast to prevailing views (especially EFL) that they are speech-like and conversational in nature. An 'additive MDA' (Berber Sardinha 2014) performed on the *Corpus of English Pop Lyrics* (LYPOP; Werner 2020) from 2001 to 2016, comprising 1,842 lyrics and 547,758 tokens, mapped the register onto Biber's (1988) original dimensions. In line with their non-conversational situational characteristics, pop lyrics usually have closest score associations with written (all dimensions) and partly with formal and informational types, such as official documents (Dimension 2) or academic prose (Dimension 4). Nevertheless, the large standard deviations exhibited by pop lyrics scores always include (Dimensions 3, 4, and 5) or overlap substantially with

conversation (Dimensions 1, 2, 6). Werner concludes that while the pop lyric register is not conversational as such it uses a range of features in such a way as to produce the impression of an imagined speech event with pseudo-dialogicity and thus a pretence at conversationality.

Proceeding from the fact that little is known about EFL academic learner writing beyond performance in the register of argumentative essays, the chapter by Larsson, Paquot, and Biber reports a new MDA to investigate register effects in EFL learner writing. The aim is to find out how it differs both from native writing and across different L1-groups and also to investigate how these findings are influenced by register. A MDA performed on a 3.5 million word corpus of native and EFL argumentative essays, research papers, and published scientific articles (all drawn from existing corpora such as the *International Corpus of Learner English* [ICLE], the *Louvain Corpus of Native English Essays* [LOCNESS] the *Varieties of English for Specific Purposes dAtabase* [VESPA], the *Louvain Corpus of Research Articles* [LOCRA], among others)[2] led to two dimensions, Dimension 1 distinguishing a personal versus a topic-focused style and Dimension 2 an evaluative style as opposed to factual description. On both dimensions the influence of register is shown to be more important than either (non-)nativeness or the specific L1 of learners. A more personal style thought to be a generic characteristic of EFL writing is only found for argumentative essays (including those of native speakers), while all writers show register awareness by adopting a more topic-focused (Dimension 1) and factual-descriptive (Dimension 2) style for research papers. Moreover, EFL learners are not a coherent group, but show significant differences between different L1 backgrounds: for example, while Norwegians prefer a more personal approach, French learners use more topic-focussed writing.

The final three papers in the collection all take a diachronic perspective. Rodríguez-Puente's chapter charts the attestations of nominalisations with nine Romance and native suffixes covering four meanings across 18 registers of Early Modern English (1500–1760), taken from the *Corpus of English Dialogues* (CED; Kytö and Walker 2006), the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME; Kroch *et al.* 2004), and the *Corpus of Historical English Law Reports*

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html

(Rodríguez-Puente *et al.* 2018). The suffixes include highly frequent *-ion*, medium-frequent *-ment*, *-ity*, *- ness*, *-age*, *-ship*, and low-frequent *-dom*, *-hood*, and *-head*. Their occurrence is significantly linked to register, with formal, writing-based and writing-purposed texts showing higher and informal, speech-related texts lower frequencies. Exceptions to this pattern are due to the type and purpose of text, with a narrative style leading to lower frequencies (e.g., the bible, fiction, or travelogue) and an instructional and persuasive outlook to higher frequencies (e.g., sermons). While nominalisations increase in most registers, the most in sermons (1640–1710), in line with the increasing nouny-ness and literate character of texts in the period, private letters and trial proceedings show a decrease. The developments are shown nicely in Figure 4, which, however, seems to be lacking the last period for the registers drama, trial proceedings, and witness depositions. The productivity of suffixes, shown by aggregation of types from the first to the last period, is most pronounced for borrowed suffixes overall, but also for *-ness*. There are generally no register effects regarding productivity, with trials again standing out and showing an unusual decline.

Degaetano-Ortlieb's contribution is an investigation into the development of the scientific register in the twentieth century, here represented by the mathematical, physical, and engineering publications of the *Proceedings of the Royal Society of London*.[3] In a bottom-up and data-driven approach based on POStrigrams and using Kullback-Leiber divergence, critical periods of change as well types of change were identified. The (early) 1920s turned out as an important period of change, followed by later stabilisation, while the crucial constructions in the registerial change all involve nominal compounds (in particular, det-N-N, N-N-prep and adj-N-N). With premodified noun phases rather than those with prepositional postmodification becoming a more distinctive use, the change also represents informational densification. Also, the rise of pure N-structures instead of N+prep structures shows not only higher informativity but also increasing specialisation.

The final contribution, by Hiltunen, deals with sub-register variation in nineteenth-century newspapers, but also with the question of how to work best with the *British Library Newspapers*[4] database in a corpus-linguistic approach. Regarding the latter, Hiltunen extracted two corpora (A and B) automatically from the database for the

[3] https://royalsocietypublishing.org/journal/rspl
[4] https://www.bl.uk/collection-guides/newspapers

purpose of data triangulation to overcome weaknesses of individual corpora. Corpus A (4.9 million words) sampled whole issues from five geographically diverse papers published during the whole century by choosing issues from two months in 10-year intervals. Corpus B (10.1 million words) was compiled from 100 texts for each of the seven sub-registers in focus extracted for every decade. Both corpora also required an Optical Character Recognition (OCR) confidence level of 90 per cent for the texts to be included. It was especially this essential criterion for linguistic accuracy which led to massive loss of data (e.g., one whole newspaper for Corpus A) and caused Corpus A to be very unbalanced. As for Corpus B, even though it was more balanced, it also showed some coverage gaps regarding sub-genres (e.g., sports, classified ads). The POS-tagged corpora were used to carry out a synchronic register analysis with selected features chosen from Biber's (1988) Dimensions 1, 2, 4, and 5, namely, private verbs, first, second and third person pronouns, past tense forms, suasive verbs, infinitives, conjuncts, and sequences of two proper nouns. All of those indicated sub-register differences, with most marked distinctions in editorials for involved, persuasive, and explicit-linkage features, and with news and sports characterised by narrative features. As an outlier, birth/death/marriages notices were only characterised by proper noun sequences.

The entire volume clearly makes for very stimulating reading, with many convincing insights and with inspirations for data sources and use, as well as for methodology. Sometimes, however, the technical details regarding the latter become somewhat overwhelming and not easy to follow for the reader —a minor criticism that applies more to papers in the first half of the volume. Another minor weakness of the volume concerns the general lack of cross-references, for which there would have been ample opportunity given the very real links between the contributions. One striking example of a missing cross-reference concerns Matthiessen's (2019: 26) map of register traditions mentioned in Chapter 6 (p. 145), which is actually reproduced in Chapter 4 in the volume (p. 88).

REFERENCES

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University. http://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf

Berber Sardinha, Tony. 2014. 25 years later: Comparing internet and pre-internet registers. In Tony Berber Sardinha and Marcia Veirano Pinto eds. *Multi-dimensional Analysis, 25 Years on: A Tribute to Douglas Biber*. Amsterdam: John Benjamins, 81–105.

Berry, Margaret. 1995. Thematic options and success in writing. In Mohsen Ghadesy ed. *Thematic Development in English Texts*. London: Pinter, 55–84.

Biber, Douglas. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Susan Conrad. 2019. *Register, Genre, and Style.* Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert and Daniel Keller. 2020. Reconceptualizing register in a continuous situational space. *Corpus Linguistics and Linguistic Theory* 16/3: 581–616.

Biber, Douglas, Jesse Egbert, Daniel Keller and Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics* 171: 20–35.

Davies, Mark. (2008–). *The Corpus of Contemporary American English (COCA): 560 Million Words, 1990-Present*. https://corpus.byu.edu/coca/

Davies, Mark. (2016–). *Corpus of Online Registers of English (CORE)*. https://www.english-corpora.org/core/

Halliday, M. A. K. 1985. *An Introduction to Functional Grammar*. London: Edward Arnold.

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)*. Department of Linguistics: University of Pennsylvania.

Kytö, Merja and Terry Walker. 2006. *Guide to a Corpus of English Dialogues 1560–1760*. Uppsala: Acta Universitatis Upsaliensis.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3: 319–344.

Matthiessen, Christian M.I.M. 2019. Register in systemic functional linguistics. *Register Studies* 1/1: 10–41.

Rodríguez-Puente, Paula, Teresa Fanego, María José López-Couso, Belén Méndez-Naya, Paloma Núñez-Pertejo, Cristina Blanco-García and Iván Tamaredo. 2018. *The Corpus of Historical English Law Reports 1535–1999 (CHELAR), v.2*. University of Santiago de Compostela: Research Unit for Variation, Linguistic Change and Grammaticalization.

Werner, Valentin. 2020. Teaching grammar through pop culture. In Valentin Werner and Friederike eds. *Pop Culture in Language Education: Theory, Research, Practice*. London: Routledge, 85–104.

Reviewed by
Claudia Claridge
University of Augsburg
Faculty of Philology and History
Universitätsstraße 10
86159 Augsburg
Germany
E-mail: claudia.claridge@philhist.uni-augsburg.de