UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Doctorado en Lingüística Inglesa

TESIS DOCTORAL

Academic discourse at university: corpus approaches to learner writing

El discurso académico en la universidad: enfoques de corpus a la escritura de estudiantes

Tesis para optar al Grado de Doctor
presentada por
Noelia Navarro Gil

Directoras
Dra. Elena Martínez Caro
Dra. Helena Roquet Pugès

Madrid,

UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE FILOLOGÍA

Doctorado en Lingüística Inglesa

Tesis doctoral en formato publicaciones

Academic discourse at university: corpus approaches to learner writing

El discurso académico en la universidad: enfoques de corpus a la escritura de estudiantes

Tesis para optar al Grado de Doctor
presentada por
Noelia Navarro Gil

Directoras
Dra. Elena Martínez Caro
Dra. Helena Roquet Pugès

Madrid

Acknowledgments

I would like to thank my thesis directors, Dr Elena Martínez Caro, and Dr Helena Roquet Pugès, for their constant guidance and support throughout the course of this doctoral thesis. Elena, thank you for accompanying me in this long journey, in which distance has not been an obstacle to many hours of stimulating conversation and inspiring discussions about language. Your constant encouragement, kindness, and insightful comments have helped me to develop as a researcher and also to become a better writer. Helena, thank you for your kindness, support, and your faith in me throughout all these years. Your enthusiasm, inspiring feedback and helpful suggestions on different research processes, such as data collection and statistics, have substantially improved the quality of my work. Thank you both for your constant availability and feedback on the content and writing of my work. I will always remember and cherish the support you have given me, and I hope I can follow your example in my academic life for many years to come.

I am also grateful to Anna Navarro and to all members of the Institute for Multilingualism at Universitat Internacional de Catalunya, whose generosity made it possible for me to attend many international conferences (e.g. ICLHE in Copenhagen, Learner Corpus Based approaches to Second Language Acquisition, in Utrecht, Writing Analytics, in Malmö) –these unforgettable experiences made a considerable impact on my development as a researcher and provided me with the confidence to pursue my goals for this PhD thesis.

My gratitude extends to all my colleagues at the Institute, Dr Mandy Deal, Monica Clua, Dr Joan Ploettner, Wendi Smallwood, Dr Jennifer Ament, Natalie Gommon, Dr Natalia Evnitskaya, Dr Mayya Levkina, Yagmur Elif, Marta Segura, and many others. I know that by mentioning you all may not be as special as it might be, but I am really grateful for your help at times that were particularly stressful during this project. I feel blessed to have had the opportunity to work with you, and to share conferences, hostels, and talks about research, articles, and read-me-for-pleasure books. A very deep thank you to Dr Júlia Barón, for your friendship, your support, your tips on journals and paragraphs, and for all the great times we have spent together. You are a source of inspiration.

To my dear parents and sister: thanks to your love, enthusiasm, and support, I was finally able to finish this thesis. I could not have done it without you. Thank you for being there every step of the way. Finally, I would like to thank my wonderful partner: for your endless love, encouragement, patience, and for starting this journey together; you helped me to set goals that seemed unreachable at the beginning; goals that, with time, became only difficult, to eventually be *just* something that it turned out I was able to do. Thank you so much. Gracias.

Noelia Navarro Gil,
Barcelona, June 2019

Abstract

Academic writing in English has often been described as a primarily reader-oriented discourse, in which the structure, objectives, and claims are made explicit and carefully framed. Second or foreign language (L2) learners often transfer part of their first language (L1) writing cultureH into their L2 texts. This is problematic because academic texts call for a high degree of *disciplinarity*: learners not only have to be aware of the conventions of the L2 regarding language use in a particular genre, but also of the academic conventions of their own discipline. The present doctoral dissertation demonstrates how corpus approaches to L2 texts can help to identify learner writing features when compared to native or expert counterparts. The four studies presented in this thesis highlight some of the linguistic challenges students face when writing in English for different academic purposes and disciplines at university, and provide pedagogical suggestions for the teaching and learning of certain linguistic features that can be useful for L2 academic writers and instructors.

Study one examined the effects of content-based language instruction (CBI) on the production of academic vocabulary in a classroom writing task. The texts were written by first-year university students enrolled in two different instruction settings, English as medium of instruction (EMI) and the same programme in their L1, over one semester. Both the materials used in class and the learner corpus were examined in order to identify the degree to which they incorporate items from three lists of interdisciplinary academic terminology, namely the Academic Vocabulary (AVL), Collocations (ACL) and Formulas List (AFL). The results indicated that the learner corpus, both L1 and EMI learners, produced more general academic and technical words after the course; EMI learners also increased their use of collocations and formulas. The benefits of CBI for acquiring academic terminology and for developing disciplinary literacy are discussed in the light of the instruction settings under study.

Study two explored the use of adversative Linking Adverbials (LAs) in the academic writing of advanced English as a Foreign Language (EFL) learners with different linguistic backgrounds. The learner corpus used in this study consisted of 50 argumentative essays, which were the final assignment of a content subject at university. These were contrasted with a native corpus: the American university students' subcorpus included in the LOCNESS corpus. Liu's (2008) list of adversative LAs was used for the analysis. Findings revealed that both non-native (NNS) and native speakers of English (NS) use similar types of adversative LAs, but NNS place them regularly in sentence- and sometimes in paragraph- initial position, which often resulted in punctuation issues and misuse. The analysis performed according to L1 yielded unexpected results in terms of preference, frequency, and placement of adversative LAs. The so-called ‘teaching effect’ is considered one of the main factors influencing the learners’ choices.

Study three investigated the use of reflexive metadiscourse (MD) in a learner corpus of bachelor dissertations (BDs) written in English by Spanish L1 undergraduates in their last year of studies in medicine and linguistics, and compared the results with an expert corpus of research articles (RAs). The results showed that overall both corpora contain similar frequencies of textual MD, but this is only true when we look at the results according to discipline. In spite of this quantitative similarity, there were cases of overuse and underuse in the learner corpus that highlight features of the BD genre on the one hand, and EFL Spanish writing on the other hand.

Finally, study four explored the use of lexical bundles (LBs) in the learner BDs and the expert RAs corpora. By focusing on the introduction and conclusion sections, the most frequent 3-, 4- and 5-word bundles were identified, to later study their types, structures, and functions. The results showed differences in the use of LBs across disciplines, genres and sections, suggesting pedagogical implications for the inclusion of LBs in the L2 writing curriculum.

This doctoral thesis presents quantitative and qualitative analyses of learner corpora that represent EFL academic writing at university. Based on the findings that emerge from the corpus-based and corpus-driven analyses of the linguistic devices explored, three main implications arise: first, many important discourse elements in academic writing, such as academic formulas, linking adverbials, metadiscourse markers and lexical bundles, are situated on the phraseological dimension of language. The results obtained support a pedagogical approach that contemplates this highly patterned nature of language. Second, results show that there are certain academic writing practices that can transfer from students’ L1 to their L2. Taking into account students’ L1 writing culture when teaching English academic writing may be useful to identify and deal with possible error-prone items. Finally, the results support the notion of academic writing as a highly genre- and discipline-specific discourse, and thus emphasize the need for more learner and expert corpus-informed pedagogical materials on academic writing at university.

Resumen

La escritura académica en inglés se ha descrito como un discurso orientado principalmente al lector, en el que la estructura, los objetivos y las afirmaciones se hacen explícitas y se enmarcan cuidadosamente. Los estudiantes de inglés como segunda lengua o como lengua extranjera (L2) a menudo transfieren parte de las convenciones de su lengua madre (L1) a sus textos en L2. Esto es problemático porque los textos académicos requieren un alto grado de *disciplinaridad*: los estudiantes no solo deben conocer las convenciones de la L2 con respecto al uso del lenguaje (por ejemplo, la gramática) en un género en particular, sino también las convenciones de su propia disciplina. La presente tesis doctoral demuestra cómo diferentes enfoques de corpus aplicados a la escritura de estudiantes en L2 pueden ayudar a identificar las características de este tipo de escritura, cuando se compara con la redacción académica de nativos o expertos. Los cuatro estudios que construyen esta tesis resaltan algunos de los desafíos lingüísticos a los que se enfrentan los estudiantes al escribir en inglés para diferentes propósitos académicos y disciplinas en la universidad y proporciona sugerencias pedagógicas para la enseñanza y el aprendizaje de ciertas construcciones lingüísticas que pueden ser útiles para escritores e instructores del inglés académico como L2.

El estudio uno examinó los efectos de la instrucción de lengua basada en contenido (CBI por sus siglas en inglés) en la producción de vocabulario académico en una tarea escrita de clase. Los textos fueron redactados por estudiantes universitarios de primer año inscritos en dos modalidades diferentes, inglés como medio de instrucción (EMI por sus siglas en inglés) y el mismo programa en la L1, durante un semestre. Tanto los materiales utilizados en la clase como el corpus de estudiantes se examinaron para identificar el grado en el que incorporan elementos de tres listas de terminología académica interdisciplinaria, específicamente las listas de vocabulario (AVL), de colocaciones (ACL) y de fórmulas (AFL) académicas. Los resultados indicaron que los estudiantes, tanto de L1 como de EMI, produjeron un mayor número de palabras académicas y técnicas después del curso; Los estudiantes de EMI también aumentaron el uso de colocaciones y fórmulas. Los beneficios de CBI para adquirir terminología académica y desarrollar la alfabetización disciplinaria se discuten a la luz de las dos modalidades estudiadas.

El estudio dos exploró el uso de las conjunciones adversativas en textos académicos escritos por estudiantes avanzados de inglés como lengua extranjera con diferentes L1s. El corpus utilizado en este estudio consistió en 50 textos argumentativos, que fueron el trabajo final de una asignatura de contenido en la universidad. Éstos se contrastaron con un corpus nativo: el subcorpus de estudiantes universitarios americanos incluido en el corpus LOCNESS. La lista de conjunciones adversativas de Liu (2008) se utilizó para el análisis. Los resultados revelaron que ambos, hablantes nativos y no nativos de inglés, utilizan categorías similares de conjunciones adversativas, pero que los estudiantes no nativos las colocan regularmente al principio de la frase y, algunas veces, también del párrafo, lo que a veces resulta en problemas de adecuación y ortografía. El análisis con respecto a las diferentes L1 mostró resultados interesantes en términos de preferencia, frecuencia y colocación de conjunciones adversativas. El llamado ‘efecto pedagógico’ se considera uno de los factores principales que pudo influenciar las elecciones de conjunciones adversativas por los estudiantes.

El estudio tres investigó el uso del metadiscurso (MD) reflexivo en un corpus de estudiantes de trabajos de final de grado (TFGs) escritos en inglés por estudiantes universitarios cuya L1 es el español, en su último año de los grados de medicina y lingüística, y lo compara con el uso del MD en un corpus de artículos de investigación (AI) en las mismas disciplinas. Los resultados mostraron que, en general, TFGs y AIs contienen frecuencias similares de MD textual, pero esto sólo es cierto cuando miramos los resultados desde una perspectiva disciplinar. A pesar de esta similitud cuantitativa, hay casos de ‘uso excesivo’ y ‘escasez de uso’ en el corpus de estudiantes, lo que destaca características del género por un lado, y de la transferencia de convenciones del español por otro.

Finalmente, el estudio cuatro exploró el uso de paquetes léxicos (LBs por sus siglas en inglés) en el mismo corpus de estudiantes (TFGs) y de expertos (IAs) en lingüística y medicina. Centrándonos esta vez en las secciones de introducción y conclusión de los textos, identificamos los LBs de 3, 4 y 5 palabras más frecuentes en el corpus, para luego estudiar sus tipos, estructuras y funciones retóricas. Los resultados mostraron diferencias en el uso de LBs entre disciplinas, géneros y secciones, lo que sugiere implicaciones pedagógicas para su inclusión en la enseñanza de la escritura académica en inglés.

Esta tesis doctoral presenta análisis cuantitativos y cualitativos de varios corpus de estudiantes que representan la escritura académica en inglés como lengua extranjera en la universidad. Tras los diferentes análisis de fenómenos lingüísticos, surgen tres implicaciones principales: primero, que muchos de los elementos importantes del discurso académico, tales como fórmulas académicas, conjunciones adversativas, marcadores de metadiscurso y paquetes léxicos, se sitúan en la dimensión fraseológica del lenguaje. Los resultados obtenidos apoyan un enfoque pedagógico que dé cuenta de estos patrones del lenguaje. En segundo lugar, los resultados muestran que hay ciertas prácticas académicas que se transfieren de la L1 de los estudiantes a su L2. Tener en cuenta la cultura de la escritura académica en diferentes L1s durante la enseñanza de la escritura académica puede ser útil para identificar y tratar elementos que son propensos a errores. Finalmente, los resultados apoyan la noción que describe la escritura académica como un discurso altamente específico en sus géneros y disciplinas, y en consecuencia se enfatiza la necesidad de utilizar materiales pedagógicos sobre escritura académica en la universidad que estén basados en corpus de estudiantes y expertos.

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ACL	Academic Collocation List
AFL	Academic Formulas List
AVL	Academic Vocabulary List
BAWE	British Academic Written English
BDs	Bachelor dissertations
BNC	British National Corpus
CBI	Content-based (language) instruction
CEFR	Common European Framework of Reference
CEUNF	Corpus of Spanish EFL students
CIA	Contrastive Interlanguage Analysis
CL	Corpus Linguistics
CLIL	Content and Language Integrated Learning
CM	Class material
COCA	Corpus of Contemporary American English
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EMI	English as Medium of Instruction
ESL	English as a second language
ESP	English for Specific Purposes
FLOB	Freiburg-London-Oslo-Bergen corpus
HE	Higher Education
ICLE	International Corpus of Learner English
ICLHE	Integrated Content and Language in Higher Education
IP	Interpersonal metadiscourse
L1	First language
L2	Second language
LAs	Linking Adverbials
LBs	Lexical Bundles
LCR	Learner Corpus Research
LFP	Lexical Frequency Profile
LIN	Linguistics
LOCNESS	The Louvain Corpus of Native English Essays
MD	Metadiscourse
MED	Medicine
MI	Mutual Information
MT	Metatextual (textual metadiscourse)
MUC	Maastricht University Corpus
NIP	Non-initial sentence position
NNS	Non-native speakers
NP	Noun phrase
NS	Native speakers
PBL	Problem-based Learning

PICAE	Pearson International Corpus of Academic English
PMW	Per million words
PP	Prepositional phrase
RAs	Research articles
SIP	Sentence initial position
SLA	Second Language Acquisition
SPE	Corpus of professional editorialists writing in English
T1	First time
T2	Second time
TL	Target language
VP	Verbal phrase

List of original publications[1]

Chapter 5: Navarro-Gil, N. (2019). The effects of a content-based language course on students' academic vocabulary production. *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education,* 2 (2), 25-42.

Chapter 6: Navarro-Gil, N., Roquet-Pugès, H. (in press). Linking or delinking of ideas? The use of adversative linking adverbials by advanced EFL learners, *Revista Española de Lingüística Aplicada.*

Chapter 7: Navarro-Gil, N. (2018). Reflexive Metadiscourse in a corpus of Spanish bachelor dissertations in EFL. *Research in Corpus Linguistics,* 6, 29-49.

Chapter 8: Navarro-Gil, N., Martínez-Caro, E. (2019). Lexical bundles in learner and expert academic writing. *Bellaterra Journal of Teaching & Learning Language and Literature,* 12 (1), 65-90.

[1] The letters of acceptance from each journal editor/s have been included in Appendix 10.

Table of Contents

1. Introduction

In the past few decades, researchers have become increasingly interested in describing how academic discourse is constructed in different disciplines and genres. Globalization and the emergence of English as a *lingua franca*, and also as the language of science and research, has made English academic discourse a requisite for publishing, and thus a basic skill for novel researchers and university students. This has, in fact, had a considerable impact in most European higher education (HE) institutions, in which the number of programmes and subjects offered in English has seen a steady rise. University students are often required to listen to (lectures), speak (presentations), read (literature), and write (assignments) at different levels of immersion, in English. In the Spanish context, globalization has also triggered the raise of English as Medium of Instruction (EMI) programmes, and/or English for Academic or Specific Purposes (EAP, ESP respectively) courses (Pérez-Vidal et al., 2018). This, together with the English B1–B2 requisite that most universities have set for students to be able to graduate[2], pose an additional challenge for non-native speakers (NNS) of the language, especially when the status of English is that of a 'foreign language', i.e. not an official language in the country. English-as-a-foreign-language (EFL) undergraduate students are not only required to learn and produce 'general English' in order to be able to obtain the B1 or B2 certificates required in bachelor degrees, but also to manage 'specialized English' or the discourse of their disciplines, in order to succeed academically.

Research that looks at the implications of teaching content through English (e.g. EMI) or teaching academic discourse in English (e.g. EAP) in HE, from different perspectives (e.g. learners, instructors, or contents used) is gaining traction. With the development of corpus linguistics, i.e. the study of authentic language in the form of electronic texts (that can come from spoken –transcribed– or written data), the way languages are perceived, e.g. in terms of grammar, lexis, structures, functions, and patterning, has changed drastically, moving from initially intuitive perceptions to more evidence-based explanations. In the advent of corpus research, academic language is no longer seen as an invariable entity, composed of individual words or types that form grammatical structures, but more as an organic construct, that adapts, changes, and evolves according to modes, registers, genres, and disciplines. This has inevitably changed, and is still changing, the way languages are taught and learnt.

[2] For some specific programmes, the B1/B2 English certificate is even an entrance requirement.

Over time, different corpus tools and methods have been put forward in an attempt to account for and capture the variability of language. In fact, corpus methods are increasingly being used in other fields and for different purposes (e.g. second language acquisition, pragmatics, sociology etc.), mainly because they can provide highly contextualized explorations of language. With this regard, one of the main contributions corpus research has made to language-related fields is the identification of recurrent word combinations (i.e. ‘formulaicity’, ‘phraseology’, ‘language patterning’ [Hunston, 2002; Meunier & Granger, 2008; Wray, 2002]) as the building blocks that are fundamental for the construction of language. These recurrent word combinations have, at the same time, been shown to reduce processing time for users and interlocutors; therefore, learning and producing language through ‘chunks’ or ‘formulas’ can certainly be more effective than learning isolated words out of context. This finding has joined two previously separate paradigms, namely lexis and grammar, and has since then changed the study of lexicography. Another important contribution of corpus research that is often mentioned in the literature is the corroboration of genre variability: academic, fiction, magazines, news, spoken events, etc., they all seem to show particular linguistic conventions (see e.g. Biber et al., 1999; Hyland, 2008b); sometimes, language practices are shared across genres. A general approach to language focuses on the latter, the shared features. A more specialized perspective, on the other hand, would focus on one genre, e.g. the academic genre, in order to describe its particular characteristics. However, while it is true that “grouping similar genres together makes the description of large numbers of texts more manageable, and enables us to make comparisons across disciplines” (Gardner & Nesi, 2013, p. 32) and each of these genres has particular characteristics (e.g. use of ‘signposting devices’ in the academic genre), they are per se still too broad to form a generalizable mass; for example, essays, articles, evaluation rubrics, and institutional e-mails could all be subgenres of the abovementioned ‘academic genre’, but these can differ vastly in the language they use, the structures they contain, and the goals they pursue. Numerous authors have advocated the specificity of language and the need for contextualizing linguistic explorations, especially if pedagogical implications are to be derived from research findings.

In the particular case of academic writing, there is a widespread agreement that specificity is key for both teaching and learning, and that corpus methods can certainly contribute to providing a more accurate and specific picture of actual language in use.

Unfortunately, the degree to which findings and pedagogical implications that emerge from corpus studies are applied later on in the classroom or used by language material developers is still relatively low (Gilquin et al., 2007; Paltridge, 2002; Römer, 2011; Springer, 2012). Furthermore, those materials that do include corpus-based evidence tend to rely solely on native corpora (Gilquin et al., 2007). In order to know, however, what items may be causing difficulties for learners, i.e. non-native speakers of a language, learner data should be explored. The present doctoral thesis aims to conduct quantitative and qualitative analyses of the academic writing produced by EFL learners at university. Three different text types that represent common assignments students face at some point during their bachelor studies, namely a classroom writing activity, a subject's final assignment, and a degree's final dissertation, have been collected and converted into different corpora for analysis. Each of these text types represent a genre family which, following the classification of university student writing developed by Gardner and Nesi (2013), display three different educational purposes and structures: the first type corresponds to the genre 'problem question' which aims "to provide practice in applying specific methods in response to simulated professional problems problem (…), application of relevant arguments or presentation of possible solution(s) in response to scenario" (2013, p. 36); The second type belongs to 'essays' which, according to the authors "demonstrate/develop the ability to construct a coherent argument and employ critical thinking skills" (2013, p. 35); and finally, the third type can be classified as 'research reports', which seek "to demonstrate/develop ability to undertake a complete piece of research including research design, and an appreciation of its significance in the field" (2013, p. 36). In addition, four specific phenomena that have been described as important linguistic devices for the successful development of academic writing in learner corpus literature were tracked: (1) general and discipline-specific academic terminology (cf. Coxhead, 2017; Durrant, 2016; Granger, 2017a), (2) linking adverbials (adversative in particular) (cf. Granger & Tyson, 1996; Liu, 2008; Rica-Peromingo, 2012) (3) metadiscourse (reflexive in particular) (cf. Ädel, 2006, 2016; Hyland, 2010; Mauranen, 2010) and (4) lexical bundles (cf. Biber et al., 1999; 2004; Biber & Barbieri, 2007; Hyland, 2008a).

The first phenomenon, academic terminology, is a particularly important aspect of academic writing. There are numerous words, collocations and phrases that can be categorized as 'academic' and that enhance the sophistication of a text. Some can be found across disciplines (e.g. *abstract, on the other hand, preliminary results*, etc.);

there are other, more technical words and expressions, which, in contrast, can only be found in specific disciplines (e.g. *perform an extraction, chief complaint, language attrition,* etc.). The former set is known as 'general' academic vocabulary while the latter is referred to as 'discipline-specific' or 'technical' vocabulary (Granger, 2017a). The effectiveness of teaching and learning academic writing focusing on one or the other is still a matter of debate. There have been numerous efforts to unify and describe general academic vocabulary in a way that could be useful to novice and non-native writers. For example, corpus-informed lists that comprise words, collocations and formulas that are shared by a wide range of academic genres have recently been developed (e.g. Ackermann & Chen, 2013; Gardner & Davies, 2014; Simpson-Vlach & Ellis, 2010). Some studies, on the other hand, have added more weight to the disciplinary specificity of academic genres and claim that the actual presence of these 'general' items in specific disciplines is relatively low (Granger, 2017a; Hyland, 2008; Hyland & Tse, 2007); they indicate that a pedagogical approach based exclusively on the former would thus be less effective than teaching and learning academic terminology that is contextualized in a particular discipline. Regardless of the approach, both general and specific academic vocabulary production can be challenging to EFL learners writing in their disciplines. Furthermore, relatively few studies have looked into the actual use and development of academic terminology (both general and specific) in learner corpora.

The second area of study deals with the use of linking adverbials (e.g. *on the other hand, alternatively, moreover*). These elements play an important role in creating discourse coherence and cohesion in academic writing. Despite the fact that most EFL courses deal with these connecting devices from very early stages, non-native learners often struggle to use them appropriately. Quite a number of studies that explore the use of connectors in academic writing have found that, compared to native or expert writers, learners may use linking adverbials in a completely different manner, e.g. in terms of types, placement and frequency (Biber et al., 1999, 2004; Granger & Tyson, 1996; Lei, 2012; Rica-Peromingo, 2012; Swales, 2002). In fact, linking adverbials that belong to the adversative category (e.g. *despite, nevertheless, in contrast*) have been found to pose the greatest challenge to learners; they are, in addition, one of the most common types of connectors in argumentative writing. Items in the adversative category can have different degrees of contrasting power (e.g. concessive –*yet*, corrective -*rather*, contrastive –*in fact*) (Liu, 2008). This and the fact that they can take different positions within a sentence (e.g. *yet* in sentence initial or medial position, preceded by a conjunction, performs a contrastive function; it would, on the other hand, turn into a time adverbial when placed, without a comma, in sentence final position) can as well pose a challenge to non-experienced or non-native writers. Texts written by learners who are not aware of these characteristics often show overuse, underuse and even misuse of adversative linking adverbials.

The third area of study is metadiscourse. Metadiscourse in academic writing refers to those linguistic markers that help writers to address two main entities: (1) the text that evolves, and (2) the reader and/or the author of the text (Ädel, 2006; Hyland, 2017). Metadiscourse differs from the ideational or propositional content of a text in that it does not add new information, but it is vital for the content to be understood. There are a wide range of devices that can qualify as metadiscourse, but they are usually grouped into two macro categories: markers that refer to the text, i.e. textual metadiscourse (e.g. *in figure 1, secondly, as mentioned previously*), and markers that refer to the writer or the reader of the text, i.e. interpersonal metadiscourse (e.g. as *I* said, *see* appendix 1, *you* may question). Academic texts that make a proper (i.e. according to specific language, culture or genre conventions) use of these types of metadiscourse markers are often found to be more comprehensible and reader-friendly. Given the fact that clear structures, careful framing of arguments and constant guidance for readers are common practices in academic texts written in English, a proper use and understanding of metadiscursive markers is of considerable importance, especially when writing long academic texts (e.g. bachelor dissertations). An additional difficulty is the fact that these metadiscursive practices can be highly discipline-specific, which means that the types and the extent to which metadiscourse markers appear in a text are dictated by the discipline and the specific genre of that text (Hyland, 2000, 2005, 2012). EFL learners are not always aware of these particularities and often fail to use metadiscursive devices appropriately.

Lexical bundles are the fourth linguistic device explored. A lexical bundle is a recurrent word combination, which can have different lengths (e.g. three, four, five words), different structures (e.g. noun-phrase, verbal-phrase, prepositional-phrase based) and perform different functions (e.g. *on the other hand* is a 4-word bundle that performs a text-orienting/transitional function) (Biber et al., 1999). Lexical bundles' quantity and diversity abound in language. Efforts have been made to look at lexical bundles from different mode, register and disciplinary perspectives (Ädel & Erman, 2012; Biber & Barbieri, 2007). Research has shown that, unfortunately, there seems to be no single pool of lexical bundles one can employ generally; quite the contrary: each mode, register, and discipline tends to use, with more or less frequency, a group of bundles for their very particular purposes. While it is true that certain lexical bundles can be found across modes, registers and disciplines, other bundles are more specific and, in order to demonstrate membership in a given community, one needs to be certain about which, how and when to use these bundles. EFL learners do not generally pick up lexical bundles from mere exposure and problems such as underuse, overuse and misuse of lexical bundles when compared to the use of these devices in native or expert writing have been described in the literature (Ädel & Erman 2012; Chen & Baker, 2010; Liu, 2012; Meunier & Granger, 2008).

In the present doctoral thesis, these four problematic areas have been explored in various academic texts written by learners at university. These represent actual L2 learner writing practices across bachelor degrees and disciplinary communities, which are linked to the teaching experiences of the author of the thesis: the teaching of different English academic writing- and research-related subjects in different educational settings and to different learner populations motivated the author to apply corpus methods and tools in order to investigate particular areas of academic writing that most learners seemed to struggle with, and for which there seemed to be neither clear consensus nor evidence-based pedagogical advice in academic writing materials. Learners are EFL writers, with different L1s (mainly European) in the first two studies; in the last two studies, on the other hand, learners were specifically Spanish L1 writers. The exploration of corpora that come from students with different linguistic backgrounds has helped the author to analyse learner features from different perspectives, i.e. academic writing produced in an international classroom setting on the one hand, and academic writing produced by one specific L1 population on the other hand. The participants were university students in their first year of studies (in studies one and two), and in their last year of studies (in studies three and four). Since writing complexity increases from one text to the other in the four studies performed (e.g. from a short writing task to a bachelor dissertation), the possibilities of analysing more complex linguistic phenomena increased accordingly (e.g. academic vocabulary is tracked in study one while metadiscursive expressions are annotated and explored in study three). In addition, texts were produced in four different bachelor degrees, namely dentistry, European studies, medicine and linguistics. Working with texts of different lengths, disciplines, and purposes has led the author of the present thesis to explore different linguistic paradigms, namely academic vocabulary in short writing tasks written by students with different L1s; the use of adversative linking adverbials by EFL learners in argumentative essays that were the final assignment of a content subject; metadiscursive devices in bachelor dissertations written in English by Spanish L1 speakers; and finally, lexical bundles in the introduction and conclusion section of these dissertations. These studies reflect both corpus-based, i.e. top-down (e.g. use of pre-defined lists of linguistic structures or validated taxonomies to confirm the researcher's hypothesis using corpus data) and corpus-driven, i.e. bottom-up (e.g. no pre-conceived ideas; exploration of items and patterns actually present in the corpus) methods (cf. Callies et al., 2014).

As we have seen, the literature advocates the need for approaching academic writing pedagogy from corpus- and discipline-based perspectives. Hence, the objective of this thesis is twofold: it seeks to explore non-native learners' performances regarding the use of four different linguistic phenomena in their academic texts written in English. Second, and after considering the need for more empirical research into second language academic writing and its pedagogical implications, this thesis aims to provide pedagogical advice on the use of these devices, after the comparison with different reference corpora. In this regard, the reference corpora used in the studies come from both large and widely known general or academic corpora, and more specialized, self-compiled corpora. These specialized corpora in particular come from English native (university students' texts) or expert (published research articles) writers. Pedagogical implications are drawn from each study, aiming to assist novice and non-native learners improve their academic writing skills regarding the use of academic vocabulary, adversative linking adverbials, metadiscourse, and lexical bundles. Furthermore, it is hoped that the findings that emerge from this thesis can be of interest to academic writing instructors and material developers, as they provide contextualized, corpus-informed and discipline-specific analyses of learner academic writing production. Finally, and in order to comply with the requirements set for compilation thesis of the doctoral program in English Linguistics at the Complutense University of Madrid, all four studies included in the present thesis have been published or accepted for publication as full original research papers that report on completed leaner corpus-based research in indexed, double-blind peer-reviewed, English-medium academic journals (see Appendix 10 for the letters of acceptance from editors).

This doctoral thesis has been structured as follows. First, Chapter 2 offers an overview of the literature that is relevant for the object of study. This chapter starts by providing defining features of academic writing, particularly in HE and EFL contexts; it continues by describing the linguistic phenomena explored in the thesis, namely academic vocabulary, linking adverbials, reflexive metadiscourse and lexical bundles, and provides their theoretical foundation. The chapter concludes by discussing the notions of disciplinary literacy and learner corpus research, and presents previous studies on these topics. Next, in Chapter 3, the overarching question established for the present thesis and the research questions and main hypotheses formulated for each study are introduced. Chapter 4 presents the methodology and the different approaches that the author has followed in order to carry out the four studies by describing the learner corpora collected, the reference corpora compiled, the software and text-analysis tools used, and the types of quantitative and qualitative analyses performed. The chapters that follow present the four studies that were conducted for this doctoral thesis.

Chapter 5 presents the first study, titled 'The effects of a content-based language course on students' academic vocabulary production', which deals with the use of academic vocabulary (words, collocations and formulas) in texts written by first-year university students enrolled in two different instruction settings (EMI and L1); texts were collected before and after a content-based language course. These texts are contrasted with widely used lists of general academic terminology and with a self-compiled list of technical terminology extracted from the class materials. Differences in the use of general and technical vocabulary according to the time of task and students' setting of instruction are explored. This study has been published in the *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education*, volume 2 (2), pages 25-42.

The second study is presented in Chapter 6. It has been titled 'Linking or delinking of ideas? The use of adversative linking adverbials by advanced EFL learners' and it explores the use of adversative linking adverbials in a learner corpus of argumentative essays written by advanced EFL learners with five different L1s, in their first year of studies. The usage patterns found in the learner corpora are compared with the usage patterns in an English-native corpus of argumentative essays written, as well, by university students. The main differences in terms of frequency, types, and position of adversative linking adverbials are discussed. This study has been accepted for publication in the journal *Revista Española de Lingüística Aplicada* (RESLA).

Chapter 7 presents the third study, titled 'Reflexive Metadiscourse in a corpus of Spanish bachelor dissertations in EFL'. It focuses on the use of reflexive metadiscourse, which is manually annotated in a learner corpus of bachelor dissertations written in English by Spanish L1 students in their last year of studies, in linguistics and medicine. An annotated expert corpus of research articles in the same disciplines is used for comparisons. Frequencies and the specific functions of textual and interpersonal markers are explored in order to uncover genre and disciplinary differences. This study has been published in the journal *Research in Corpus Linguistics* (RICL), volume 6, pages 29-49.

The fourth study, titled 'Lexical bundles in learner and expert academic writing' is presented in Chapter 8. In this last study, lexical bundles are extracted from the same learner and expert corpus of bachelor dissertations and research articles respectively, in order to compare their most frequent lengths, types, structures and functions. L2 learner writing features and preferences according to the academic communities explored are discussed. This article is published in the *Bellaterra Journal of Teaching & Learning Language and Literature* (BJTLLL), volume 12 (1), pages 65-90.

Following these four studies, Chapter 9 summarizes the main findings and addresses the research questions formulated. It also discusses the main implications and the contributions made to the fields of second language writing and corpus linguistics. Chapter 10 concludes the thesis and offers directions for future research.

2. Literature review

This chapter begins with an introduction to academic writing, on the one hand, and academic writing in EFL at university more specifically, on the other hand (Sections 2.1 and 2.2), in which the concepts of academic terminology, linking adverbials, metadiscourse and lexical bundles are defined and described. Section 2.3 introduces the concept of disciplinary literacy. After this, Section 2.4 presents and revises previous literature on corpus linguistics, and zooms in on learner corpus research to describe some of the findings that emerge from these studies.

2.1 Academic writing

Compared with other genres, academic discourses "embody the social negotiations of disciplinary inquiry, revealing how knowledge is constructed, negotiated and made persuasive" (Hyland, 2004, p. 3). In the form of writing, when a text "anticipates the knowledge that its readers will bring to it, the questions they will implicitly ask, and tailors its content and form accordingly" (Clippinger & McDonald, 1983, p. 730), it enhances its persuasive style and models *good* academic writing. The presence of certain 'discourse clues' such as particular word combinations, connectives that signal text structure, or items that indicate text purpose, seem to mark a text as 'academic'. Academic writing can, however, take alternative forms in different academic genres (e.g. essays, articles, reports, evaluation rubrics, etc.), which can have unique conventions in terms of length, structure and communicative purpose. But not only different text forms imply different linguistic practices; diverse academic disciplines, apart from dealing with different topics, also have their own way of presenting information. As Hyland (2004, p. 3) observes:

> Scholarly discourse is not uniform and monolithic, differentiated merely by specialist topics and vocabularies. It is an outcome of a multitude of practices and strategies, where what counts as convincing arguments and appropriate tone is carefully managed for a particular audience.

For novice writers, the fact that scholarly or academic discourse (and academic writing in particular) follows different conventions regarding e.g. vocabulary use, text structure, or rhetorical moves, and that this set of norms varies according to specific genres and disciplines, can pose a considerable challenge. Also creating difficulties for novice writers is the fact that "words take on additional meanings as a result of their regular co-occurrence with other items" (Hyland & Tse, 2007, p. 246); hence, there is no "pool of semantically equivalent candidates" (ibid, p. 246) writers can draw on for their linguistic choices. Academic writing conventions considerably differ from other ways of speaking or writing, and require users to adapt to these norms if they want to *fit in* or become accepted *insiders* in a given academic community; this can be particularly difficult for writers operating in a language that is not their own. One of the fundamental goals of academic discourse is that of creating and disseminating knowledge in order to contribute to scientific progress in our society. Learning how to navigate and produce scientific knowledge seems therefore a basic academic skill for students in higher education, regardless of their field of studies.

The emergence of English as a *lingua franca*, and the internationalization of many of today's universities, has had consequences not only at a pedagogical level, but also in academic publishing, in which the number of English-medium publications is steadily increasing, and represent almost 90% of the journal literature in some disciplines (Thompson Corp., 2019). Knowledge of academic English is thus fundamental for current researchers and university students alike in order to be able to understand and share scientific findings in ever-growing international contexts. In Spain for example, out of the 1.3 million students enrolled in bachelor degree programmes (both public and private institutions) in the academic year 2017-2018, almost five percent (i.e. 63,266 students) were international students, mostly from other European countries; the percentage of international students is even higher in Master degrees (20.8%) and in PhD programmes, in which one in four students is international (Ministerio de Ciencia, Educación y Universidades, 2019). In many of these institutions, EAP courses are offered in order to help students develop their academic writing skills in English.

Quite a number of studies have highlighted, however, some limitations found in EAP pedagogy (Airey, 2018; Flowerdew, 2002; Gilquin, Granger, & Paquot, 2007; Hyland, 2009a; Römer, 2011; Springer, 2012). There are three recurrent issues: first, the fact that EAP courses tend to offer a general view of academic writing, following a 'one-size-fits-all' approach which often fails to notice differences between academic genres and disciplines. In addition, possible transfer issues of particular L1 populations are not usually dealt with in the classroom. Second, and also related to the first limitation is the fact that EAP courses are commonly taught by language experts, "who may or, what is more likely, may not be experts in the specific discourse they teach"

(Römer, 2011, p. 209). This can be problematic since academic writing, as previous literature in the field has shown, is far from being a "unitary mass" (Hyland & Tse, 2007, p. 247), and thus, linguistic practices and strategies should be taught and learnt in agreement with the specific discipline one is dealing with. Finally, it has also been noted that only few materials designed to help students improve their writing skills used in EAP courses are based on evidence that considers actual language use –i.e. corpus-based (see e.g. Barlow & Burdine, 2006, Biber et al., 1999; Tono, 2011). In addition, these corpus-informed materials mostly rely on English native speaker or expert performances only (e.g. published articles), and do not consider learner data. As we will see in Section 2.4, learner corpora, i.e. naturally occurring language produced by non-native speakers of a language, can be highly relevant for EAP and L2 teaching in general, as it enables researchers, instructors, and/or material developers to identify those aspects of a language that are more difficult for L2 students (e.g. Nesselhauf, 2005). In the following section, an overview of studies that explore academic texts produced by native, expert or learner writers is given in order to illustrate the benefits writing research and corpus-based data in particular hold for the teaching and learning of academic writing.

2.1.1 Writing research typology

A relatively modest amount of writing research has been conducted into the nature of text-based studies that take the final product as an object of study –although some ethnographic studies have also looked into writing as a process, for example, by using different versions or drafts of the same text (Lee et al., 2015), or comparing the feedback received by peers or teachers with the text's final version (Anson & Anson, 2017; O'Sullivan, & Chambers, 2006). Polio (2001) proposes a taxonomy that groups writing research into nine different categories.[3] The corpus works that are relevant for the current thesis are comprised under the last four categories:

(1) Studies on lexical features: these studies normally look at how varied or rich a text's vocabulary is. In order to measure lexical variety or richness, type/token ratios are used. Also, the *keyness* index or the *keyword* function many text-analysis software offer can help to see words that are unusually salient and distinctive of a particular text type. Lu (2012) has developed a very useful and freely accessible web-based tool called

[3] The nine categories mentioned in Polio's (2001) are: (1) revision, (2) content, (3) mechanics, (4) overall quality, (5) linguistic accuracy, (6) lexical features, (7) complexity, (8) coherence and discourse features, and (9) fluency.

Lexical Complexity Analyzer (LCA)[4], which allows for lexical comparisons of texts (using both single or batch modes) and looks at different indices such as lexical sophistication (measured in terms of mean length of words, and/or number of academic words), lexical variation (number of different words or type/token ratio), lexical diversity (measured in terms of lexical word variation) and lexical density (operationalized in terms of lexical words –nouns, verbs, adjectives and adverbs– and compared to total tokens). There are numerous studies on lexical sophistication (e.g. Hyland & Tse, 2007; Hyland, 2008; Lee & Chen, 2009; Lu, 2012; Sugiura et al., 2007, among others) that compare learner competence to a corpus that represents English native writing. This type of analysis is, however, not without its problems. For example, some researchers have cautioned that findings on lexical sophistication depend on text length: the longer the text, the more likely it is that words are repeated (Meunier, 1998, p. 32). As a result, recent learner corpus studies include more than one lexical variable, use normalized values per 1,000 words, or divide the total number of types by the square root of the total number of tokens (also known as *Guiraud's index*). Another caveat concerning lexical features is that analytical measures do not always reflect all the complexity of language: measures of sophistication for example are often based on single-word analyses, and this vision does not account for Sinclair's idiom principle according to which: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (1991, p. 110). Language is therefore made up of word combinations, and the fact that words acquire meanings from their context calls for a more inclusive form of analysis when looking at lexical features such as lexical complexity or lexical sophistication; moving beyond the analysis of isolated words has led researchers to discover certain aspects of language development that are situated in the phraseological dimension.

Studies on phraseology have gained traction in the last decade (e.g. Hyland, 2012a; Liu, 2008; Paquot, 2017; Rica-Peromingo, 2012) and have brought to the fore certain features of learner writing that are intrinsically related to the L2 teaching and Second Language Acquisition (SLA) paradigms. For example, in her corpus-based study, Paquot (2017) compares phraseological complexity measures with traditional measures of complexity (i.e. lexical and syntactic complexity) in order to see if the

[4] Accessible from: http://aihaiyang.com/software/

former could be used to describe L2 performance in a similar or a more accurate manner. She found that phraseological analysis did help to distinguish between proficiency levels in the learner corpus, while no measure of lexical complexity did. Phraseological analyses account for context and this undeniably provides a more natural and broader picture of the language use. In the present thesis, study one deals with the production of academic words, collocations and formulas in academic texts, and study four examines the use of lexical bundles (of three, four and five words) in L2 learner academic writing. These studies could therefore be included in the typology of writing research that focuses on lexical and phraseological features of language.

(2) Studies on complexity: this type of writing research typology contemplates different measures, such as mean length of sentence, mean length of t-unit, number of clauses per sentence, or the proportion of sophisticated (i.e. academic or technical) words. As mentioned above, Lu's (2010) lexical and syntactic complexity analyzers[5] (LCA and L2SCA respectively) can help to automate the analysis of written samples of the English language, using different measures proposed in first and second language acquisition literature. There are several studies that have used Lu's L2SCA and LCA to explore learners' syntactic and lexical complexity (e.g. Ai & Lu, 2013; Lu, 2010, 2011; Lu & Ai, 2015, Paquot 2017, Polio & Yoon, 2017). There are, however, other ways of looking at writing complexity. Demol and Hadermann (2008) for example provide a fine-grained analysis of the degree of packaging, integration and ellipsis of texts written by French and Dutch learners in their L1 and L2, and confirm that "the use of syntactically more complex structures means that the tasks of planning, organizing and abstracting become increasingly important and represent a considerable cognitive burden for the learner" (2008, p. 258). Another study that is relevant in this category is Hannay and Martínez-Caro's (2008) study on Spanish and Dutch learners' construction of the theme zone, i.e. "complex of constituents up to and including the subject and its non-restrictive postmodifiers in the first declarative main clause of the sentence" (2008, p. 228). Thematic material can serve as a backgrounding device, as well as a foreteller of the message to come. The texts written in English by Dutch learners in their study showed a more accurate exploitation of the theme zone than their Spanish counterparts. When the theme zone is exploited correctly and it conveys ideas in a reader-friendly manner, students' texts exhibit a more complex and proficient degree of both linguistic

[5] Accessible from: http://www.personal.psu.edu/xxl13/downloads/l2sca.html

and discourse competence. Another way of measuring writing complexity is by looking at the proportion of sophisticated terminology (i.e. technical or academic words) in a text in comparison with its non-academic (or more general) vocabulary (see e.g. Coxhead & Nation, 2001; Coxhead, 2017; Laufer, & Nation, 1995). In the present thesis, study one analyses the use of academic terminology in texts written by EFL learners, before and after a content-based language course. This study could also be regarded as a study of language complexity.

(3) Studies on coherence and organization: these can involve the analysis of metadiscursive features such as textual markers, or as defined by Intaraprawat and Stefferson, "those facets of a text which make the organization of the text explicit" (1995, p. 253). Also within this typology, there are numerous studies on the production of linking adverbials in academic writing by learners' with different L1s (e.g. Granger, 1996; Leńko-Szymańska, 2008; Liu, 2008; Prommas & Sinwongsuwat, 2011; Rica-Peromingo, 2012). Research has shown that learners often rely on a small range of these items, which tend to be overused when compared to native or more advanced peers. These widely used words have been called 'lexical teddy bears' (Hasselgren, 1994) in the SLA literature and their high frequency is often triggered by the usually decontextualized emphasis given to certain devices, such as connectives, in ESL classrooms. In order to avoid this 'teaching effect', Granger suggests that "learners should not be presented with lists of 'interchangeable' connectors but instead taught the semantic, stylistic and syntactic behaviour of individual connectors, using authentic texts" (2004, p. 135). Studies two and three included in the present thesis, which deal with the use of linking adverbials and reflexive metadiscourse respectively, can be regarded as studies of coherence and organization.

(4) Studies on fluency: although more characteristic of spoken than written language, text-based studies on fluency involve an evaluation of how native-like or reader-friendly a text is; from a CAF perspective (i.e. complexity, accuracy and fluency) (Wolfe-Quintero, Inagaki, & Kim, 1998, p. 14), fluency is regarded as the ability to write more words and structures in less time, and is thus measured by looking at the speed at which a text is produced (e.g. number of words per minute); although fluency is still difficult to define (see Springer, 2012, p. 23 for a detailed discussion on this matter), an attempt is Callies' (2008) work on Advanced Learner Variety (ALV). He alludes to five typical features of learner writing that can contribute to foreign-soundingness, i.e. the perception of a text as being written by a non-fluent

speaker/writer, namely: (1) the overuse of high frequency vocabulary (i.e. ‘teddy bears’, safe formulae), (2) the overuse of a limited number of prefabs (e.g. certain academic formulas such as *on the one hand* and *on the other hand*), (3) a higher degree of personal involvement (e.g. *I think, in my opinion*), (4) stylistic deficiencies such as an overly spoken style (e.g. the use of phrasal verbs in academic writing) or the mixture of formal and informal marks in the same text (e.g. use of contracted forms), and (5) the use of different discourse structures to organize information (e.g. unusual theme zones, misplaced connectors, etc.).

Given the complexity of the fluency parameter in L2 writing, and the different writing conventions across genres and disciplinary communities, studies that look at fluency call for a high degree of contextualization. In other words, in order to see if L2 writers are fluent, a comparison with the desired target is necessary. For example, to measure fluency in university students’ writings about the use of *stem cells*, they could be contrasted with texts written by more advanced students (e.g. Master or PhD students) or experts (e.g. published articles) in the same discipline, in order to analyse the type of high-frequency vocabulary, collocations, degree of personal involvement, or the use of structuring and organizing devices that are typical of that particular genre and discipline. Repetition of certain formulas, or an infrequent author involvement, which can at first strike the linguist investigator as foreign-sounding practices or denote a lack of fluency, could be a perfectly accepted practice in a specific discipline or genre; we will explore this notion of disciplinary literacy more in depth in Section 2.3. The four studies that were carried out for the present doctoral thesis have looked at fluency in different ways: studying the use of general and discipline specific academic terminology (including the comparison with academic vocabulary lists, and the same tasks written by native speakers) in study one; analysing the use of connective devices, such as adversative linking adverbials in argumentative writing and comparing it to the production of these devices by native student writers of a similar text type in study two; exploring metadiscourse in learners’ bachelor dissertations and comparing it to articles published in the same disciplines in study three; and finally, by extracting lexical bundles from the introduction and conclusion sections of learners’ texts and contrasting their types, structures and functions with the bundles found in analogous experts’ texts in study four.

2.2 EFL Academic writing at university

As an increasing number of professions require high levels of qualifications, more and more students enter and complete HE programmes. In the case of Spain, this trend continues to increase. In fact, statistics show that 32.3% of the total population (aged 18-24) was enrolled in different HE programmes in the academic year 2017-2018 (Ministerio de Ciencia, Educación y Universidades, 2019). Due to globalization, and to internationalization programmes of most universities, university students represent, more than ever, a mixture of different linguistic, cultural and social backgrounds. As a consequence, EMI has emerged as a common pedagogical practice in many universities, with the intention to cater for both the increasing linguistic necessities of a globalized society, and the multicultural profile of today's classrooms (Pérez-Vidal et al., 2018). In order to navigate academic learning and be able to contribute to the current scientific progress, communication is key. As Hyland points out, "complex social activities like educating students, demonstrating learning, disseminating ideas and constructing knowledge, rely on language to accomplish" (Hyland 2009, p. 1). As we have seen, academic communication in English is, however, not a uniform and invariable entity that can be taught and learnt straightforwardly. As languages develop, so does academic discourse, and this forces instructors and students alike to adapt quickly to the literacy demands of their discipline of studies; this is particularly difficult for students who are users of English as a foreign or second language. In addition, although there are different amounts of EMI exposure students can receive (e.g. full immersion, semi-immersion, etc.), one recurrent concern regarding EMI pedagogy is that EMI instructors may not feel comfortable dealing with language issues or correcting students' linguistic mistakes in their content classes (Airey, 2011; Ha & Hyland, 2017); as a consequence, there may be a lack of emphasis on certain linguistic phenomena that are important for the development and communication of content knowledge in English in their disciplines.

As previous research has pointed out, knowledge of academic discourse is necessary for students' successful educational performance (Airey, Lauridsen, Räsänen, Salö, & Schwach, 2017; Csomay & Prades, 2018). Prioritizing words and expressions that learners need later on in their subjects can nevertheless be a key teaching problem. Using a corpus that contains specific discourse types might help L2 learners to "find it easier to develop both their receptive and productive skills when confronted with the most common lexical items of a language" (Römer 2011, p. 208). In addition, scholars like Biber et al. (1999), Granger (2017a), Hyland (2008a), Hunston (2002), and Römer (2010) have recurrently emphasized the phraseological nature of language. Teaching and learning the most frequent word combinations in a given discipline could foster fluency, accuracy and idiomaticity (Römer, 2011), and thus providing novice and L2 learners with an integrated view of phraseological items used in context seems therefore reasonable.

L2 learners' production can show developmental, pedagogically-induced, or transfer-related issues (Granger, 2004). As we will see in Section 2.4, analysing learners' output in comparison with native-speaker or expert data can help to reveal characteristics of learner language. As research on second language writing has shown, certain writing conventions can also differ considerably in two different languages. For example, academic writing in Spanish has been described as a slightly writer-oriented discourse (Hinds, 1987), in which it is often the reader who has to make an effort to understand the writer's purpose and structure of a text. In academic writing in English, in contrast, structures, purposes and contents have to be clearly framed –always anticipating reader's needs– in order to be persuasive (Hyland, 2008a). L2 writers not only need to adapt to the conventions of their own discipline –which, as we will see in the next section, vary from discipline to discipline (Hyland, 2008; Hyland & Tse, 2007); they are also expected to acquire a new set of strategies in order to develop their L2 academic writing skills, which is often in addition to developing academic writing skills in their own L1 (Airey, 2018). And, "[b]ecause learners, as a rule, have a limited repertoire of expressions at their disposal to fulfil a particular rhetorical function, they tend to rely on a few items only, which they use over and over again, to the detriment of other, perhaps salient expressions" (Gilquin et al., 2007, p. 16).

Foreign-sounding writing (e.g. overuse of high frequency vocabulary, underuse of elaboration, etc. [e.g. Callies, 2008; Springer, 2012]) is frequently found in learner academic texts. In fact, some of the most common problems reported on in the literature are recurrently situated in the phraseological dimension (Granger, 2017a). Learner corpora analyses have helped to determine whether these characteristics are due to L1 transfer, different developmental stages, or instead, teacher-induced (Granger, 2004). In the following paragraphs, a description of the linguistic features analysed in the four studies included in this thesis, namely academic terminology, adversative linking adverbials, reflexive metadiscourse, and lexical bundles, is provided and complemented with relevant findings that emerge from corpus-based literature.

2.2.1 Academic words, collocations, and formulas

In academic writing, academic terminology and expressions are a very important aspect of the knowledge construction process. Concepts such as *research, objective* and *methodology* can indicate procedures and structures that make up scientific writing and can characterise a text as ‘academic’. These words can be found in many academic texts, regardless of their particular disciplines (e.g. biology, history, nursing). However, not only general or ‘interdisciplinary’ vocabulary is important for writing academically; technical or discipline-specific words also play an important role (cf. Granger, 2017b). For example, the term *scaffolding,* which may seem a fairly common term, has completely different meanings and significance in particular academic communities such as education and architecture (i.e. in education, *scaffolding* refers to an instructional method in which teachers gradually reduce assistance, so that learners can develop their autonomy; in architecture, *scaffolding* refers to a temporary structure used on the outside of a building under construction). These keywords or technical words are used by specialists in the field, and their presence in academic texts can denote authors’ membership and level of expertise. The use of academic words (both general and discipline-specific) can help writers to situate their work in the academic realm and particularly within their specific disciplinary domains. But academic language is indeed highly patterned (Römer, 2010), and so far we have referred to ‘single’ or ‘isolated’ words only. Academic writing contains sequences of recurrent word combinations such as collocations (e.g. *basic assumption, careful analysis*) and formulas (e.g. *in other words, at the same time*) that can be general but also discipline-specific (e.g. *risk factor* and *the prevalence of* can be commonly found in medical writings). These can be difficult for learners, even at advanced levels, as shown by Nesselhauf (2005) who uncovered idiosyncratic collocation uses in a learner corpus, and examined the relationship between EFL pedagogy and the use of linguistic corpora.

Novice writers who need to develop academic writing skills in a foreign language, but that also have to learn the disciplinary conventions of their own academic community in order to succeed academically, may find vocabulary learning an arduous and daunting task. Recently, three vocabulary lists that contain frequent words, collocations and formulas shared across disciplines have been developed, namely 1) the Academic Vocabulary list (AVL) (Gardner & Davies, 2014), 2) the Academic Collocation List (ACL) (Ackermann & Chen, 2013), and 3) the Academic Formulas List (AFL) (Simpson-Vlach & Ellis, 2010). These lists can help learners to acquire general academic vocabulary and also help instructors to prioritize the most pedagogically relevant items. In order to learn and teach discipline-specific vocabulary, on the other hand, the compilation of an expert corpus of specific readings and research articles in the discipline studied is recommended, to later explore and/or extract frequent vocabulary (keywords), collocations and formulas that are present in that corpus. In study one, both types of academic vocabulary are tracked in a learner corpus to see if students improve their lexical and phraseological sophistication after a course that provided instruction on both types of vocabulary.

2.2.2 Adversative linking adverbials

One specific type of linguistic device that is particularly common in academic writing and that helps to create discourse cohesion is linking adverbials (henceforth LAs). LAs such as *moreover*, *in conclusion*, and *on the other hand* indicate the relationships between different pieces of information (i.e. addition, summation, and contrast respectively). The latter category, contrast, also called 'adversative' (Biber et al., 1999; Liu, 2008), abound in argumentative writing (Granger & Tyson, 1996). In academic writing in English, building persuasive arguments by contrasting ideas, exposing pros and cons of a given topic, or presenting opposing views to a particular notion, is a common practice (Pérez-Llantada, 2011) for which the use of adversative LAs is necessary.

Not all LAs that belong to the adversative category mark the same kind of contrastive relationship, however. Liu (2008), drawing on previous classification frameworks, such as Biber et al.'s (1999) and Celce-Murcia and Larsen-Freeman's (1999), provides a comprehensive list of adversative LAs, and differentiates four broad categories, namely concessive, contrastive, correction and dismissal, that help to classify adversative LAs according to their contrastive power and rhetorical function more accurately. Inexperienced writers have been shown to overuse a small group of these LAs (e.g. *however, nevertheless, despite*), and this is often due to the emphasis given to high frequency LAs in academic writing instruction or in general language courses (Granger & Tyson, 1996). Study two deals with the use of adversative LAs in argumentative learner writing and uncovers certain learner writing features that are worthy of pedagogical attention.

2.2.3 Reflexive metadiscourse

Another broad linguistic category that helps academic writers to organize the ideational content of their texts and to present it coherently and persuasively to readers is what has been called metadiscourse (MD). MD often differentiates between two main spheres: textual and interpersonal MD. The first category includes those expressions or markers (they can be words, collocations or formulas) that refer to the evolving text, such as endophoric markers, code glosses, or linking devices, each comprising different subcategories in which markers perform important discourse functions such as adding, contrasting, enumerating, reformulating, etc. The second category, i.e. interpersonal, includes markers that refer to the author/s of the text (self-mention) or the readers (directives) (Toumi, 2009). Some authors also include stance markers such as hedges and boosters in the second category (Hyland, 2017). Ädel (2006) and Mauranen (1993) on the contrary, focus on 'reflexive' markers that reflect processes that take place in the evolving text only (as opposed to showing author's stance, as this would reflect experiences and opinions from the real world), and see stance markers as a different category, outside of MD. In the present thesis, the reflexive perspective of MD is studied.

The presence of textual and interpersonal markers, and their different categories and subcategories can be as well highly culture-, genre- and discipline-specific. Research has shown that writers of medical articles do not use textual markers to the same extent as writers of linguistic articles, for instance (Hyland, 2005). The nature of the topics presented (i.e. language vs. medicine) and the different conventions of each disciplinary community (e.g. emphasis on crafting persuasive arguments, or on showing objective research procedures) can trigger different metadiscursive choices. Once again, understanding MD practices in both a foreign language *and* in a new discipline can be challenging for L2 writers. Study three explores reflexive MD in the academic writing of L2 learners in their last year of studies in two different disciplines, and compares it to an expert corpus of published research articles. The differences found between genres and disciplines can provide pedagogical aid to both L2 learners and instructors on the use of MD practices in academia.

2.2.4 Lexical bundles

As can be seen, recurring sequences of word combinations are the chunks or puzzle pieces that put academic discourse together. Another way of looking at academic formulas is by performing a corpus-driven approach: e.g. extracting repeated sequences of 3, 4, or 5 words from particular texts in order to identify the most common formulas in that specific genre or discipline. These formulas are also known as lexical bundles (henceforth LBs) (Biber et al., 1999) and encompass expressions that perform different discourse functions, and that are often part of academic vocabulary, linking adverbials, and metadiscourse markers lists.

LBs are normally identified by setting minimum frequency and range cut-offs (e.g. bundles that appear at least 3 times in 5 texts) in order to avoid including features of individual writers. Often, LB studies eliminate bundles that refer to the topic being discussed (e.g. *second language acquisition*), and also merge overlapping bundles (e.g. *due to the fact, due to the fact that*) so as to prevent inflated results. LBs grammatical structure is often classified according to four broad groups: noun-phrase based, prepositional-phrase based, verbal-phrase based, and other types (Chen & Baker, 2010) and their rhetorical functions are also divided into three main categories: research-oriented, text-oriented, and participant-oriented (Biber, Conrad, & Cortes 2004; Cortes 2004; Hyland 2008). For example the bundle *the use of*, which tends to be very frequent in academic writing, has a 'noun phrase with *of* phrase' structure, and performs a 'research-oriented procedural' function. Classifying bundles structurally and functionally and looking at their frequencies in different texts can help to reveal genre and disciplinary conventions. Study four focuses on the use of LBs in learner and expert academic writing and describes the most frequent bundles in each subcorpus, the bundles that they have in common, and the structures and functions that characterise each type of writing.

2.3 Disciplinary literacy

As we have seen, evidence-based analysis of linguistic corpora have made two important contributions to the field of academic writing: (1) the description of unique features –specially at the phraseological level- of academic discourse, such as the inseparability of lexis and grammar, and the patterning of language through recurrent word combinations (see e.g. Biber et al. 1999; Biber & Barbieri, 2007; Cortes, 2004; Gilquin et al., 2007; Römer, 2011). And (2) the variability across academic genres and disciplines (see e.g. Flowerdew 2002; Hyland, 2009a). This disciplinary variability has been explored from different perspectives such as comparing similarities and differences across languages (e.g. Moreno, 1997), disciplines (e.g. Cortes, 2004; Green

& Lambert, 2018; Hyland, 2002), genres (e.g. Hyland, 2008b), and even across different sections of the same texts (e.g. Biber & Finnegan, 1994; Bondi, 2010; Sheldon, 2018); a common pedagogical implication derived from these studies is the need for specificity when teaching academic writing.

Disciplinary literacy involves not only understanding but also being able to (re)produce the discourse of a discipline: in academic writing, this means presenting information, framing arguments, making suggestions, or showing stance in such way that experienced writers that belong to the same community will find familiar and convincing. As Hyland (2009b, p. 6) puts it:

> Writing as a member of a discipline involves crafting texts in a way that insiders can see as 'doing biology' or 'doing sociology' and this both restricts how something can be said and authorizes the writer as someone competent to say it. In other words, students learn what counts as good writing through an understanding of their discipline and the conventions and genres regarded as effective means for representing knowledge in that discipline.

In his corpus-based research for disciplinary specificity, Hyland (2009a) analyses four important features of academic writing, namely hedges (e.g. modal verbs), self-mention (e.g. personal pronouns), citation (i.e. reporting verbs), and transitions (e.g. linking adverbials), across disciplines and genres. Drawing on hard-soft discipline comparisons (e.g.: humanities and social sciences vs. science and technology) he discovered that these features have different frequencies and behave in dissimilar ways depending on the discipline. For example, when measuring the frequency and choice of reporting verbs, Hyland found that in fact "engineers *show*, philosophers *argue*, biologists *find* and linguists *suggest*." (2009, p. 11). Disciplinary literacy thus goes beyond correct grammar or native-likeness, and establishes itself in the paradigm of sophistication and peer-likeness. As Hyland and Tse (2007, p. 245) point out, "all disciplines adapt words to their own ends, displaying considerable creativity in both shaping words and combining them with others to convey specific, theory-laden meanings associated with disciplinary models and concepts". EMI teachers, EAP instructors and university students need to be aware of these conventions and practices in order to be able to participate as 'disciplinary insiders' (Jiang & Hyland, 2017) in academic discourse effectively.

To gain disciplinary discourse effectiveness in an L2, even in an L1, explicit instruction and intensive reading and writing practices are often needed (Hinkel, 2002, 2003). As Springer (2012, p. 217) aptly puts it, some features of academic writing are "so subtle that many teachers are not aware of them on a conscious level, [so] it cannot be expected that learners simply pick these up from wide exposure". EAP courses, often due to financial and logistic reasons, tend to focus on general academic discourse rather than exploring specific disciplines (Flowerdew, 2002; Römer, 2011). While it is true that there are some general academic words, collocations and formulas that can be found across disciplines (see e.g. Ackermann & Chen, 2013, for a list of academic collocations; Gardner & Davies, 2014, for a list of academic vocabulary, and Simpson-Vlach & Ellis, 2010, for a list of academic formulas), some studies have implied that the actual frequency and usability of this general terminology in different disciplines, EAP materials, and in students' papers in particular, tends to be quite low (Cortes, 2004; Csomay & Prades, 2018; Durrant, 2016; Durrant & Mathews-Aydınlı, 2011; Hyland, 2009a; Wood & Appel, 2014); providing leaners with highly contextualized examples of academic language in use in order to help them develop their disciplinary literacy seems therefore reasonable.

An additional problem is that, as mentioned previously, materials and textbooks that are designed to help students improve their academic writing skills contain tips and recommendations that are not always corpus-based nor are discipline-specific (see Gilquin et al., 2007, for a detailed discussion on this matter). There is also a tendency in these materials to unify and present academic writing as a homogeneous set of rules, oversimplified and applicable to all disciplines. As Hyland and Tse (2007, p. 247) caution, "the discourses of the academy do not form an undifferentiated, unitary mass, as might be inferred from such general lists as the AWL [Academic Word List, Coxhead, 2000], but constitute a variety of subject-specific literacies". Corpus analyses allow us to observe disciplinary variation and can enable instructors to provide learners with informed advice based on actual language use. Compiling a corpus of texts (from a written or transcribed source) that represent good models of writing (or speaking) in a particular discipline can also be a reliable and cost-effective way of obtaining accurate examples of linguistic features that are salient in a given discipline. In the next section, the notion of corpus linguistics and different corpus methods are presented so that general and discipline-specific writing features of the academic discourse produced by learners and experts can be explored accurately.

2.4 Corpus linguistics and learner discourse

Corpus Linguistics (henceforth CL) studies naturally occurring language in the form of electronically stored texts or corpora. Sinclair defines corpus as "a collection of pieces of language text in electronic form, selected according to external criteria, to represent, as far as possible, a language or language variety as a source of data for linguistic research" (2005, p. 16). As languages are constantly evolving, corpus-based studies can help linguists to keep track, record and explore up-to-date language use, while providing "large amounts of natural language examples" (Römer, 2011, p. 206). For Leech (1992, p. 106) CL is an "'open sesame' to a new way of thinking about language"; this is because corpus approaches have opened new avenues for research and have served to identify new linguistic phenomena in different language-related fields such as language teaching, discourse analysis, phraseology, pragmatics, or second language acquisition.

Although many corpus-based approaches consider frequency a key feature to identify language patterns, it is often not the only type of investigation performed; many studies follow a mixed-method approach, and provide qualitative interpretations of the data analysed as well (as is the case of the present thesis). CL's ultimate goal is thus to uncover features of language use and language structure relying on both quantitative and qualitative perspectives (Biber, Conrad, & Reppen, 1998), as we will see in Section 2.4.2.

Dictionaries and grammar books benefitted immensely from the incorporation of corpus-based work in the early days. This empirical and evidence-based methodology allowed for comparisons between different language varieties (e.g. American English vs. British English) and also between different registers (e.g. fiction vs. academic writing) that changed both the lexis and syntax paradigms (Granger, 2002). CL has also affected the second/foreign language (L2) teaching and second language acquisition (SLA) fields: applications of corpus tools (e.g. software packages) and corpus methods, i.e. "the analytic techniques that are used when we work with corpus data" (Römer, 2011, p. 206) can help instructors (and material developers) to make informed decisions on what linguistic items should be taught/learnt first because they are the most relevant ones in a given field or discipline, and when to teach them according to each particular developmental stage. A more direct applicability of corpora in the L2 teaching realm is the use of corpora in the classroom to show how something is used in context. In order to know, however, which aspects of a language are more or less difficult for learners, or what linguistic items can be error-prone, one needs to analyse learners' linguistic output.

The analysis of learners' authentic production of a second/foreign language is called Learner Corpus (also Computer Learner Corpus) (Granger & Tyson, 1996). Comparing texts written by non-native speakers of a language with another corpus (e.g. texts written by native or more advanced speakers of the same language) can help to uncover unique learner writing characteristics. A method of analysis that has helped to identify these features is the Contrastive Interlanguage Analysis (CIA) method (Granger, 1996), which allows for quantitative and qualitative comparisons between different non-native speaker (NNS) texts, but also between NNS and native speaker (NS) data. The CIA method has recently been revised and what was once called "native language" is now called "reference language varieties" in order not to discriminate between language varieties, since, as Granger aptly puts it "there are a large number of different valid reference points against which learner data can be set" (2015, p. 17). In fact, it has been shown that L2 learners not only have their own distinctive developmental problems (which can also be characteristic of particular L1 populations, and thus be indicative of possible transfer issues), but also share certain difficulties with novice native writers (Gilquin et al., 2007). Therefore, comparing a learner corpus with another corpus consisting of high-scoring native students' compositions (such as the British Academic Written English corpus –BAWE, or the American university student's subcorpus included in LOCNESS) can help to uncover these particularities. Quite a number of studies have brought to light common features that are recurrent in learner writing, such as lack of register awareness, overuse or underuse of certain items, phraseological mismatches, and/or pragmatic misuse, that call for more explicit attention in the classroom (e.g. Ai & Lu, 2013; Gilquin et al., 2007; Lee & Chen, 2009; Rica-Peromingo, 2012).

Because of the technical and methodological processes learner data analyses imply, together with a desire for more quantifiable findings, "corpus-linguistic methods have established themselves as among the most powerful and versatile tools to study language acquisition, processing, variation, and change" (Gries & Newman, 2013, p. 257); they also hold tremendous potential for L2 academic writing research. Over the last few decades, learner corpus-based studies approached through CIA have sprouted up in order to contribute to SLA research, connecting these two previously disparate fields (Granger, 2002, 2004). However, a first step in gaining reliability in corpus studies resides in the corpus building process.

2.4.1 Corpus building principles

As we have seen, corpus studies can yield remarkable results on language use. Corpora, and learner corpus in particular, must nevertheless be built on the basis of strict design criteria to control for all the variables (e.g. individual differences, task setting, etc.) that can affect language production (see e.g. Biber, Conrad, & Reppen, 1998; Granger & Tyson, 1996; Granger, 2002; Gries, 2014; Sugiura et al., 2007). Sinclair (2005) developed ten basic principles for compiling representative and well-controlled corpora for linguistic purposes:

1) Authenticity: “the contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise” (Sinclair, 2005, p. 1). This first principle shows how important it is that corpora are selected using external criteria only and not on the basis of internal linguistic features; selecting texts according to their ‘communicative function’ ensures generic comparability. Linguistically controlled or semi-naturalistic tasks (e.g. asking writers to use the second conditional structure to answer a question) influence the writer’s or the speaker’s output and this would not represent natural language use. In this respect, it has often been implied that when a corpus is produced out of a language-teaching context, it is not *entirely* natural (Granger, 2002). While it is true that instructors impose topics, time and word limits, the fact that a learner, aiming to become a proficient writer of English, writes under time/topic constrictions is not very far from reality: professional writers, pressed to meet editors’ requirements, also have to deal with tight deadlines, imposed topics and word limits. Texts not produced for linguistic analysis (Gries, 2014) that result from ‘authentic classroom activity’ (Granger, 2002, p. 4) can therefore qualify as authentic corpora.

2) Representativeness: “corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen” (Sinclair, 2005, p. 2). This principle is somehow related to principles number 6 –sample size– and number 10 –homogeneity: a corpus is often regarded as representative if it includes many texts (size) that belong to the same community of writers (homogeneity) (e.g. written by learners that share the same L1, that have a similar L2 proficiency, or that belong to the

same class). However, and as we will see in these forthcoming principles, 'big' can be good but not for all purposes. If, for example, the analysis of certain linguistic items involves manual exploration or annotation, the probability of working with a corpus of hundreds of texts is often low; this does not necessarily mean that the corpus is less representative (Noble, 2010). In addition, and given the internationality, and thus multiple linguistic and cultural identities, of today's classrooms, finding texts written by learners that are homogeneous in terms of L1 and L2 proficiency is often not possible nor realistic.

3) Contrast: "only those components of corpora which have been designed to be independently contrastive should be contrasted" (Sinclair, 2005, p. 3). According to Sinclair, comparability is often taken for granted during corpus building processes. This implies that corpus builders need to make sure that the language varieties or linguistic structures they are contrasting are indeed comparable. The concept of language target still is a question of debate: as Gilquin et al. (2007, p. 11) indicate, "[s]ome learners may have native (or native-like) writing as target, while others may consider English as a *lingua franca* the ideal target. These factors, combined with the discipline and L1 specificities (…) result in a great diversity which may be quite difficult to reconcile". Native norm is a common target in L2 writing research, but one needs to assess whether it is legitimate to compare learner's interlanguage against an ideal native norm. There are also some caveats when performing cross-study comparisons. For example, 'advanced' or 'academic discourse' may not mean exactly the same level of proficiency, or contain the same disciplines in two different studies; failing to provide a fine-grained distinction between for instance, articles of literature or articles of linguistics, or essays and dissertations, which may differ in the language they contain, the authors' expertise, or the communicative purpose they represent, may reduce the comparability of the results (see Springer, 2012, for a detailed discussion on genre vagueness). Size must also be taken into account when comparing two corpora of different lengths, and provide normalized results (e.g. by 1,000 words) when necessary.

4) Structural criteria: "criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination" (Sinclair, 2005, p. 5). According to Sinclair (ibid), establishing texts' selection criteria is the first step in building a trustable corpus. Mode (written or spoken), domain

(academic or popular), genre (e.g. review, article, dissertation), or language variety (e.g. British English) are some of the most common criteria in corpus building. For example, Granger and Tyson (1996) found ‘comparability’ a fundamental characteristic of learner corpora to be representative. In order to build the widely known International Corpus of Learner English (ICLE), they set four main criteria: first, they included untimed texts written by learners who probably made use of external resources. In this respect, as exposed by Ädel (2008), task setting (i.e. time available) and intertextuality (i.e. access to secondary sources) can have an impact on writing style, and can profoundly influence the linguistic output. In fact, in Petch-Tyson’s study (1998) on writer/reader visibility, an interesting pattern emerged: whether an essay was timed or not had an effect on the presence and proportion of certain involvement features; task setting and intertextuality are therefore important conditions that need to be considered when working with learner corpora. The second criterion set for ICLE was that only learners’ final product (output) was analysed (as opposed to different versions of the same text); these texts consisted of academic essays of an argumentative nature. Third, size was also accounted for, and texts no longer than 500 words were excluded. And finally, all texts were produced by a learner population in a similar stage, that is, they were all university students. Setting clear structural criteria helps to achieve representativeness in a corpus.

5) Annotation: “any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications” (Sinclair, 2005, p. 5). According to this principle, tags or annotations should be stored separately from the original corpus or raw text. Since there is no unified tagging system yet (e.g. there exist different ways to annotate metadata, different nomenclature for part-of-speech or error taggers, parsers, etc.) and as each researcher tends to annotate her/his corpus according to specific research purposes, it is difficult, therefore, to reuse or recycle corpora. However, Sinclair (ibid) suggests that by storing tags and raw text separately, the problem of corpora reusability can be solved, and many more corpora could be made accessible to other researchers.

6) Sample size: “samples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get to this target as possible. This means that samples will differ substantially in size” (Sinclair, 2005, p. 7). In this regard, Sinclair explains that there was a tendency to think that equal text length in corpora was needed in order to perform a reliable analysis. This

practice is no longer justifiable, since "the integrity and representativeness of complete artifacts is far more important than the difficulty of reconciling texts of different dimensions" (ibid, p. 6). When looking at frequencies, however, it is important to normalize the results since, as mentioned earlier, text length can have an effect on word frequencies. In the case of the corpora included in this thesis, native, expert and learner texts represent complete and unedited texts that vary in length, and normalized results are provided when necessary.

7) Documentation: "the design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken" (Sinclair, 2005, p. 8). Following this principle can help researchers to prove that there is nothing wrong with their corpus in case they get unexpected results. In addition, any corpus study that provides a well-documented description of its data will, at the same time, favour cross-study comparisons. This can be done by recording specific details about the participants' linguistic and educational background (e.g. information about their L1/s, L2 level of proficiency, other L2s, etc.) as well as by explaining the selection process of the texts that compose the corpus (i.e. set of criteria mentioned earlier).

8) Balance: "the corpus builder should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components" (Sinclair, 2005, p. 9). According to this principle, a balanced corpus should include equal proportions of written and spoken, general and specialized, and formal and informal language in it. However, this principle refers only to large corpora representing general language use, such as the Corpus of Contemporary American English (COCA), or the British National Corpus (BNC) (for a survey on general and specialized corpus resources see McEnery, Xiao, & Tono, 2006).

9) Topic: "any control of the subject matter in a corpus should be imposed by the use of external, and not internal, criteria" (Sinclair, 2005, p. 10). According to this principle, only external criteria should be taken into account when compiling a corpus, such as the selection of texts according to disciplines (e.g. linguistics, medicine), genres (e.g. essays, reviews, articles), etc. On the other hand, no or little control should be exerted over the topic or the linguistic structures used in the text –as this would represent a somewhat less natural use of the language. Nesselhaulf (2005, p. 128)

provides a scale of naturalness, and helps to distinguish between four different degrees of naturalness; a text can be fully natural, product of teaching process, controlled task, or scripted, depending on how the task has been designed.

10) Homogeneity: "a corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided" (Sinclair, 2005, p. 14). While it is important to group texts that belong to the same, or at least similar genres or disciplines written by a similar type of writer in order to achieve homogeneity and to provide results that are representative of a particular community, the avoidance of 'rogue' texts is somehow a matter of debate in learner corpus research. Learner texts may contain errors and unless they are identified and properly tagged beforehand, they may go unnoticed. Unidentified language errors can have effect on certain linguistic analyses if errors are not the subject of study –for example, if a text contains misspelled words, these words will not appear in corpus-based frequency searches using a text-analysis tool.

These ten principles were taken into consideration when compiling the different corpora for the studies presented here. A final point worth mentioning is that, even though a corpus "may not capture all the patterns of the language, nor represent them in precisely the correct proportions" (Sinclair, 2005, p. 2), if it has been well designed, it can certainly help corpus linguists to explore, describe and understand language.

2.4.2 Quantitative and qualitative corpus analyses

As we have seen, corpus building processes are of paramount importance to achieve reliable results in corpus studies. There are, as well, other elements that affect corpus research procedures: research questions, for example, play an important role in determining the type of analysis that will be performed. Depending on which come first, the corpus data or the research questions, the researcher may begin by exploring the data itself without any preconceived ideas of what can be found, i.e. corpus-driven approach; in contrast, if questions are set beforehand, then a list of pre-selected items (e.g. list of connectors) may be tracked in the corpus to examine their frequencies and the ways they behave in different contexts (i.e. corpus-based approach) (see Callies et al., 2014 for a detailed differentiation between corpus-based and corpus-driven approaches). Regardless of the method, a corpus can be looked at through quantitative and/or qualitative perspectives. Regarding the former, Gries (2014) alludes to four common corpus-linguistic methods that holistically describe quantitative corpus analyses: (1) Frequency lists: these are used in order to see the number of hits of a particular element in a given text. Most text-handling software packages, e.g. Voyant (Sinclair & Rockwell, 2016) or AntConc (Anthony, 2018), can highlight vocabulary that is particularly salient in a given corpus. The creation of wordlists is also useful to see the total number of words (tokens) or different words (types) in a text. Calculating the type/token ratio (number of different words divided by the number of total words) can be helpful to see e.g. how diverse a writer's lexis is. (2) Dispersion: this type of analysis is used to display where in a corpus an element occurs and how often. This is useful if we track e.g. stance devices, to see if they appear at the beginning (introduction) or at the end of the text (conclusion) more often; comparing dispersion plots of texts that belong to different fields can help to uncover disciplinary conventions (e.g. amount of stance devices in the introduction vs. the conclusion sections). (3) Collocation, colligation, and collostruction: these are used to analyse around which other elements an element occurs and how often. We can benefit from this approach when tracking the use of e.g. academic collocations, discourse markers, or lexical bundles, which at the same time can help us to measure how sophisticated or complex a text is, for example. (4) Concordancing: this type of analysis enables researchers to see exactly what comes before and after a tracked element, allowing for a more proper interpretation –it accounts for the most contextualized form of analysis. Using normalized values when comparing texts of different lengths, and including descriptive or inferential statistics (such as the *t-test*) can enhance the reliability of the quantitative analysis. The four studies included in this thesis have made use of these quantitative resources in their analysis.

Quantitative methods abound in the corpus research literature, but there are also qualitative studies on corpus, and studies that include both perspectives and follow mixed-methods approaches. In fact, Biber, Conrad, and Reppen (1998) point out that researchers should not rely on quantitative data alone when exploring corpora, but rather on a multi-method technique by which each quantitative result involves a functional interpretation. Frequency is a very reliable source to analyse corpora, but it is also true that, in terms of linguistic analysis, one cannot regard language as discrete forms that can be counted. Words and expressions may carry different meanings and perform different functions in different contexts; certain types of analyses (e.g. exploring the functions of lexical bundles) call for a more qualitative exploration of the context in which the searched items occur. As Granger notes, “studying the lexical realization of other discourse feature types will uncover differences that cannot be identified by quantitative analysis alone” (2004, p. 135). In this sense, the trend of exploring patterns of language use has beaten the Chomskyan view of language as an “unordered list of all lexical formatives” as cited in McCarthy (2006, p. 9), to turn into a more innovative exploration of the language in motion, identifying expressions and phraseological units that characterise particular registers (Biber et al., 1999; McCarthy, 2006; Rica-Peromingo, 2012). For example, when analysing quantitative features such as overuse and underuse in learner corpora, these alone are descriptive and not prescriptive terms –in other words, they describe when certain items have been used more or less frequently in the learner corpus than in the control corpus– they do not, in any case, imply that items in these categories always require pedagogical attention. In order to know whether the characteristics discovered will be “selected for pedagogical action or ignored depends on a variety of features, including learner needs, teaching objectives and teachability” (Granger, 2009, p. 22). Following these parameters, qualitative explorations of the data under study have been performed in the four studies that are presented in this thesis.

3. Research questions and hypotheses

In this chapter, the main objectives, research questions and hypotheses formulated for the studies are presented. There is an overarching question that connects the four studies that were carried out for the present doctoral thesis: Taking disciplinary literacy and the phraseological dimension of language into account, how can corpus approaches contribute to identifying L2 learners' academic writing features regarding four different linguistic phenomena that are important for successful academic writing at university? The authors' motivation to apply corpus approaches to learner writing in order to find distinctive usage patterns that differ from the use of the same devices in comparable texts written by native or expert counterparts inspired this question. The main aim of this doctoral thesis is therefore to contribute to the body of research that studies second language writing and disciplinary literacy through corpus linguistics, and to serve as a useful pedagogical resource for both instructors and L2 learners regarding the effective teaching and learning of four different linguistic devices that are important for the construction of academic discourse in English at university.

In study one, the use of general and technical academic vocabulary is analysed before and after a content-based instruction (CBI) course. The main research question guiding this piece of research is: What effect does the CBI course have on students' academic vocabulary production? It is hypothesized that there would be a higher production of academic vocabulary in the texts written after the course (T2) as a positive effect resulting from the instruction received. Additionally, a more sophisticated use of academic vocabulary is expected from one of the groups (in the EMI setting), compared to the other group (in the L1 setting), due in part to a higher exposure to the target language in an academic context. The extent to which learners have used academic words, collocations and formulas is also investigated.

In study two, the use of adversative linking adverbials (LAs) in argumentative essays written by EFL learners with different L1s is explored and compared to argumentative essays written by English L1 students. There are two main research questions: 1) How do English NNS undergraduate writers use adversative LAs in terms of frequency, placement and types compared to English NS undergraduate writers? Based on previous studies on LAs, EFL learners are expected to use more LAs than their native counterparts. This is due to the fact that LAs are frequently taught in English language courses; these connectors are commonly presented as long lists of items (often out of context), and learners who use them in their texts tend to be awarded higher marks (Granger & Tyson, 1996; Granger, 2004; Lei, 2012; Rica-Peromingo, 2012; Wray, 2002). The second question formulated for this study is: 2) How do learners with different L1 backgrounds use adversative LAs when compared to one another? It is expected that the group of learners with Romance L1s (i.e., French, Italian, Spanish) and the group with Germanic L1s (i.e., Dutch and German) would show similar frequencies and usage patterns when compared to one another in their groups, as a possible influence from their mother tongue.

Study three investigates the presence of reflexive metadiscourse (MD) in bachelor dissertations (BDs) written in English by Spanish L1 learners, and compares it to the use of these devices by expert writers of research articles (RAs) in the same disciplines. The main questions formulated for this study are: 1) How do Spanish L1 learners use reflexive MD markers when writing in academic English compared to an expert corpus of RAs?, and 2) Are there any differences in the use of MD across disciplines? Frequency rates of two main categories of MD markers (i.e. textual and interpersonal) are calculated, to later look at interdisciplinary (linguistics vs. medicine) and writer status (learner vs. expert) variation. It is hypothesized that students would use some MD markers more frequently (e.g. textual markers in order to provide structure to their texts) and less frequently (e.g. interpersonal markers to address and engage readers) than experts, possibly pointing towards learner writing (such as transfer from their L1) or particular genre (university project) characteristics, and that some MD markers would be present only in linguistics or only in medicine texts, as a possible consequence of interdisciplinary variation.

Finally, study four looks at the use of lexical bundles (LBs) in BDs and RAs. The main questions guiding this study are: 1) How do Spanish L1 learners use LBs in the introduction and conclusion sections when writing in academic English, compared to an expert corpus of RAs?, and 2) How are these LBs used in terms of structure and function? Learners are expected to use similar bundles in their BDs, regardless of their discipline, particularly due to rigorous genre conventions (i.e. specific BD guidelines and requirements), and the canonical structure and functional purposes of the introduction and conclusion sections in this academic genre. It was also expected that the use of bundles would differ from published RAs in the same discipline due to authors' different expertise, and that linguistics and medicine writers' use of LBs would also differ in terms of frequency and functions (e.g. more text-orienting *vs.* more research-orienting bundles), possibly due to different disciplinary conventions.

4. Methodology

In this chapter, a description of the different learner corpora used and their collection processes are provided in Section 4.1. This is followed by Section 4.2, which comments on the reference corpora employed. Finally, Section 4.3 presents the text-analysis tools that were utilized for the different analyses, and summarizes the quantitative and qualitative approaches followed in each of the studies.

4.1 Learner corpora

In order to investigate different linguistic phenomena in L2 learner academic writing, three different types of text produced at university were collected:

1) Writing activity (see Appendix 1): a writing activity was designed with the purpose of measuring students' use of academic vocabulary before and after a content-based language course that is taught in the first year of a Dentistry bachelor degree, in study one. The exact same activity was performed twice and it was integrated as a classroom activity. Students were given a short questionnaire to indicate their age, gender and L1. Moreover, students received a score using Friedl and Auer's (2007) rating scale for assessment of a writing task and qualitative feedback from their instructors on this activity at the end of the course. The activity comprised four questions which could elicit different types of academic language. Only those texts that were longer than 150 words (texts contained 300 words on average) and that appeared at both T1 and T2 were used for the study (N=56).

2) Final assignment: in study two, a collection of argumentative essays written in English by advanced EFL students with different L1s was gathered. These texts were the final assignment for a content subject in the students' first year of studies at a European Studies bachelor degree, but it also served as a final project for a short Academic Writing Skills course, which was taught during the same semester. Although these texts were generally shorter than a research article (3,000 words per text on average), they contained a scientific structure (IMRaD) and were argumentative in nature. Students peer-reviewed these texts with the help of a class-made checklist taking into account the notions covered during the course, and the final version was collected for research. Only those texts that were successful in both the content and the language subject were included for the analysis (N=50). Moreover, in order to look at the use of adversative linking adverbials according to L1, students' mother tongue was documented.

3) Bachelor degree's final project: in studies three and four, the learner corpus consists of a collection of bachelor dissertations written in English by Spanish L1 learners in their last year of studies in linguistics and medicine. These texts are longer than the learner corpora collected previously (5,198 words on average) and their structure and contents resemble much more those of academic research articles in the same disciplines. Students filled out a questionnaire in order to indicate their L1, proficiency level, and other data to complete their linguistic background. As in the case of the abovementioned corpora, only BDs that were successful (obtained a final score of 'C+' or more) were included for the analysis (N=20). Reflexive metadiscourse and the use of lexical bundles were explored in these texts.

4.2 Reference corpora

In order to see whether the different linguistic phenomena explored in the learner corpora (i.e. academic vocabulary, adversative linking adverbials, reflexive metadiscourse, and lexical bundles) were different or similar, in both quantitative and qualitative terms, to the presence of these items in academic texts written by native speakers of English or professional writers, three different corpora were used for comparisons:

1) Academic vocabulary lists: in study one, three well-known lists of general academic vocabulary (i.e. terminology present across disciplines extracted from native corpora) were used in order to measure the extent to which the use of academic vocabulary increased or decreased after a CBI course. It is important to note, however, that these list do not represent a unitary corpus itself: they are three different corpus-derived resources that are tracked in the learner corpora in study one. With the intention of exploring a wider range of academic expressions, – or in other words, to avoid looking at single words only– sequences containing one (vocabulary), two (collocations), and three or more words (formulas) were explored. The expressions included on the Academic Vocabulary List (Gardner & Davies, 2014), the Academic Collocation List (Ackermann & Chen, 2013), and the Academic Formulas List (Simpson-Vlach & Ellis, 2010) (i.e. a total of 6,090 entries) were tracked in the learners' texts, to later compare their frequencies at T1 and T2. The texts of five native speakers of English that were enrolled in the same CBI course were also collected for comparative purposes. The abovementioned lists of academic terminology had been compiled following rigorous methods of word and expression retrieval drawing on well-

known written academic corpora (e.g. BNC, COCA, PICAE). In addition, these lists have been widely recognized in previous studies that measure lexical complexity and sophistication in L2 academic texts (Durrant, 2016; Green & Lambert, 2018; O'Loughlin, 2012; Webb & Nation, 2017; Wood & Appel, 2014).

2) Louvain Corpus of Native English Essays (LOCNESS): in study two, 176 argumentative essays included in the American university students' corpus, which is part of the LOCNESS corpus and was provided by Centre for English Corpus Linguistics of the Université Catholique de Louvain, were used for comparisons. These texts were selected because they had been written by an equivalent type of learner (i.e. participants in LOCNESS were aged 18-21 and were university students), and they were comparable in terms of text type (i.e. untimed academic argumentative texts written using reference tools). LOCNESS has also been widely used in previous literature that looks at the use of connectors in L2 writing and that explores L1 transfer (e.g. Granger & Petch-Tyson 1996; Leńko-Szymańska, 2008; Prommas & Sinwongsuwat, 2011; Rica-Peromingo, 2012).

3) Expert corpus of research articles (RAs): in studies three and four, an expert corpus of RAs in linguistics (N=25) and medicine (N=25) was compiled. These articles were published between 2010 and 2018 in peer-reviewed, English-medium academic journals. With regards to the authors' nativeness, since there was a mixture of both English L1 and L2 authors, they can be considered examples of English as a *lingua franca* in academia. In addition, these texts were chosen because they approximate the learner corpora in terms of discipline and topic, and because they represent good models of writing. Using an expert corpus of research articles as a reference corpus is an acknowledged practice in corpus studies of L2 academic writing (cf. Granger, 2017b; Hyland, 2008b, 2014).

4.3 Tools and analyses

The use of five text-analysis software and tools was necessary in order to perform the different corpora analyses presented in the studies:

1) AntConc (Anthony, 2018): this freeware corpus analysis toolkit has been used in all four studies in order to search lists of words (e.g. Liu's 2008 list of adversative LAs) and to calculate frequency rates. AntConc has also been used to calculate the total number of tokens and types in each corpus. In some cases (e.g. study two), the 'sort' function (case sensitive, punctuation, and keyword in context) was used

to observe when a searched term was used in initial, medial or final position. In addition, the function 'cluster n-gram' in AntConc was used to extract repeated sequences of words (3-, 4-, 5-word bundles) with a minimum frequency and range cut-off, in order to identify phraseological patterns in a particular corpus (see e.g. study four). AntConc was especially useful to perform qualitative analyses of concordance lines in order to see the searched term in context in all four studies.

2) Collocate 1.0 (Barlow, 2004): this web-based tool has been used in study one to extract the most frequent collocations in a collection of texts using the Mutual Information statistical test.

3) R (RStudio, 2012): this open-source statistical analysis software has been used, specifically with the Quanteda package, in order to track lists of words and expressions that belonged to different lists in a number of texts, in study one.

4) TagAnt (Anthony, 2015): this freely available software has been used to tag words according to their part of speech, and also to count the exact number of sentences present in each text, in study two.

5) Voyant tools (Sinclair & Rockwell, 2016): this web-based text reading and analysis environment has been used to calculate the total number of tokens and types, and to see the most frequent words in different corpora, in study one.

In some cases, these tools have helped to perform corpus-based analyses: for example, in studies one and two, already compiled lists of terms, namely academic vocabulary lists and a list of adversative linking adverbials, were tracked in the corpora in order to see if the text contained these specific words or expressions and their frequencies (hits). This type of analysis is advantageous in that it allows for comparisons of large quantities of items in corpora of a bigger size (e.g. 6,090 entries were tracked in 112 texts in study one). The downside of this method is that items that are actually present on the texts but that are not included on the lists will never be found. On the other hand, corpus-driven (i.e. data-driven, text-centred [Callies et al., 2014]) approaches have also been followed in some of the studies, and have provided a more accurate picture of the usage patterns of particular linguistic devices in the corpus. For example, the analyses in studies three and four are based on linguistic items that were provided by the corpus itself –e.g. in study three, all text were carefully read, and reflexive MD markers present in the texts were manually tagged; in study four, lexical bundles of different lengths and with different structures and functions were extracted and evaluated relying on computer techniques. These corpus-driven methods provided a deeper understanding of the linguistic phenomena studied. The problem is, however, that as these analyses are often more time-consuming, the data analysed is consequently smaller (e.g. 20 complete BDs were manually coded in study three, and a list of 218 lexical bundles were classified in study four).

The quantitative analyses of frequency included in all four studies have also been complemented with different qualitative analyses of the linguistic phenomena studied. For example, different educational tracks (e.g. students' use of academic vocabulary according to their setting of instruction, in study one), positioning (e.g. placement of linking adverbials, in study two), categories (e.g. of metadiscourse markers, in study three), structure, and function (e.g. of lexical bundles, in study four) have been explored. These qualitative approaches gave the author a deeper view of the choices learners and experts made and how these patterns affected the whole structure of the text; in sum, they significantly improved the author's understanding of these linguistic devices and offer a broader picture of the use of these items in academic writing. The four chapters that follow present the four studies that were carried out for this doctoral thesis.

5. The effects of a content-based language course on students' academic vocabulary production

5.1 Introduction

Academic discourse refers to the "ways of thinking and using language which exist in the academy" (Hyland, 2009b, p. 1), and it plays a fundamental role in developing students' understanding of any discipline (Ha & Hyland, 2017). Knowledge of academic discourse implies acquiring a multifaceted set of language skills that may involve presenting research, be it through speech or writing, interacting with peers and experts, and also navigating a variety of academic genres and tasks successfully. These are some of the demands universities impose on students at the various levels of higher education (HE). As an increasing body of recent research shows, knowledge of academic discourse is of paramount importance for students' successful educational performance (Airey et al., 2017; Csomay & Prades, 2018; Granger, 2017a; Webb & Nation, 2017) as it allows them to show their skills in applying, analyzing and evaluating knowledge in their field of studies appropriately. And yet, what is considered 'appropriate' may vary from discipline to discipline (Hyland, 2008b; Hyland & Tse, 2007) and from one linguistic context to another (e.g. Moreno, 1997). English-as-a-Foreign-Language (EFL) students at English-medium universities are also expected to develop an academic discourse in the target language (TL), and this is sometimes in addition to developing their L1 academic discourse as well, which can be difficult to achieve, and can trigger transfer-related issues (Airey, 2018).

In the particular case of academic writing, specific terminology and formulaic language play an important role in knowledge making, not only because they carry the ideational weight of the text, but also because they portray disciplinary conventions. However, the fact that specialized vocabulary may account for 10% to 30% of the words in an academic text (Coxhead & Nation, 2001) can pose a challenge to novice EFL readers and writers: as Hinkel points out "learners will generally not pick up even more obvious characteristics of academic writing by mere exposure" (2003, p. 297). An additional difficulty is the fact that academic discourse differentiates between two types of discourse: 1) *discipline-specific* discourse, that is, those words and expressions that are related to content knowledge and that differ from discipline to discipline (in the dentistry field, we could find e.g.: *enamel, partial restoration, scaling and root planning*), and 2) *general academic* discourse, i.e. terminology and expressions used across different academic contexts and that can be found in a wide range of disciplines (e.g. *evaluated at baseline, qualitative analysis, significant differences*) (Granger, 2017). The effectiveness of teaching and learning academic discourse by focusing on the former or the latter type is currently under debate.

Some studies, on the one hand, claim that academic discourse can be highly discipline-bound in nature (Granger, 2017; Hyland, 2008; Hyland & Tse, 2007) in that each academic discipline operates within very specific conventions and specialized discourses, which significantly reduces the effectiveness of learning *generic* academic vocabulary only. Recent efforts have been made to create discipline-specific lists of vocabulary, such as the nursing academic word list (Yang, 2015), or the medical academic word list (Lei & Liu, 2016). On the other hand, there is another line of research which claims that a generic core of linguistic devices across disciplines does exist, and that, given the cross- and inter-disciplinary nature of most studies and tasks EFL learners are exposed to, its pedagogical relevance is warranted (Durrant, 2016; Simpson-Vlach & Ellis, 2010). In fact, several lists of general academic terminology, of different lengths and breadths, have been created drawing on large academic corpora, such as the British National Corpus (BNC), and the Corpus of Contemporary American English (COCA), and have been designed using advanced methods of word retrieval (see Gardner & Davies, 2014 for a comprehensive description), as we will see in Section 3.2. There is also an increased tendency for vocabulary lists compilers to move away from the analyses of isolated words, and study longer strings of words instead, also referred to as 'formulaic language' (Cortes, 2004; Hyland, 2008; Wood & Appel, 2014), that can be more (e.g. *collocations*) or less (e.g. *formulas* or *lexical bundles*) idiomatic (Durrant, 2016; Granger, 2017; Paquot, 2017).

While some educational settings usually have content teachers teaching technical terms and English teachers focusing on general academic vocabulary (Green & Lambert, 2018), there is an approach that can provide EFL learners with opportunities to learn both *general academic* and *disciplinary* discourse in context, namely Content-Based (language) Instruction (henceforth CBI). CBI programmes, and more specifically the 'adjunct model', are parallel language courses instructed by a language specialist in collaboration with content specialists (see e.g. Roquet et al., forthcoming) that go hand in hand with other content subjects in the same programme. Section 5.2 provides more information about this type of communicative language teaching, as it is part of the context of this investigation.

Academic and disciplinary literacies are increasing areas of research. Several studies have focused on academic discourse, be it generic or technical, to explore how it is deployed in textbooks students are exposed to (Green & Lambert, 2018), by expert writers in research papers (Cortes, 2004; Hyland, 2009a), other academic texts in EAP courses (Wood & Appel, 2014). However, few studies have explored the actual use and the development of academic discourse in texts written by EFL students longitudinally. When comparing two texts written by the same learner over time, a more frequent and varied presence of academic terminology can indicate a developing fluency in the academic discourse (Gee, 1991) and therefore the emergence of a disciplinary voice (Hyland, 2009a).

Taking a corpus-based approach, the present study adopts an innovative analytical approach in that it explores the occurrence of general academic discourse, not only in terms of words, but also in terms of collocations and formulas, in a learner corpus of EFL student writing over time. In order to provide a more inclusive analysis of academic discourse, it also explores lists of words, collocations, and formulas that are specific to the discipline studied (i.e. dentistry), using vocabulary lists extracted from the class materials. The objective of the present study is to see the effects of a one-semester CBI course on students' production of academic discourse by means of a pre- and post-test design. First, this study gives an overview of previous literature on CBI and the effects on written academic discourse at university level. Second, corpus-based research on academic discourse is reviewed and a description of three lists of academic language, namely the AVL, the AFL and the ACL is given, to later expose the research questions and hypotheses. Next, Section 5.4 describes the context, the data collection and the analytical procedures. Results regarding academic language coverage in the class materials, and usage by learners according to their setting of instruction are discussed in Section 5.5. Finally, conclusions, pedagogical implications, and recommendations for further research are given in Section 5.6.

5.2 Content-based language instruction

CBI is a form of communicative language teaching in which language is used as a real means of communication. In the literature, three different models of CBI have been described: 1) in the *theme-based* model, a language specialist usually focuses on different topics to teach language and it is typical of language schools or courses for adult learners; 2) in the *sheltered model*, on the other hand, it is a content specialist who teaches her/his subject in a more student-centered manner to ESL students, providing comprehensive input; and finally, 3) in the *adjunct model*, a language specialist together with a content specialist develop a language course in which linguistic structures and specialized terminology are made visible to students (see Richards & Rodgers, 2014, and Stryker & Leaver, 1997 for a more detailed description of these models). The CBI adjunct model can equip students with transferable linguistic skills to perform successfully in parallel content subjects so that they can be better prepared (Römer 2009, 2011). However, while these three models use authentic material for language learning, the latter integrates language and content in a more contextualized manner by using the subject as a background for language learning. As Richards and Rogers point out, "people learn a second language more successfully when they use the language as a means of acquiring information, rather than as an end in itself" (2014, p. 209) and this is particularly one of the advantages CBI offers (Brinton et al., 2003).

There are three other types of communicative language teaching that have become popular and widespread practices in the past few decades –specifically English for Academic Purposes (EAP), English as Medium of Instruction (EMI), and Content and Language Integrated Learning (CLIL). The extent to which these models focus more or less on the language can vary greatly, depending on the educational level these are implemented in, the discipline, and also the instructor. In order to provide a clearer definition of these constructs, Airey (2016, p. 73) has created the "language-content continuum" which reflects the orientation of these three educational approaches with respect to their learning outcomes, as shown in Figure 1 below:

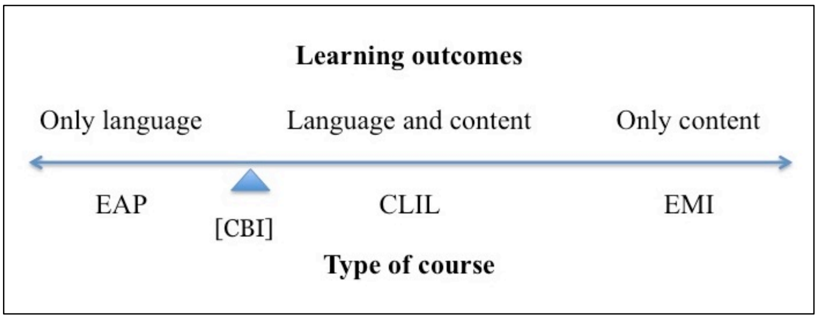


Figure 1: the language-content continuum (adapted from Airey, 2016)

CBI could be placed towards the "language-driven" side on the continuum, since it addresses specific language needs. The main difference with the first type of course, however, is that CBI programmes are 'parallel courses', often obligatory, closely associated with other subjects (as is the case with the course explored in this study), and they are therefore more content-oriented than EAP. EAP courses are normally offered outside the curriculum, and can be thus less discipline-oriented (Airey, 2016, p. 74). In this regard, Ha and Hyland (2017, p. 35) observe that "EAP teachers (…) often lack the specific field knowledge to develop suitable teaching materials about technical vocabulary and often feel vulnerable in this area". Recent corpus-based studies have also implied that the materials used in EAP courses seem to contain a low coverage of general and specific academic vocabulary (Durrant, 2016; Wood & Appel 2014), which can be unfavourable to EFL learners.

CLIL, or the acronym used in higher education –ICLHE–, programmes, on the other hand, are integrated as part of the curriculum, i.e. are subjects *per se*[6] (as opposed to the 'parallel' status of the adjunct model) in which "curricular content is taught through the medium of a foreign language" (Dalton-Puffer, 2011, p. 183) pursuing a dual objective, i.e. teaching and learning the language *and* the content. Studies that explore the impact of CLIL instruction on learners' linguistic gains have often implied that CLIL students tend to outperform non-CLIL students in terms of fluency, lexical and syntactic complexity, and test scores (see e.g. Dalton Puffer, 2011; Lasagabaster, 2008; Pérez-Vidal & Roquet 2015; Ruiz de Zarobe, 2011). However, while CLIL programmes have become very popular in primary and secondary education in recent years, they are uncommon in tertiary education (Airey, 2018). This may be due to the fact that CLIL requires instructors to have a dual expertise in both the content and the language, and the effort required to redesign teaching materials to meet this dual objective at university level can discourage HE instructors to implement and offer CLIL subjects in their programmes.

The last type of instruction –EMI– that has gained traction in the European HE arena in recent years (Ament & Pérez-Vidal, 2015; Smit & Dafouz, 2012) is an educational approach in which content specialists teach content through English. It serves various purposes, two of which are: 1) to offer internationalization at home for national students, and 2) to satisfy the needs of the ever-growing population of

[6] This may not be the case in all institutions, however. In Belgium, for example, many schools offering CLIL subjects have to offer as well equivalent subjects in French or Dutch (Meunier & Van Goethem, 2017).

international students (Pérez-Vidal et al., 2018). There are different amounts of EMI exposure (e.g. full immersion, semi-immersion) that can increase or decrease as the degree goes on. However, one of the recurrent concerns of EMI instructors is that they do not feel comfortable correcting students' linguistic mistakes (Airey, 2011; Ha & Hyland, 2017), so no/little attention is paid to language learning. As a result, the impact EMI can have on the development of specific domains of the L2 competence can be rather limited.

There is a fourth approach that has not been included in Airey's (2016) model, namely English for Specific Purposes (ESP). In fact, ESP could also be placed towards EAP, near CBI on the continuum. While we can see links between CBI and ESP in that they both pay attention to highly specialized discourse (e.g. Business English, English for Lawyers) and are tailored courses to meet specific language needs, the latter derives from a more traditional English language teaching approach, and is thus less content-oriented than CBI. In addition, the type of materials used in ESP courses tend to be adapted for EFL learners, and can be less authentic than in CBI adjunct model courses, in which authentic materials (e.g. research articles, case studies) are provided, and are often connected to other content subjects (Brinton, 1993).

In conclusion, the value of L2 generic and disciplinary academic discourse may not be sufficiently exploited in the abovementioned types of instructions in HE, and CBI may overcome this shortcoming. As Roquet et al. (forthcoming) indicate, CBI can help to narrow the linguistic gap between EMI and L1 students in terms of syntactic gains. There are some studies that have looked into learners' lexical and morphosyntactic gains (Roquet et al., in press) and overall performance (Dafouz et al., 2014; Hernández-Nanclares & Jiménez-Muñoz, 2017) when exposed to CBI approaches in tertiary education. However, there are few studies that analyse EFL learners' lexical sophistication through the production of academic vocabulary after a CBI course. The present study aims to address this gap of knowledge.

5.3 Disciplinary literacy and academic vocabulary lists

5.3.1 Disciplinary literacy

The development of disciplinary literacy, meaning the "ability to appropriately participate in the communicative practices of a discipline", is often one of the primary goals of any degree programme (Airey, 2011, p. 3). Finding a "credible disciplinary voice" (Jiang & Hyland, 2017, p. 14) can moreover allow students to relate to their professional community "in ways that seem familiar and engaging" (Hyland, 2005, p. 71). However, disciplinary discourses can be highly context-sensitive, and therefore, not only the language, but also the mode (e.g. written vs. spoken), the genre (e.g. lecture vs. textbook) and even the type of task (e.g. essay vs. research paper) can influence linguistic choices. The same words may have different frequencies, collocations, and different meanings in different disciplines –consider the use of *attrition* in linguistics and in dentistry, for example. In a corpus-based study on academic writing across disciplines, Hyland found that "it turns out, in fact, that engineers *show*, philosophers *argue*, biologists *find*, and linguists *suggest*" (2009, p. 183). It is clear then, that to succeed in any discipline, EFL learners need to engage in its discourses; as Gee aptly put it, "all writing is embedded in some Discourse" (1989, p. 11; *Discourse* with capital D was used in the original source).

While it is true that "successful academic writing is more than just using a thesaurus and filling a paper with fancy sounding words" (Csomay & Prades, 2018, p. 108), using terms and expressions (e.g. *dental anxiety, we explore, on the other hand*) that are used in a particular discipline, and in the academia in general (interdisciplinary), can help EFL learners increase the sophistication of their texts. Some studies have shown that expert writers tend to rely more on collocations than novice writers, and that this use of fixed expressions is often considered a marker of proficiency and fluency in academic writing (Granger, 1998; Nesselhauf, 2005). For this reason, measuring the degree to which EFL learners produce academic vocabulary can determine: 1) their proficiency level and 2) their linguistic development. Laufer and Nation (1995) introduced the concept of Lexical Frequency Profile (LFP) that measured variation (number of different words), density (number of content words), and sophistication (number of academic words) in a given text. The present study explores writing sophistication in a learner corpus. Both the fact that students need general academic vocabulary due to the interdisciplinary nature of their programmes, but also discipline-specific vocabulary to fully develop their disciplinary voice has motivated the author of the present study to investigate the extent to which students draw on general academic vocabulary on the one hand, and more specific vocabulary of their discipline on the other hand.

5.3.2 Academic vocabulary lists

While there is an ample range of corpus-based studies that have performed lexical analyses using academic vocabulary lists (Coxhead, 2017; Durrant, 2016; Laufer & Nation, 1995) in academic writing, the study of academic collocations and formulas is yet to be exploited. As Granger points out, "phraseology is now recognised as a major component in general L2 learning and teaching. In the specialised field of academic literacy, however, the phraseological dimension has yet to establish itself as a core facet" (Granger, 2017a, p. 22). Academic language can be highly patterned (Römer, 2011) and thus analysing EFL learners' phraseological profile can also help researchers to uncover new learner writing features. In fact, Cortes (2004) found that university students in her study rarely used target bundles in their texts, compared to professional writers in biology and history, even though they were exposed to these expressions in their readings.

There are three corpus-based lists that have recently been developed using large academic corpora, text analysis tools, and different statistical tests to retrieve 1) words, 2) collocations, and 3) formulas:

1) The Academic Vocabulary List (AVL) (Gardner & Davies, 2014) draws from a 120 million words academic subcorpus of nine disciplines (mostly research papers) from the COCA corpus, and contains 3,015 lemmas (e.g. *system, social, however*). I support the authors' view that the AVL reflects academic words more accurately than the Academic Word List previously developed by Coxhead (2000), since it pulls from a larger and more recent corpus; also, the fact that the list is lemma-based, and part-of speech tagged makes it more relevant for EFL teachers and learners. The full list is available at: https://www.academicvocabulary.info

2) The Academic Collocation List (ACL) (Ackermann & Chen, 2013) comprises 2,468 cross-disciplinary academic collocations extracted from the 25 million words Pearson International Corpus of Academic English (PICAE) (e.g. *academic writing, brief overview, crucial factor)*. The list is available at: https://pearsonpte.com/organizations/researchers/academic-collocation-list/)

3) The Academic Formulas List (AFL) (Simpson-Vlach & Ellis, 2010) draws from different corpora, such as the Michigan Corpus of Academic Spoken English

(MICASE), and BNC for spoken academic English, and Hyland's 2004 corpus of research articles for written academic English, totalling 4.2 million words. The AFL contains 607 most frequent formulaic sequences of 3-, 4- and 5-grams, subdivided into academic spoken English (e.g. *be able to, this is the, you can see*), academic written English (e.g. *on the other hand, due to the fact that*), and a core list with formulas that are common in both academic written and spoken English (e.g. *in terms of, at the same time, from the point of view).* Combining Mutual Information (MI) scores, frequency, and manual scoring by experts, Simpson-Vlach and Ellis (2010) also created a metric called "formula worth teaching" and included formulas organized into discourse-pragmatic categories (e.g. 'contrast and comparison': *as opposed to*). The AFL can be found in Simpson-Vlach and Ellis' (2010, p. 37).

Nowadays there are useful software packages such as the 'AntWord Profiler' (Anthony, 2014), the 'Web Vocabprofile' (Cobb, 2002), the 'Wordandphrase' (Davies, 2012), or the 'Lexical Complexity Analyzer' (Ai & Lu, 2010) that can help to identify academic language and measure sophistication in a given text; these tools, however, look at isolated words, very often classifying them according to their frequency band, calculating type/token ratios, or using one pre-set list of academic words. In the present study, however, not only single words from the AVL, but also academic collocations and formulas from the ACL and AFL respectively have been tracked in order to calculate coverage in the CBI course materials, and usage in the EFL learners' texts, analysing the proportion of academic vocabulary compared to non-academic vocabulary (see Section 5.3 for more information about the analytical procedure). Table 1 summarizes tokens and types of each list, and the total number of words.

Table 1: the AVL (Gardner & Davies, 2014), ACL (Ackermann & Chen, 2013) and AFL (Simpson-Vlach & Ellis, 2010) tokens and types

	AVL	**ACL**	**AFL**
No. Lists	1	1	3
Tokens	3,015	2,468 (entries); 4,936 tokens	607 (entries); 2,025 tokens
Types	3,015	1,302	330
Total tokens		9,976	

With the aim of measuring the extent to which students have incorporated academic and disciplinary discourse from the CBI materials they were exposed to during the course, three research questions have been formulated:

1) To what extent do the materials used for the CBI course include general academic vocabulary? A relatively high coverage would be expected due to the academic nature of the subject.

2) What effect does the CBI course have on students' academic vocabulary production? It is hypothesized that there would be a more frequent use of academic vocabulary in the texts written after the course (T2) as a positive effect resulting from the instruction received in the CBI course.

3) Whose vocabulary has benefitted the most from CBI: EMI or L1 students'? Texts produced in the EMI setting are expected to show a somewhat higher production of academic vocabulary than texts in the L1 group, due, in part to a higher exposure to the target language in an academic context in their studies.

5.4 Data and methodology

5.4.1 Context

The study took place at the Dentistry Faculty of a Spanish university. The Dentistry bachelor degree is a five-year degree programme that offers two parallel settings of instruction called the "English track" and the "L1 track". In the former, all courses in the first two years of the degree are taught through EMI, equalling 600 EMI hours per academic year. On the other hand, in the latter setting, courses are taught in Catalan or Spanish throughout the degree. Regardless of the instruction setting, there are three courses, namely *English for Dentistry 1* (first year), *English for Dentistry 2* (fourth year), and *English for Dentistry 3* (fifth year) that are taught in English. We will focus only on *English for Dentistry 1*, since this is the course in which the research was carried out.

5.4.2 The CBI course

English for Dentistry 1 is a one-semester course (60 hours of class time) for first-year students enrolled in the Dentistry degree. This course follows an CBI 'adjunct model' approach, in which the instructors, native and non-native speakers of English who are certified language specialists, have been trained in the content of the course through collaborations with the dentistry department and pursue language learning objectives, which are intrinsically linked to the disciplinary content of other subjects taught in the same year. The course includes reading and listening activities aimed at providing students with the language competences necessary to understand and present basic aspects of dental research in English. In terms of content, the course explores different types of research (e.g. experimental vs. non-experimental), as well as common study design features (e.g. randomized, controlled, blinded trials), and pays attention to high-frequency dental terminology related to oral health conditions, dental instruments, and the most common treatments. Apart from quizzes and exams, one of the main projects consists in replicating a population study in which students carry out a survey on different areas of the dentistry field, compare the results with the original study, and present it orally to the class. As for materials, the language specialists developed a student dossier that contains readings (e.g. academic abstracts from published articles, texts on dental conditions and different types of research, practical explanations on how to write academic abstracts, dental histories, or present research orally), activities that were regularly done in class (e.g. comprehension questions on the abstracts, exercises to practice writing the sections of an abstract, turning informal language into formal and more appropriate expressions, dental vocabulary matching exercises, etc.), and finally lectures on different topics related to dental health with the support of PowerPoint slides (e.g. dental anxiety, differences between *abfraction*, *abrasion* and *attrition*, a randomized controlled trial on the effects of herbal tea on enamel, etc.). These classroom materials, i.e. the student dossier and the PowerPoint presentations, have been used in the analysis.

5.4.3 Participants

The participant sample comprises 56 first-year students enrolled in the Dentistry degree. There are two different groups: 1) students enrolled in the "English track" instruction setting –we will refer to these as the EMI group (N=26)–, and students who have most courses in Catalan or Spanish (except for the *English for Dentistry 1* course, which is taught in English in both settings) –we will refer to these as the L1 group (N=30)[7]. This

[7] All texts, regardless of the setting of instruction in which they were produced (i.e. EMI or L1), were originally written in English.

sample reflects the internationality of the university: data comes from both male (N=18; 32.1%) and female (N=38; 67.8%) students, aged 18-23, from seven different mother tongue backgrounds: Spanish-Catalan (46.4%), French (23.2%), Arabic (10.7%), English (8.9%), German (7.1%), Greek (1.7%), and Russian (1.7%). All participants (with the exception of native speakers of English) were given a level test (the SIMTEST developed by Universitat Autònoma de Barcelona)[8] at the beginning of the academic year to assess their proficiency level in English according to the Common European Framework of Reference (CEFR). Three different levels were found: A2 (N=6, 11.7%), B1 (N=14, 27.4%), and B2 (N=32, 62.7%). The two settings of instruction had similar spreads of proficiency levels.

5.4.4 Instrument and data collection

With the intention of collecting data to analyse students' academic writing performance before and after the CBI course, a writing task was developed and included as a classroom activity (see Appendix 1). It comprised four questions:

- Look at the following image, and respond to the questions about it below:
    1. Describe the scene shown in the image. What do you think has just happened?
    2. Write a possible dialogue among the people shown in the image.
- Answer the following two questions (write in paragraph style):
    3. What would you do in this situation?
    4. How could you determine whether your approach is the best one for this situation?

These questions were formulated in order to prompt the use of different types of academic language; for example, questions 1 and 2 could make the student use more descriptive and discipline-specific language such as clinical vocabulary, dental conditions, and/or doctor-patient communication, while questions 3 and 4 could make the student use more cross-disciplinary academic language to show critical thinking, stance, and/or refer to scientific evidence.

[8] https://simtest.uab.cat/simtest/

This study has adopted a longitudinal pre-test post-test design over one semester, including two data collection times: the exact same writing task was done in class twice, i.e. at the beginning (week 1=T1) and at the end of the *English for Dentistry 1* course (week 17=T2). The instructors allowed 20 minutes for the task completion. The texts were then collected and manually typed in order to create the corpus of learner writing. Only those texts present at both T1 and T2, and that contained more than 150 words, were included for the analysis.

5.4.5 The corpus

For the purposes of this study, two different corpora were compiled:

1) The learner corpus: it consists of 112 texts written by first-year dental students (33,854 total words) collected before and after the *English for Dentistry 1* course. These texts fall into three main subcorpora: EMI students' writings (n=42), L1 students' writings (n=60), and English native speakers' (NS) writings (=10). The NSs are first-year dental students enrolled in the EMI setting; since these students have attended the *English for Dentistry 1* course, and, at the time of the study, had been exposed to the same academic input for two semesters, their texts have been included in the analysis for comparative purposes.

2) The class material corpus: three subcorpora were created in order to differentiate between pedagogical (instructions) and more discipline-oriented language (readings and lectures), as has been done in previous studies (O'Loughlin, 2012; Wood & Appel, 2014).[9] The class material corpus represents a substantial part of the content and language input students have been exposed to. The procedure to create the class material corpus involved two important steps: (1) converting the student dossier and the PowerPoint slides, together with the instructors' notes, into raw txt. files, in which tables, figures, images, etc. were removed from the text; and (2) classifying these materials manually in order to create three different subcorpora: the *reading input* subcorpus, which consists of all the abstracts, academic texts, theoretical concepts and explanations present in the students' dossier; the *supplementary input* subcorpus, which

[9] This was motivated by the intention to explore both *generic* and more *specific* academic discourse in EFL learner writing. At first, I considered using an already compiled and validated academic list of discipline-specific terminology, as I did for the general academic discourse, and I found Lei and Liu's (2016) 'Medical Academic Vocabulary List' (MALV). However, this wordlist is based on specialist areas and journals that do not include dentistry, and was therefore discarded.

contains all the exercises, comprehension questions, and instructions that are also included in the dossier; and finally, the *listening input* subcorpus, which consists of the PowerPoint slides used in class and the instructors' notes used for these PowerPoint presentations –as these were the only source of *listening* input.[10] This class material corpus contains 56,708 words in total. Tables 2 and 3 show the total number of texts, tokens, and types in each corpus.

Table 2: The learner corpus

	EMI		**L1**		**NS**	
Time	T1	T2	T1	T2	T1	T2
No. Texts	21	21	30	30	5	5
Tokens	8,297	8,405	7,406	5,935	1,884	1,927
Types	1,089	1,161	1,010	890	570	585
Mean text length	395	400.2	246.8	197.8	376.8	385.4
Total tokens	16,702		13,341		3,811	

Table 3: The class material corpus

	Reading Input	**Supplementary Input**	**Listening input**
No. Texts	1 (dossier)	1 (dossier)	21 (presentations)
Tokens	19,789	20,570	16,349
Types	3,382	3,226	2,452
Total tokens		56,708	
Total types		5,484	

5.4.6 Dentistry-specific lists

Additionally, and in order to see to what extent students drew on discipline-specific vocabulary they were frequently exposed to through the class materials, three additional lists were generated: first, a *vocabulary* list was created using Voyant (Sinclair & Rockwell, 2016) to identify the most frequent words in the class material corpus. Second, Collocate 1.0 (Barlow, 2004) was used to automatically extract the most frequent collocations in the corpus by means of the Mutual Information (MI) test. Third, AntConc (Anthony, 2018) was used to identify recurrent word combinations (i.e. formulas) of 3, 4 and 5 words, with a minimum frequency of 10 hits, in order to create a

[10] This division was made for comparative purposes only so that coverage of the generic academic lists could be looked at separately.

formulas list. A manual screening of these lists was required in order to merge plural words (e.g. *patient, patients*), and to eliminate overlapping formulas (e.g. *the case of, in the case of*) in order to prevent inflated results. In addition, for the creation of these lists, no distinction was made with regards to the *reading, listening* or *supplementary* part of the materials, as these were naturally integrated in the course; in other words, in a normal class, the professor would use a PowerPoint to present the content (*listening*), after that, students would often read an abstract (*reading*), to later answer comprehension questions (*supplementary*). Nevertheless, and as it could be anticipated, there were some items from the class material lists that coincided with items from the general academic lists, specifically 139 items: 86 words (e.g. *condition, abstract, anxiety*), 2 collocations (*experimental research, increased risk*), and 51 formulas (e.g. *associated with the, the relationship with*). As previous studies have pointed out, the boundaries between general and disciplinary academic discourse are difficult to operationalize and often overlap (Green & Lambert, 2018; Paquot, 2010). In the analysis, however, only 45 of these duplicated items were found in the learners' texts (i.e. 27 words and 18 formulas); since they represent both general academic and discipline-specific discourse, I decided to keep them on –and count them for– both lists. Table 4 indicates tokens and types and total number of words for the vocabulary (VL), collocations (CL), and formulas (FL) lists derived from the class material corpus (see Appendix 2 for the top-50 entries in each of these lists).

Table 4: lists from class material corpus

	VL	CL	FL
No. Lists	1	1	1
Tokens	279	454 entries/ 908 tokens	499 entries/ 1,597 tokens
Types	279	300	349
Total tokens		2,784	

5.4.7 Tools and analysis

The web-based text reading and analysis environment Voyant tools (Sinclair & Rockwell, 2016) was used to calculate the number of tokens and types of the different corpora. Subsequently, R package (Rstudio, 2012) –Quanteda–[11] was used in order to

[11] https://quanteda.io

track the occurrence (i.e. frequency and range) of items from the various lists explored (i.e. AVL, ACL, AFL) in the corpora, determining 1) the extent to which the class material corpus includes items from these vocabulary lists, and 2) the proportion of both general and discipline-specific academic vocabulary in the learners' text. Finally, two statistical non-parametric tests, i.e. Mann Whitney U and the Wilcoxon signed rank test, were performed in order to detect if there were significant differences across groups (EMI, L1) and times (T1, T2) respectively. The analyses that follow are based on mean usage (%) per text, which means that text size does not affect the results.

5.5 Results and discussion

In this section I first explore the coverage of academic vocabulary in the class material corpus, to later examine the proportion of general and discipline-specific academic vocabulary in the learner corpus across times and instruction settings. All the examples given have been taken from the various corpus analyses. Results show significant links between time and an increased used of academic words and other items from the lists. The effects of the CBI course on students' academic vocabulary production according to their setting of instruction are discussed.

5.5.1 Academic vocabulary coverage in the class material

The class material (CM), considered as a single corpus, offers different levels of coverage for the academic vocabulary (AVL), collocations (ACL) and formulas (AFL) lists, which range from 5.5% to 64.7%. As Table 5 shows, the list that is more extensively represented in the CM corpus is the AFL core (i.e. formulas that are frequent in both spoken and written academic English), which may confirm the blend of pedagogic and disciplinary discourse included in the materials. In particular, the listening input subcorpus –that is, the PowerPoint presentations used by the instructors and their notes– contains a slightly higher number of items from the AFL lists in general than the reading and supplementary input subcorpora: the speaking notes allow the instructors to deliver student-centered explanations, mostly through formulas (e.g. *this type of, in other words, an example of*), while the slides often display written disciplinary content (e.g. *evidence*, *the effects of, factors such as*). Additionally, the second most broadly covered list in the CM corpus is the AFL written (45.5%) (e.g. *to determine whether, with regard to, carried out by*) followed by the AVL (32.8 %) (e.g. *study, research, data*) and the AFL spoken (21.5%) (e.g. *as you can see, let's look at,* *this kind of*). Curiously enough, the CM only includes 135 of the 2,468 collocations present in the ACL (e.g. *collect data, experimental research, casual relationship*).

Table 5: coverage of AVL, ACL, and AFL in the class material corpus

	AVL	ACL	AFL		
			Core	*Written*	*Spoken*
Reading input	22.1%	2.7%	38.2%	24.0%	10.0%
Listening input	17.5%	2.0%	39.1%	25.0%	11.5%
Suppl. input	22.7%	2.2%	38.6%	17.5%	10.5%
Class material[12]	32.8%	5.5%	64.7%	45.5%	21.5%

Table 6: raw frequencies and types of AVL, ACL, and AFL items in the CM

			AFL				
Lists and length	**AVL** 3015	**ACL** 2468	*Core* 207	*Written* 200	*Spoken* 200	**Total types**	**Total freq.**
Reading input	666	65	79	48	20	878 (14%)	3,409
Listening input	529	49	81	50	23	732 (12%)	3,145
Suppl. input	686	53	80	35	21	875 (14%)	3,617
Class material 56,708 tokens	989	135	134	91	43	1,392 (25.3%)	10,171 (18%)

In terms of frequency, the materials as a whole offer variety and repetition, as can be seen in Table 6: the results show that almost 20% of the tokens in the CM could be classified as academic (i.e. belong to any of the lists explored). In other words, a student who has read the texts, listened to the lectures, and performed the tasks in the supplementary material, would have encountered 1,392 different interdisciplinary academic words, at least 10,171 times, over the one-semester course. This input is richer in items from some lists (e.g. AFL core, AFL written, AVL) rather than others (e.g. ACL), but it still shows that there can be a useful set of academic vocabulary frequent across disciplines, which somehow contrasts with what other studies have suggested (Ha & Hyland, 2017; Hyland & Tse, 2007). This input to academic discourse would be

[12] The percentages of the three categories (reading, supplementary and listening) shown in the first three rows of the table do not total that of the class material; this is because these subcorpora were looked at separately (as separate texts). For example, if the collocation *additional information* appears in both the Listening Input subcorpus *and* in the Reading Input subcorpus (no mater how frequently –only types are taken into account in order to calculate coverage) these are counted separately, and this is what the first three rows show. On the other hand, when looking at the class material as a whole, even though *additional information* appears in both the reading/listening input subcorpora, it counts as one type, so the percentages in this last row represent the exposure of general academic vocabulary the CM as a single corpus (reading, supplementary, listening) provides students with.

even greater if we took into account the technical vocabulary typical of the dentistry field (e.g. *gingiva, maxillary, temporomandibular joint syndrome*), and also words of general meaning that may have academic meaning in the corpus (e.g. *pain, patient, tooth*). Whether this exposure has been sufficient to make students use academic words and expressions in their texts, and also the production of these items before and after the CBI course will be analysed in the next section.

5.5.2 Academic discourse in the learner corpus

In general terms, academic language represents between 20.1% and 26.9% on average of the tasks written by learners, which could be considered between the normal range for academic texts (10%-30%) described by Coxehead & Nation (2001). In addition, the extent to which items from the general academic lists (AVL, ACL, AFL) and from the discipline-specific lists (VL, CL, FL) have been used varies depending on instruction setting and the time of the task, as can be seen in Table 7; the texts written by English native speakers (NS) have been included for comparative purposes. Results show that texts written by EMI and NS students contain a higher percentage of academic language at T2, whereas the opposite tendency occurs in the case of L1 students, in which the average decreases slightly at T2. Interestingly, results also show that the use of academic language in general increases by 9.1% at T2 for EMI students, which is even greater than the increase found in the NS texts (6%). As can be seen, L1 texts at T2 contain slightly fewer words that pertain to the academic lists (-2.6%) however.

Table 7: Academic language usage in the learner corpora at T1 and T2

		EMI			**L1**			**NS**	
	T1	**T2**	**Var.***	**T1**	**T2**	**Var.**	**T1**	**T2**	**Var.**
AVL	3.6%	3.8%	7.1%	2.5%	2.5%	0.6%	5.0%	6.0%	19.2%
ACL	0.1%	0.0%	-66.2%	0.0%	0.0%	-100.0%	0.0%	0.2%	Inf
AFL core	0.7%	0.6%	-17.4%	0.7%	0.6%	-20.8%	1.0%	0.2%	-74.6%
AFL written	0.2%	0.2%	4.4%	0.2%	0.2%	-22.4%	0.3%	0.6%	79.3%
AFL spoken	0.6%	0.6%	5.6%	0.5%	0.3%	-34.2%	0.6%	0.6%	-6.9%
Vocabulary	8.3%	9.2%	11.8%	7.8%	8.2%	5.0%	8.7%	9.6%	10.5%
Collocations	7.7%	8.5%	9.9%	7.9%	7.4%	-6.8%	8.5%	8.6%	0.6%
Formulas	0.9%	1.0%	15.1%	0.8%	0.8%	-0.8%	1.1%	1.1%	0.1%
TOTAL	**22%**	**24.1%**	**9.1%**	**20.6%**	**20.1%**	**-2.6%**	**25.3%**	**26.9%**	**6.4%**

*Variation: comparison between T1 vs. T2; green cells indicate increase in the academic vocabulary production; red cells indicate decrease.

Regarding the production of specific items from the eight lists explored, discipline-specific vocabulary (max. 9.6% NS at T2 – min. 7.8% L1 at T1) (e.g. *dental, pain, examination*), collocations (max. 8.6% NS at T2 – min. 7.4% L1 at T2) (e.g. *dental treatment, oral cavity, oral hygiene*), and words belonging to the AVL (max. 6% NS at T2 – min. 2.5% L1 at T1 and T2) (e.g. *approach, important, need*) represent the most popular academic items used by learners, regardless of their setting of instruction, their speaker status, and the time of the task. The AVL was the second most represented list in the CM; the exposure provided by the CM may explain why students have included some of these items in their texts, while it also supports the usefulness of the AVL for pedagogical purposes.

As for the remaining lists, there seems to be no or very low frequency of items from the CM formulas list (max. 1.1%) (e.g. *to the dentist, to make sure, to assess the*), the AFL lists (max. 1%) (e.g. *we can see, you need to, in order to*), or the ACL (max. 0.2%) (e.g. *facial expression, clearly identified, positive impact*). Figure 3 illustrates the presence of the eight lists in the learner corpus, according to setting of instruction and time. As can be seen, vocabulary and formulas extracted from the CM, and items from the AVL are the most frequent ones in the texts on average, and also the ones that present greater variability at T2 (a more noticeable increase). Items from the remaining lists that have been scarcely produced by learners in the writing task appear at the bottom.

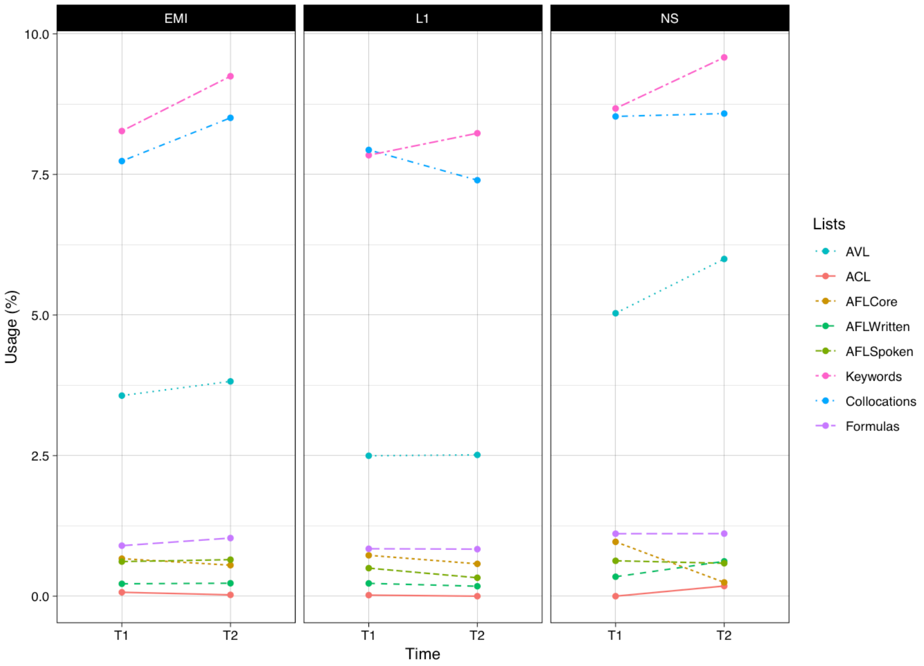


Figure 2: presence of the AVL, ACL, AFL, and CM in the learner corpora at T1 and T2

As has often been reported in the literature, it is not the technical vocabulary but the general academic words that pose greater difficulties for learners (Durrant, 2016; Granger, 2017a). In this study, learners have used words and collocations that were more technical and were present in the materials more often than interdisciplinary academic vocabulary (except for the AVL), which corroborates the previous statement. This finding supports previous studies that emphasise a high degree of specificity and technicality in the vocabulary of the disciplines (Ha & Hyland, 2017). On the other hand, the results show that the frequent exposure to academic formulas provided by the materials has not been enough for students to use them in their texts, which is in line with Cortes' (2004) and Wood and Apple's (2014) findings on high exposure and low production of formulas by university students. Even though the AFL core list was the most represented list in the CM (it covered almost 40%), these items have been barely used by the learners (max. 1%). Therefore, these findings stress the need for more explicit pedagogical attention to the use of academic formulas in general, and to dentistry-specific formulas in particular (e.g. *risk factor for, tooth surface loss, oral health care*). On the other hand, collocations from the ACL were barely used by the learners (only 4 items); these collocations were also very scarcely covered by the class materials (max. 2%), which may explain the low presence of these items in the learner corpus. The ACL seems to be, at least in the case of this CBI course in dentistry, less pedagogically relevant than other lists of interdisciplinary academic vocabulary (such as the AVL).

In terms of improvement, i.e. an increased number of academic words, collocations, and formulas in the texts written after the course, Table 7 above shows how more discipline-specific words have been produced at T2 on average by all groups of learners (EMI, L1, NS). The average production of these keywords seems to be even higher for the EMI group –almost 12% more keywords on average than at T1. This general increase of discipline-specific vocabulary production for all groups could be due to the emphasis given to those words throughout the materials, and to the explicit teaching of vocabulary in the CBI course, which may be pointing as well towards an important short-term benefit of the CBI adjunct model. In addition, the reiterated encounters with this specialized lexicon EMI learners may have had in other subjects of the degree during that semester may also explain the greater increase we see in this group. Another list that seems to be present in all texts, and with a greater presence at T2, is the AVL. This may indicate that all learners, regardless of their instruction setting, have improved their academic lexicon, and have started to use interdisciplinary academic words in their texts, shaping their conceptual knowledge, and starting to develop their academic voice.

In order to know if the differences found are significant across times and groups, two non-parametric statistical tests were performed: i.e. Mann Whitney U test and Wilcoxon signed rank test, as shown in Table 8. In terms of the intra-group comparison, no statistically significant differences were found between T1 and T2, except for the decrease in collocations in the L1 group (*p*= .01). On the other hand, inter-group comparisons between EMI and L1 groups at T1 and T2 show statistically significant differences for the use of AVL items, both at T1 (*p*= .04) and T2 (*p*= .01), the AFL spoken items at T2 (*p*= .00), and vocabulary and collocations from the CM at T1 (*p*= .00).

Table 8: inter- and intra-group comparisons across time (tests for significance value)

	Mann Whitney U test				**Wilcoxon Signed Ranked test**			
	EMI vs. L1 (T1)		**EMI vs. L1 (T2)**		**EMI** **T1 vs. T2**		**L1** **T1 vs. T2**	
	Z	*p*	Z	*p*	Z	*p*	Z	*p*
AVL	421	.04*	447	.01*	84	.28	223	.85
ACL	368	.09	345	.09	20	.35	3	.37
AFL core	295.5	.71	339	.65	137	.47	274	.40
AFL written	330	.76	371	.19	50	.77	69	.62
AFL spoken	364	.34	464	.00*	97	.53	149	.10
Vocabulary	491	.00*	555.5	4.2	60	.09	274	.10
Collocations	469.5	.00*	566	1.6	70	.19	335	.01*
Formulas	432	.02	517	9.3	44	.39	221	.11

Note: value is significant if p< .05
*Significant

5.5.3 CBI course effects

In general, EMI students seem to have benefitted the most from CBI instruction, as they show a greater production of general academic (AVL) and discipline-specific words, collocations, and formulas at T2, which are statistically significant. Since the task was timed, having a look at the number of tokens produced in that time, and the relationship between tokens and types, can provide an approximate idea about the learners' fluency and sophistication in writing respectively (see Lu, 2012 for an examination on lexical richness): while L1 students produced shorter (and less varied) texts (246 tokens/ 123 types at T1 vs. 197 tokens/ 108 types at T2 on average), EMI students wrote longer (and more varied) texts at T2 (395 tokens/163 types at T1 vs. 400 tokens/177 types at T2 on average), as can be seen in Table 2 in the previous section. In terms of length and variation, NS texts vary slightly at T2 (376 tokens/ 183 types at T1 vs. 385 tokens/ 189 types at T2 on average).

In order to observe the use of academic vocabulary at T1 and at T2 from a more qualitative perspective, three sample answers to question four of the writing task were extracted from the learner corpus:

(1) [11_FR_L1_T1][13]: I think that if my approach is good, the baby won't cry or won't look afraid/stressed/sad; he will not move as if he was in danger and he would be calm; so I think that my approach will have an immediate impact on the baby and he will respond (in a good or bad way) to what I do to him and the way I do it; usually a parent is present so the mother or father will see what you do and maybe give you clues to approach the baby positively;

[11_FR_L1_T2]: **Firstly**, I think that body language and also face expressions are a very good way to **analyze** other people's feelings, so in this situation, if the baby stays calm, relaxed, if he is not breathing super quickly, not sweating or anything else and if moreover he is obersving [sic] me with attntion [sic], I will know that my approach is not so bad. **Moreover**, I would try to talk with him and so I can see how he answers to me, if he is still shy or not, if he says positive things. Obviously, if the baby is crying or shouting. [sic] I will know that I didn't do enough to make him feel good-at-ease, and it's a failure, because as a

[13] Nomenclature indicates: students' identification number, mother tongue, setting of instruction, and time of task.

health professional, you are supposed to **be able to** use **psychology** with young **patients** and one of the **technique** [sic] that can be used to **evaluate** my approach is definitively using **questionnaires**; for young children.[sic] I would use a questionnaire very simple, with a few words, and smileys to evaluate their feelings, and for the parents a more **complete** questionnaire.

(2) [08_GR_EMI_T1]: To determine that my approach is the most appropriate one I would ask for a follow up appointment after the initial visit. I would evaluate if the instructions and the treatment method were effective for the patient, if they weren't then I would change the plan of action and request another follow up appointment. If the treatment was successful however I would ask the patient or the guardian if the patient is young to call me and report any complications that might arise.

[08_GR_EMI_T2]: In order to determine if my approach is the correct one I have to do some things. **Firstly** I would ask other dentists that I know what is their approach and then compare it to mine. If mine is very far off from all their approaches then I must be doing something wrong. Then I would read **dental literature, case studies** and **experiments** on what is the correct way to illustrate to the child and parent how the child should brush his teeth. The **articles** need to be **peer-reviewed** in order to get transparent and **rigorous results** that I can then trust. Trustworthy articles and **techniques** are really **important otherwise** my **dental work** would be compromised. **Furthermore** I could go to conferences and **observe** techniques from educated and well known [sic] dentists that will improve my technique and approach.

(3) [02_NS_EMI_T1]: There is no quantifiable way of judging which approach is the best. If at the end of the day the tooth is fixed and the child is as calm and happy as he can be in a dentist's office, I would say that it was a successful approach.

[02_NS_EMI_T2]: There is no definitively correct way of teaching someone, especially someone from an age group as characteristically versatile as children, how to brush their teeth. **However**, what should be present in all dentists is that they should not use **medical** vocabulary, **instead** substituting them with child-friendly words so that the patient doesn't get overwhelmed. **Also**, the dentist should be warm and welcoming, so that the child doesn't feel stressed or afraid, will be receptive to what the dentist is saying, and will look forward to dentists' visits as much as possible.

In examples (1) (2) (3) we can see how students have used new words and expressions at T2 (highlighted in bold) that were not present in their T1 and that belong to some of the lists explored. It is interesting to note, for example, the use of connectors to structure the answer at T2 in (1) and (2), the reference to evidence-based literature to contrast approaches to dental practice in (2), or a more elaborate and reader-oriented answer that includes examples and recommendations, as well at T2, in (3).

Furthermore, if we look at the data sample through individual variation plots, interesting results arise. In Figures 3 and 4 each line represents one student in the sample; red lines in the left column show students whose texts include fewer academic words and expressions from the lists explored at T2; green lines in the right column represent on the other hand students who have produced more academic vocabulary at T2. The distance between lines represent the percentage these academic words have with respect to the total number of words in their texts.

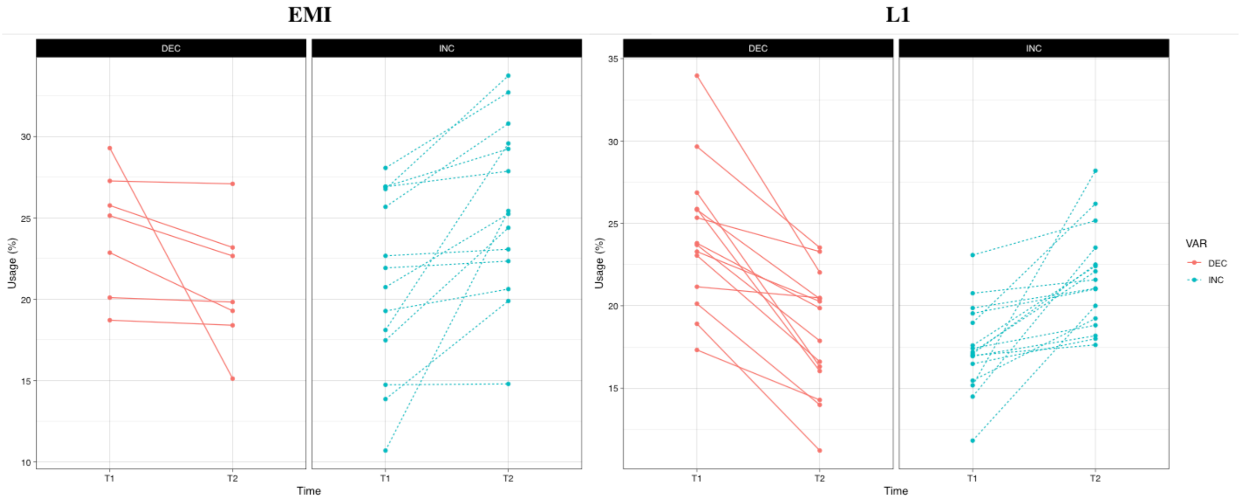


Figure 3: individual variation in EMI and L1 at T1 and T2 (decrease vs. increased use of academic language)

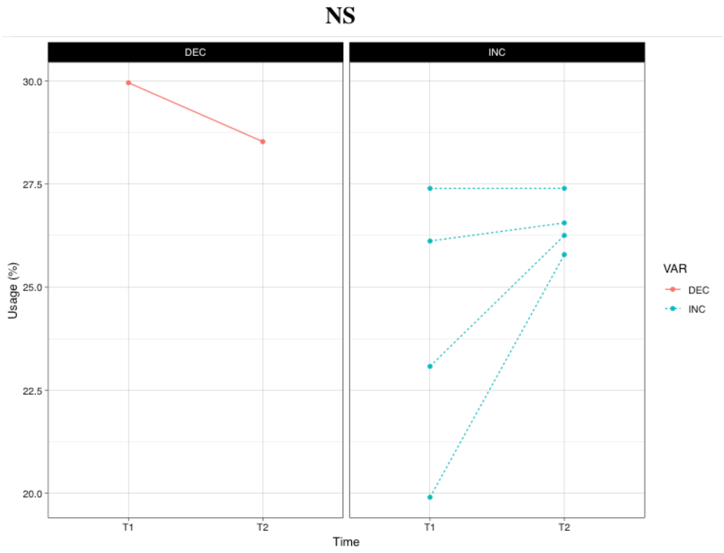


Figure 4: individual variation in NS at T1 and T2 (decrease vs. increase)

As can be seen in Figure 3, there is a notable difference between EMI and L1 texts in terms of increase. In EMI, 14 out of 21 students (66%) produced more academic words at T2, and the ones who did not, remained almost the same as in T1 (all except one, whose use of academic words dropped considerably). In the L1 group, only half of the students improved, more specifically 16 out of 30 (53%), and the other half seem to have produced noticeably fewer academic words than in the T1. Regarding the distance between the lines, the EMI group shows a greater heterogeneity, whereas students in the L1 group seem to have performed more similarly in terms of percentage of academic discourse. These individual variation plots clearly display a greater improvement for the EMI group. Finally, and as can be seen in Figure 4, NS texts contain a higher percentage of academic discourse at T2 overall (note how the usage percentage is higher) and this production is greater in 4 out of 5 students, which shows that not only L2 learners but also native speakers of English also benefit from discipline-specific CBI approaches, which corroborates both the need and the usefulness of such training.

5.5.4 Limitations and further research

A potential limitation of the present study is that the data collected represent six months of exposure to CBI. Repeating the study at the end of the next academic year would allow us to explore these learners' academic vocabulary development and retention more accurately. This would for example allow us to investigate why students, in some cases, used fewer items from some lists at T2 –discipline-specific collocations in the L1 group, and formulas from the AFL in the NS group, in particular, which merits further investigation.

In addition, even though the CBI course materials covered a high percentage (in some cases) of items from lists of general academic terminology (e.g. AVL, AFL), exposure alone did not have much effect on students' production of those items (especially for the production of formulas or lexical bundles), which corroborates the need for explicit instruction and the inclusion of writing tasks so students can improve their academic writing abilities. Further analyses could be carried out on the extent to which such explicit instruction has an effect on students' written production.

This study would have also benefitted from the use of parallel corpora (i.e. texts written in the students' L1) to compare the amount of academic vocabulary that transfers from L1 to L2 and vice versa. In addition, this study has investigated the use of academic words, collocations and formulas from validated corpus-informed lists, and from the materials used in class. However, compiling a corpus of expert writing in dentistry and measuring the most frequently used academic vocabulary in actual research papers could also be pedagogically relevant. Finally, analysing the extent to which a higher or lower proportion of academic terminology in a text correlates with higher or lower syntactic complexity or with higher or lower scores would also be something worth investigating.

5.6 Conclusion

This study has sought to measure the effects of a CBI course on students' academic vocabulary production, and has examined three research questions. First, the coverage of academic vocabulary from three different lists (namely the AVL, the ACL, and the AFL) has been calculated for the materials used in the course. The results show that the CM offers substantial coverage of the academic language present in these lists. Almost 65% of the formulas included in the AFL core are provided by the CM, which highlights the academic and pedagogical nature of these materials: this partially confirms the initial hypothesis, which was that the CM would provide a substantial coverage of the academic terminology included on the lists. However, the author did not expect to find such a limited presence of items from the ACL, which certainly affected students' exposure, and could explain the low presence of these items in the learners' texts. Materials used for CBI should offer variety and repetition, not only in terms of academic words and formulas, but also in terms of academic collocations.

Secondly, the effects of the CBI course on students' academic language production were measured by performing a pre-/post-writing task. Results show that texts contain more discipline-specific vocabulary at T2 on average, and also more items from the AVL. This finding might suggest that CBI instruction, and the adjunct model in particular, could be beneficial for both generic and discipline-specific vocabulary learning and production in the short term (one semester).

Finally, academic discourse improvement according to two settings of instruction (i.e. EMI and L1) was analysed. The findings show that EMI have produced more discipline-specific vocabulary, collocations and formulas than their counterparts in the L1 group, and also more general academic vocabulary; this difference has been statistically significant. This confirms the hypothesis that more exposure to the target language in an academic context would have created more opportunities for direct/ incidental learning of academic terminology for the EMI group; the more widely spaced input of the L1 group may account for the differences found. These findings corroborate the need for more pedagogical attention to academic collocations and formulas in particular, adapted to the needs of different learner populations.

Corpus-informed resources could help instructors of CBI programmes to select and prioritize certain vocabulary items, and this selection might include both, or progress from, more general (interdisciplinary) academic vocabulary to more technical, discipline-specific vocabulary. Academic vocabulary is just one aspect of the quality of writing, but it can provide a foundation for schemata development: if students are able to understand and use the terminology of a particular subject, they will very likely understand its theoretical concepts as well. This is of particular importance for EFL instructors and learners, since being aware of the different forms and usages of academic terminology in the disciplines can help them face the challenge of teaching and developing academic literacy in an L2.

6. Linking or delinking of ideas? The use of adversative linking adverbials by advanced EFL learners

6.1 Introduction

Corpus-based studies approached through contrastive analyses, and topics such as the use of connectors in English non-native and native compositions have been receiving much attention lately (e.g., Chen, 2006; Conrad, 2000; Liu, 2008; Peacock, 2010; Shaw, 2012). Connective devices (*on the other hand, alternatively, moreover*) play an important role in building discourse cohesion and in showing the author's intention in both written and spoken registers, and although most educators of English as a second or foreign language (ESL/EFL) devote time to teach these units at some point, non-native learners often struggle to use them appropriately (Biber, 2004; Granger & Tyson, 1996; Lei, 2012). The present study explores the use of adversative linking adverbial in argumentative essays written by advanced EFL learners with different language backgrounds. Our findings can add to the body of knowledge that describes the use of adversative linking adverbials in learner corpora and give pedagogical suggestions for the teaching and learning of these connective devices. First, corpus-based studies that compare the use of linking devices in the academic writing of English non-native students with English-native speakers or expert writers are reviewed. Sections 6.3 and 6.4 constitute the theoretical background of the study, in which the taxonomy of adversative linking adverbials, the underpinnings of learner corpus research, and the research questions formulated for this study are presented. Section 6.5 describes the context of the study, the comparability of the corpora, and the data analysis. Both quantitative and qualitative results of frequency and use of adversative linking adverbials are reported in Section 6.6 and later discussed in Section 6.7. Finally, some pedagogical suggestions for the teaching and learning of linking adverbials through corpus-informed materials are given in Section 6.8, which concludes the paper.

6.2 Literature review

Linking adverbials have received alternative names in literature: 'conjunctive' (Halliday & Hasan, 1976, p. 228), 'logical connectors' (Celce-Murcia & Larsen-Freeman, 1999, p. 519), 'discourse markers' (Bell, 2010, p. 1912), 'conjunctive adverbials' (Chen, 2006, p. 113), and 'linking adverbials' (Biber, Conrad, & Leech, 2002, p. 389). Although these terms refer to a similar linguistic construct, some authors integrate both adverbials and conjunctions in their analysis (the first four), while others (the last two) include items that belong to the adverbial category only (see Liu, 2008, p. 493 for a detailed discussion). We have adopted the latter, since it provided a systematic basis for analysis and a comprehensive list of the *adversative* type. Linking adverbials (henceforth LAs) and adversative LAs in particular, are one of the most common types of cohesive devices in academic prose (Biber et al., 2002). For example, *however* is the number one linking adverbial and it occurs over 100 times per million words in the academic register (ibid p. 393).

Non-native students that write academic texts in English (e.g. argumentative essays) seem to show preference for certain LAs that do not always coincide with the native or the expert's choice (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber, Conrad, & Cortes, 2004; Granger & Tyson, 1996; Rica-Peromingo, 2012; Swales, 2002). But not only disparity in the selection of LAs has been reported: an overuse of a limited set of these units when compared to a reference corpus has also been observed (Lei, 2012).

Corpus-based analyses have helped to bring to light learner language features that deserve a more qualitative look. For example, Granger and Petch-Tyson's (1996) study on connector usage in the French L1 subcorpus of the International Corpus of Learner English (ICLE) showed how not only overuse and underuse, but also misuse of certain connectors occurs when compared to an English native corpus, the Louvain Corpus of Native English Essays (LOCNESS), and to another comparable learner corpus with a different L1 –German. Apart from transfer-related issues, the authors found that French L1 learners seem not to recognize semantic properties and stylistic restrictions (e.g., too formal or too informal) most connectors have. This lack of sensitivity was attributed to a lack of proper training and exposure to LAs in context, which is also known as 'teaching effect'.

Similarly, Chen (2006) explored the use of 'conjunctive adverbials' in the final dissertations of Taiwanese MA students. In her study, both sentence-based and word-based calculations were made in order to detect over and underuse of these units when compared to a control corpus of professional writers. Surprisingly, each calculation provided different results: LAs were slightly overused by NNS under a word-based calculation, and the opposite tendency was reported (experts produced more LAs than NNS) when sentences were the basis for calculation. According to the author, the mean sentence length of the latter was longer than the ones produced by NNS, which could account for the discrepancies found. In the present study, both calculations will be made.

Leńko-Szymańska (2008) used three different corpora of argumentative texts in order to investigate 'connectors' in academic writing: LOCNESS as the novice native writers corpus; the International Corpus of Learner English (ICLE) as the non-native novice writers corpus; and the Freiburg-London-Oslo-Bergen corpus (FLOB), as the expert corpus of professional writers. The author found that there was a general overuse of connectors in both NNS and NS corpora in contrast with the professional writing corpus; surprisingly, this overuse seemed to be characteristic of novice writing (either native or non-native) rather than a distinctive feature of L2 production. In the present study we have performed a similar analysis of novice academic writing in English produced by NNS and NS writers in order to explore the use of adversative LAs. In addition, and as we will see in Section 6.6.2, two expert corpora (i.e. BCN and COCA) will be consulted in order to provide a third comparison for some cases of overuse found.

Another study that also investigated the production of 'discourse connectors' in argumentative compositions written by Thai EFL students compared to LOCNESS is that of Prommas and Sinwongsuwat (2011). The authors concluded that there was not a noticeable overuse of connectors in the NNS compositions, as in Leńko-Szymańska's (2008), and that there existed in fact an agreement on the top-5. It was in the placement of these connectors inter-clausually that Thai learners differed from the native production, however. The positioning of adversative LAs will also be analysed in the current study.

L1 transfer, or the preference for certain linguistic devices in a second language (L2) that resemble those in the L1, can also affect the choice of linking devices: the analysis of 'multi-word' linking adverbials in argumentative writing performed by Rica-Peromingo (2012) revealed that NNS students with different L1s from the ICLE corpus, and from a self-compiled corpus of Spanish EFL students (CEUNF) use multi-words units (e.g. *first of all, in spite of, on the contrary*) more often than native speakers in LOCNESS and a corpus of professional editorialists writing in English (SPE). The preference for these units and the differences in frequency compared to English native students and experts was explained through an L1-transfer scope. In the present study, different linguistic families will be taken into account when comparing the frequency and use of adversative LAs in the learner corpus.

The last study that is relevant to this paper is Lei's (2012). After exploring Chinese doctoral dissertations, the author found that in general learners used LAs more often than professional writers, but also that they relied on a more limited set. Interestingly, the LAs that were the most problematic for the learners were the ones belonging to the *adversative* category, which were in fact underused. The author attributed the over, under and misuse found to the students' lack of register awareness and the inappropriate advice from some textbooks and language teachers who sometimes provide long lists of connectors labeled as equivalents, without showing contextualized examples or giving proper instruction.

These corpus-based studies have shown how ESL/EFL learners with different levels of proficiency and different linguistic backgrounds had problems when using connective devices in academic texts such as argumentative compositions and essays. After exploring the use of different categories of LAs (i.e., additive, adversative, casual and sequential; see Liu, 2008, for a comprehensive list) some studies have reported that LAs belonging to the *adversative* category pose a challenge to most learners, even to the most advanced ones (e.g., Celce-Murcia & Larsen-Freeman, 1999; Granger & Tyson, 1996; Lei, 2012). Undoubtedly, more emphasis should be given when teaching ESL/EFL learners how to use LAs with regards to form, meaning, position and register, using corpus-informed materials. Some pedagogical recommendations will be presented at the end of this paper.

6.3 The importance of adversative linking adverbials

With the aim of exploring discourse patterns that grant the required textual coherence and cohesion in academic writing, and that are discipline- and genre- specific, this paper focuses on the use of LAs, in particular the ones with an *adversative / contrastive* function. Argumentative composition, a common type of assignment in many university degrees, is one of the forms of academic writing that comprises a larger number of LAs in general (Biber et al., 1999; Liu, 2008) and contrastive/adversative LAs in particular (Granger & Tyson, 1996; Rica-Peromingo, 2012). According to Celce-Murcia and Larsen-Freeman (1999), students do not seem to have much problem when using time (*after*), location (*where*) or manner (*as*) logical connectors, but concessive connectors (such as *however, yet*, and *though*) deserve a more complex analysis. Often, these connectors convey an inferential meaning: when reader and writer do not share background knowledge, the latter needs to make his/her intentions as transparent as possible, and inferential connectors help to avoid unnecessary periphrasis and wordiness. This can be seen in example (1) taken from Celce-Murcia & Larsen Freeman (1999, p. 528):

(1) Barbara isn't in town. The reason should be clear to you: it's because David isn't here.
Barbara isn't in town. *After all,* David isn't here.

The writer's actual intention could be misunderstood, or even lost if connectors are not used appropriately. As Biber et al. (2002, p. 391) note, adversative adverbials "mark some kind of contrast or conflict between information in different discourse units. Some of these adverbials clearly mark contrasting alternatives […] other adverbials mark a concessive relationship". And yet, the categorization of connectors as *concessive* is too general to fully describe their function. In order to find an accurate definition and a complete list of adversative linking adverbials (LA), we used Liu's (2008, pp. 514-515) comprehensive list of adversative LAs, which was based on Celce-Mucia and Larsen Freeman's framework for classification (1999, pp. 530-532) and other grammar books (e.g. Biber et al., 1999). Concessive connectors are now comprised in a broader category called *adversative* linking adverbials, and are divided into four categories:

1. Proper adversative/Concessive (e.g., *however*)
2. Contrastive (e.g., *on the other hand*)
3. Correction (e.g., *rather*)
4. Dismissal (e.g., *in any case*); (for a complete list, see Appendix 3)

Regarding the placement of these adverbials, initial position seems to be the most common one (Biber et al., 2002, p. 394). In the academic register, however, LAs placed in medial and final position are also found and account for approximately 40% and 20% of the occurrences respectively (Biber et al., 1999, p. 891); LAs placed in medial and final position require a distinctive punctuation and give the sentence a different emphasis. Therefore, we will also explore the differences between the two corpora regarding the placement of adversative LAs within the sentence in the present study.

For inexperienced writers, adversative LAs can be difficult to master: writers need to be aware of the different types of 'contrastive power' each of these adversative LAs has, use them only when necessary, and place them correctly within the sentence.

In addition, adversative connectors can be divided into subordinating conjunctions that work at clause level (e.g., *although*), or conjunctive adverbials, that work at sentence level (e.g., *however*), but learners are not always aware of this peculiarity. We tracked the use and frequency of the latter, conjunctive adverbials, with just one exception: two conjunctions were also included, namely *although* and *even though*. The reason behind this decision is that, even though the use of conjunctions is normally not too complex in comparison with that of the LAs (Liu, 2008), and in contrast with other adversative conjunctions (e.g. *unless, while, whereas*), we found that the frequency and use of these two items in particular was somewhat unusual in the learner corpus, and thus worth exploring. The complete list of adversative LAs can be found in the appendix (either 3 or 4).

6.4 Corpus linguistics, learner corpora and contrastive methods

Corpus Linguistics and Contrastive Interlanguage Analysis are the methodological frameworks of the present study. Corpus Linguistics (CL) is a branch of linguistics that helps us make empirical linguistic observations. As languages are constantly evolving, corpus-based analyses are useful to keep record, explore and provide founded statistics on up-to-date language use (Sinclair, 2005). The advantageous use of computers, together with the creation of very practical text analysis software (e.g., AntConc, Voyant, WordSmith tools), affords us data analysis, storage, and most importantly, saves much time. Of course, a competent use by the human resource is still needed, since the ability to discover and explore new linguistic phenomena belongs to the functional analysis and the approach used by the researcher.

Biber et al. (1998, p. 4) stressed that the criteria adopted by the corpus compiler during the compilation process are of considerable importance: the fact that, for example, a researcher should always explore authentic texts, belonging to a previous written text, or a transcribed set of oral language, but always produced in a natural communicative setting. If we wanted to analyse language produced by non-native speakers of a language in a classroom setting, something we may question is whether this type of production would be 'natural' enough according to the criterion described above. In this regard, Granger (2002) indicates that when the corpus is produced out of a language-teaching context is not entirely natural. She describes, however, different 'degrees of authenticity', such as "being gathered from the genuine communications of people […] or resulting from authentic classroom activity" (2002, p. 5). The corpora used in this study could be regarded as the latter since it consists of students' own compositions, which were the final assignment for a real content subject, as will be later described in Section 6.5.1. This and the fact that the texts were not written with the intention of being included into a corpus, makes them *authentic* texts.

The type of research that explores students' production of a second/foreign language has been called Learner Corpus Research (LCR), and was originated in the 1990s. Granger, co-founder of the International Corpus Learner English (ICLE) project in Louvain, has conducted remarkable investigations on learner corpora and provides valuable insights in her multiple publications (Granger & Tyson, 1996; Granger, 1996, 1997, 2002, 2004, 2015). Corpus-based Second Language Acquisition (SLA) research is gaining momentum (Granger, 2004, 2015). Questions such as why students, although receiving the same input, show wide differences in their development, or to what extent the students' mother tongue (L1) interferes when using a second/foreign language (L2), are being answered through corpus-based analyses of learner data informed by solid theories of SLA (see Granger, 2009, 2015, 2017; Paquot, 2017).

LCR often follows a contrastive method called Contrastive Interlanguage Analysis (CIA) (Granger, 1996, 2015). This contrastive approach allows researchers to make quantitative and qualitative comparisons between different NNS groups, but also between NNS and NS data. In the case of the present study, the first type of comparison (NNS – NNS) will let us observe the different choices students made when using adversative LAs, and see if these are somehow distinctive of their linguistic background (L1s) or otherwise. The second type of comparison (NNS – NS) will help us to detect variance in the use of adversative LAs, showing the extent to which learners differ from the native corpus. Hence, with the aim of exploring the use of adversative LAs in the academic writing of EFL learners with different L1s following a CIA approach, two research questions have been formulated:

1. How are adversative linking adverbials used in argumentative essays written by advanced EFL learners compared to native students, in terms of frequency, placement and taxonomy (i.e. concessive, contrastive, correction, dismissal)?

Hypothesis 1: NNS are expected to use more LAs than their native counterparts, due, in part, to a possible lack of proper training –e.g., give out a long list of connectors and make students use them in their writing to get a higher mark (Granger & Tyson, 1996; Granger, 2004; Lei, 2012; Rica-Peromingo, 2012; Wray, 2002). Learners are also expected to overuse high-frequency linking adverbials (e.g., *however*) since they may feel more confident when using these familiar, widely-used words that were acquired in early stages of L2 learning, often called 'lexical teddy bears' in SLA literature (see Hasselgren, 1994).

2. How do advanced EFL learners with different L1 backgrounds (Dutch, German, French, Italian and Spanish) use adversative LAs in argumentative essays when compared to one another? In particular, are there differences in the use of adversative LAs between groups of students with Romance and Germanic languages?

Hypothesis 2: the group of learners with Romance L1s (i.e., French, Italian, Spanish) and the group with Germanic L1s (i.e., Dutch and German) are expected to show similar frequencies and usage patterns when compared to one another in their groups – as a possible influence from their mother tongue, students may use linking adverbials that are similar in their mother tongue more often than those which are not (Rica-Peromingo, 2012).

6.5 Methodology

6.5.1 Context and corpus compilation

For the purposes of this study, we compiled an original learner corpus, consisting of 50 argumentative essays written by first-year undergraduate students with different L1s, i.e. 10 Dutch, 10 German, 10 French, 10 Italian and 10 Spanish, produced at Maastricht University (henceforth Maastricht University Corpus--MUC). These texts were collected at the end of a short *Academic Writing Skills* course (12 hours) in the European Studies degree of the same university. This short course provided insights into how to write introductory and concluding sections, how to distinguish formal from informal language and how to hedge appropriately when writing academically. The use of connectors was, however, not included in the syllabus. It is important to mention the fact that the course followed a Problem-Based Learning approach (PBL), through which the responsibility to learn was entirely transferred to the students. PBL required the role of group discussion leaders, volunteering students who managed each tutorial; this course also required students to peer-review their drafts with the help of a rubric elaborated in class. The texts included in this corpus contain an average of 2,900 words (148,960 words in total), and deal with the topic of 'media misrepresentation of different conflicts', that was adapted by the students. A sample of titles include (2) (3) (4):

(2) Russia's Press Coverage on the Syrian conflict.
(3) American media misrepresentation of the Iraq war.
(4) Misrepresentation of the Middle East Peace Summit at Camp David.

The students worked on these argumentative texts in the *Academic Writing Skills* course and these papers served as a final assignment for both this and another content subject: *Research skills*. These texts were evaluated by language specialists that focused on linguistic aspects, and also by content specialists who assessed the study design and content of these texts. Only texts that were satisfactory (given a good pass grade) in both subjects were included in the corpus.

The learner corpus has been contrasted with a native corpus, specifically the American university students' corpus, consisting of 176 argumentative essays written by native students (149,574 words in total). These essays are part of LOCNESS corpus that was provided by the Centre for English Corpus Linguistics of the Université Catholique de Louvain. The argumentative genre was chosen for both corpora, since adversative LAs tend to be used more notably in an argumentative genre rather than in narrative essays (Rica-Peromingo, 2012). However, not only did genre need to be comparable for this analysis but also the type of learners involved had to be comparable: as Stefan Gries (2013, p. 2) cautions, "some threats to the reliability and validity of our studies [learner corpus-based studies] have to do with the degree to which we can conflate and compare different learner corpora and/or native speaker comparison corpora". The criteria established in the development of ICLE to achieve *reliability* and *comparability* (Granger & Tyson, 1996, p. 18) were taken into consideration for both the corpus selection and compilation in the present study, as follows:

a) *equivalent type and similar stage of the learner.* Participants in the MUC corpus were first-year students, between 18-21 years of age, and they held TOEFL iBT 90, IELTS 6.5, or a B2 level of English certified according to the Common European Framework of Reference (CEFR). Similarly, most students in LOCNESS were aged 18-21, and they were all native speakers of English from five American Universities.

b) *Comparable text type.* Both the learner (MUC) and the native corpus (LOCNESS) contain untimed argumentative essays written in an academic context using reference tools.

6.5.2 Data and procedure

The students' compositions were all converted to plain .txt files, in which cover pages, tables, figures, notes, references and large quotations were deleted, in order to analyse learners' raw text. The frequency of each adversative linking adverbial in each corpus was generated, manually analysing each hit to avoid counting a search word that would fit a different category or part of speech (e.g.: *still* in medial position would not function as an adversative adverbial, but as a time adverbial, meaning *even now*). Then, frequency rates were calculated and the most significant differences on the production of adversative LAs were studied closely. Although both the learner and native corpora contained a similar number of total words (nearly 150,000w), LOCNESS texts varied in length (845w average) compared to the texts included in MUC (2900w average). For this reason, and following Chen's (2006) recommendation, both normalized values per 1,000 running words and per 1,000 sentences were calculated and added to the tables. The total number of words and sentences in each corpus are shown in Table 9.

Table 9: the learner and the reference corpus

	LOCNESS (NS)	MUC (NNS)
No. Texts	176	50
Tokens	149,574	148,068
Types	19,198	10,084
Sentences	7,972	6,846
Mean Sentence Length	18.76	21.62

The freeware corpus analysis toolkit AntConc was used to concordance Liu's (2008) list of adversative LAs in each corpus (MUC vs. LOCNESS) following a corpus-based approach. AntConc was also used to see the total number of words (tokens) and types in each corpus. The *sort* function (case sensitive, punctuation, and keyword in context) was used to observe when the LA was used in initial, medial or final position. Finally, TagAnt was used to tag the texts and count the exact number of sentences.

Table 10: the learner corpus by L1 group

	DUTCH	GERMAN	FRENCH	ITALIAN	SPANISH
Tokens	30,108	35,433	31,383	26,040	25,104
Types	4,084	4,684	4,693	3,764	3,785
Sentences	1,516	1,620	1,526	1,076	1,108
Mean Sentence Length	19.8	21.8	20.5	24.4	22.6

As can be seen in Table 10, German L1 learners were the ones who produced the longest texts while Spanish learners produced the shortest ones. Italian L1 learners have the longest mean sentence length (24.4 words per sentence) and all learners in general produced longer sentences than the native writers, as shown in Table 9. If we look at lexical diversity, the ones who used the highest number of different words were the French L1 learners. However, if we look at the general statistics shown in Table 9, we can see how the native (NS) texts contain almost twice as many different words as the non-native (NNS) (19,198 vs. 10,084 types respectively). In the literature, it has been suggested that a type/token ratio may be misleading if the length of the texts varies greatly (i.e., types decreasing with increasing length). All NNS texts were therefore analysed together, in order to compare both corpora, now of a similar size, in terms of types and tokens.

6.6 Results

In this section, the quantitative and the qualitative results of the use of adversative LAs in both corpora are presented. NS and NNS are used to refer to the native and non-native (learner) speaker corpus respectively. First, the overall differences in frequency, placement, and top-5 adversative LAs between NS and NNS are given in Section 6.6.1. Second, the most overused and underused adversative LAs in the corpora are shown in Section 6.6.2. Next, a second analysis according to L1 groups is presented in Section 6.6.3. To gain accuracy, the quantitative results have been normed by 1,000 running words and by 1,000 sentences in both corpora. As for the qualitative results, we will zoom in on five cases of misuse in Section 6.6.4 by analysing some of the most overused items in the learner corpus.

6.6.1 Overall frequency of adversative LAs

In general, NNS have used more adversative LAs than NS (737 vs. 655), as can be seen in Table 11. There is also a clear difference in the placement of these adverbials: initial position is more frequent in NNS (448), in contrast with the usual medial position of the native writers (344)[14]. It is true that LAs can occupy different positions in a sentence, but particularly this tendency for NNS to place linking adverbials in initial position has been described in previous studies: "initial position, or ISP as they call it, is the most common position for all L2 writers […] and L1 writers used the NIP [non-initial position] significantly more than L2 writers" (Field & Yip, 1992, p. 22). Placing LAs in different positions also implies, however, a functional difference: a study carried out by Salera called *The mobility of Certain Logical Connectors* (1978), described in Celce-Murcia and Larsen-Freeman's (1999, p. 536), suggests that a writer can place adversative connectors in a sentence-initial position in order to emphasize a fact, contrary to the expectation that was raised by the preceding sentence. This connector can, at the same time, be moved to a medial position to alleviate this stress. An overuse of the first frame (i.e.: placing connectors in initial position) throughout a text could therefore have an overwhelming effect on the reader. Table 11 presents the overall and the normed values:

Table 11: adversative linking adverbials in LOCNESS and MUC: frequency and position

Type	**LOCNESS (NS):** 149,574W			**MUC (NNS):** 148,068W		
	Total hits	**Initial**	**Medial**	**Total hits**	**Initial**	**Medial**
Proper Adversative Per 1000 words Per 1000 sentences	**395** (2.64) (49.55)	227	168	**469** (3.17) (68.51)	327	142
Contrastive	**106** (0.71) (13.30)	40	66	**129** (0.87) (18.84)	76	53
Correction	**120** (0.80) (15.05)	31	89	**87** (0.59) (12.71)	13	74
Dismissal	**34** (0.23) (4.26)	13	21	**52** (0.35) (7.60)	32	20
TOTAL	**655** (4.38) (82.16)	**311**	**344**	**737** (4.98) (107.65)	**448**	**289**

[14] The final position search produced very low hits in both the NS and the NNS corpora (14 and 10 total hits respectively) with only 3 out of 30 adverbials being used in that position: *however*, *though* and *anyway*. As there were no significant differences in the way both NNS and NS used final position, this section has been removed from the tables.

As can be observed, after exploring four categories of adversative LAs, both NS and NNS show preference for *proper adversative*: NNS 469 hits (63%) with 327 placed in initial position, and NS 395 hits (60%), and a more balanced placement of LAs, with 227 in initial position. Both NS and NNS coincide with the production of *dismissal*, as the least frequent category (e.g., *anyway, after all, still*). In the academic texts analysed in Liu's (2008), professional writers used *dismissal* (10% of total LAs used in the academic register) slightly more often than the NS and the NNS writers in our study (4-6%). Dismissal LAs (e.g. *after all, still, in spite of*) help the authors provide counter arguments to refute information, and since the writers in both our corpora were all undergraduate students, they may lack the confidence needed to perform such an action in their texts. On the other hand, the *proper adversative* category comprises very popular LAs such as *however*, which may in part account for the high frequency counts in this category.

Table 12: Top-5 adversative linking adverbials. Total hits and normed values

LOCNESS (NS)				**MUC (NNS)**			
LAs	Raw	Word	Sentence	LAs	Raw	Word	Sentence
However	175	1.17	21.95	However	178	1.20	26.00
Instead	62	0.41	7.78	Although*	73	0.49	10.66
Rather	56	0.37	7.02	Nevertheless	54	0.36	7.89
Yet	52	0.35	6.52	Rather	44	0.30	6.43
Although*	51	0.34	6.40	Yet	41	0.28	5.99
TOTAL	**366 (55%)**				**390 (52%)**		

*subordinating conjunction

Table 12 shows the top-5 most frequently used adversative LAs in both corpora. In general, both MUC and LOCNESS show similar patterns of use and these top-5 account for more than half of all the LAs produced. *However* is the most popular one in both corpora: NS 175 hits (1.17), 106 placed in initial position, and NNS 178 hits (1.19), with 144 in initial position. This popularity could be due to the fact that *however* is a conjunctive adverbial used almost generically as a marker of difference, and it does not need any syntactic structure, other than a following comma when placed in sentence-initial position. In contrast, non-native writers' use of *nevertheless* differs greatly from the English-native writers': the correct use of *nevertheless* requires a special logical sequence in the text, which many students failed to convey by using it interchangeably with *however*. Besides, *nevertheless* is practically non-existent in the native corpus (NS hits: 3 / NNS hits: 54), which points towards a case of overuse. *Instead* is the only item in the NS top-5 that does not appear on the NNS list. Of all instances of *instead* (102), sentence-medial *instead of* + *gerund* (33) as an adverbial subordinator (e.g.: "*instead of attempting to combat", "instead of integrating their thoughts"*) was the most frequent one in both corpora. *Instead* performs a corrective function that allows writers to present an alternative to the preceding idea (e.g. "reading about the struggles in the lives of ethnic Americans should not bring strife. *Instead*, ethnic American literature allows us to see the struggles…"); NNS writers in our corpus may not feel confident enough to provide readers with complementary information –or may lack the linguistic resources to do so–, which may explain the less frequent use of this adverbial (learners have used *instead* almost half as often).

6.6.2 Overuse and underuse of adversative linking adverbials

Overuse and underuse patterns have been identified setting a 0.01 difference in the case of the word-based calculation and 0.20 in the case of the sentence-based measurement (the mean sentence length of both corpora is 20 words). This cut-off figure is in agreement with Chen's (2006), and Lei's (2012) identification of over and underuse from a quantitative perspective. Also, it is important to mention that we use the terms 'over' and 'underuse' in a purely quantitative sense, i.e., to show a higher or lower frequency of certain adversative LAs. It does not, in any case, imply that the learners deviate from the norm established by the native corpus, mainly because the native corpus used here consists of novice writing as well. A total of 5 items have been identified as overused by the learners, as can be seen in Table 13.

Table 13: Overused adversative linking adverbials (NS vs. NNS total and normed values)

LOCNESS (NS)				**MUC (NNS)**			
LAs	Raw	Word	Sentence	LAs	Raw	Word	Sentence
Nevertheless	3	0.02	0.38	Nevertheless	54	0.36	7.89
Although*	51	0.34	6.40	Although*	73	0.49	10.66
In contrast	4	0.03	0.50	In contrast	28	0.19	4.09
Despite	6	0.04	0.75	Despite	28	0.19	4.09
Nonetheless	0	0.00	0.00	Nonetheless	20	0.14	2.92
TOTAL	**64**				**203**		

*subordinating conjunction

The most overused item was *nevertheless*, with a difference of 20.8 times more per 1000 sentences compared to the native corpus. A more qualitative look shows that this item was also stylistically and syntactically misused, which will be explained in the discussion. These items were evenly distributed in the learner corpus, regardless of the students' L1, which points to a possible teaching effect. Only one discrete item, *in contrast*, appears more predominantly in one linguistic group: 15 of 28 instances of *in contrast* belong to the Dutch L1 subcorpus.

Table 14: Underused adversative linking adverbials (NS vs. NNS total and normed values)

LOCNESS (NS)				**MUC (NNS)**			
LAs	Raw	Word	Sentence	LAs	Raw	Word	Sentence
Actually	38	0.25	4.77	Actually	17	0.11	2.48
Instead	62	0.41	7.78	Instead	40	0.27	5.84
Though	35	0.23	4.39	Though	18	0.12	2.63
Anyway	9	0.06	1.13	Anyway	1	0.01	0.15
TOTAL	**144**				**76**		

Regarding underuse, four items were identified as underused, as shown in Table 14. *Actually* was the most underused linking adverbial by the learners in MUC compared to LOCNESS. In this case, we did not consider it a case of misuse: *actually* is most commonly used in spoken rather than written contexts, as shown in the Corpus of Contemporary American English (COCA) (spoken 51% vs. academic 9%) and in the British National Corpus (BNC) (spoken 67% vs. academic 1.7%) (Liu, 2008). *Anyway* is a similar case: it was also underused by learners in MUC. *Anyway* is not commonly used in the academic registers either, as shown in both COCA and BNC (2.9% and 2% respectively), so we may say that native learners in LOCNESS have used a few linking adverbials that belong to a spoken register more frequently than the NNS. In this regard, the topics of some of the texts in LOCNESS could have triggered the use of such items: whereas the texts comprised in MUC deal with the same topic (i.e. "media misrepresentation of different conflicts"), topics in LOCNESS vary (e.g. US government, Portrayal of Women in fashion magazines, Violence on television, Recycling). The lack of uniformity in the topics covered in LOCNESS may have affected writers' linguistic choices to some extent. Also, despite being native speakers of the language, students may need specific training to develop their academic discourse skills –this could explain why items such as *anyway* and *actually* could be found more predominantly in the native corpus.

Interpretations of overused and underused items, as discussed in learner corpus literature (see Leech, 1998; Granger, 2009), need to be given with caution. In the case of the current analysis, certain items identified as over or underused can indeed be so and denote a lack of diversity in the students' repertoire, over-reliance on popular items, L1 transfer, or a potential teaching effect. However, comparisons with two different native corpora that contain expert (e.g., COCA and BNC) and student writing (LOCNESS) have proven useful to discard possible misconceptions (e.g., tell students they need to use *actually* or *anyway* more frequently in their texts). In this regard, it is important to say that not every case of over or underuse found in a corpus requires pedagogical attention. Writing is a complex construct and there are many different variables that can affect the writers' choice. It is difficult to know if certain patterns apply only to the subjects who participate in a corpus or to a bigger community, instead. Increasing the corpus size and recording learner metadata can help researchers decide which findings can or cannot be generalized, and which ones are worth devoting class time to. These suggestions will be explained further in the discussion section.

6.6.3 Use of adversative linking adverbials by L1 groups

We turn now to our second research question, which expressed the intention to look at the results according to different linguistic backgrounds. The analysis of the NNS corpus according to L1 groups has revealed some interesting results, as can be seen in Figure 5. After normalizing values, French L1 students seem to be the ones using the highest number of adversative LAs (5.5 times per 1k words), followed by German L1 students (5.3) and Spanish L1 students (5.4).

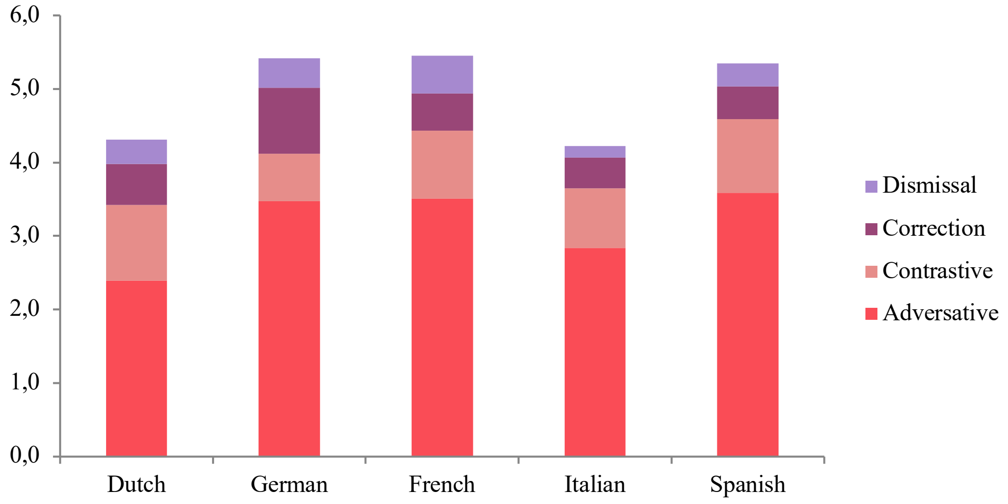


Figure 5: use of adversative LAs by NNS, normed values per 1k words.

There seems to be a general agreement on the use of *proper adversative/concessive* as the most frequent category among students, whereas LAs belonging to the *dismissal* subcategory were the least used by all of them. As we can see, category preferences according to individual L1 groups are quite homogeneous, and do not present significant variances when compared to overall group results. Table 15 presents the overall results.

Table 15: Adversative LAs by NNS: frequency, normed values and prevalent position

	DUTCH		**GERMAN**		**FRENCH**		**ITALIAN**		**SPANISH**	
	Total	**Pos.**	**Total**	**Pos.**	**Total**	**Pos.**	**Total**	**Pos.**	**Total**	**Pos.**
Adversat. Per words Per sentences	**72** (2.3) (47.4)	51 (init)	**123** (3.7) (75.9)	91 (init)	**110** (3.5) (72.0)	66 (init)	**74** (2.8) (68.7)	51 (init)	**90** (3.5) (81.2)	68 (init)
Contrast.	**31** (1) (20.4)	20 (init)	**23** (0.6) (14.2)	12 (med)	**29** (0.9) (19)	16 (init)	**21** (0.8) (19.2)	17 (init)	**25** (1) (22.5)	13 (med)
Correct.	**17** (0.5) (11.2)	14 (med)	**32** (0.90) (19.7)	26 (med)	**16** (0.51) (10.4)	15 (med)	**11** (0.4) (10.2)	10 (med)	**11** (0.4) (9.9)	9 (med)
Dismiss.	**10** (0.3) (6.6)	8 (init)	**14** (0.4) (8.6)	8 (med)	**16** (0.51) (10.4)	10 (init)	**4** (0.1) (3.7)	3 (init)	**8** (0.3) (7.2)	5 (med)
TOTAL Per words Per sentences	**130** (4.3) (85.7)	82 (init)	**192** (5.4) (118.5)	116 (init)	**171** (5.5) (112)	93 (init)	**110** (4.2) (102.2)	72 (init)	**134** (5.3) (120.9)	85 (init)

We hypothesized that learners with the same linguistic family (i.e. Romance L1s and Germanic L1s) would show similar frequencies and usage patterns. Unexpectedly, we found some discrepancies among linguistic families: students with Romance languages i.e., French, Italian and Spanish do not show similar patterns in the use of different types of LAs: French and Spanish L1 learners have produced the highest number of LAs in their texts, whereas Italian L1 learners contain the lowest production of LAs in almost all subcategories. The analysis of the texts written by students with Germanic languages, i.e., Dutch and German, has also produced different results: German L1 learners used all four types of adversative LAs much more frequently than Dutch L1 learners, specially the ones in the adversative category. All learners in MUC, regardless of their L1 background, placed adversative LAs predominantly in sentence-initial position (55%-65%).

6.6.4 Difficulties in the use of LAs: zooming in on five cases of misuse

We have performed a qualitative analysis by carefully exploring all cases of overuse and underuse in the learner corpus. This has served to uncover certain structures that may be missing from the students' repertoire, or are being used too much and therefore need more scaffolding or explicit instruction in the classroom. As Celce-Murcia and Larsen Freeman pointed out, "not all languages distinguish three classes of connectors in the way English does" (1999, p. 526). These 'three classes' refer to the categorization of connectors as coordinators (*but*), subordinators (*although*) and conjunctives (*however*). These classes are factually so similar that students may consider them as grammatically equivalent, what may result in punctuation and syntax issues. Examples from MUC have been selected and described in the sections below in order to illustrate five important misuses of adversative LAs.

Nevertheless

Nevertheless was the number one overused item in the learner corpus. *Nevertheless* is not necessarily followed by a comma and 42 out of 54 the NNS hits do. Students may have used *nevertheless* and *however* interchangeably. For instance, in both (5) and (6) *nevertheless* was placed at the beginning of a paragraph[15]. In (7) we see an example of misuse, as the writer uses it to add information:

[15] One shared characteristic among non-native students is that they place many adversative LAs at the beginning of a paragraph. In academic writing, paragraphs are expected to be unified –i.e., each paragraph should deal with only one issue (Wilkinson & Hommes, 2010). If an adversative connector, whose main function is to contrast, correct, or dismiss a previous idea, is placed at the beginning of the paragraph, this unity is not fulfilled.

(5) [IW][16] *Nevertheless*, up until then, the government had never intervened.

(6) [DW] In the NYT, *nevertheless*, the frame of anti-Americanism was still strong and presented people in the Middle East as (former) supporters of terrorism, due to their culture and religion.

(7) [FW] That is why they invaded Iraq. *Nevertheless*, another reason might be considerd [sic] too.

Nevertheless means "in spite of a fact that you have just mentioned" (Longman Dictionary of Contemporary English) and while *however* can indicate both concession and contrast, *nevertheless* indicates only concession, demanding a logical connection between the two sentences. Sometimes, students use *nevertheless* in place of *however*, and this function is not accomplished. *Nevertheless* is frequently placed sentence initially, but it can also be placed in the middle of the sentence; instances of the latter were not found in our corpus. Examples of sentence-medial use of *nevertheless* included in BNC are (8) and (9):

(8) It was a predictable, but *nevertheless* interesting, fact.

(9) Thus we can talk of a local government system which is different from a central government system but *nevertheless* interacts with it.

Although

Although was the second most overused item by the learners. No special punctuation is needed when *although* appears in the structure [*although* main clause + subordinate clause], and only when this structure is inverted, a comma is required [e.g., main clause +, *although* subordinate clause]. However, *although* is never separated by a period or semicolon since adverbial subordinators work at clause level. (10) and (11) are two examples of misuse:

[16] Dutch L1 Writer (DW); German L1 Writer (GW); French L1 Writer (FW); Italian L1 Writer (IW); Spanish L1 Writer (SW).

(10) [DW] It can be stated that the media coverage of this conflict was basically the same around the world. *Although* an important aspect of the paper that can be concluded is that the conflict was represented in different ways.

(11) [SW] Public accusations were instantly fired-off against Iraq; *although* intelligence officials of the 9/11 Commission Report quickly determined that there was no evidence linking Iraq and the terrorist attacks.

In contrast

In contrast was the third most overused item in the learner corpus. *In contrast (to/with)* is used when two subsequent topics are different in at least one respect (Biber et al., 1999). Sometimes, *in contrast* is confused with *in other words* or *on the contrary*, as can be seen in (12) and (13) respectively.

(12) [DW] How can you provide reporters from avoiding some facts? Or *in contrast*, how can you be sure that these reporters do not add some information flattering for themselves?

(13) [GW] Concerning the other misrepresentation of Saddam Hussein's alleged relationship to Al-Qaeda terrorists, evidence was not discovered either. *In contrast*, Al-Qaeda aimed at dispossessing secular rulers such as Saddam Hussein.

In contrast to shows an idea or concept that differs from or is at variance with something else (Biber et al., 1999). The following example (14) has been taken from the BNC to illustrate this function:

(14) *In contrast to* that of personal property, the range of landed incomes bears (…)

According to the Oxford Dictionary of English*, on the contrary* is used to “intensify denial, suggesting that the opposite is the case”, as in (15), taken from COCA.

(15) It had no negative effect on nerve function; *on the contrary*, it provided functional recovery.

However

However is the most frequently used adversative LA in the corpora explored (also in COCA and BNC). Even though it was not over or underused by the learners, we have found some instances of misuse. *However* is usually preceded by a period or semicolon and it is often placed in SIP followed by a comma. *However* can also fall in the middle of a clause, but it will always be preceded or followed by commas. Non-native writers sometimes overlook this punctuation and syntactic requirement, producing run-on sentences as in (16):

(16) [FW] These figures show that the majority in Syria are the Sunni Arabs, *however*, the Alawites, a minority group, is in power.

On the other hand

Although *on the other hand* had similar frequency counts in both the NNS and NS corpora, most instances in the NNS corpus followed the structure *On the one hand – On the other hand (15 of 24 hits).* Native writers, on the contrary, did not use this structure at all. *On the other hand* expresses two contrasting qualities of the same subject, but this function is not accomplished when *on the other hand* is used at the beginning of the paragraph as in example (17). Example (18) shows how the student used the *on the one hand – on the other hand* structure but ends with an unfinished sentence:

(17) [FW] *On the other hand,* I will show how the process of social and political change from a country which has been a dictatorship for over 40 years towards democracy engenders automatically new conflicts.

(18) [GW] English media demonstrated that the justification given by the government for advocating armed forces in Iraq was changing all the time. *On the one hand*, it is the weapon of mass destruction excuse. *On the other hand* is the human rights abuses of Saddam Hussein.

The fact that NNS writers used *on the other hand* mostly in SIP and very often preceded by *On the one hand* is an example of how rote learning can have a negative effect on students' production. The students must have learnt this structure by heart, as something rigid that cannot be changed, while the native corpus shows that it is in fact unusual to find this structure in an academic text, if at all. For example, a quick search of *on the other hand* in the academic register of COCA, displays 8145 total hits, 45% of which were used in SIP, and only 31% of the total hits were preceded by *on the one hand*.

Many of the disparities found in the NNS corpus occurred repeatedly in several texts, regardless of the writer's linguistic background. These findings of misuse are not poorly dispersed but rather evenly distributed throughout the learner corpus, which possibly point towards a lack of accuracy when teaching and learning linking adverbials. This 'teaching effect' on LAs acquisition and production has also been reported on in previous learner corpus-based studies (see Granger & Tyson, 1996; Lei, 2012; Leńko-Szymańska, 2008; Rica-Peromingo, 2012).

6.7 Discussion

This corpus-based study has yielded interesting results on the frequency of use as well as the usage patterns of adversative LAs in two different learner corpora. One of the initial hypotheses stated for the present study was that (1) non-native students would use more adversative LAs than their native counterparts and they were expected to rely on a limited set of frequently used adversative LAs. This hypothesis has partially been supported: MUC contains a slightly higher number of adversative LAs compared to LOCNESS, but the two corpora seem to share the same type of most frequently used linking devices, i.e., *proper adversative/concessive*. Regarding preference for specific items, the top-3 LAs produced by learners in MUC were: *however, although* and *nevertheless*, whereas the top-3 used by the native writers in LOCNESS was: *however, instead,* and *rather*. If we contrast this results with Liu's (2008), we see that the top-3 adversative LAs with frequency of 50 and above per million words in the BNC were: *however, yet,* and *nevertheless* (the subordinating conjunction *although* was not included in Liu's list, so we will not refer to it here). Surprisingly, in the native corpus of the present study, *nevertheless* got very few counts – to be more precise: 3 of the 403 proper adversative/concessive LAs used. Liu's study, however, was performed across different registers (speaking, academic, fiction, news and other). If we focus only on the academic register, which is the most comparable one for this study, *nevertheless* accounts for only 7% of the total hits. The top-3 items in the academic register of the BNC are therefore: *however, yet,* and *in fact,* which is similar to what we found in our analysis.

The frequency and preference for specific units by learners in MUC, however, contrasts with the NNS learners' in previous studies, specifically those in Granger & Tyson's (1996) and Lei's (2002). In the former, learners from ICLE underused adversative connectors that change the direction of an argument, such as *however, yet, though*, and *instead,* compared to LOCNESS. The authors attributed this underuse to the difficulty these adverbials pose for unskilled writers. In Lei's (2002), the adversative category was also underused by doctoral EFL students: *However* was highly underused while *actually* was one of the most overused items in the learner corpus. *However* and *actually* were precisely the most and the least used items by the students in MUC respectively. These results do not coincide with our findings: in terms of frequency and preference, learners in our corpus have produced LAs to a more similar extent compared to native writers in LOCNESS than previous studies. Here, we must note that, while learners in MUC have a certified B2 or advanced level of English, "advanced" in the ICLE corpus refers to "university students of English, usually in their third or fourth year of study, who therefore make relatively few morphosyntactic errors" (Granger & Tyson, 1996:18), so this slight difference in the English proficiency level of the participants may have accounted for the discrepancies found.

In terms of placement, learners in MUC have shown a strong preference for LAs in sentence-initial position (SIP). This finding is in line with previous research: Granger and Tyson (1996, p. 24) also found that there was a significant overuse of LAs in SIP in ICLE, specifically in the French and the Chinese subcorpora (e.g.: 68% and 87% of the uses of *however* were in SIP respectively compared to the 49% of the control corpus). Similarly, Yvette Field and Yip Lee Mee Oi (1992) found that Cantonese L1 students used more linking adverbials than Australian native students and that they placed these connective devices at the beginning of the sentence. As it was previously mentioned, placing connectors at the beginning of the sentence, when balanced, is an accepted praxis in academic writing (Biber et al., 2002). The problem arises when they are continuously placed in sentence-initial position, and especially when they are placed at the beginning of paragraphs: in this case, the text may evoke a sense of redundancy and disunity. This finding has important pedagogical implications for the teaching and learning of linking devices in academic writing.

It was also hypothesized that (2) the group of students with Romance languages (French, Italian, Spanish) and the group with Germanic languages (Dutch, German) would show similar frequency and usage patterns (e.g., syntactic placement) in the production of adversative LAs as a possible transfer-related factor. With regards to the Romance L1 group, French and Spanish L1 learners have produced a similar number of adversative LAs, being the French L1 learners who have produced the highest number of LAs in general; Italian L1 learners, however, have produced the lowest quantity of LAs of all groups. As for the Germanic L1 group, Dutch and German L1 learners do not coincide with the frequency of use of adversative LAs either. Therefore, since there are no similarities according to the linguistic families included in MUC, we can say that our second hypothesis has not been fully supported. All learners, however, coincide with the most preferred category (*proper adversative*), the least used category (*dismissal*) and the predominant position within the sentence (SIP).

All things considered, it is difficult to determine whether or not these choices have been influenced by the students' L1. Preference for certain adversative linking adverbials is not significantly different among these L1 groups. This fact could have been originated by the type of methodology these learners were exposed to: although the Academic Writing Skills course did not include explicit information on the use of connectors, it followed a PBL methodology that required peer-reviewing each piece of writing, looking at textual and paragraph organization, and language accuracy in general. This means that students received feedback from at least two different students, and given the multicultural and international scope of the course, it may have eased the overuse or underuse of certain LAs.

6.8 Conclusion

This paper has analysed the frequency and use of adversative LAs in two corpora, consisting of argumentative compositions of both English native and non-native university students. A second analysis among non-native students has been conducted in order to explore the production of LAs according to groups of different L1s. Findings suggest that non-native students use adversative LAs slightly more often than their native counterparts. However, it is not in the quantity but in the placement and functionality of these adverbials that the difference is remarkable. In this regard, the NNS strong preference for sentence-initial position, especially when they place adversative LAs at the beginning of paragraphs, prevents the unity and coherence of the text. Under a more qualitative perspective, cases of misuse have also been identified for five particular units. This information should be of interest to (language) teachers and learners, and to developers of materials on academic writing.

The results obtained from the second analysis of the non-native corpus by L1 groups showed that the over-, under-, and misuse of certain adversative LAs applies to all leaners regardless of their L1. Both, the fact that they shared a context (in terms of age, year of study, discipline, genre, and university), and that there is a need for more corpus-informed teaching materials on the use of connectors in academic writing, may explain these results. The results of this paper could serve as guidance to help learners use adversative LAs more accurately. It can also help teachers to decide which adversative LAs can be taught first, since they are the most frequently used by English native speakers, and how to address common problems learners face when writing academically.

The findings, drawn from the quantitative and qualitative analyses of this paper, also carry important pedagogical implications for the teaching and learning of linking devices in academic writing: LAs are not optional décor writers can or cannot add to their texts; these units really help readers understand where the writer stands. Students tend to believe that the more connectors they use in their writing, the higher their mark will be, what very often produces an unnatural use of LAs only to achieve 'surface logicality' (Lei 2012: 268) and also foreign-sounding ideas in their texts (Granger & Tyson, 1996; Leńko-Szymańska, 2008; Rica-Peromingo, 2012). As Granger (2004, p. 135) cautioned, "learners should not be presented with lists of 'interchangeable' connectors but instead taught the semantic, stylistic and syntactic behaviour of individual connectors, using authentic texts". Using corpus-based pedagogical tools in the classroom (e.g.: COCA corpus' website, or a self-compiled corpus of more advanced peer writing, published articles in the same discipline, or other bibliographical references) would enable students to see how connectors are used in context (i.e., different disciplines and academic genres), what their frequencies are, what word they collocate with, their position within a sentence, and their pragmatic and discursive function (Bennett, 2010). But not only L1 data can solve doubts on the use of LAs: teaching students how to build a corpus using their own texts, and how to perform an analysis in order to identify differences or similarities with their peers, will probably motivate them to keep practicing and reflect on their writing.

It is important to conclude by exposing the limitations of this study and giving some suggestions for further research. The results of this study could be improved by using a larger corpus; our corpora consisted of almost 150,000 words each, but may not be large enough to make generalizations on the use of LAs in EFL learner's academic writing. While descriptive analyses were carried out to calculate total and normalized frequencies for the adversative LAs explored in the corpora, inferential statistics such as the T-test could be used to investigate whether the differences found were statistically significant. In addition, this study could be improved by applying a classical Contrastive Analysis of parallel corpora (i.e.: students' production in their L1), to study interlanguage more accurately. In this sense, a (semi) longitudinal study would also be interesting to observe the learners' language development, and see whether the production of adversative LAs becomes more similar to the native or the expert production at a higher level (e.g.: C2). Recording and analysing different variables (e.g.: cognitive, educational, social, and cultural background of the participants) can also be useful to assess the implications of the study. Also, the reference corpus used consisted of texts written by American students; in order to see if there are differences in the use of LAs, it will be necessary to examine other varieties of English (Australian, Canadian, Singaporean, etc.). Finally, the use of professional writing as the reference corpus (e.g., published research articles) would help to decide if over, under, and misuse of certain adversative linking adverbials characterises non-native writing or instead, all novice writers.

7. Reflexive Metadiscourse in a corpus of Spanish bachelor dissertations in EFL

7.1 Introduction

Metadiscourse (hereafter, MD) is an umbrella term used in discourse analysis to describe a range of linguistic elements that, deliberately used by the writer, helps readers to navigate successfully through a text. For example, *'our study', 'see Table 3',* or '*in other words'*, signal 1) authorial involvement, 2) an awareness of the reader, and 3) an awareness of the evolving text, respectively. Reader-oriented texts, i.e. those that contain metadiscursive markers to help readers "organise, classify, interpret, evaluate and react" to the ideas presented (Vande Kopple, 1985, p. 83) are found to be more convincing, comprehensible, and more likely to be remembered (Crismore & Vande Kopple, 1997). In this paper I employ a reflexive model of MD (Ädel, 2006; Mauranen, 1993) to study MD features in the academic writing in English of Spanish undergraduate students of medicine and linguistics.

Hyland (2008b, p. 548) points out that, "compared with many languages, academic texts in English tend to be more explicit about structure and purposes, to be less tolerant of digressions, to be more cautious in making claims, and to use more sentence connectors". For learners of English as a foreign language (EFL), as well as for novice writers, MD markers that help to achieve some of the writing goals mentioned above may be difficult to acquire, and may even go unnoticed when reading a text (Low, 1996; Hyland, 2010). Learners' academic written production that lacks metadiscoursal devices can come across as too direct, digressive, and sometimes unconvincing (Hinds, 1987; Montaño-Harmon, 1991). In contrast, an appropriate use of MD markers is often related to text quality, enhanced readability, and even higher grades (Cheng & Steffensen, 1996; Dafouz, 2003; Hyland, 1998; Intraprawat & Steffensen, 1995; Lee & Deakin, 2016; Noble, 2010). Becoming acquainted with the many forms and functions MD markers can have is therefore of paramount importance for academic language writers.

As in most European universities, undergraduate students in Spain are required to write a bachelor dissertation (BD) at the end of their studies, and many of them do it partly (e.g. abstract, viva) or entirely in English. BDs are a major piece of scholarly work that allows students to adopt a scientific approach to explore a topic in depth and present it to experts in the field (i.e. a supervisor, and the examining committee), and it is the academic project that most resembles a research paper. Academic writing courses, textbooks, or style guides are sometimes provided to guide learners through this writing process. Many have argued, however, that these often take a 'one-size-fits-all' approach, and group all needs, failing to notice differences across disciplines (Hyland, 2008b; Springer, 2012). Moreover, some textbooks provide conflicting advice about the extent to which writers can intrude into their texts (Hyland, 2001, 2002), and since MD tends to be considered secondary to the main objective, i.e. presenting information, little instruction on MD is provided (Martín-Laguna & Alcón, 2015). EFL learners often fail to use sufficient metadiscursive markers, and may not be aware of the contribution these elements make to the full understanding of the text, or the differences between their L1 and L2 disciplinary discourses (Hyland, 2000, 2005, 2012). To date, there are few studies that explore reflexive MD in Spanish EFL academic writing across disciplines (cf. Mur-Dueñas, 2011; Pérez-Llantada, 2010). Corpus-based and corpus-driven studies that explore the MD dimension in EFL texts produced by Spanish undergraduate students are therefore needed.

The present study seeks to analyse the frequency and types of reflexive MD markers in a learner corpus of EFL Spanish undergraduates' BDs, with the intention of highlighting rhetorical conventions of this genre, and L2 writers' linguistic features regarding the use of MD. The corpus has been analysed from a discipline variable, exploring BDs in medicine and linguistics, and also from a writer status variable, comparing the learner corpus with an expert corpus of published research articles (RAs) in the same discipline. I hope that the results of the present study will shed light on the use of reflexive MD in EFL academic writing, and stress the importance of teaching MD to L2 writers taking into account their specific discipline. This study also presents pedagogical implications, relevant for academic writing teachers who wish to equip their students with language-, genre-, and disciplinary-sensitive metalinguistic devices. Finally, the present paper provides a systematic basis for the analysis of reflexive MD markers in BDs and RAs, useful to design pedagogical material on MD that is corpus-informed and genre-sensitive.

7.2 A view on the trajectory of metadiscourse

In applied linguistics, the term *metadiscourse* was first coined by Harris in 1970, but the concept gained traction with Williams' (1981) work, who defined it as "discourse about discourse" (1981: 47), or "writing about writing, […] whatever does not refer to the subject matter being addressed" (1981, p. 212). Since its conception, there has been a distinction between *meta*discourse and *primary* discourse (i.e. propositional content) (Crismore, 1989; Sinclair, 1981; Vande Kopple, 1985; Williams, 1981). As aptly described by Toumi (2009, p. 66):

> [Metadiscourse] marks the writer's awareness of the current text as text or as language, of him/herself as writer, and of the potential reader as reader of this text. Metadiscourse supports propositional content, but remains separate from it. It is the means by which propositional content is made coherent, legible and persuasive to the reader in accordance with the writer's intentions.

In writing, metadiscursive elements can make reference to three dimensions: (1) the evolving text (e.g. *in figure 1, secondly, as mentioned previously*), (2) the writer of the text (e.g. as *I* said, *we* found, *our* study), and/or (3) the imagined reader (e.g. *see* appendix 1, *you* may question, *we* will see how); these categories are not exclusive and markers can refer to one or more of these dimensions at the same time (Toumi, 2009). In some cases, the second and third categories (i.e. writer and reader) are merged into one category only, called 'interpersonal' (Ädel, 2006; Bondi, 2010; Dafouz, 2003; Halliday, 1973; Mauranen, 2010; Toumi, 2009) or 'interactional' (Hyland, 2005).

Since the early days of MD (Crismore, 1983; Crismore & Farnsworth, 1990; Vande Kopple, 1985; Williams, 1981; see Toumi, 2009 for a comprehensive review) three differences have been made: (1) metadiscourse from ideational content; (2) textual from interpersonal elements; and (3) reflexive from attitudinal MD. The first difference, as mentioned earlier, has been the starting point of the discipline: distinguishing *metadiscursive* elements from the *ideational* content of the text. The characters in bold in (1) illustrate this difference:

> (1) This can be accounted for two different principles**:** a weak one **–** ***also known as*** linguistic relativity**–** and a strong one (…) (example taken from the learner corpus).

Also known as, and two punctuation marks: *colon* and *dashes* in this example, do not add content, but help the writer to (1) give an explanation of the two principles (colon), (2) add a commentary or aside (dashes), and (3) provide a different term, perhaps a more scientific one, for one of the principles (*also known as*). Even though differentiating MD markers from content may seem an easy task to perform, in some cases there is no such a clear distinction. Consider, for example, the use of the deictic marker *Here* shown in (2):

> (2) The other would be represented by a case in which commodity prices fall by the full extent of the degree of cost-cutting involved in technological progress. *Here* the effect on real wage rate is very simple to analyze.
>
> (Example taken from Toumi, 2009, p. 70)

It is ambiguous if the deictic marker *Here* refers to the current text (e.g. in this study), which would qualify as MD, or to the content (e.g. in that context or situation) in which case it could not be coded as MD. Examples like this make the nature of MD itself difficult to delimit, and, as frequently described in literature, *fuzzy* (Hyland, 2017). One of the main contributions of Ädel (2006) and Mauranen (1993) is a set of criteria to help identify reflexive MD. This set of criteria has been taken into account in the present analysis and will be described in Section 7.4.

The second difference, *textual* and *interpersonal*, gave birth to what have later been called 'broad' and 'narrow' approaches (Ädel, 2006; Mauranen, 1993; Toumi, 2009). A broad approach to MD explores and includes both textual (e.g. *in section 1, in other words, in contrast*) and interpersonal (e.g. *we* can see, *our* study, *note* that) categories (Hyland, 2004). A narrow approach, in contrast, will focus on textual categories only (Dahl, 2004; Mauranen, 1993). However, this distinction has also been a source of disagreement. Some rhetoricians claim that, since all MD elements in some way or another take the reader into account –be it textual or interpersonal– the limits between the interpersonal and textual categories are also fuzzy (Hyland & Tse, 2004) and propose a broader and more inclusive interpersonal perspective of MD called 'interactive', whose main representative is Ken Hyland (2017: 20). In this regard, Mauranen (1993) and Ädel (2006) distinguish two models of MD: the 'reflexive' model, also known as 'non-integrative' (Ädel, 2006; Ädel & Mauranen, 2010), and the 'non-reflexive' or 'integrative' model. These models are an attempt to bridge the gap between textual and interpersonal MD: markers to refer to the *text*, the *writer*, and the *reader* are included in both of these models; this conceptualization also helps restrict the fuzzy notion of MD (Ädel 2016), and share the idea that the main rhetorical strategy of MD is that of achieving *persuasiveness.* As Dafouz (2003, pp. 32-33) aptly puts it, "metadiscourse categories, both textual and interpersonal, ultimately intend to convince readers of the validity of the arguments presented in the text […] it is the perfect combination of these two elements that makes a text persuasive".

The third and last difference, *reflexive* and *attitudinal* MD, is what separates the two models mentioned above: the 'interactive' approach includes the category of *stance* as a unit of analysis, i.e. makers that show the writer's *attitude*, express certainty (such as boosters) or doubt (hedges) (e.g*., fortunately, clearly, might*). The 'reflexive' approach, on the other hand, excludes stance and focuses on the *reflexive* aspect of language, i.e. items used exclusively to refer to the finite world of the evolving text; stance is a non-reflexive feature of language because it reflects the state of mind of the writer, as an experiencer of the real world (Toumi, 2009). However, a tendency of reflexivity and stance to co-occur in academic or professional writing has been described in the literature (Dafouz, 2003; Mauranen, 2010) and often labelled as 'discourse collocations' (Mauranen, 2010) (e.g.: *our paper* has *clearly* shown). A view that defends a reflexive approach to MD comes from Mauranen, who argues that "if we opt for a very broad, embracing notion of metadiscourse [e.g. including stance, hedges, or boosters], we risk losing sight of its collocability and interaction with other discourse phenomena" (2010, p. 37). The reflexive model adopted in this study afforded the researcher a narrower approach to MD which, together with the text-internal criterion, facilitated the identification and selection of MD markers in the corpora. The taxonomy of reflexive MD used, together with the identification and tagging system will be described in Section 7.3.

These different approaches, broad and narrow, and interactive and reflexive, not only differ in the categories they explore, but also in the methodology they apply. There are two types of methodology that are often used in MD research, namely 'thin' and 'thick' (Bondi, 2010). The first one is a corpus-based approach that consists in predefining a list of terms to be analysed (e.g. comparing the frequency and types of MD markers between two corpora). It allows for cross-linguistic, cross-disciplinary and cross-generic comparisons of large corpora. The downside of the thin method is that potentially metadiscursive items present in the texts but not included on the list will never be found (e.g. Crismore et al., 1993; Hyland, 2005; Vande Kopple, 1985). The thick approach, on the contrary, relies on a corpus-driven methodology. In this contextualized form of analysis, the elements explored are based on and set by the data (i.e. no predefined list of terms). The main difference is that, as the analysis is mostly done manually, that is, discovering and tagging markers actually present in the data, the units of analysis are often smaller (e.g. one category of MD markers such as 'self-mentions') than in the thin method (e.g. Ädel, 2006; Bondi, 2010; Mauranen, 1993; Pérez-Llantada, 2010). I have adopted a mixed-method approach (i.e. thin and thick) by which each reflexive MD marker, belonging to a predefined set of categories (e.g. 'endophoric markers'), actually present in the texts has been manually tagged, to later calculate frequency counts for all the elements found.

Due to the fact that the quantity of elements that qualify as MD vary from one model to the other (e.g. *stance, hedges* and *boosters* would be included in the analysis of 'interactive' MD, but excluded in a 'reflexive' approach to MD), the estimates about average proportion and range of MD markers in a given genre and discipline vary greatly in the literature: for example, following an interactional model, Hyland (2005) reported that 1 every 15 words in RAs was metadiscursive (an average of 370 occurrences per paper), and 1 every 21 words in postgraduate dissertations (Hyland 2010); Pérez-Llantada (2010), in contrast, reported that the quantity of reflexive metadiscourse represents a very low proportion compared to ideational content.

7.3 Taxonomy of reflexive MD markers

The current study follows a reflexive model of MD drawing on Mauranen (1993) and Ädel (2006). Previous taxonomies have been taken into account as a point of departure, but some adjustments have been made in order to render the proposed taxonomy more applicable for the RA and the BD genres. I have explored metatextual (MT) and interpersonal (IP) markers in both the learner and the expert corpus across disciplines (linguistics and medicine). These categories were manually analysed and tagged in the texts as shown in Table 16:

Table 16. Reflexive Metadiscourse: categories, subcategories, examples and tags

Category	Subcategory	Example	Tag
Metatext			_MD_MT_
References to the text			_RT_
	Full text	*this study/ the current paper/ our article*	_FT
	Part of the text	*this section/ Appendix A/ in this chapter*	_PT
	Semiotic modes	*Table 1/ this diagram/ Fig.*	_SM
Endophoric markers			_EN_
	Anaphoric	*Aforementioned/ as previously discussed/ as noted above*	_AN
	Cataphoric	*The following/ as follows/ next paragraph*	_CA
	Deictic	*Here/ now/ so far*	_DE
Code glosses			_CG_
	Reformulators	*i.e./ that is,/ in other words*	_RE
	Exemplifiers	*e.g./ for instance/such as*	_EX
	Parentheticals	*(inaccurate) translations/ in a degenerative (vs. naïve) environment*	_PA
	Colons	*in the data:/ three reasons:*	_CL
	Semicolons	*Pandora's box; hence/ FI hours; however*	_SC
	Dashes	*categorical difference –i.e., between writer and the audience/ paradigm of three pillars -- scaffolds, cells, signals --*	_DA
Linking Devices			_LD
	Additive	*in addition/ also/ furthermore*	_AD
	Contrastive	*however/ in contrast/ nevertheless*	_CN
	Consecutive	*therefore/ as a result/ thus*	_CO
	Organizers[17]	*firstly/ second/ third*	_OR
	Topicalizers	*regarding/ as for/ with respect to*	_TO

[17] In order to qualify as MD, these elements must function text-internally (i.e. signal transition in the world of discourse) and not text-externally (refer to real processes: e.g. *second*, we added the solution, and *then,* we removed the lid) (Ädel, 2006; Mauranen, 1993).

Category	**Subcategory**	**Example**	**Tag**
Interpersonal			MD_IP
Writer oriented			-WO
	Self-mention	*I/ our/ (exclusive) we/ the researcher/ the author/*	_SF
Reader oriented			_RO
	Directives	*See/ consider/ cf.*	_DI
	Rhetorical questions[18]	*if L2 proficiency alone cannot account for the incorrect meaning components, what are other possible explanations?*	_RQ
Participant oriented			_PO
	Inclusive we[19]	*Let's have a look/ as we can see/ if we take*	_IW

Examples of how MD markers were tagged are given in (3) and (4), in which the code MED (short for medicine) or LIN (linguistics) indicates the discipline, and BD (short for Bachelor Dissertation) or RA (Research Article) indicates the subcorpus the example belongs to:

(3) MED_BD02: *(e.g. see Appendix 1)*
Tagged text:
(e.g._MD_MT_CG_EX see_MD_IP_RO_DI Appendix_MD_MT_RT_PT 1) _MD_MT_CG_PA
Tags stand for:
*e.g.*_Metadiscourse_Metatext_Code Gloss_Exemplyfing
*see*_Metadiscourse_Interpersonal_Reader-oriented_Directive
*Appendix_* Metadiscourse_Metatext_Reference to text_Part of the text
*)*_Metadiscourse_Metatext_Code Gloss_Parenthetical

(4) LIN_RA02: *For instance, let us take the PV show up with the following meaning sense distribution:*
Tagged text:

[18] Research questions are excluded here.

[19] Only those cases in which inclusive 'we' is used to refer to 'you and me', i.e. the author and the reader of the text, qualify as reflexive MD. As a rule of thumb, Noble (2010) suggests that those instances in which 'we' can be replaced by the term 'people' or 'anyone', as it is overtly general, do not qualify as MD.

> *For instance_MD_MT_CG_EX , let us_MD_IP_PO_IW take the PV show up with the following_MD_MT_EN_CA meaning sense distribution:_MD_MT_CG_CL*
>
> Tags stand for:
>
> *For instance*_Metadiscourse_Metatext_Code Gloss_Exemplifying
> *Let us*_Metadiscourse_Interpersonal_Participant Oriented_Inclusive We
> *The following*_Metadiscourse_Metatext_Endophoric_Cataphoric
> *:*_Metadiscourse_Metatext_Code Gloss_Colon

This reflexive model excludes stance markers (e.g. hedges and boosters) and also intertextual references (e.g. reporting verbs). As was mentioned earlier, the set of criteria developed by Ädel (2006) and Mauranen (1993) to help identify reflexive MD markers, namely (a) explicitness or self-awareness, (b) contextuality, (c) current text and (d) writer and reader, was applied during the selection process as follows:

a) *explicitness or self-awareness*: to qualify as reflexive MD, the writer had to make explicit reference to (a) the on-going text, to (b) her/himself as the writer, and (c) the reader of the text.

b) *contextuality*: according to this criterion, the rhetorical function of each MD marker refers only to its immediate discourse context (Ädel, 2010). Thus, all items were analysed in context to count reflexive elements only (e.g. the isolated word *author* could refer to the author of the text, to the author of any other text, or *authors* in general).

c) *current text:* from a reflexive perspective, the connection with the real world –e.g. propositional content, personal judgments and opinions, or intertextuality– does not qualify as reflexive MD. Only those markers that refer to the evolving text were counted.

d) *writer and reader:* only references to the writer and reader as immediate participants of the current text, and not as experiencers of the real world, qualify as MD (see e.g. specifications for "inclusive we" mentioned previously).

After the identification process, 230 reflexive markers belonging to 21 different categories were found and tagged (see Appendix 8 for a complete list of markers).

7.4 Corpus-based studies on metadiscourse

Regarding the use of MD in academic writing, four main patterns have been found in the literature: 1) EFL learner writers tend to underuse certain categories of MD markers when compared to experts or native writers in the same discipline. Devices that signal authorial confidence such as 'self-mention' or 'elaboration' (Springer, 2012), or refer to the evolving texts, such as 'endophoric markers' (e.g. in the *following* section), and 'reader-oriented markers' (e.g. *see* table 3) are often underused, which has been attributed to students' "inexperience in structuring big texts" (Burneikaite, 2008, p. 45), and to having a "low audience-awareness" (2008: 45) possibly due to a lack of exposure and explicit learning of MD markers. Also, 2) several studies contrasting the use of MD in L1 and L2 English, and L1 Spanish in different disciplines (medical sciences, social sciences, and humanities), genres (research articles, textbooks, newspaper opinion articles), and contexts (international vs. national journals) suggest that texts in English are likely to contain quantitatively more MD (especially 'logical markers', 'code glosses', 'adversative connectors', and 'self-mentions') than texts in Spanish (Dafouz, 2003; García-Negroni, 2008; Moreno, 1997; Mur-Dueñas, 2011; Pérez-Llantada, 2010). 3) Differences are also found from an interdisciplinary perspective: research on the use of MD shows how different disciplinary communities have different conventions of MD. In fact, texts belonging to humanities (e.g. linguistics) are likely to contain quantitatively more MD devices than other disciplines (e.g. medicine) (Hyland, 2001). This difference has been attributed to the need of human sciences to elaborate claims more, since they are often based on qualitative methods (Hyland 2010); the nature of the topics itself –i.e. language being the subject matter of the linguistic discipline, also accounts for the discrepancies found (Salas, 2015). Finally, 4) L2 learners or novice writers who do not use MD markers in their texts accurately (i.e. may not be aware of their disciplinary community conventions) tend to produce less persuasive, and thus, less successful texts: positive correlations between high-scoring essays and a higher frequency and range of MD devices have been found in the literature (Intaraprawat & Steffensen, 1995; Noble, 2010). Teaching MD explicitly seems to be both educationally and statistically significant in that learners improve their texts' quality and achieve higher scores (Cheng & Steffensen, 1996).

As we have seen, academic writing is community situated (Hyland, 2005, p. 142), and therefore not only language (e.g. English), but also discipline (e.g. medicine), mode (e.g. written), genre (e.g. research articles), and even part of the text (e.g. introduction) play a role in the choice of metadiscursive practices (Bondi, 2010; Dafouz, 2003; Hyland, 2012b; Hyland & Tse, 2004; Mur-Dueñas, 2011; Pérez-Llantada, 2010; Salas, 2015). Writers who conform to specific linguistic and disciplinary conventions, express ideas clearly and persuasively, and create a balanced textual persona that sounds familiar and convincing to their readers, are more likely to succeed in the scientific communication realm (Intaraprawat & Steffensen, 1995; Ivanič, 2004). Hence, in order to investigate the production of reflexive MD by EFL undergraduate learners in different disciplines, three research questions have been established:

1. To what extent do Spanish undergraduate students use reflexive MD markers when writing in academic English (bachelor dissertation)? The frequency rates of all reflexive MD markers found in the corpora will be calculated, to later explore the different categories used.
2. Are there any differences across disciplines? We will look at interdisciplinary variation in the corpus (i.e. BDs and RAs in medicine and linguistics).
3. Is there overuse or underuse of reflexive MD markers when compared to an expert corpus of RAs? This analysis will help us to identify possible learner features in this particular academic genre.

7.5 Methodology

7.5.1 Data collection

In order to carry out an interdisciplinary analysis of reflexive MD markers in medical and linguistic academic texts, two corpora were compiled, namely (1) a learner corpus of 20 BDs written in English by Spanish undergraduate students in linguistics and medicine from two Spanish universities (103,971 words) and (2) an expert corpus of 50 RAs published in medical and linguistic academic journals (see Appendix 5 for the list of journals) to match the discipline and (roughly) the topic of the BDs (258,223 words). The articles chosen for the compilation of the expert corpus were, in some cases, part of the bibliographical references of the students' BDs. In other cases, they were chosen because they were highly cited articles in the field. The texts in both the learner (BDs) and the expert corpus (RAs) varied in length; for this reason, normalized values per

1,000 running words were calculated and added to the tables. The total number of texts, tokens and types in each corpus are shown in Table 17.

Table 17. The learner and the expert corpus

	BDs		RAs	
Discipline	Linguistics	Medicine	Linguistics	Medicine
No. Texts	10	10	25	25
Tokens	65,180	38,791	177,041	81,182
Types	5537	4656	9853	7553
Average text length	6518	3879	7081	3247
Total words	**103,971**		**258,223**	

7.5.2 Data analysis

I carefully read and scanned all reflexive MD markers in each text (see tagging system and identification criteria in Section 7.3) and only relevant examples –that is, reflexive and text-internal– were coded.

This corpus-driven or, as previously described, 'thick' method (Bondi, 2010) used to retrieve instances of MD gave me a deeper view of the choices learners made, how textual and interpersonal interactions were realized, the most prevalent types of MD markers in each discipline, how they were distributed, and how these patterns may have affected the whole structure of the text. Subsequently, the corpus analysis software AntConc was used to concordance all the different categories (searching by code, e.g.: _MD_MT_RT_FT). Frequency rates were calculated, and the most remarkable differences on the use of reflexive MD markers were carefully studied.

7.6 Results and discussion

The results of the analysis of reflexive MD are reported on as follows: first, the overall differences between BDs and RAs production across disciplines are given. Second, the frequency counts of metatextual (MT) and interpersonal (IP) categories, and their subcategories in each corpus are presented. Finally, a second and more qualitative analysis across disciplines (i.e. linguistics vs. medicine) and writer status (learner vs. expert) is performed to explore cases of overuse and underuse –these terms are used in a quantitative sense, that is, to refer to the highest or lowest differences in frequency when comparing the learners' and the experts' production– to finally draw some pedagogical implications.

7.6.1 Overall frequency of reflexive MD markers

The overall frequency results of the two main types of MD markers (i.e. textual and interpersonal) across disciplines is shown in Table 18 and illustrated in Figure 6. Appendix 6 presents global results of all MD categories and subcategories explored, and it provides both raw and normalized results.

Table 18: Total production of reflexive MD in BDs and RAs according to discipline (per 1000 words)

	LIN		**MED**	
	BDs	**RAs**	**BDs**	**RAs**
Total MT	32	32.5	25.4	24.6
Total IP	3.2	6.2	6.3	4
Total MD	35.2	38.8	31.7	28.6
Total MD %	3.5%	3.8%	3.1%	2.8%
Avg. markers per text	229	275	123	92

The analysis of textual markers in both corpora reveals that both learners and experts have used MD to a similar extent. However, this is only true if we look at the texts according to discipline (linguistics and medicine), which suggests that disciplinary conventions do play an important role in the choice of MD practices. These global results support frequency findings across disciplines reported in the literature (e.g. Hyland, 2001; Hyland & Tse, 2004; Hyland, 2010; Salas, 2015). As can be seen, linguistics contains more MD markers in general (RAs 38.8, BDs 35.2) than medicine (RAs 28.6, BDs 31.7). In fact, medical RAs contain the lowest amount of MD markers in all five measures: total metatext (MT), total interpersonal (IP), total metadiscourse (MD), percentage of MD (%), and average markers per text. It is interesting to note, however, that BDs in medicine contain a higher frequency of interpersonal markers than any other subcorpus in this study, being almost twice as frequent as in the medical RAs; this points towards a case of overuse that will be explored further in Section 7.6.5.

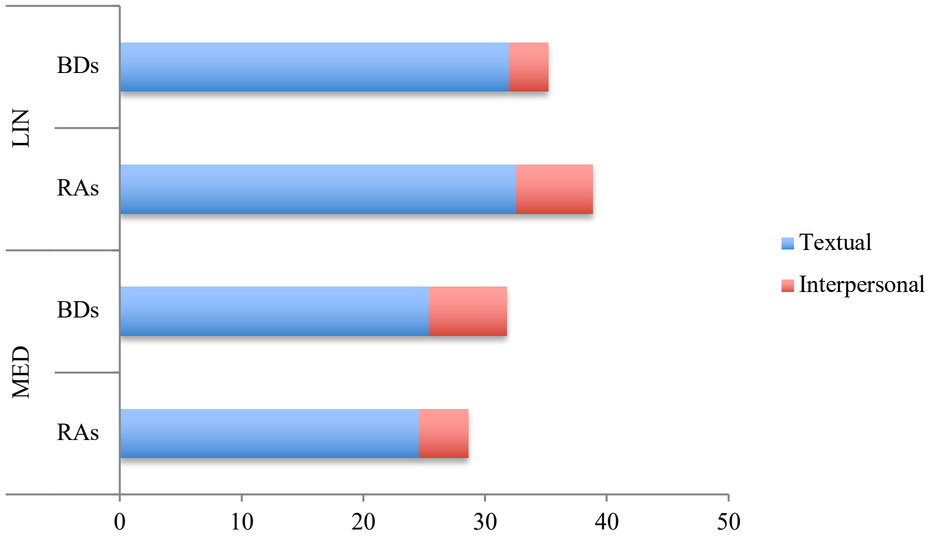


Figure 6. Textual and interpersonal MD in linguistics and medicine

In Figure 6, we can see how textual markers have been used much more frequently than interpersonal markers (there are many more subcategories that belong to textual MD, which partly explains why); this finding is also in line with previous research (Dafouz, 2003; Hyland, 2001; Hyland 2010; Pérez-Llantada, 2010; Salas, 2015). It is interesting, however, to remark how both BDs and RAs in linguistics, and BDs and RAs in medicine have used textual MD to practically the same extent compared to one another (at least numerically). This could very well suggest that learners in this corpus are aware of the textual MD practices of their discipline, perhaps thanks to the exposure to RAs for their BD preparation. Another possible explanation is the fact that many of these textual markers (e.g. use of connectors, exemplifiers, reformulators) are often taught in English language instruction in secondary or tertiary education, so EFL students may feel more confident when using them. In spite of this quantitative similarity, there is nevertheless an interesting difference in the choice of makers within this category, which may reflect that the learning of these markers was not genre- or discipline-specific; this will be explored further in Section 7.6.2.

Regarding the use of interpersonal markers, the learner corpus has yielded somewhat unexpected results: while BDs and RAs seem to agree in their use of textual markers according to discipline, the use of interpersonal markers varies greatly in all four subcorpora, as illustrated in Figure 6. BDs in linguistics have used half as many interpersonal markers (3.2) as the RAs (6.2), and the opposite tendency occurs in BDs in medicine (6.3) compared with the RAs in the same discipline (4.0). Although it is difficult to find the exact reason for these differences, a possible explanation could be related to the fact that BDs and RAs have different audiences: a BD displays knowledge to a supervisor and the evaluating committee, while RAs display knowledge to peers of more or less the same expertise. Mauranen (2001, p. 209) hypothesized that "those in a dominant position in any speech event will use more reflexive expressions". However, this is only true for the linguistic subcorpora, and not for the medicine subcopora, in which the learners have produced more MD in general than RA authors. In any case, I believe that the lack of explicit teaching on the use of writer, reader, and participant-oriented mentions in different disciplines may account for this quantitative difference. Let us have a closer look at each of these categories (textual and interpersonal) across disciplines in order to see these differences in more detail.

7.6.2 Textual metadiscourse

Table 19 displays the categories and subcategories that belong to textual MD. The most significant differences in each subcategory are explained below.

Table 19. Frequency of reflexive metatext in BDs and RAs (normed per 1000 words)

	BDs		**RAs**	
	LIN	**MED**	**LIN**	**MED**
Reference to the text				
Full text	1.76	1.39	1.81	1.22
Part of the text	1.96	1.16	1.10	0.60
Semiotic modes	1.21	0.80	2.68	3.07
TOTAL RT	**4.94**	**3.35**	**5.59**	**4.89**
Endophoric markers				
Anaphoric	1.38	0.46	0.90	0.59
Cataphoric	0.81	0.67	1.19	0.64
Deictic	0.20	0.00	0.75	0.02
TOTAL EN	**2.39**	**1.13**	**2.83**	**1.26**
Code Glosses				
Reformulators	2.12	1.01	2.50	1.34
Exemplifiers	2.38	0.80	3.93	1.64
Parentheticals ()	4.08	7.53	3.49	5.57
Dashes (–)	0.34	0.00	0.23	0.16
Colons (:)	2.95	2.55	1.56	0.65
Semicolons (;)	1.03	0.62	1.22	1.68
TOTAL CG	**12.89**	**12.50**	**12.93**	**11.04**

Linking Devices				
Adding	2.42	2.60	1.59	1.88
Constrasting	4.66	2.81	4.43	2.82
Consecutive	1.69	1.50	1.96	1.13
Organizers	2.33	1.16	2.19	1.39
Topicalizers	0.68	0.36	1.04	0.18
TOTAL LD	**11.78**	**8.43**	**11.22**	**7.42**
TOTAL METATEXT	**32.00**	**25.42**	**32.57**	**24.60**

Reference to text

RAs in linguistics have included more references to the text (5.6) than any other subcorpus, followed by BDs in linguistics (4.9). According to these findings, authors in the field of linguistics tend to refer to the full text (e.g. our *paper*), and to parts of the text (e.g. the next *section*) more often than authors of other disciplines. On the other hand, RAs in medicine seem to contain more references to semiotic modes (e.g. see *figure* 1); to be more precise, there is an average of 10 references to semiotic modes per paper (*figure* is the no. 1 semiotic mode in medicine RAs), whereas in the medical BDs corpus, there is an average of 3 references per text. Learners in this corpus do not refer to their semiotic modes (tables, figures, diagrams) as often as the RA authors. We will return to cases like this in Section 7.6.3.

Endophoric markers

As shown in Table 19, the linguistic subcorpora contain more endophoric markers than medicine. There is, however, a notable difference: BDs in linguistics have used anaphoric markers (e.g. *as mentioned previously*) more frequently (1.3) than the RAs (0.9). In contrast, RAs have used cataphoric markers (e.g. *as follows*) to tell the reader to look forward in the text, more often: cataphoric markers help foreground upcoming material, so the reader knows what is next, and where to find that information. The frequent use of anaphoric markers by learners in linguistics (average of 9 anaphoric references per text) may have made some parts of their texts a bit redundant. Another important observation here is the fact that the medicine subcorpus (both RAs and BDs) contain very few –or practically none– deictics (e.g. *here, now*).

Code glosses

Markers in this category are the most popular ones in the corpus. Exemplifiers (e.g. *for instance*) and parentheticals (e.g. *(see table 2)*) abound in all four subcorpora. The former is one of the most frequent MD subtypes in RAs and BDs in linguistics (3.9 and 2.3 per 1000 words respectively). Authors of this discipline tend to provide the reader with many examples in order to illustrate their points. *Such as*, *e.g.,* and *for example* are the top-3 markers that help authors exemplify in their texts (see Appendix 7 for a list of the top-3 textual and interpersonal markers in each subcorpus). The latter, parentheticals, is one of the most frequent markers in the medical BDs. Learners have used parentheticals to refer the reader to different sections in their text, or to specify the type of variable they have used, as in example (5) below:

(5) MED_BD01: Measured trough Charlson Comorbidity Index (Charlson/Deyo version*)_MD_MT_CG_PA* with data figuring in the clinical course (see Annex IV*) _MD_MT_CG_PA*. This variable will be categorized (…)

In the case of colons, they have been used more frequently in the BDs (LIN 2.9, MED 2.5) than in the RAs (LIN 1.5, MED 0.6), and they often appear after the cataphoric marker *the following,* preceding examples or lists of concepts, as in (6):

(6) LIN_BD04: Some examples of epistemic modality are*: _MD_MT_CG_CL* "We may/might lose the elections / They must have won the elections"

In contrast, semicolons have been used much more frequently in medical RAs (1.6), especially before *and*, *however,* and *therefore,* as illustrated in (7):

(7) MED_RA05: It may be presumed that physicians prescribe statins to patients who suffered more severe obesity*;_MD_MT_CG_SC* therefore, statin users could have been more likely to develop diabetes and diabetic complications.

Regarding the use of dashes, they were only found in BDs and RAs in linguistics, in particular before -*also known as, -and,* and *–thus;* authors used single (–) double (--), or even triple (---) dash at the beginning, and sometimes also at the end of the commentary, as can be seen in (8) and (9):

(8) LIN_BD08: This can be accounted for two different principles: a weak one –*_MD_MT_CG_DA* also known as linguistic relativity– and a strong one –*_MD_MT_CG_DA* also known as linguistic determinism–.

(9) LIN_RA13: In the same vein, the Pidgin uses full-NPs to signal anaphoric --*_MD_MT_CG_DA* and thus by logical inference (22a) also cataphoric discontinuity.

Linking devices

Two subtypes –adding and contrasting– were the most popular ones in the corpus across disciplines. BDs and RAs in linguistics contain 11.7 and 11.2 linking devices per 1000 words respectively, whereas BDs and RAs in medicine contain notably fewer markers in this category (8.4 and 7.2 respectively). Within the linking devices category, contrastive markers are more frequent than additive markers, especially in BDs in linguistics (4.6) –almost twice as many as in medical BDs (2.8). *However* is the number one contrastive marker in all corpora, followed by *therefore* and *thus*. On the other hand, the most popular additive marker is *in addition*, followed by *moreover* and *furthermore*. It is also worth mentioning that there are two subtypes, i.e. organizers –illustrated in example (10), and topicalizers –in (11), that mainly appear in the linguistic corpus only. *In terms of, in the context of* and *with respect to* are the top-3 topicalizers in the corpora:

(10) LIN_BD03: *First_MD_MT_LD_OR,* an overview on what ToM means (…). *Then_MD_MT_LD_OR*, different theories on which elements of language foster ToM development are explained (…). *Finally_MD_MT_LD_OR*, the view of those who deny the role of language (…)

(11) LIN_RA12: *With respect to_MD_MT_LD_TO* vocabulary acquisition from a supportive reading context, the results showed that providing explicit clues can result in relatively high lexical gains (…)

7.6.3 Overuse and underuse of textual markers

If we look at the total production of textual MD according to discipline, as we did earlier, we see that BDs and RAs in linguistics (32 and 32.5) and BDs and RAs in medicine (25.4 and 24.6) contain quite a similar amount of textual markers. However, when we look in more detail at the type of MD markers used in each category, important differences emerge. It is interesting to note here that, in the case of textual MD, all cases of overuse and underuse are found in both subcorpora of BDs, regardless of their discipline, which could highlight learner-writing features as opposed to conventions of different disciplines, in this case.

First, BDs in general refer to parts of the text (e.g. in this *section*) more often (LIN 1.9, MED 1.1) than RAs (LIN 1.1, MED 0.6). This finding contrasts with Burneikaite (2008) who found that EFL learners in fact underused endophoric markers, producing somewhat unstructured texts. It could be argued that learners in this corpus have a higher audience-awareness: they indicate and inform the reader, perhaps too often, about the different sections of their texts. In contrast, however, the BDs do not include as many references to semiotic modes (LIN 1.2, MED 0.8) as the RAs (LIN 2.6, MED 3), even though they did include tables and figures in their dissertations. This could suggest that learners do not guide the reader enough through the semiotic modes presented in their texts; it is up to the reader, in some cases, to understand and analyse the information presented. This could be indicative of transfer from their L1 (Spanish), a slightly more reader-responsible writing style (Hinds, 1987), and thus, worthy of pedagogical attention. Regarding the use of exemplifiers, BD writers seem to have underused this type (LIN 2.3, MED 0.8) compared to RA writers (LIN 3.9, MED 1.6). Students may lack confidence, or may not know enough so as to give examples about certain topics. It is also possible that, having a supervisor who knows well the topic as the intended reader of their text, students may not feel the need of giving many examples in their dissertations. Another difference found in the analysis concerns the use of colons: BDs have used colons much more frequently (LIN 2.9, MED 2.5) than authors of the RAs (LIN 1.5, MED 0.6). In addition, and with regard to semicolons, it is important to mention that medical BDs contain very few semicolons (if at all) (0.6), which contrasts with the use of semicolons in published RAs in the same discipline (1.6). Colons and semicolons are important typographical devices that introduce reformulations or examples. This finding suggests that learner writers need to revise the use of these two punctuation marks in academic writing.

Finally, and again, in the learner subcorpora, additive markers are used much more frequently in the BDs (LIN 2.4, MED 2.6) than in the RAs (LIN 1.5, MED 1.8). Spanish L1 writers of English have preferred to "add" ideas to their argument to achieve credibility, which is a common practice in academic literature written in Spanish, rather than including pros and cons of the discussed topic, or contrasting findings and different perspectives on the matter, which is a common practice in academic literature written in English. This finding is in line with previous studies (Dafouz, 2003; Pérez-Llantada, 2010) that suggest writers may retain part of their Spanish L1 writing style when writing in English. More pedagogical attention should therefore be given to culture-driven preferences in general, and to the use of linking devices in academic texts in particular.

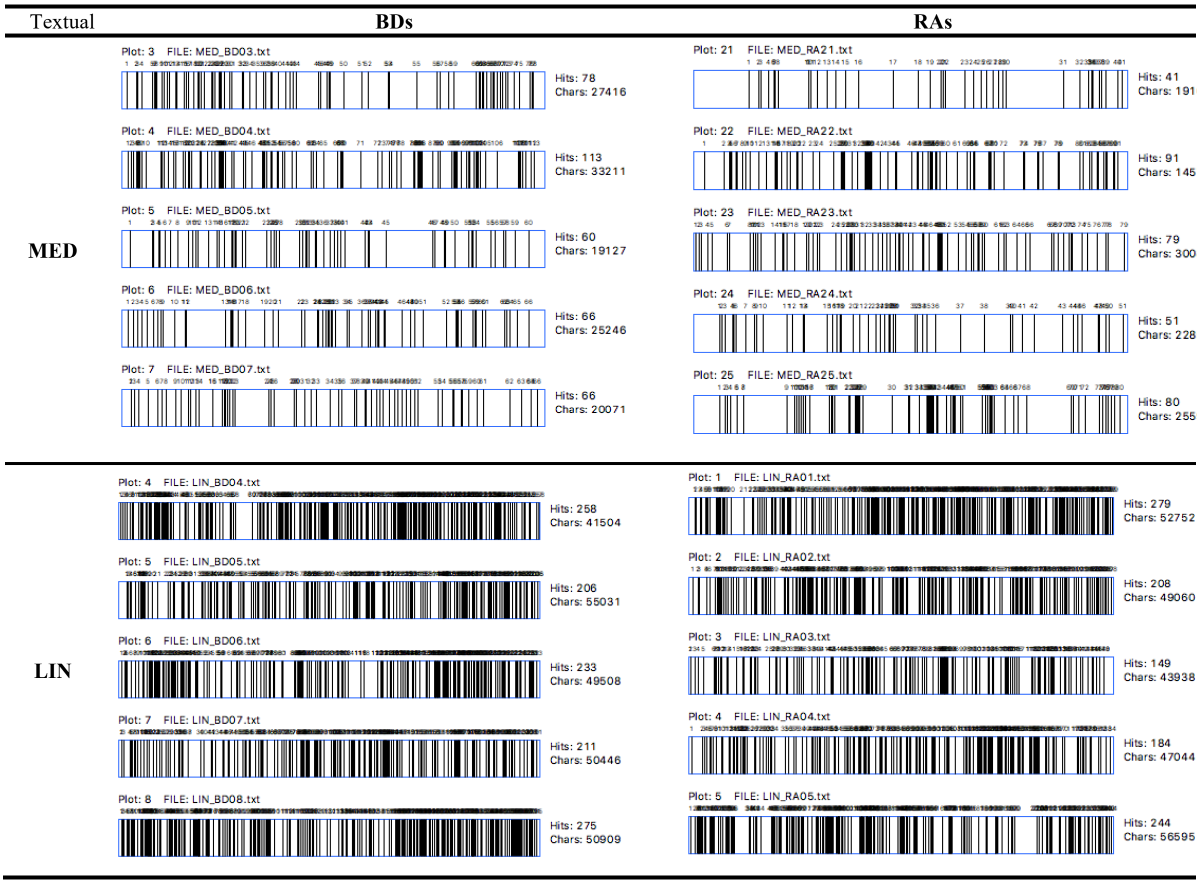


Figure 7. Concordance plot of the use of textual markers

Figure 7 shows where exactly textual markers (of all subtypes) occur along the texts and how frequently. Five random texts of each subcorpus have been selected to illustrate the plot. We can observe how there is a similar dispersion (distribution of vertical lines) of textual markers across potentially different sections (e.g. introduction, method, conclusion), and texts (e.g. LIN_BD04, MED_BD09, MED_RA14), but also a different density (thicker lines represent higher frequency) according to discipline. This frequency and distribution of textual markers contrasts very much with the use of interpersonal markers, which can be seen in Figure 8, at the end of the next section.

7.6.4 Interpersonal reflexive metadiscourse

Turning now to the use of interpersonal MD markers, we can see some remarkable differences: as illustrated in Table 20, and as mentioned earlier, BDs in medicine contain the highest frequency of interpersonal markers –especially self-mention (6.3 markers per 1000 words)– than any other subcorpus analysed in this study.

Table 20. Frequency of interpersonal markers in BDs and RAs (normed per 1000 words)

	BDs		RAs	
	LIN	**MED**	**LIN**	**MED**
Writer oriented				
Self-mention	1.83	5.13	4.10	3.82
Reader oriented				
Rhetorical Questions	0.05	0.00	0.07	0.00
Directives	0.43	0.90	1.13	0.17
TOTAL RO	**0.48**	**0.90**	**1.20**	**0.17**
Participant oriented				
Inclusive we	0.89	0.34	0.98	0.04
TOTAL INTERPERSONAL	**3.19**	**6.37**	**6.27**	**4.03**

If we look closely at each category, writer-oriented (i.e. self-mention) markers are more popular than reader- or participant-oriented markers in all subcorpora, regardless of their discipline: *we* is the most preferred marker of self-mention –even for single-authored texts, which sometimes serves a hedging purpose (Hyland, 2001)–, followed by *our*, *I* (only in the case of linguistics) and *us* (see Appendix 7). Also important to note here is that *directives* are much more frequent in RAs in linguistics (1.1), compared to the other subcorpora, as can be seen in Table 20; the way authors prefer to direct to the reader is by means of the imperative *see* as in (12):

(12) MED_RA11: one that does not in itself create the potential for contamination of the environment in which it is used (*see_MD_IP_RO_DI* Box 1 for experimental details).

As for rhetorical questions (see e.g. (13) and (14)), they were infrequent in general, and only present in BDs and RAs in linguistics:

(13) LIN_BD05: What does this mean*?_MD_IP_RO_RQ* According to the interpretations provided before (…)

(14) LIN_RA22: What lessons for syllabus design can one draw from these findings*?_MD_IP_RO_RQ* As far as prepositional postmodifiers are concerned (…)

7.6.5 Overuse and underuse of interpersonal markers

In the case of interpersonal markers, it is important to mention that the cases of overuse apply to medical BDs only, whereas the cases of underuse apply to BDs in linguistics: medical BD writers have produced self-mention (15), directives (16), and inclusive we (17) (5.1, 0.9, and 0.3 respectively) much more frequently than RA writers (3.8, 0.1, and 0.04 respectively), and in some cases, even more frequently than BDs and RAs in linguistics.

(15) MED_BD02: *We_MD_IP_WO_SF* expect to observe the existence of additional benefits, not explained by the weight loss alone, (…)

(16) MED_BD07: *Take_MD_IP_RO_DI* the high number of atypical squamous cells of unknown significance (…) detected by Pap test for instance, (…)

(17) MED_BD05: about the incidence of schwannoma as reference, *we_MD_IP_PO_IW* could find that the incidence of the vestibular schwannoma (VS) has been stabilized (…)

With regard to underuse, self-mention appears much less frequently in linguistic BDs (1.8) than in RAs (4.1). In addition, directives also seem to have been underused by BD writers in linguistics (0.4) compared to RAs (1.1). These findings thus have an important implication, namely that undergraduate students need more explicit instruction on the use of interpersonal markers taking into account their specific field of study.

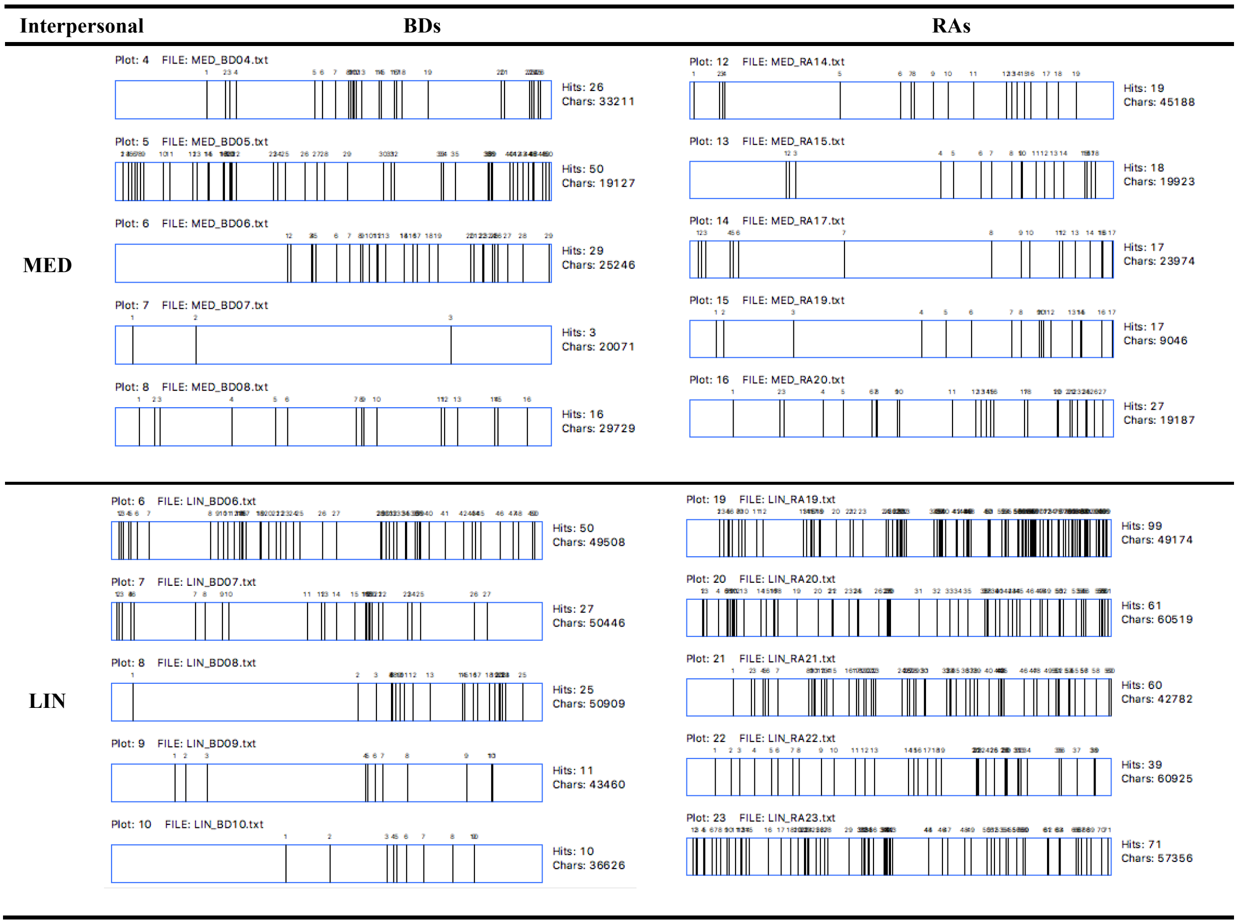


Figure 8. Concordance plot of the use of interpersonal markers

Figure 8 shows the concordance plot of interpersonal markers. Two interesting points arise here: first, the density is clearly much less prominent than that of textual markers (see Table 20 above); and second, the dispersion of these markers in the BDs and RAs is different: if we take a closer look at the plot, we can see how RAs have used interpersonal markers (especially self-mention) mostly at the beginning and towards the end of the text (which could represent the introduction or methods, and discussion or conclusion sections); such a pattern cannot be found in the BDs, in which interpersonal markers have been used elsewhere. In addition, and in terms of density of interpersonal markers, it differs in both BDs and RAs, in both disciplines, so we can say that learners' use of interpersonal markers does not approximate the use of these by experienced writers.

7.6.6 Summary

This corpus-driven study has yielded results on the frequency as well as the usage patterns of reflexive MD markers produced by learners and expert writers. One of the first objectives of the present study was to find out the extent to which Spanish undergraduate students use reflexive MD in their academic texts. The results show that MD represents an average of 3.1% (BDs in medicine) and 3.5% (BDs in linguistics) of the total texts. Comparisons with the expert corpus show that overall learners have used MD to a similar extent (MD represents 2.8% in RAs in medicine, and 3.8% in RAs in linguistics). We may therefore say that EFL Spanish undergraduate students have produced reflexive MD to an appropriate extent in terms of frequency. The second objective was to detect differences across disciplines: the analysis shows that BDs and RAs in linguistics contain more MD in general than BDs and RAs in medicine (except for self-mention markers in medical BDs). This result supports previous findings reported in the literature about different conventions of MD across disciplines (Hyland, 2001; Hyland, 2010; Mur-Dueñas, 2011; Salas, 2015). Finally, the third objective was to see if there were any differences according to writer status –i.e. learners vs. experts. The analysis reveals that there is an extremely similar frequency of textual MD in both BDs and RAs, which suggests learners in this corpus are aware of their readership and have guided their readers appropriately through their texts. However, comparisons with the expert corpus has also allowed me to find cases of overuse and underuse of certain MD markers, which surprisingly apply to the entire learner corpus, regardless of their discipline. Moreover, and in terms of interpersonal markers, we saw that BDs neither approximate the use of *self-mention, inclusive we,* or *directives* in RAs, nor are they comparable to one another. Some of these findings could be indicative of a different genre (e.g. BDs display knowledge to a supervisor); they could also denote a more reader-responsible writing style, as a culture-driven preference (L1 transfer), or even be due to the conflicting advice on the use of *self-mention* devices in academic writing textbooks or provided by different supervisors.

7.7 Conclusion

This paper has analysed the density and range of reflexive textual and interpersonal MD markers present in two corpora, namely a learner corpus of BDs written in English by Spanish undergraduate students in two different disciplines (linguistics and medicine), and an expert corpus, consisting of RAs published in English-medium academic journals. The quantitative and qualitative analysis performed shows that overall, BDs and RAs in the same disciplines contain a similar amount of textual MD markers, which may indicate EFL Spanish undergraduate students are aware of the textual MD conventions of their community of practice, at least in terms of frequency of use. Under a closer look, however, a qualitative analysis shows that BD learner writers have used references to parts of their text, colons, and additive linking devices significantly more often than expert writers. On the other hand, learners seem to underuse references to their semiotic modes, exemplifiers, and semicolons. These cases of overuse and underuse of textual MD markers are present in both corpora, regardless of their discipline, which may highlight features of the BD genre on the one hand, and of EFL Spanish undergraduate students' writing on the other hand. Regarding the use of interpersonal MD, the learner corpus in this study has yielded interesting results: learners' use of interpersonal markers does not approximate that of more experienced writers at all: BDs in linguistics seem to underuse self-mentions and directives compared with RA writers, and the opposite tendency occurs in BDs in medicine, in which writers have referred to themselves and engaged the reader much more frequently than RA writers. And neither does their use of interpersonal markers approximate one another in the same genre (BDs). These cases of overuse and underuse are therefore worthy of pedagogical attention.

It is important, however, to expose the limitations of this study. The first limitation is related to the corpus size (362,194 total words), and the number of participants (20 undergraduate students): the manual analysis and tagging of MD markers in the corpora was very time-consuming and did not allow me to include more texts in the corpus; using a larger corpus would certainly help to make these findings more representative and generalizable. Second, the comparison of MD markers across corpora was done from a word-level scope (normalizing values per 1,000 words). It has been argued, however, that T-units may be a more appropriate basis for calculating density than words, since MD markers typically have a clause-level scope (Intaraprawat & Steffensen, 1995). Calculating the mean length of T-units in the corpus and using it as a basis for comparison between two corpora could provide different results. In a similar vein, including inferential statistics (such as the *t-test*) could enhance the reliability of the quantitative analysis. This study has also looked at *inter*disciplinary variation (linguistics vs. medicine), but not at *intra*disciplinary variation (e.g. texts on second language acquisition vs. texts on learner corpus research). Performing an intradisciplinary analysis to explore the differences in the use of MD within texts in the same discipline, but on different topics, would be something worth investigating. Finally, this study could be improved by using a classical contrastive analysis of parallel corpora (e.g. BDs in L2 English and in L1 Spanish) to study interlanguage, which could help to detect transferred MD practices from an L1 more accurately.

The results of this study have attempted to shed light on the types and frequency of reflexive MD makers in two somewhat similar genres (BDs and RAs) across two vastly different disciplines (linguistics and medicine). This analysis has also provided a comprehensive list of 230 MD markers in 21 different subcategories that may be of interest to EFL learner writers, and also to academic writing teachers and material developers, who are interested in teaching the use of MD in these two specific disciplines. To conclude, the findings of the present research corroborate the need for more explicit teaching and corpus-informed materials on MD: more pedagogical attention should be given to MD and its different categories, especially in EFL academic writing, taking into account the writers' L1, genre, and field of study, so that MD practices are taught and learnt in agreement with each cultural and disciplinary community.

8. Lexical bundles in learner and expert academic writing

8.1 Introduction

Over the last few decades, numerous corpus analyses have brought to the fore the fact that language is highly patterned (Hunston, 2002; Römer, 2010; Sinclair, 2005). Sequences such as *additional information* or *is one of the main*, especially common in particular registers, are 'ready to use' chunks, "stored and retrieved whole[s] from memory at the time of use" (Wray, 2002, p. 9) rather than generated item-by-item. These pre-fabricated units have been shown to facilitate production for authors and also save processing effort for readers and listeners (Nattinger & DeCarrico, 1992).

Collocations (see Ackermann & Chen, 2013; Nesselhauf, 2005), idioms (see Grant & Bauer, 2004), or lexical bundles, also called formulas, clusters, or chunks (see Biber et al., 1999; Cortes, 2004; Hyland, 2008a), are some of the different subsets studied in phraseology (Granger & Paquot, 2008; Meunier & Granger, 2008).

Lexical bundles (henceforth LBs) were first identified by Biber and colleagues (Biber & Conrad 1999, Biber et al., 1999) and have been defined as "the most frequently recurring sequence of words" (Biber & Barbieri, 2007, p. 264), as well as "important building blocks of discourse" (p. 270). The identification of LBs in corpus studies has been primarily based on corpus-driven approaches of frequency and range, following the pioneering lexical bundle approach developed by Biber, Conrad, and Reppen (1998). In order to qualify as a lexical bundle, a sequence needs an occurrence of at least 20 or 40 times per million words (Biber & Barbieri, 2007; Chen & Baker, 2010; Cortes, 2004). Range of dispersion (i.e. the number of texts in which the bundle appears) is normally set at 3 or 5 texts or 10% of the texts in the corpus (Hyland, 2008a). This criterion is used to guard "against idiosyncratic uses by individual speakers or authors" (Biber & Barbieri, 2007, p. 268).

Structurally, less than 5% of LBs represent complete structural units (Biber et al., 1999, p. 991), and are commonly used to bridge phrases (e.g. *in the case of*) or clauses (e.g. *I want to know*). Even though LBs are not structurally complete, they have been shown to perform major functions in discourse. They can also occupy different positions in a text. According to Hoey (2005, p. 13), lexical items "are primed to occur in or avoid certain positions within the discourse", which Hoey calls 'textual colligation', another feature that facilitates text processing and production.

Textual colligation analyses can help to reveal the interaction between positioning of LBs and discourse functions. In particular, there are two main sections of academic texts which tend to be highly conventional and contain certain LBs that help to accomplish rhetorical moves: these are the introduction and conclusion sections. Lexical items in these sections respond to genre and discipline conventions, since they reflect recurrent communicative purposes of a particular community. According to Bondi (2010, p. 99), "the ethos of the discipline –what the community considers appropriate methodology and relevant objectives– may have an impact on language choice". For example, finding the bundle *our study has shown* –which normally occurs in the conclusion section– earlier in the text (e.g. in the methods section) may strike the expert reader as an unusual practice and denote immaturity or foreignness on the part of the writer (Mur-Dueñas, 2011; Sheldon, 2018).

Each mode (e.g. written), genre (e.g. student essay), register (e.g. formal), and discipline (e.g. medicine) tends to "employ a distinct set of lexical bundles, associated with [its] typical communicative purposes" (Biber & Barbieri, 2007, p. 265). Thus, there seems to be no "single pool of lexical bundles" (p. 265) writers or speakers can draw on. In order to demonstrate membership in a given community, authors need to successfully use the LBs that are typical of that genre and discipline (Ädel & Erman, 2012). Writers who lack experience or exposure to the target language in a particular register may not choose the most appropriate expressions, and will not easily be accepted as an 'insider' of that community (Durrant & Mathews-Aydınlı, 2011; Hyland, 2008a; Wray, 2002). Unfortunately, knowledge of phraseology does not seem to be something innate: it is indeed far from being a "language universal skill" (Pérez-Llantada, 2014, p. 85). Due to their quantity and diversity, L2 and novice writers may find LBs difficult to acquire and master (Liu, 2012); in this respect, problems such as underuse, overuse, or misuse (both structural and functional) of certain bundles have been reported in the literature (see Ädel & Erman 2012; Chen & Baker, 2010; Meunier & Granger, 2008).

The present study aims to further the understanding of phraseology in learner writing by exploring the use of LBs in the introduction and conclusion sections of bachelor dissertations (BDs) written in English by Spanish L1 university students in linguistics and medicine. In order to compare the frequency of form, structure, and function of these bundles, an expert corpus of research articles (RAs) in the same disciplines is used as the reference corpus. The comparisons will be made from both a quantitative point of view –applying a corpus-driven approach to identify bundles in the learner and the expert corpus– and a qualitative approach –classifying the bundles structurally and functionally in both corpora. This study hopes to contribute to the body of research that studies phraseology in academic writing, and to serve as a useful pedagogical resource for L2 learners of English who are trying to accommodate to the conventions of these specific disciplines.

8.2 Literature review

Among the numerous studies on LBs over the last decades, we find comparisons of different populations (e.g. native *vs.* non-native speakers or students *vs.* experts [Ädel & Erman, 2012; Appel & Wood, 2016; Chen & Baker, 2010; Durrant & Mathews-Aydınlı, 2011; Hyland, 2008a]), genres (e.g. RAs *vs.* textbooks [Bondi, 2010; Römer, 2010]), disciplines (e.g. soft and hard sciences [Byrd & Coxhead, 2010; Cortes, 2004; Hyland, 2008a; Liu, 2012]), registers (e.g. written *vs.* spoken [Biber & Barbieri, 2007]), languages (e.g. academic Spanish *vs.* academic English [Pérez-Llantada, 2014]), and different sections of a text (e.g. introduction and conclusion [Bondi, 2010; Sheldon, 2018]).

One recurrent finding is that English L2 writers' use of LBs does not always approximate the use by expert or native writers in terms of frequency, form, and function. For example, the masters and PhD candidates' writings explored in Hyland (2008a) seemed to contain more impersonal clusters (i.e. avoiding stance), and more clusters in general compared to RA writers. The author suggests that less proficient writers rely on word combinations more often than expert writers. This finding contrasts with Durrant and Mathews-Aydınlı's (2011) study, in which student essays showed a lower production of formulas compared to RAs; differences regarding functional moves were also found. The authors suggest that the lack of attention paid to different genres and disciplines in academic writing education may account for these differences.

Another interesting finding in the literature in relation to our study is English L1 students' greater and more varied use of LBs, especially in structures such as unattended *this*, existential *there*, hedging and negations, as compared to that of L2 university students: L2 students' texts contained learner writing characteristic features, such as anticipatory *it* which, coupled with some informal lexical choices (e.g. *it is easy to*), pointed at register difficulties (see Ädel & Erman, 2012). In terms of functionality, L1 writers used stance more frequently than L2 writers. Interestingly, stance is one of the functions that differed the most among RA writers of the different languages (Spanish L1, English L2, and English L1) and disciplines studied in Pérez-Llantada

(2014) and in Sheldon (2018): English L2 writers were found to transfer some of their L1 (Spanish) rhetorical practices into their L2 writing, which made their texts less interactional.

In order to investigate the use of LBs by Spanish L1 undergraduate learners writing in English in two different disciplines (i.e. linguistics and medicine) and sections (i.e. introduction and conclusion) in comparison with their expert-writer counterparts, three research questions were established in this study:

1. What are the most common lexical bundles in the introduction and conclusion sections of L2 learners' BDs in linguistics and medicine?
2. How are these lexical bundles used in terms of structure and function?
3. To what extent does the use of lexical bundles approximate or differ from published RAs in the same discipline?

8.3 Data collection

In order to carry out a quantitative and qualitative analysis of LBs in academic writing, two corpora were compiled: (1) a learner corpus of BDs in linguistics and medicine written in English by Spanish L1 undergraduates in their last year of studies, and (2) an expert corpus of RAs in the same disciplines published in English-medium and peer-reviewed academic journals. The introduction and the conclusion sections of each text were extracted and saved as raw .txt files for their separate analysis. Table 21 describes the number of texts, tokens, types, and paragraphs per genre, discipline and section.

Table 21. The learner and the expert corpora

	BDs		RAs	
Discipline	Linguistics	Medicine	Linguistics	Medicine
Intro no. texts	10	10	25	25
Tokens	5,724	9,063	17,722	11,535
Types	1,409	2,367	3,057	2,717
Avg. words intro.	572.4	906.3	708.8	461.4
Avg. paragraphs	3.1	3.9	2.9	1.2
Concl. no. texts	10	10	25	25
Tokens	4,703	4,555	15,214	14,679
Types	1,370	1,353	2,771	3,005
Avg. words concl.	470.3	455.5	608.5	587.1
Avg. paragraphs	2.5	3.5	3.5	1.2
Total words	10,427	13,618	32,936	26,214

8.3.1 Extraction, filtering, and classification of lexical bundles

In the present study, a corpus-driven approach was adopted in order to retrieve LBs from the corpora –i.e. no previous assumptions were made with respect to the LBs' form or function, and no pre-defined list of bundles was used. The function 'cluster n-gram' in AntConc (Anthony, 2018) was used to extract LBs from the introduction and conclusion sections of the corpora. In terms of length, even though the 4-word scope is the most researched length in LB studies (Ädel & Erman, 2012), other studies suggest that many recurrent word combinations come in as 3-word bundles (Simpson-Vlach & Ellis, 2010); as a result, we decided to adopt a more inclusive approach and explore 3-, 4- and 5-word bundles in the texts. As for frequency, given the relatively small size of the corpora, the frequency cut-off was set at a minimum of 20 times per million words. In addition, a dispersion range of three texts, which represent three different writers, was set; the selection of these cut-off criteria was based on previous corpus studies (Ädel & Erman, 2012; Biber & Barbieri, 2007; Chen & Baker, 2010). It is important to note that when a bundle appears only on one of the lists, it does not mean that this specific bundle was not used at all by writers in the other subcorpora; as Ädel and Erman aptly put it, "it simply means that the frequency and dispersion criteria were not met in the other group's material" (2012, p. 85).

Once the LBs were automatically retrieved, manual filtering was required in order to eliminate undesired 'noise' that could affect the comparability of the multidisciplinary corpora –i.e. context-dependent bundles– and that could also inflate the results –i.e. overlapping bundles. To deal with the first type, context-dependent bundles such as *second language acquisition, native and non-native speakers, stem cells management* were manually eliminated from the lists. The second type, overlapping bundles, involved combining sequences such as *as a result* and *as a result of,* in which *of* appears in brackets (e.g. *as a result (of)*). Frequency, range, number of grams (i.e. number of words in the sequence), and section (introduction and/or conclusion) in which each bundle appeared were explored.

With regards to the grammatical structure of LBs, we initially followed Biber et al.'s (1999, pp. 1014-1024) classification, which distinguishes 12 structural categories for LBs in academic prose. After revising this and the taxonomy they provide for conversation, we present a taxonomy of 15 categories with four broad structural groups: 'noun phrase-based', 'prepositional phrase-based', 'verbal phrase-based', and 'other' bundles, following Chen and Baker (2010, p. 34), which can best integrate the LBs found in our data. The NP-based bundles include noun phrases, with or without post-modifier fragments (e.g. *the risk of*, *the most prevalent).* PP-based bundles refer to those starting with a preposition plus a noun-phrase fragment (e.g. *of this paper*, *in addition to).* The VP-based broad category is the largest group, integrating nine different structures, all containing a verb component –or an introducing element of a clause (e.g. *it is a*, *can be used to*, *to do so).* Different structural patterns are included here, such as subject + predicator structures, other verb phrase combinations, such as those followed by a noun-phrase or prepositional-phrase fragment, those containing a passive verb, anticipatory *it* structures, and patterns with the clause-introducing elements *that* and *to.* This structural classification involved manual revision and classification of all bundles according to their structures (e.g. *the study of* was categorized as 'noun phrase with *of*-phrase fragment').

For the functional classification, on the other hand, we followed previous taxonomies (Biber, Conrad, & Cortes 2004; Cortes 2004; Hyland 2008a) and classified all bundles into three main categories and their subcategories:

1) Research-oriented –also called referential in other models (e.g. Biber et al., 1999): LBs in this category help writers to situate, contextualize and describe their research. There are four main subcategories: 1) location (e.g. *at the beginning, at the university*), 2) procedure (e.g. *the use of the, the purpose of*), 3) quantification (e.g. *a part of, one of the most*), and 4) description (e.g. *the size of the, the nature of the*).

2) Text-oriented –also called discourse organizers (Biber et al., 1999): these LBs are concerned with the structure of the text and the interrelations established between the ideas presented. There are four main subcategories: 1) transitions (e.g. *on the other hand, in contrast to the*), 2) resultative (e.g. *as a result, due to the fact that*), 3) structuring (e.g. *in the next section, in this study*), and 4) framing (e.g. *with respect to, in the case of*).

3) Participant-oriented: LBs in this category show writers' attitudes towards the ideational content and address readers directly or indirectly. It comprises two main categories: 1) stance (e.g. *may be due to, are likely to*), and 2) engagement (e.g. *as can be seen, it should be noted*).

This functional classification was complex not only because the categorization involves subjectivity, but also because some LBs can perform more than one function (Liu, 2012). A concordance analysis was performed in order to see the extended context of certain bundles that seemed multifunctional. For example, *the basis of* is a 3-word bundle that can act as a research-oriented descriptive bundle, as in (1)

(1) Findings from such a study can form the basis of learner-relevant form-focused instruction. (LIN_RA01_I)[20]

But, when this sequence is part of the 4-word bundle *on the basis of,* it can mark a text-oriented resultative relationship, as in (2)

(2) Other linguistic accounts differentiate the two forms on the basis of information status, particularly in terms of topic. (LIN_RA15_I)

For those cases in which the authors could not agree on the categorization, even after analysing their extended context, previous literature that included examples on LBs and their functional categories was consulted (Cortes, 2004; Hyland, 2008a; Pérez-Llantada, 2014). These structural and functional classifications allowed us to better understand the use of LBs in the corpora studied.

8.4 Results and discussion

The results of the analysis of LBs are reported on as follows. First, the most frequent LBs in the introduction and conclusion sections of BDs and RAs in medicine and linguistics are explored. Convergent bundles (i.e. those bundles that appear on more than one list) are then presented. Finally, a second and more qualitative analysis of the structures and functions of bundles is presented, exploring the similarities and differences found in the corpora.

[20] In the examples, the following abbreviations are used: MED (short for medicine) or LIN (linguistics) indicates the discipline, BD (short for Bachelor Dissertation) or RA (Research Article) indicates the genre, and I (short for introduction), or C (conclusion) indicates the section in which the LB was found. The number is the identification number assigned to each text.

8.4.1 Frequency and convergence of lexical bundles in the corpus

There are a total of 218 different bundles in the corpus as a whole (for the full list, see Appendix 9) with a total frequency of 1,151 hits, which represents around 4.5% of the tokens in the corpus. The most frequent bundle is *the use of* with a raw frequency of 85 counts, which equals more than 1000 times per million words (pmw) in our corpus. Moreover, *the use of* appears in all genres and disciplines explored in this study, so it could be regarded as a core or convergent bundle, following Pérez-Llantada's (2014) nomenclature. It is noteworthy to mention that *the use of* appears in the conclusion section of the corpora 50 out of 85 times, clearly indicating a preference for the last sections of a text. RAs in linguistics (37) and in medicine (21) are the genres that contain more hits of *the use of*, very often paired with other nouns (*questions, tools, English, other alternatives, somatic stem cells*). This bundle seems to help writers to display results, as in (3) or limitations, as in (4).

(3) Trends for the social science fields indicate a reduction in <u>the use of</u> these informal features. (LIN_RA04_C)

(4) Another limitation was <u>the use of</u> asymptomatic microembolic signals as a surrogate marker. (MED_RA02_C)

The second most frequent bundle in the corpus is *in order to*, with a raw frequency of 62 counts, i.e. about 750 pmw. By contrast to *the use of*, this bundle appeared in the introduction sections of the texts more often, in particular, 39 out of 62 times. Taking into account the total number of words in each corpus, BDs in linguistics show a predominant use of this bundle (22 raw hits) followed by RAs in linguistics (24), BDs in medicine (12), and medical RAs (6). Different procedure verbs such as *address, determine, provide, show, solve, facilitate,* and *gain* are used after this bundle. *In order to* can help writers to emphasize the study's main objective or justification, as in (5) and (6) respectively.

(5) This study aims to analyse comprehension and production of false friends in students of English in a C1 level classroom <u>in order to</u> *explore* the influence of their mother tongue (L1) on a second language (L2). (LIN_BD10_I)

(6) Moreover, DTC's low prevalence requires the participation of a high number of medical centers in order to *obtain* a representative sample of patients. (MED_BD09_I)

The third most frequent bundle is yet another core bundle present in all subcorpora: *as well as* (43 hits). *As well as* appears more frequently in the introduction sections (24 times), and rather than just adding new information, this bundle helps writers to focalize and frame the ideas presented, as in (7) and (8):

(7) FN is a dimeric glycoprotein that is found in plasma as well as in the extracellular matrix (ECM) of various tissues (MED_RA03_I)

(8) Conclusions will be drawn to justify the analyzed usages of discursive strategies as well as the historical and social consequences that can derive from them. (LIN_BD02_I)

The use of, *in order to* and *as well as* are also included on Biber et al.'s (1999) list of the most common 3-word bundles in academic prose. These three bundles appear as well in the academic formulas list developed by Simpson-Vlach and Ellis (2010), and are in the top-200 'formulas worth teaching' (ranking 29, 4 and 5 respectively), which underlines their pedagogic relevance.

In terms of length, 3-word bundles were the most frequent in the corpus (85.7% of the total bundles), while 4- and, especially, 5-word bundles were scarcely used (10.2% and 3.9% respectively). This finding was similarly reported on in previous studies, such as Biber et al.'s (1999, p. 994), who found that 3-word bundles were much more frequent in academic prose (over 60,000 times pmw) than 4-word bundles (which occur over 5,000 pmw).

If we look at each subcorpus separately, in particular, we will find some interesting patterns. As can be seen in Table 22, BDs in medicine and linguistics have produced almost the same quantity of LBs in the introduction and conclusion sections (conclusions were a bit shorter in this genre compared to the introduction, which partially explains why they contain half the amount of LBs as introductions); this seems to point at a shared quantitative feature in the use of LBs between texts of two different disciplines but that belong to the same genre (BDs). This is only true, however, for the learner genre; RAs show a vastly different use of LBs in terms of frequency: even though there are the same number of texts, with similar tokens for both introduction and conclusion sections, articles in linguistics contain almost three times more LBs than medical articles.

Table 22. Lexical bundles used in the learner and the expert corpus

	BDs		RAs	
Discipline	Linguistics	Medicine	Linguistics	Medicine
LBs intro.	23*	25	74	17
LBs concl.	10	8	73	23
Total LBs	33	33	147	40
Total freq.	156	131	674	190
N-grams	3-w (24)	3-w (30)	3-w (125)	3-w (38)
	4-w (5)	4-w (2)	4-w (17)	4-w (2)
	5-w (4)	5-w (1)	5-w (5)	5-w (0)

*all values are raw counts

This finding has been supported by previous literature on LBs in academic writing across disciplines (Hyland, 2008a; Liu, 2012) and points towards a disciplinary difference: research suggests that soft-knowledge disciplines very often emphasize interpretative language in order to present persuasive arguments, compared to hard-knowledge disciplines, that tend to be more impersonal in their methods and discussions. The linguistic items that allow writers to achieve this objective are, more often than not, part of recurrent word combinations (e.g. *it is important to, has the potential to, it can be argued that, are likely to, seems to be, it should be, needs to be*), which can explain the prominent LB occurrences in linguistic RAs. Hyland (2008a) reported that less mature writers had used LBs more often. This finding contrasts with our results, but only for one of the two disciplines: BDs in medicine do contain more LBs than RAs in the same discipline (3.3 *vs.* 1.6 bundles on average per text); particular characteristics of the BD genre with regards to its audience –for example, that of being an academic final assignment in which students need to show and convince their supervisors (as a superior entity) that they have acquired certain knowledge– can contrast with published RAs in which authors present information to peers (of more or less the same expertise) and could account for this quantitative difference.

Adopting another perspective, the comparison of all LBs lists has yielded an inventory of 35 shared bundles. Some of these bundles are shared in the introduction and conclusion section of the same subcorpus, but some are also shared between genres (BDs, RAs), disciplines (linguistics, medicine), and some of them appear in all lists, regardless of their genre or discipline, what we call core bundles. These 35 bundles are the best candidates for general academic writing education and, supporting Pérez-Llantada (2014, p. 88), this inventory "might indicate that the writers have memorized these language sequences and routinized them in their writing practices". Table 23 shows convergent bundles in the corpora:

Table 23. Convergent LBs found in the corpora

	LIN BD intro	LIN RA intro	MED BD intro	MED RA intro
Discipline*	*in order to, in this paper, it has been, the fact that, the use of, there is a*		*as well as, in order to, the prevalence of, the risk of, the use of*	
Genre	in order to, the use of	a number of, as well as, in order to, the use of, there is a	N/A	N/A
	LIN BD concl	LIN RA concl	MED BD concl	MED RA concl
Discipline	*as well as, in order to, the fact that, the use of, this study has*		*the results of*	
Genre	in order to, one of the	as well as, in this study, the current study, the present study, the use of, there is a	N/A	N/A
	LIN BD	LIN RA	MED BD	MED RA
Core LBs	*in order to, the use of, as well as*			
Intro/ concl.	in order to, it has been, the fact that, the use of	as well as, based on the, differences in the, in order to, in relation to, in terms of, in this paper, in this study, of the most, some of the, the current study, the fact that, the importance of, the number of, the present study, the role of, the use of, there is a, understanding of the	in order to	a number of, as well as, the presence of, the prevalence of, the use of, there is a

*The 'discipline' row shows LBs that are shared between LIN and MED, regardless of the genre. The 'genre' row, on the other hand, shows LBs shared between BDs, and between RAs only. That is why each row displays two cells, and not four.

As Table 23 shows, there are more LBs shared by discipline (linguistics shares 11 bundles, and medicine shares 6, in both introduction and conclusion sections) than by genre (BDs share 4 bundles, and RAs share 11). The fact that BDs, despite having noticeably fewer tokens than RAs, share more bundles with their respective discipline in a published genre than with their learner counterparts indicates the important role disciplinary conventions play in academic writing.

If we look at specific bundles, as previously mentioned, *the use of* (85 hits), *in order to* (62) and *as well as* (43) are core bundles shared across all corpora in our study. Hyland (2008a, p. 12) found a total of 5 core bundles across four disciplines (*on the other hand, as well as the, in the case of, at the same time,* and *the results of the*), which is somewhat similar to our results. In terms of bundles that appear in both the introduction and conclusion section of BDs and RAs, there are a total of 23 different bundles, 19 of which appear in the introduction and conclusion sections of RAs in linguistics; these items can be a useful resource for L2 writers of academic English. Convergent bundles not only vary in their grammatical structure but also in the discourse functions they perform, as we will see in the next section.

8.4.2 Structures and functions of lexical bundles in the corpus

Table 24 below shows the frequency of LBs per structure across genres and disciplines, taking the four broad groups and the 15 structural categories into consideration, and provides one illustrative example for each category. An important caveat to understand the discussion of the findings that follows is that the frequencies given refer to the *type* of bundles used and not to the number of times each bundle type was used (*raw frequency*).

Table 24. Frequency of LBs per structure: overall figures per genre and discipline (%)

LBs structures		BDs		RAs		
		LIN	MED	LIN	MED	Example
NP-based	Noun phrase with of-phrase	30.3	33.3	30.6	35.0	*the use of*
	Noun phrase with other post-modifier	9.1	3.0	6.8	5.0	*the fact that*
	Other noun phrase (fragment)	0	3.0	3.4	5.0	*the present study*
PP-based	Prepositional phrase with embedded of-	3.0	-	7.4	5.0	*in terms of*
	Other prep. phrase (fragment)	12.1	18.1	17	10	*of the most*
VP-based	Pronoun/noun phrase (fragment) + be	12.1	12.1	5.4	10	*there is a*

	Noun phrase (frag.) + verb phrase (except copula be)	12.1	3.0	2.0	2.5	*this study has*
	Verb phrase with active verb	-	3.0	2.0	-	*seems to be*
	Verb phrase + noun phrase fragment	-	9.1	2.0	-	*has the potential to*
	Verb phrase + prep. phrase fragment	-	-	2.7	-	*refer to the*
	Passive verb + prep. phrase fragment	6.0	3.0	3.4	12.5	*based on the*
	Anticipatory it + verb phrase/adjective phrase	-	-	2.7	-	*it can be argued that*
	(Verb phrase +) that-clause fragment	-	-	5.4	-	*that they are*
	(Verb/adjective +) to-(clause) fragment	9.1	6.0	6.8	10.0	*in order to*
Other	Other expressions	6.0	6.0	2.0	5.0	*as well as*
	Total	100	100	100	100	

As can be seen, there is a clear prevalence of NP-based bundles over the rest of structural categories in all corpora. This prevalence is especially evident in the expert corpus, in both linguistics and medicine (both with a total frequency of more than 40%), over the second most common group of structures, the VP-based bundles. The PP-based categories rank in the third position in all four subcorpora. It is worth looking at specific rather than general structural categories to obtain a more realistic and clarifying picture of the findings. Of all 15 categories, the most common structure overall is the noun phrase with *of*-phrase, representing in all cases more than 30% of all categories, with the highest frequency in the medicine RAs (35%). In particular, we found a total of 78 bundles with this structure, with a raw frequency of 375 –that is, LBs belonging to this category account for 32% of the total frequency of LBs in the corpus as whole. Biber et al. (1999) indicate that as much as 70% of the most common bundles usually consist of a noun phrase with an *of*-phrase fragment. The prevalence of this structure has also been found in previous studies on LBs (Chen & Baker, 2010; Hyland, 2008a; Liu, 2012). As it could be expected given its high raw frequency, *the use of* is the most frequent bundle in this category (62 hits), with a higher presence in medicine RAs (21 hits). Other common examples are *one of the* (13 hits), *the analysis of (the)* (11 hits), and *the risk of* (11 hits). Examples (9), (10), (11) and (12) illustrate some of the most common LBs in this category:

(9) As the analysis of the selected linguistic features has illustrated, both adverbials and empty adjectives have been slightly more frequent in men's weblogs. (LIN_BD04_C)

(10) Disciplinary vocabulary also remains one of the most challenging areas. (LIN_RA12_I)

(11) That does not exclude the possibility of bias to the point where it is non-existent but it is an attempt to attenuate its effect. (MED_BD08_C)

(12) The use of different BMI reference values produced different prevalence estimates for the overweight category in the different populations. (MED_RA10_C)

The second most common structure is the other prepositional phrase, that is, bundles introduced by a preposition, excluding those with an embedded *of*-phrase; common LBs in this category are *of this paper, according to the, in this study,* and *of the most*. We noticed above that LBs tend to be incomplete structural units; when they can be used as potentially complete units, these tend to act as discourse signalling devices (Biber et al., 1999, p. 999). The category of other prepositional phrases is one of the two which may integrate these complete structural units: see the examples from our corpus *between the two, as a result, on the other hand, in this (present) study/paper, on their own*; the other is the category of other noun phrases, e.g. *the present/current study, the following three.*

We have already mentioned particular examples of bundles which are especially recurrent in our corpus. One instance is *in order to*, which we consider a *to*-clause fragment (rather than a prepositional-phrase pattern; cf. Pérez-Llantada, 2014, for instance), and partly explains the relatively high frequency of the (verb/adjective +) *to*-clause structural pattern in all subcorpora. In addition, our data show two further common structures of bundles in specific subcorpora. One of them is the passive verb (+ prepositional phrase) with a higher use in the medicine RAs, exemplified by bundles such as *is associated with*, *have been proposed*, and *can be used to*, which interestingly are all found in the conclusion section of these texts. The impersonal nature of the passive construction seems to fit well with the medicine discipline, in which writers allegedly attempt to hide authorial interpretation more than their linguistics counterparts. This finding supports disciplinary differences on structural categories reported on in Hyland (2008a, p. 11). The other structural category that shows a higher frequency than in other corresponding subcorpora is the noun phrase + verb phrase in BDs in linguistics. Examples of these bundles are *paper aims to*, *this paper will focus on* and *this study has.* We may hypothesize that this higher use is due to the emphasis placed on these non-agent text subjects in the teaching of academic discourse to university students.

A general tendency emerging from the figures represented in Table 24 relates to the variation in the use of LB structures. In this respect, RAs in linguistics show the greatest proportion of variation, as the only subcorpus illustrating all 15 categories. This subcorpus presents a rich range of different structural types of bundles, some of them of a more elaborated nature than in the learner corpus: e.g. the NP-based bundles *a growing interest in, our understanding of, body of research, avenues for future research,* and the VP-based bundles *has the potential to, play an important role in.* Compared with this wide range of bundles, BDs in linguistics exhibit a less illustrative choice, with seven structural categories not represented, which can be explained by the less proficient writing skills of these authors. In the medicine corpora overall, however, the choice of bundles is definitely less varied. Curiously enough, medicine RAs show a much lesser degree of variation and representativeness in the use of LB structures, even though they belong to the same genre as their linguistics counterparts. It is difficult to say why this might be, but disciplinary variation and the topic of linguistic articles itself (language) could account for the discrepancies found.

The analysis of LBs according to discourse function has also revealed interesting insights. Table 25 provides an overview of the LB functions across genres and disciplines. As can be seen, bundles with text-oriented functions are prevalent over the other two types in general. The second most common type of bundle are those performing research-oriented functions. The comparison between these two functional categories, however, provides an interesting disciplinary distinction: whereas in linguistics there is a significant difference in frequency between the text-oriented and research-oriented functions in both learners and experts, and a particularly high use of text-oriented bundles (over 50%) in BDs, in medicine, on the other hand, the figures are closer between these two functions, and in medicine BDs they are exactly the same. This is (partly) in line with Hyland (2008a, p. 14), who found a greater use of bundles with a referential function in the hard sciences to the same use in the soft-knowledge fields (i.e. linguistics), providing to the former "a greater real-world, laboratory-focused sense to writing", and thus emphasizing the empirical over the interpretative, as seen above. The more evident prevalence of text-oriented bundles in linguistics would also agree with this picture.

Table 25. Frequency of LBs per function: overall figures per genre and discipline (%)

Subcorpus	Research-oriented	Text-oriented	Participant-oriented	Totals
Linguistics BD	39.3	54.5	6.0	100
Linguistics RA	37.4	46.7	15.9	100
Medicine BD	42.4	42.4	15.1	100
Medicine RA	40.0	42.5	17.5	100

LBs with a participant-oriented function are far less frequent in our data with frequencies around 15%, except for the linguistics BDs, where the figure drops to only 6%. This underuse of participant-oriented bundles in our Spanish L1 writers agrees with findings in other studies that have noted an avoidance of stance bundles in learners in comparison with English L1 authors (see Hyland, 2008a, p. 19; Pérez-Llantada 2014, p. 91; Sheldon, 2018, p. 34). Pérez-Llantada (2014) notes that Spanish-speaking learner writers in English avoid personal markers to a greater extent than the corresponding expert writers of academic discourse. Our results also point to a potential lack of confidence on the part of the linguistics learners to express their stance and subjectivity.

In order to turn now to a more detailed analysis, we present Table 26 below with the figures of bundle types for the specific discourse functions included in each of the broad functional categories just mentioned. As with the discussion of the structure of bundles, a first thing to note is the greater and richer variety of functions in the linguistics RAs, with all ten categories represented in the table, in comparison with the other three subcorpora.

Table 26. LBs functions and their subcategories (%)

LBs functions		BDs		RAs	
		LIN	MED	LIN	MED
RES	Location	6.0	-	4.7	-
	Procedure	15.1	3.0	13.6	7.5
	Description	6.0	18.1	6.8	25.0
	Quantification	12.1	21.2	12.2	7.5
TEX	Transitions	-	3.0	4.0	2.5
	Resultative (inferential)	9.1	21.2	8.1	10.0
	Structure (identify/focus)	27.2	-	21.7	20.0
	Framing	18.1	18.1	12.9	10.0
PAR	Stance (probability, evidentiality, attitude)	6.0	15.1	14.9	17.5
	Engagement	-	-	1.0	-
Total		100	100	100	100

Concentrating on the most important functional category, that of text-oriented bundles, we see a clear preference for the structuring type in linguistics, and especially in linguistic BDs. Although the expert writers in medicine also exhibit an important use of this category, their learner counterparts, by contrast, make no use at all of these bundles, clearly preferring bundles with a resultative/inferential function instead, as will be discussed below. Structuring bundles, having an identifying and focusing meaning, allow writers to draw the reader’s attention to a particular idea in the text, and to intensify the force of their arguments. Linguistics experts have used structuring bundles in their conclusions more often, a practice which contrasts with their learner counterparts. These functional categories of bundles tend to be expressed by NP-based (common examples include *the aim of*, *the importance of* and *the current study*), as in (13) and (14), or VP-based structures (*aim of this paper is*, *this paper will focus on*, *there is a* and *that they are*), as in (15). The word *aim*, as noun or verb, is a recurrent one in bundles with this function.

(13) The aim of the present paper is to study the preference for the use of one-word verbs to multi-word verbs (LIN_BD09_I)

(14) This observation is consistent with the importance of cell-cell and cell-matrix contact in the activation of fibroblasts. (MED_RA25_C)

(15) This qualitative study has offered a general overview of those discourse functions which academic speech and writing have in common and those for which there is a marked difference in distribution. (LIN_RA05_C)

As just mentioned, resultative bundles are fairly common (21.2%) in medical BDs, by comparison with the other three subcorpora (with less than half this frequency), and by contrast, no instance of the structuring function was found. Interestingly, these writers have placed almost all their resultative bundles in the conclusion sections, as illustrated in (16) and (17). Other common bundles with this function are *the conclusion that*, *as a result of,* and *due to the fact that.*

(16) (…) call for the involvement of mental health professionals in the Emergency Room in order to offer a more complete evaluation of patients once medically stabilized. (MED_BD08_C)

(17) The results of this study demonstrate a need to distinguish at least two separate age-groups (…) (MED_BD10_C)

A final point worth mentioning in relation to the text-oriented category is that framing is more frequent in the learner corpus than in the expert data, exemplified by bundles such as *according to*, *related to the* and *as well as*. The greater need for these learners to situate and establish links between non-linear arguments with respect to others may have a genre-specific explanation; academic writing instruction may emphasize this writing strategy over others.

In research-oriented bundles, the second most important functional category, an interesting tendency arises: whereas the medicine data overall favour bundles contributing to the description of research objects, especially in RAs, linguistics favours the procedural bundles. This is not entirely surprising, considering the nature and object of study of each of these academic texts. And thus, whereas in medicine the description of the 'real-world' problem (medical conditions, clinical studies, etc.) is of great importance to their studies, in linguistics texts it is important to show the procedures of the research methods and demonstrate a certain ability in explaining how the research has been conducted. Both functions, i.e. method and procedure, are overwhelmingly often expressed by a NP-based bundle and very frequently by the noun phrase with *of*-phrase. Common bundles of description from the medicine texts are *the prevalence of, the presence of, the risk of,* and from the VP-based pattern, *it is a/the*. To express procedure, the most commonly used bundle is, by far, *the use of.* Other common bundles expressing procedure are *(the) analysis of (the)*, *the role of*, *the ways that,* and from the VP-based group of bundles, *can be used to*. Description and procedure bundles are exemplified in (18) and (19) respectively:

(18) Musculoskeletal disorders represent a relevant part of global morbidity and have an important impact on the prevalence of chronic diseases. (MED_RA03_I)

(19) This paper has tried to provide an accurate analysis of the English language in terms of lexical and grammatical parameters (…) (LIN_BD07_C)

A final insight from the group of research-oriented bundles is the high proportion of bundles with the meaning of quantification in medicine BDs, with respect to the other three corpora, and which again are mainly from the NP-based group of bundles. Examples include *the rest of, of the most* and *the most prevalent*.

The final category, participant-oriented bundles, mostly covers stance markers expressing opinion rather than facts, and may indicate degree of probability and epistemic meaning, on the one hand, or be part of the so-called 'other stance markers' (see Cortes, 2004, p. 209), on the other, which include LBs with evidential meaning, indicating the source of the information (e.g. *recent studies have*, *have been proposed*). The former type, the most common one, tends to be expressed by a recurrent set of structural categories, namely anticipatory *it*-constructions containing an evaluative element (*it is true that, it would be interesting to, it can be argued that*), bundles with modal or semi-modal verbs (*should not be, seems to be*), epistemic adverbs, notably *likely* (*are likely to, is likely to be*), and other bundles expressing stance (*still in its infancy*, *has the potential to*). It is worth mentioning that stance can also be expressed in other ways than 3-, 4- and 5-word bundles, and that our study refers only to stance expressed in these sequences. Interestingly, stance is more common in the conclusion sections of the BD genre, whereas RAs contain more bundles of this type in their introduction sections: persuading readers from the very beginning through evidential and epistemic bundles seems to characterise more confident writing. Finally, engagement is almost non-existent in our corpus with only one bundle, namely *our understanding of,* used in the conclusion section of RAs in linguistics.

8.5 Conclusion

This paper has analysed the use of LBs in the introduction and conclusion sections of learner and expert academic writing in linguistics and medicine. The quantitative and qualitative analysis performed in order to explore the frequency, structures and functions of LBs has yielded interesting results: LBs are very useful devices for the construction of discourse, but they behave in dissimilar ways in different disciplines and genres.

Regarding frequency, of the 218 bundles retrieved, 3-word bundles were more frequent in all subcorpora; of these, *the use of, in order to,* and *as well as* stand out as the most popular LBs. BDs in linguistics and medicine have produced a similar quantity of LBs in both sections, whereas RAs vastly differ in their frequency of use of LBs, which points towards a disciplinary difference. When comparing the learner and the expert corpus, on average, BDs in medicine contained more LBs than RAs in the same discipline, and the opposite tendency was found for linguistic BDs, which contained fewer LBs than their expert counterparts. In addition, a list of 35 convergent bundles was found, which can be a pedagogically useful resource for general academic writing. This quantitative analysis was complemented by qualitative analyses of structure and function which, after manual classification and revision of concordance lines, provided a more comprehensive picture of LB usage.

In terms of structure, both learner and expert writers favoured NP-based bundles; the structure noun-phrase with *of*-phrase was by far the most frequent one in all corpora. BDs and RAs also agreed on the second most common LB structure: other prepositional phrase, which allowed writers to include frequent discourse signalling devices in their texts. The main difference, however, lies in the greater structural variation of the LBs used by experts in linguistics; LBs in medical RAs, and in the learner genre, were definitely less varied. Finally, with regards to function, LBs performing text-orienting functions were the most prevalent in all subcorpora. The second group, LBs with research-oriented functions, was more popular among medicine expert writers, who seem to emphasize the empirical over the interpretative. The last function, participant-oriented, was the least represented one; this low frequency is especially marked in BDs in linguistics, which points towards a case of underuse.

Additionally, while learners placed stance markers mostly in the last section of their texts, expert writers showed a preference for the use of stance in their introduction sections. Placement of LBs in particular sections of a text is yet another important feature that depicts writers' academic literacy. On the other hand, the lack of structuring bundles in medical BDs, and their recurrent use of resultative bundles also calls for explicit pedagogical attention. Disciplinary differences were also found regarding the prevalence of descriptive bundles in medicine, and of procedural bundles in linguistics; disciplinary conventions and the object of study of each of these texts could account for the discrepancies found.

The present study has some limitations worthy of mention. The first one is a methodological limitation: in order to extract sequences of words automatically, our retrieval method only included LBs that were fixed in nature; that is, our lists do not include variable bundles or bundles with open slots (e.g. *in section (...)*, *up to (...) %*, *to a (...) extent*). This method therefore does not capture LBs in their entirety. Including this type of permutations (e.g. using the ConcGram function in Wordsmith tools) could have helped to show a more comprehensive picture of LBs in academic writing (see O'Donnell et al., 2012). Another methodological limitation has to do with the fact that the learner corpus had not been error-tagged, which could have somehow affected the number of LBs extracted (i.e. if there were typos in particular words that were part of LBs, the software did not retrieve them). All texts included in the learner corpus, however, were successful BDs evaluated by their supervisors and the evaluating committee, so the probability of containing numerous typos is unlikely. Using a larger learner corpus would also have made the findings more representative. This study could also be improved by applying inferential statistics such as the *t-test* in order to investigate whether the differences found were statistically significant. In addition, our analysis has looked at the use of LBs in the introduction and conclusion section of academic texts, as these sections tend to be the most conventional ones in these particular genres. Analysing LB positions, not only with regards to sections but also with regards to paragraphs or sentences, would be interesting (see Römer, 2010). Finally, when comparing our findings across previous studies that utilized corpora of different lengths and breadths, it was difficult to accurately match the results. This limitation has also been attested to by Chen and Baker (2010, p. 43), who claim that "it is virtually impossible to find different corpora, of exactly the same size composed of the same number of texts, for direct comparison"; therefore, the cross-study comparisons included in this paper have to be regarded with caution.

Our analysis has provided a comprehensive list of 218 different bundles that may assist L2 learners to accommodate their academic writing to their specific discipline and genre. The results underline the importance these expressions have in order to write successful academic texts and to achieve disciplinary competence. As it has been shown, even though LBs are very frequent in language, mere exposure is often not enough for the acquisition and mastery of these devices in academic writing. Our findings therefore emphasize the need for more explicit teaching of LBs, always through corpus-informed materials, in agreement with the discipline and the genre studied.

9. General Discussion

In this chapter, a summary of the results obtained from the four studies is presented and the main research questions formulated are answered in Section 9.1. This is followed by a discussion in Section 9.2, in which the main implications drawn from each study and the contributions made to the fields of second language writing and corpus linguistics are presented.

9.1 Summary of the findings

In study one, the main research question was: What effect does the CBI course have on students' academic vocabulary production? The main findings showed that the materials students were exposed to in class provided variety and repetition of general and discipline-specific academic terminology. In terms of improvement, i.e. an increased number of academic words, collocations, and formulas in the texts written after the course, results showed that more discipline-specific words were produced at T2 on average by all groups of learners, which could indicate an important short-term benefit of the CBI approach. In general, more discipline-specific terminology (words, collocations and formulas) was used after the course, compared to general academic terminology, except for words from the Academic Vocabulary List, which ranked third on the list of most used items, underlining its pedagogical usefulness. In the literature, the fact that technical vocabulary does not pose as many difficulties for learners as general academic vocabulary has been reported (Durrant, 2016; Granger, 2017), which seems to be in agreement with the findings from this study.

Apart from showing a higher frequency of discipline-specific vocabulary, the vocabulary that could be categorized as 'interdisciplinary' also increased at T2, regardless of the students' setting of instruction. This highlights another important benefit of CBI programmes, namely that being exposed to this particular approach can enhance students' terminology, both from a general and a discipline-specific perspective.

As for general academic formulas, the results showed that even though students were exposed to academic formulas quite frequently throughout the materials (these covered more than 40% of the Academic Formulas List), this exposure was not enough to trigger students' production of academic formulas (a maximum of 1% of the list was covered in the learner corpora). With regards to the use of general academic collocations, they were the least used category in the learners' texts. This could be due to the fact that the Academic Collocations List was barely covered by the class material itself. More explicit pedagogical attention to recurrent word combinations such as formulas and collocations in CBI programmes seems therefore needed.

In terms of setting of instruction, EMI texts contained a higher percentage of academic language overall at T2, whereas the opposite tendency was found in the L1 texts. Moreover, the increase in certain categories (such as general and discipline-specific words) is much more noticeable in the EMI group than in the L1 group; these differences were statistically significant. EMI students seem to have benefitted the most from CBI instruction; this may be due to the fact that EMI students had more frequent encounters with general and specialized lexicon in other subjects of the degree during that semester. Even though CBI instruction seems to have had quite positive effects on students' production of academic language, these findings corroborate the need for more corpus-informed materials in order to help CBI instructors to select and prioritize certain items of the academic discourse so that all kinds of terminology (not only single words, but also collocations and formulas) can be taught, learned and practiced properly.

In study two, one of the research questions was: How do English NNS undergraduate writers use adversative linking adverbials in terms of frequency, placement and types compared to English NS undergraduate writers? The findings showed that NNS had used slightly more adversative LAs than their NS counterparts, and also that there was a clear difference in LA placement: initial position seems to have been preferred by NNS writers, in contrast with the predominant medial position found in NS texts. Also in terms of frequency, NNS and NS writers coincided with the most popular LA, *however*, mostly placed in initial position. With regards to LA categories, both NNS and NS writers showed a similar preference for the 'proper adversative' category (the fact that *however* belongs to this category and had very high frequency counts could explain this predominance) over the other three categories (i.e. 'contrastive', 'correction' and 'dismissal'). When the items were looked at separately, however, it was found that certain items had a noticeable higher or lower frequency in the NNS corpus when compared to the NS texts. For example, *nevertheless* was used almost 21 times more by NNS writers, compared to the NS use. Regarding underuse, *actually* was the most underused item by NNS writers compared to the NS corpus. A further comparison with two other reference corpora, namely BNC and COCA, showed that this item was in fact not typical of academic writing. Hence, the low frequency of *actually* in the NNS texts was not regarded as a case of underuse that required pedagogical attention. A qualitative analysis of all the items that were used with different frequencies revealed as well stylistic and syntactic misuses that were explored further in the NNS corpus. In sum, even though the NS corpus represented novice writing as well, and thus the cases of over- or underuse had to be taken with caution (i.e. not understood as *deviation* from the norm), these comparisons were still useful to uncover certain types that could be missing from the students' repertoire or that needed scaffolding in the classroom. Explicit instruction of adversative LAs in order to avoid the so called 'teaching effect' (see Granger & Tyson, 1996; Lei, 2012; Leńko-Szymańska, 2008; Rica-Peromingo, 2012), or, in other words, when NNS writers misuse certain structures mainly due to the provision of long list containing items out of context for their instruction, is emphasized by this study.

The second research question established for this study was: How do learners with different L1 backgrounds use adversative LAs when compared to one another? The results showed that there seemed to be a general agreement on the categories used, i.e. 'proper adversative' as the most, and 'dismissal' as the least used categories by all NNS writers, and also on the placement of these adverbials in sentence initial position. These preferences were quite homogeneous regardless of the writers' L1. It was hypothesized that learners with the same linguistic family (i.e. Romance L1s and Germanic L1s) would show similar frequencies and usage patterns of adversative LAs within each group. However, some discrepancies were found in terms of frequency: French and Spanish L1 writers produced the highest number of LAs, while Italian L1 writers contained the lowest production. Dutch L1 and German L1 writers' use of LAs did not coincide either: German L1 learners had used all four types of LAs much more frequently than Dutch L1 learners did. This was somehow contrary to what was expected. The fact that all NNS learners were exposed to the same type of methodology, i.e. PBL –which also required peer-reviewing each piece of writing– could have eased possible L1 writing features transferred to the L2. Although further investigations are needed to confirm this hypothesis, the PBL approach and peer-reviewing practices seem beneficial for gaining homogeneity in writing and reducing possible L1 transfer issues in international classrooms.

In study three, one of the research questions was: How do Spanish L1 learners use reflexive MD markers when writing in academic English, compared to an expert corpus of RAs? The results showed that BDs in linguistics and medicine contained a similar amount of MD in their texts (3.5% and 3.1% respectively), which was somewhat similar to what was found in the expert corpus (MD represents 2.8% in RAs in medicine, and 3.8% in RAs in linguistics). It can be said therefore that in terms of frequency, and compared to experts, learners used MD to an appropriate extent. A more qualitative analysis of the types of reflexive MD, however, revealed interesting learner writing features that are worthy of pedagogical attention. In the case of textual MD, findings showed cases of over- and underuse of certain markers that applied to BDs in both disciplines: references to text (e.g. *in this section*) and additive markers (e.g. *moreover*) were used much more frequently in the BDs than in the RAs. In contrast, BDs did not include as many references to semiotic modes (e.g. *in Table 4*), or exemplifiers (e.g. *in other words*). Regarding the use of interpersonal MD, cases of overuse were found in medical BDs, such as the production of self-mention, directives, and inclusive we, which were noticeably more frequent than what was found in the expert corpus. In contrast, the production of interpersonal markers by BD writers in linguistics was lower compared to their expert counterparts. Different communicative purposes of genres (e.g. BDs displaying knowledge to a supervisor, while RAs do so to peers of more or less the same expertise), L1 transfer (a more reader-responsible writing style of Spanish academic texts), and even the conflicting advice on the use of authors' involvement features such as self-mention in academic writing textbooks or provided by different supervisors could explain the results.

The second research question formulated was: Are there any differences in the use of MD across disciplines? The analyses performed did detect some differences across disciplines: linguistic BDs and RAs contained more MD markers in general compared to medical BDs and RAs, which was in line with previous corpus studies that explore MD practices across different disciplines (Hyland, 2001; Hyland, 2010; Mur-Dueñas, 2011; Salas, 2015). The nature of the contents itself (i.e. language being the subject matter of linguistics BDs and RAs) could explain the differences found.

In study four, one of the research questions investigated was: How do Spanish L1 learners use lexical bundles in the introduction and conclusion sections when writing in academic English, compared to an expert corpus of RAs? The analyses revealed that BDs featured a similar number of LBs in their introduction and conclusion sections, regardless of their discipline, which could point at a quantitative similarity of this genre. RAs, on the contrary, showed a vastly different production of LBs in terms of frequency: RAs in linguistics contained 3 times more LBs than medical RAs overall, and had almost the same number of LBs in both the introduction and conclusion sections. Medical RAs contained a generally low frequency of LBs, and they were more predominantly used in the introduction section. This finding points at a disciplinary difference between the academic communities studied. From the 218 different LBs retrieved, 3-word bundles represented the most frequent bundle length (in particular *the use of, in order to,* and *as well as*), and 35 bundles were shared by all subcorpora (i.e. convergent bundles). In this regard, there were more bundles shared by discipline than by genre, which emphasizes once again the important role academic communities play in academic writing.

This study also answered the question: How are these lexical bundles used in terms of structure and function? The results showed that NP-based bundles, and in particular the structure noun-phrase with *of*-phrase fragment, was the most predominant structure of all LBs used. One of the main differences lay in the greater variation of structures found in RAs in linguistics. In terms of function, LBs with a text-orienting function were the most popular in all texts. LBs with research-orienting functions were particularly more popular among medicine RA writers, who seemed to highlight empirical procedures over interpretative arguments. Finally, LBs with participant-orienting functions were the least represented ones, especially in BDs in linguistics, which points to a possible case of underuse. The placement of bundles in the introduction and conclusion sections according to function was also explored and resulted in interesting differences, such as the placement of stance markers in the conclusion sections of BDs, whereas RAs, involving readers since the very beginning in their texts, tended to place stance markers in the introduction section, which could be seen as a persuasive technique typical of this genre.

Finally, one of the main objectives of this doctoral thesis was to see how corpus linguistics methods could contribute to identifying L2 learner writing features. The study has brought to light a number of findings that show how corpus-based and corpus-driven approaches applied to learner writing helped to uncover learner features in academic writing. From the application of learner and expert corpora compilation methods, to the selection and careful investigation of linguistic devices that were found problematic in previous literature on learner writing, the present doctoral thesis offers an exploration of different writing assignments learners faced at university. The use of text-analysis software to calculate frequencies and to see items in context, and also the manual annotation of some texts have allowed the author to find noticeable differences in the learner corpora that call for more pedagogical attention. The next section presents the main implications drawn from the different studies and also the contributions this thesis has made to the fields of corpus linguistics and second language writing.

9.2 Implications and contributions of the study

The findings drawn from the quantitative and qualitative analyses performed in the different studies carry three important pedagogical implications. First, as was previously mentioned, language, and written academic language in particular, is highly patterned. Important discourse elements that grant coherence and cohesion to academic texts, such as linking adverbials (e.g. *on the other hand*), metadiscursive elements (e.g. *as we can see*), or various lexical bundles (e.g. *the use of*), as well as general and specific academic terminology (e.g. *randomized controlled trials*), are situated on the phraseological dimension of language. Therefore, the present thesis encourages a phraseological approach to academic language teaching and learning, especially when it comes to the accommodation of learners' writing to certain academic disciplines. For example, instead of creating glossary lists of isolated words (often based on intuition) for students to learn and memorize, a compilation of a specialized corpus of e.g. different readings, the digitalized textbook, or academic articles, is recommended as a class practice. As we did in study four, students could look for the most frequent 3-, 4-, 5-word bundles in these texts, or create keyword lists; these would reflect the most salient, and thus pertinent, keywords or bundles used in that specific genre and discipline, and most importantly, students would be able to see the searched items in context through concordance lines and explore e.g. placement in the sentence or the rhetorical functions performed by these items.

Second, the findings have suggested that there are some practices that can transfer from the students' L1 when writing in an L2. For example, studies three and four focused on Spanish L1 learner writing. Certain practices such as a more frequent use of additive markers to construct persuasive arguments, or an underuse of markers that refer to semiotic items in the text, or that frequently engage readers in the conversation, have been found to be more characteristic of Spanish rather than English academic writing practices. The transference of these conventions into texts can strike the expert reader as being written by an 'outsider' to both academic literacy in general and to the discipline in particular. The exploration of a learner corpus containing texts written in English by Spanish L1 writers has been deemed useful to identify practices that seem to belong to a specific L1 community, and that can be thus addressed in the classroom.

Finally, the results also support the notion of academic writing as a highly genre and discipline-specific discourse. Novice and L2 learner writers struggle to conform to academic conventions and they generally do not pick up specific academic writing conventions from mere exposure. For example, the formulas explored in study one were barely produced by students after the course, even though they were highly frequent in the materials provided. This indicates that exposure alone did not work for the acquisition and mastery of certain linguistic devices and corroborates the need for explicit instruction. Similarly, in study two, a more qualitative look at learners' production of adversative LAs revealed that there were cases of overuse and misuse that call for pedagogical attention. Studies three and four also showed generic and disciplinary differences on the use of MD and LBs, some of which could be due to a possible 'teaching effect' or conflicting advice from textbooks and/or academic writing materials. This thesis supports the use of corpus-informed materials, and learner-corpus informed resources in particular, in order to help instructors select and prioritize certain linguistic practices that should be taught and learnt in agreement with each genre and discipline studied.

A final comment here is related to the teaching of academic writing: considering specific L1 conventions and common problems when writing in an L2 can be advantageous in that these can be pinpointed and addressed more directly and concretely in the classroom (e.g. lack of stance markers and participant-oriented bundles in BDs written in English by Spanish L1 students explored in studies three and four). However, many of today's university classrooms represent a different reality: they tend to be heterogeneous classrooms of an international nature that include students with different cultural, sociological, and linguistic backgrounds. Therefore, teaching academic writing from the perspective of one L1 in particular would not be practical nor feasible. In order to include all types of languages, and sometimes even different proficiencies in the classroom, the teaching of academic writing should be made disciplinary relevant and not L1 relevant. One way to get round this difficulty is the use of corpus-based pedagogical tools in the classroom. Compiling a corpus of the literature in the students' own discipline, or searching for patterns in the academic register of larger, well-designed corpora, e.g. using COCA's or the BNC's website, can enable both instructors and learners to study how any linguistic phenomenon is used in context.

Also, compiling a learner corpus of students' own texts can help them to identify possible cases of over- and underuse and to find differences and similarities with their peers. These corpus-based methods may require previous training on corpus-research skills, but they hold tremendous pedagogical potential in that they can raise students' language awareness, while triggering discovery learning, and helping them to "clarify, give priorities, reduce exceptions and liberate the creative spirit" (Sinclair, 1997, p. 38).

A recurrent concern that has been highlighted in the literature is the little pedagogical impact corpus-based research often has (Römer, 2011); according to Flowerdew, many of the corpus findings are not "applied directly to pedagogy and tend to remain at the level of implications" (2001, p. 366). The four studies carried out for this doctoral thesis have attempted to contribute to second language writing and corpus linguistics research by producing several pedagogically useful resources that can be interesting for L2 university writers who need to accommodate their writing to a new discipline or genre. It has also employed different corpus analyses that can inspire further research on other linguistic phenomena. For example, study one showed how CBI programmes, and the adjunct model in particular, can be beneficial for improving learners' writing skills, especially when it comes to learning general and discipline-specific vocabulary in the short term. In addition, study two, apart from highlighting different frequencies of adversative LAs in NNS and NS writing, provided explicit guidance on specific LAs (e.g. *nevertheless, in contrast, however*) regarding their contrastive power and their placement within the sentence, in order to prevent misuse. Following a corpus-driven approach that involved manual annotation of 70 complete texts, study three offered a list of 230 textual and interpersonal MD markers that were classified into 21 subcategories. This list may be of interest to L2 learner writers and to academic writing teachers or material developers who deal with the use of MD in specific disciplines and genres. Finally, and also through a corpus-driven approach that involved extraction and manual classification of LBs, study four provided a comprehensive list of 218 different bundles with their respective structural and functional description. This list could aid L2 writers to accommodate to their specific discipline and genre, and reinforces the importance word combinations have in order to write successful, smoothly flowing academic texts, and to acquire disciplinary literacy. Although much work still remains to be done, these four studies are an attempt to bridge the gap between corpus research findings and pedagogical practice, and support the use of corpus tools in the classroom.

10. Conclusion

The present doctoral thesis has sought to investigate how corpus approaches could serve to identify learner-writing features. With this intention, four studies were carried out. First, the use of general and discipline-specific terminology in a writing task produced by L2 first-year students before and after a CBI course was analysed. It then looked at the use of adversative LAs in longer argumentative texts also written by L2 first-year students, and compared it to the use of these devices by English-native students. Metadiscursive practices in even longer texts, i.e. bachelor dissertations, produced by L2 last-year students in medicine and linguistics, were then explored and compared to an expert corpus of published research articles in the same disciplines. Finally, all lexical bundles present in the introduction and conclusion sections of these texts (BDs and RAs) were extracted, counted and classified structurally and functionally.

The different corpus approaches employed have served to identify important learner writing features from which some pedagogical implications were drawn: first, it was found that learners produced more academic vocabulary (both general and discipline-specific) after a CBI course, which shows a possible short-term benefit of this type of instruction at university. However, CBI seemed to have little or no effect on the use of general academic formulas and collocations, which remained the same after the course. More pedagogical attention to formulaic sequences in academic writing was therefore emphasized. An analysis of the texts according to students' setting of instruction (i.e. EMI or L1) also showed that the EMI group's use of academic terminology after the course was significantly higher, compared to their L1 counterparts, who did not show such an improvement. A greater exposure to the target language in an academic context experienced by the first group could account for the differences found.

Second, it was found that the use of adversative LAs in argumentative essays of NNS and NS writers were comparable in terms of frequency. A more qualitative look at the types and categories of LAs showed however that writers did not always agree with the LAs choice, especially regarding the placement of these items in sentences and paragraphs. Different cases of overuse (e.g. *nevertheless*) underuse (e.g. *actually*) and misuse (e.g. *in contrast*) were found and explored in detail. Due to the fact that the NS corpus belonged to university students, and the widespread concern that these texts do not always model good writing practices, further comparisons with expert corpora (such as the academic subcorpora included in BNC and COCA) were deemed necessary in order to compare frequencies and uses of different items and to determine if the cases of overuse and underuse needed pedagogical attention. The exploration of LA usage according to two different linguistic families also provided interesting results: there were differences between languages of the same family, and all subcorpora shared general preferences for certain categories and placement of LAs. The fact that these students, despite having different L1s, received the same type of instruction and peer-reviewed the texts in class could have eased any possible differences in terms of adversative LAs production. Peer-review practices and the PBL approach seem to have had a positive effect in reducing possible L1 transfer issues.

Third, the analysis of textual and interpersonal markers in learner BDs and expert RAs revealed that the differences were not only due to a different writer status (learner vs. expert) and/or genre (such as the avoidance of interpersonal markers in linguistic BDs, or the predominance of additive markers in all BDs in general); they were also indicative of different disciplinary conventions (e.g. the overall predominance of textual MD in the linguistic subcorpora, compared to medical texts).

And fourth, the extraction of bundles of different lengths from the introduction and conclusion sections of these texts (BDs and RAs), and the classification according to their main structures and functions, similarly revealed preferences that could denote writers' immaturity (e.g. a much less varied number of types, structures and functions in the bundles used by BD writers), and also highlight practices of different academic communities (e.g. high frequency of research-oriented bundles in the medical texts vs. the predominance of textual-oriented bundles in the linguistic subcorpora). The findings presented in this doctoral thesis reinforce the usefulness of corpus methods, which applied to L2 texts and together with the use of comparison corpora, have been useful to find particular L2 writing features that are worthy of pedagogical attention.

Interesting avenues for further research include the use of parallel corpora, i.e. student's production in their own L1, in order to see if the different usage patterns are due to L1 transfer or otherwise; this would allow researchers to study interlanguage more accurately. In addition, longitudinal studies of learner writing would also be optimal in order to observe learners' academic writing development and to identify exactly when their writing practices approximate those of the (native or expert) target production. Also, the learner corpora explored in the present thesis belonged mainly to students with European L1s, and was contrasted to English produced by American students and writers publishing in English-medium journals (mainly British); including other varieties of understudied L1 (e.g. Indian, Singaporean) and L2 English (e.g. Czech, Turkish) in future studies would also be useful in order to provide a broader picture of how different L1 backgrounds and cultures can have an effect on academic writing, especially in international classrooms. Looking at *intra*disciplinary variation (academic texts in the same discipline but on different topics, such as *orthodontics* vs. *endodontics*) would also be interesting in order to analyse frequencies and usage patterns of different linguistic devices. A final suggestion for future corpus studies is the creation of a database in which to share the different corpora compiled, so as to make it accessible for other teachers, learners, and researchers alike. Having a unified set of guidelines for the compilation, annotation and tagging of corpora would make cross-study comparisons much easier and more accurate.

It is hoped that the findings on academic discourse and corpus approaches to learner writing that have been presented in this thesis can be useful to future academic writing learners, instructors and material developers, and that this thesis has made a contribution to the fields of second language writing and corpus linguistics. Finally, this thesis reinforces the need and benefit of using corpus-informed materials in the classroom and supports academic writing instruction from genre and discipline perspectives.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL)–A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, 12 (4), 235-247.

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. John Benjamins, Amsterdam/Philadelphia.

Ädel, A. (2008). Involvement features in writing: do time and interaction trump register awareness? In Gilquin, G., Papp, S., and Díez-Bedmar, B. (Eds.) *Linking up contrastive and learner corpus research*, 35-53. Brill Rodopi.

Ädel, A, & Mauranen, A. (2010). Metadiscourse: diverse and divided perspectives. *Nordic Journal of English Studies,* 9 (2), 1-11.

Ädel, A., & Erman, A. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purpose,s 31* (2), 81-92.

Ädel, A. (2016). Just to give you kind of a map of where we are going: A taxonomy of Metadiscourse in spoken and written Academic English. *Nordic Journal of English Studies,* 9 (2), 69-97.

Ai, H., & Lu, X. (2010). A web-based system for automatic measurement of lexical complexity. Paper presented at "the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10)". Amherst, MA. June 8-12.

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Díaz-Negrillo, A., Ballier, N. and Thompson, P. (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, 249-264. Amsterdam/Philadelphia: John Benjamins.

Airey, J. (2011). The disciplinary literacy discussion matrix: A heuristic tool for initiating collaboration in higher education. *Across the disciplines*, 8(3). Retrieved from https://wac.colostate.edu/docs/atd/clil/airey.cfm [Octorber, 2018]

Airey, J. (2016). EAP, EMI or CLIL?. In Hyland, K., & Shaw, P. (Eds.) *The Routledge handbook of English for academic purposes*, 95-107. London: Routledge.

Airey, J., Lauridsen, K. M., Räsänen, A., Salö, L., & Schwach, V. (2017). The expansion of English-medium instruction in the Nordic countries: Can top-down university language policies encourage bottom-up disciplinary literacy goals? *Higher education*, 73 (4), 561-576.

Airey, J. (2018 November). Disciplinary Literacy. Presentation at the round table "How are languages best learned? Study abroad, immersion (CLIL/EMI) and formal instruction". Pompeu Fabra University. Barcelona.

Ament, J. R., & Pérez-Vidal, C. (2015). Linguistic outcomes of English medium instruction programmes in higher education: A study on economics undergraduates at a Catalan University. *Higher Learning Research Communications*, 5 (1), 47-68.

Anson, I. G., & Anson, C. M. (2017). Assessing peer and instructor response to writing: A corpus analysis from an expert survey. *Assessing Writing*, 33, 12-24.

Anthony, L. (2014). AntWordProfiler. Tokyo, Japan: Waseda University. URL *http://www.laurenceanthony.net/* [September 8, 2018].

Anthony, L. (2015). TagAnt (Version 1.2.0) [Macintosh OS X]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Anthony, L. (2018). AntConc (Version 3.5.7) [Macintosh OS X]. Tokyo, Japan: Waseda University. *http://www.laurenceanthony.net/* [September 8, 2018].

Appel, R., & Wood, D. (2016). Recurrent word combinations in EAP test-taker writing: Differences between high-and low-proficiency levels. *Language Assessment Quarterly, 13* (1), 55-71.

Barlow, M. (2004). Collocate 1.0. Collocation extraction software. Houston TX: Athelstan. URL: *http://www.michaelbarlow.com* [September 8, 2018].

Barlow, M., & Burdine, S. (2006). American phrasal verbs (CorpusLAB Series). Houston, TX: Athelstan.

Bell, D. M. (2010). Nevertheless, still and yet: Concessive cancellative discourse markers. *Journal of Pragmatics*, 42 (7), 1912-1927.

Bennett, G. R. (2010). Using corpora in the language learning classroom: corpus linguistics for teachers. *University Of Michigan Press*.

Biber, D. (2004). Lexical bundles in academic speech and writing. In B. Lewandowska-Tomaszczyk (ed.) *Practical Applications in Language and Computers. PALC* 2003, 165-178. Frankfurt am Main: Peter Lang.

Biber, D., & Finegan, E. (1994). Intra-textual variation within medical research articles. In Oostdiijk, N. & de Haan, P. (Eds.), *Corpus-based Research into Language* (pp. 201-221). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. *Language and Computers 26*, 181-190.

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

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Halow, England: Pearson/ Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics,* 25 (3), 371–405.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes 26* (3), 263–286.

Bondi, M. (2010). Metadiscursive Practices in Introductions: Phraseology and Semantic Sequences across Genres. *Nordic Journal of English Studies,* 9 (2). 99-123.

Brinton, D. (1993). Content-based Instruction and ESP: Same or Different? *TESOL Matters*, 3 (4), 9.

Brinton, D., Snow, M. A., & Wesche, M. B. (2003). *Content-based second language instruction*. University of Michigan Press.

Burneikaite, N. (2008). Metadiscourse in linguistics master's theses in English L1 and L2. *Kalbotyra*, 59 (3), 38-47.

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL, 5* (5), 31-64.

Callies, M. (2008). Easy to understand but difficult to use? Raising constructions and information packaging in the advanced learner variety. In Gilquin, G., Papp, S., and Díez-Bedmar, B. (Eds.), *Linking Up, Contrastive and Learner Corpus Reasearch*, 201-226.

Callies, M., Díez-Bedmar, M. B., & Zaytseva, E. (2014). Using learner corpora for testing and assessing L2 proficiency. In Leclercq, P., Edmonds, A., and Hilton, H. (Eds.), *Measuring L2 proficiency: Perspectives from SLA*, 71-90.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The Grammar Book: An ESL/EFL Teacher's Course*. 2nd ed. Boston, MA: Heinle & Heinle.

Chen, W. Y. C. (2006). The use of conjunctive adverbials in the academic papers of advanced Taiwanese EFL learners. *International Journal of Corpus Linguistics*, 11(1), 113-130.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14* (2), 30–49.

Cheng, X., & Steffensen, M. (1996). A technique for improving student writing. *Research in the teaching of English,* 30 (2), 149-181.

Cobb,T. (2002). Web Vocabprofile. An adaptation of Heatley, Nation & Coxhead's (2002) Range. UQAB Canada. URL *http://www.lextutor.ca/vp/* [September 8, 2018].

Conrad, S. (2000). Will Corpus Linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34, 548–560.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: examples from history and biology. *English for Specific Purposes, 23*, 397-423.

Coxhead, A. (2000). A new academic word list. *TESOL quarterly*, 34 (2), 213-238.

Coxhead, A., & Nation, P. (2001). The specialised vocabulary of English for academic purposes. *Research perspectives on English for academic purposes*, 252-267.

Coxhead, A. (2017). *Vocabulary and English for Specific Purposes research: Quantitative and qualitative perspectives*. Routledge.

Crismore, A. (1989). *Talking with Readers: Metadiscourse as Rhetorical Act*. Peter Lang, New York.

Crismore, A., & Farnsworth, R. (1990). Metadiscourse in popular and professional science discourse. In Nash, W. (Ed.) *The Writing Scholar: Studies in Academic Discourse*, 118-136. Newbury Park, CA: SAGE.

Crismore, A., Markkanen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students. *Written Communication*, 10, 39-71.

Crismore, A., & Vande Kopple, W. (1997). Hedges and readers: effects on attitudes and learning. In Markkanen, S. et al. (eds) *Hedging and Discourse: Approaches to the Analysis of a Pragmatic Phenomenon in Academic Texts*, 83-114. Berlin: Walter de Gruyter.

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, 33, 100-118.

Clippinger, J. H., & McDonald, D. D. (1983, August). Why good writing is easier to understand. In *Proceedings of the Eighth international joint conference on Artificial intelligence*-Volume 2 (pp. 730-732). Morgan Kaufmann Publishers Inc..

Dafouz, E. (2003). Metadiscourse revisited: a contrastive study of persuasive writing in professional discourse. *Estudios Ingleses de la Universidad Complutense,* 11, 29-52.

Dafouz, E., Camacho, M., & Urquia, E. (2014). 'Surely they can't do as well': A comparison of business students' academic performance in English-medium and

Spanish-as-first-language-medium programmes. *Language and Education*, 18 (3), 223–236.

Dahl, T. (2004). Textual metadiscourse in research articles: a marker of national culture or of academic discipline? *Journal of Pragmatics,* 36 (10), 1807-1825.

Dalton-Puffer, C. (2011). Content and language integrated learning: From practice to principles? *Annual Review of Applied Linguistics*, 31, 182-204.

Davies, M. (2012). Wordandphrase. Brigham Young University. URL: https://www.wordandphrase.info [September 8, 2018]

Demol, A., & Hadermann, P. (2008). An exploratory study of discourse organisation in French L1, Dutch L1, French L2 and Dutch L2 written narratives. *Language and Computers,* 66 (1), 255-282.

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes, 30* (1), 58-72.

Durrant, P. (2016). To what extent is the Academic Vocabulary List relevant to university student writing? *English for Specific Purposes*, 43, 49-61.

Field, Y., & Yip L. M. O. (1992). A comparison of internal conjunctive cohesion in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal,* 23 (1), 15-28.

Flowerdew, L. (2002). Corpus-based analyses in EAP. In J. Flowerdew (Ed.), *Academic Discourse,* 95-114. Harlow: Longman.

Fraser, B. (2009). An account of discourse markers. *International Review of Pragmatics,* 1, 293-320.

Friedl, G., & Auer, M. (2007). Erläuterungenzur Novellierung der Reifeprufungsverordnung fur AHS, lebende Fremdsprachen (Rating scale used for assessment of the writing task). Wien/St. Pölten: BIFIE.

García-Negroni, M. (2008). Subjetividad y discurso científico-académico, Acerca de algunas manifestaciones de la subjetividad en el artículo de investigación en español. *Revista signos,* 41 (66), 5-31.

Gardner, D., & Davies, M. (2013). A new academic vocabulary list. *Applied Linguistics*, 35(3), 305-327.

Gardner, S., & Nesi, H. (2013). A classification of genre families in university student writing. *Applied Linguistics*, 34 (1), 25-52

Gee, J. P. (1989). Literacy, discourse, and linguistics: Introduction. *Journal of education*, 171 (1), 5-17.

Gee, J. P. (1991). What is literacy? In Mitchel C., and Weiler, K. (Eds.), *Rewriting literacy: Culture and the discourse of the other,* 3-11. New York: Bergin & Garvey.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6 (4), 319-335.

Granger, S. (1996) From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg and M. Johansson (Eds.), *Languages in Contrast. Text-based cross-linguistic studies*, pp. 37–51. Lund: Lund University Press.

Granger, S. (1997). Identifying the syntactic and discourse features of participle clauses in academic English: native and non-native writers compared. In Aarts, J., de Mönnink, I. and Wekker, H. (Eds), *Studies in English Language and Teaching*, 185-198. Rodopi: Amsterdam & Atlanta.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes,* 15, 19-29.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. *Phraseology: Theory, analysis, and applications*, 145, 160.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung and S. Petch-Tyson (eds) *Computer Learner Corpora. Second Language Acquisition and Foreign Language Teaching*, 38–51. Lund: Lund University.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor and T. Upton (Eds) *Applied Corpus Linguistics: A Multidimensional Perspective*, 123-145. Rodopi: Amsterdam

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective,* 27–49. Amsterdam: John Benjamins Publishing.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer, (ed.), *Corpora and Language Teaching,* 13-32. Benjamins: Amsterdam & Philadelphia.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research,* 1(1), 7-24.

Granger, S. (2017a). Academic Phraseology: A key ingredient in successful L2 academic literacy. *Oslo Studies in Language,* 9 (3), 9-27.

Granger, S. (2017b). Learner Corpora in Foreign Language Education. In S. Thorne & S. May (Eds.), *Language, Education and Technology,* 1–14. Cham: Springer International Publishing.

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics, 25* (1), 38-61.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes*, 35, 105-115.

Gries, S. (2013). Statistical tests for the analysis of learner corpus data. In Díaz-Negrillo, A., Ballier, N., & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*, 287-309. Amsterdam & Philadelphia: John Benjamins.

Gries, S. & Newman, J. (2013). Creating and using Corpora. In R. Podesva and D. Sharma (Eds.), *Research Methods in Linguistics*, 257-288. New York: Cambridge University Press.

Ha, A. Y. H., & Hyland, K. (2017). What is technicality? A Technicality Analysis Model for EAP vocabulary. *Journal of English for Academic Purposes,* 28, 35-49.

Halliday, M. A. K. (1973). *Explorations in the Functions of Language*. London: Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hannay, M., & Martínez-Caro, E (2008). Thematic choice in the written English of advanced Spanish and Dutch learners. *Linking Up, Contrastive and Learner Corpus Research*, 227-253

Harris, Z. (1970). *Papers in structural and transformational linguistics*. Dordrecht, Holland: D. Reidel.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal Of Applied Linguistics,* 4 (2), 237-258.

Hernández-Nanclares, N., & Jiménez-Muñoz (2017). English as a medium of instruction: Evidence for language and content targets in bilingual education in economics. *International Journal of Bilingual Education and Bilingualism*, 20 (7), 883–896.

Hinds, J. (1987). Reader versus writer responsibility: a new typology. Connor, U. & Kaplan R. (eds) *Writing across Languages: Analysis Of L2 Text,* 141-152. Reading. MA: Addison-Wesley.

Hinkel, E. (2002). *Second Language Writers' Text: Linguistic and Rhetorical Features.* New York: Routledge.

Hinkel, E. (2003). Simplicity without elegance: features of sentences in L1 and L2 academic texts. *TESOL Quarterly* 37 (2), 275-301.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text,* 18 (3), 343-382.

Hyland, K. (2000). *Disciplinary Discourses: Social Interactions in Academic Writing*. Harlow: Longman.

Hyland, K. (2001). Humble servants of the discipline? Self-mention in research articles. *English for Specific Purposes,* 20, 207-226.

Hyland, K. (2002). Authority and Invisibility: Authorial Identity in Academic Writing. *Journal of Pragmatics,* 34 (8), 1091-1112.

Hyland, K. (2004). *Disciplinary Discourses: Social Interactions in Academic Writing*. University of Michigan Press.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: a reappraisal. *Applied Linguistics*, 25 (2), 156-177.

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. Continuum, London.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: a reappraisal. Applied Linguistics, 25 (2), 156-177.

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

Hyland, K. (2008a). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, 27 (1), 4-21.

Hyland, K. (2008b). Genre and academic writing in the disciplines. *Language teaching*. 41 (4), 542-562.

Hyland, K. (2009a). Writing in the disciplines: research evidence for specificity. *Taiwan International ESP Journal,* 1 (1), 5-22.

Hyland, K. (Ed) (2009b). *Academic Discourse: English in a Global Context*. Continuum Discourse Series. London

Hyland, K. (2010). Metadiscourse: mapping interactions in academic writing. *Nordic Journal of English Studies,* 9 (2), 125-143.

Hyland, K. (2012a). Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150-169.

Hyland, K. (2012b). *Disciplinary Identities.* Cambridge: Cambridge University Press.

Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics,* 113, 16-29.

Intaraprawat, P., & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing,* 4 (3), 253-272.

Ivanič, R. (2004). Discourses of writing and learning to write. *Language and Education*, 18 (3), 220-243.

Jakobson, R. (1980). *The Framework of Language.* Michigan Studies in the Humanities, Michigan.

Jiang, F. K., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes*, 46, 1-14.

Lasagabaster, D. (2008). Foreign language competence in content and language integrated learning. *Open Applied Linguistics Journal*, 1, 31–42.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics,* 16(3), 307-322.

Lee, D. & Chen, S. (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, 281-296.

Lee, J., Yeung, C. Y., Zeldes, A., Reznicek, M., Lüdeling, A., & Webster, J. (2015). CityU corpus of essay drafts of English language learners: a corpus of textual revision in second language writing. *Language Resources and Evaluation*, 49 (3), 659-683.

Lee, J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing,* 33, 21-34.

Leech, G. (1992). Corpora and theories of linguistic performance. *Directions in corpus linguistics,* 105-122.

Leech, G. (1998). Learner corpora: what they are and what can be done with them. In Granger S. (Ed) *Learner English on Computer*, 14-20. London & New York: Addison Wesley Longman.

Lei, L. (2012). Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes*, 11 (3), 267-275.

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, 22, 42-53.

Leńko-Szymańska, A. (2008). 'Non-native or non-expert? The use of connectors in native and foreign language learners' texts.' *Acquisition et interaction en langue étrangère,* 27, 91-108.

Liu, D. (2008). Linking adverbials: An across-register corpus study and its implications. *International Journal of Corpus Linguistics*, 13, 491-518.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes, 31* (1), 25-35.

Low, G. (1996). Intensifiers and hedges in Questionnaire items and the lexical invisibility hypothesis. *Applied Linguistics,* 17 (1), 1-37.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15 (4), 474-496.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45 (1), 36-62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal,* 96 (2), 190-208.

Martín-Laguna, S., & Alcón, E. (2015). Do learners rely on metadiscourse markers? An exploratory study in English, Catalan and Spanish. *Procedia-Social and Behavioral Sciences,* 173, 85-92.

Mauranen, A. (1993). *Cultural Differences in Academic Rhetoric: A Textlinguistic Study*. Frankfurt: Peter Lang.

Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In Simpson, R. & J. Swales (Eds.). *Corpus Linguistics in North America: Selections from the 1999 Symposium*. Ann Arbor, MI: University of Michigan Press. 165-178.

Mauranen, A. (2010). Discourse reflexivity - a discourse universal? The case of ELF. *Nordic Journal of English Studies,* 9 (2). 13-40.

McCarthy, M. (2006.) *Explorations in Corpus Linguistics*. Cambridge: Cambridge University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-Based Language Studies. An Advanced Resource Book*. London and New York: Routledge.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins Publishing.

Meunier, F. (1998). Computer Tools for the Analysis of Learner Corpora, in S. Granger (Ed.) *Learner English on Computer*. London: Addison Wesley Longman. 19-37

Meunier, F., & Van Goethem, K. (2017 March). Correlating linguistic performance with cognitive, educational and socio-affective variables in SLA Dutch and English CLIL in French-Speaking Belgium. Paper presented at the conference "Learner Corpus Based approaches to Second Language Acquisition". Utrecht University. Utrecht.

Ministerio de Ciencia, Educación y Universidades. (2019). Avance de la Estadística de Estudiantes Universitarios (EEU) Curso 2017-2018. Retrieved from: http://www.educacionyfp.gob.es/servicios-al-ciudadano-mecd/estadisticas/educacion/universitaria/estadisticas/alumnado/2017-2018_Av.html

Montaño-Harmon, M. (1991). Discourse features of written Mexican Spanish: current research in contrastive rhetoric and its implications. *Hispania*, 417-425

Moreno, A. (1997). Genre constraints across languages: causal metatext in Spanish and English RAs. *English for Specific Purposes,* 16, 161-179.

Mur-Dueñas, P. (2011). An intercultural analysis of metadiscourse features in research articles written in English and in Spanish. *Journal of Pragmatics, 43* (12), 3068-3079.

Nation, P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus* (Vol. 14). Amsterdam: John Benjamins Publishing.

Nevertheless. (2018). Longman Dictionary of Contemporary English. Retrieved from: https://www.ldoceonline.com/dictionary/nevertheless

Noble, W. (2010). Understanding metadiscoursal use: Lessons from a 'local' corpus of learner academic writing. *Nordic Journal of English Studies,* 9 (2), 145-169.

O'Donnell, M. B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory, 8* (1), 73-101.

O'Loughlin, R. (2012). Tuning in to vocabulary frequency in coursebooks. *RELC Journal*, *43* (2), 255-269.

O'Sullivan, Í., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, 15 (1), 49-68.

On the contrary. (2018). Oxford Dictionary of English. Retrieved from: https://en.oxforddictionaries.com/definition/on_(or_quite)_the_contrary

Paltridge, B. (2002). Thesis and dissertation writing: an examination of published advice and actual practice. *English for Specific Purposes*, 21, 125-143.

Paquot, M. (2010). *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New-York, Continuum.

Paquot, M. (2017, March). *Phraseological complexity in English writing by French EFL learners*. Paper presented at the workshop Learner Corpus based approaches to Second Language Acquisition, Utrecht, The Netherlands.

Peacock, M. (2010). Linking adverbials in research articles across eight disciplines. *Iberica,* 20 (20), 9-34.

Pérez-Llantada, C. (2010). The discourse functions of metadiscourse in published academic writing: issues of culture and language. *Nordic Journal of English Studies,* 9 (2), 41-68.

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

Pérez-Vidal, C., & Roquet, H. (2015). The linguistic impact of a CLIL Science programme: An analysis measuring relative gains. *System*, 54, 80-90.

Pérez-Vidal, C., López-Serrano, S., Ament, J., & Thomas-Wilhelm, D. J. (Eds.) (2018). *Learning context effects: Study abroad, formal instruction and international immersion classrooms* (EuroSLA Studies 1). Berlin: Language Science Press.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.) *Learner English on Computer*: New York. Addison-Wesley Longman. 107-119.

Polio, C. (2001). Research methodology in second language writing research: the case of text-based studies. In Silva, T. & Matsuda P. K. (Eds.), *On Second Language Writing*. Mahwah: Lawrence Erlbaum. 91-116.

Polio, C., & Yoon, H. (2017, March). Exploring multi-word combinations as measures of linguistic accuracy in second language writing. Paper presented at the workshop: *Learner Corpus based approaches to Second Language Acquisition*, Utrecht, The Netherlands.

Prommas, P., & Sinwongsuwat, K. (2011). *A comparative study of discourse connectors used in argumentative compositions produced by Thai EFL learners and English-native speakers*. Proceedings of the 3rd International Conference on Humanities and Social Sciences; Songkhla, Thailand, April 2nd, 2011. Faculty of Liberal Arts: Prince of Songkla University.

Rica-Peromingo, J. P. (2012). Corpus analysis and phraseology: transfer of multi-word units. *Linguistics and the Human Sciences*, 6, 321-343.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Römer, U. (2009). English in academia: Does nativeness matter. Anglistik: International *Journal of English Studies*, 20 (2), 89-100.

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction* 3 (1), 95-119.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual review of applied linguistics,* 31, 205-225.

Roquet, H., Vraciu, A., Nicolás-Conesa, F. (in press). The effect of amount of exposure to English medium instruction programmes in higher education: The case of morphosyntax. In L. Marqués-Pascual and A. Cortijo-Ocaña (eds): *Second and Third Language Acquisition in Bilingual Contexts*. (pp. -). Delaware: Juan de la Cuesta.

Roquet, H., Vraciu, A., Nicolás-Conesa, F., & Pérez-Vidal, C. (forthcoming). Integrating Content and Language in Higher Education: examining the effects on language gains. *International Journal of Bilingual Education and Bilingualism.*

RStudio (2012). RStudio: Integrated development environment for R [Computer software]. Boston, MA. URL http://www.rstudio.org/ [September 8, 2018].

Ruiz de Zarobe, Y. (2011). Which language competencies benefit from CLIL? An insight into applied linguistics research. In Ruiz de Zarobe, Y., Sierra, J. M., and Gallardo del Puerto, F. (Eds): *Content and foreign language integrated learning: Contributions to multilingualism in European contexts*. Bern: Peter Lang, 129-154

Salas, M. D. (2015). Reflexive metadiscourse in research articles in Spanish: Variation across three disciplines (Linguistics, Economics and Medicine). *Journal of Pragmatics,* 77, 20-40

Shaw, P. (2012). Linking adverbials in student and professional writing in literary studies: what makes writing mature. In Hunston, Pecorari & Charles (Ed.) *Academic Writing.* New York: Continuum Publishing Corporation. 215-235.

Sheldon, E. (2018). Dialogic spaces of knowledge construction in research article Conclusion sections written by English L1, English L2 and Spanish L1 writers. *Ibérica: Revista de la Asociación Europea de Lenguas para Fines Específicos (AELFE), 35*, 13-40.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31* (4), 487-512.

Sinclair, J. (1981). Planes of discourse. In Rizvi, S. (Ed.) *The Two-Fold Voice: Essays in Honour of Ramesh Mohan*. Salzberg: Salzberg University Press. 70-89.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford. Oxford University Press.

Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora*, 27–39. London, UK: Longman.

Sinclair, J. (2005). Corpus and text: basic principles. In M. Wynne (Ed.) *Developing Linguistic Corpora: a Guide to Good Practice* (pp 1-16). Oxford: Oxbow Books. Available online from http://ota.ox.ac.uk/documents/creating/dlc/

Sinclair, S., & Rockwell, G. (2016). Voyant tools. URL: *http://voyant-tools.org/* [September 8, 2018].

Smit, U. & Dafouz, E. (2012). Integrating content and language in higher education. An introduction to English-medium policies, conceptual issues and research practices across Europe. In *AILA Review,* 25, 1-12.

Springer, P. (2012). *Advanced learner writing* (1st ed.). Amsterdam: Vrije Universiteit.

Stryker, S. B., & Leaver, B. L. (Eds.). (1997). *Content-based instruction in foreign language education: Models and methods*. Washington, D. C.: Georgetown University Press.

Sugiura, M., Narita, M., Ishida, T., Sakaue, T., Murao, R., & Muraki, K. (2007). A Discriminant Analysis of Non-native Speakers and Native Speakers of English. *Corpus Linguistics Conference Proceedings*. University of Birmingham.

Swales, J. M. (2002). *English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Thompson Corp. (2019). ISI Web of Knowledge [Website]. Accessible from: *http://scientific.thomson.com/isi/*.

Tono, Y. (2011). TALC in action: Recent innovations in corpus-based English language teaching in Japan. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning*, 3–25. London, UK: Continuum.

Toumi, N. (2009). A model for the investigation of reflexive metadiscourse in research articles. *Language Studies Working Papers,* 1, 64-73.

Vande Kopple, W. (1985). Some exploratory discourse on metadiscourse. *College Composition & Communication,* 26, 82-93.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

Wilkinson, R., & Hommes, J. (2010). *A Guide to Academic Writing Skills*. 2nd ed. Maastricht, The Netherlands: School of Business and Economics. Maastricht University

Williams, J. M. (1981). *Style: Ten Lessons in Clarity and Grace*. Glenview, IL: Scott, Foresman.

Wolfe-Quintero, K., Inagaki, S., Kim, H. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Second Language Teaching & Curriculum Center: University of Hawai'i at Manoa.

Wood, D. C., & Appel, R. (2014). Multiword constructions in first year business and engineering university textbooks and EAP textbooks. *Journal of English for Academic Purposes*, 15, 1-13.

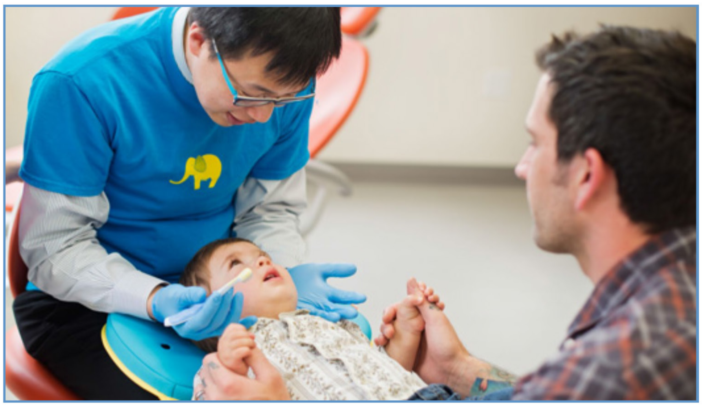
Wray, A. (2002). *Formulaic Language and the Lexicon.* Cambridge: Cambridge University Press.

Yang, M. N. (2015). A nursing academic word list. *English for Specific Purposes*, 37, 27-38.

Appendix 1. The writing task

Name and surname:
First language:

Look at the following image, and respond to the questions about it below:



1. **Describe the scene shown in the image. What do you think has just happened?**

2. **Write a possible dialogue among the people shown in the image.**

Answer the following two questions (write in paragraph style):

3. **What would you do in this situation?**

4. **How could you determine whether your approach is the best one for this situation?**

Image source: https://northseattledds.com/fear-anxiety/

Appendix 2. Top-50 most frequent words (VL), collocations (CL) and formulas (FL) in the class material corpus

	VL	CL	FL
1	dental	lateral incisor	of dental anxiety
2	study	conventional black	of the study
3	oral	root canal	according to the
4	research	carried out	the patient's
5	health	diabetes mellitus	type diabetes mellitus
6	treatment	tuc counseling	a case report
7	teeth	attachment loss	have respiratory problems
8	patient	lung cancer	the patient was
9	intervention	surface loss	what are the
10	group	black tea	the number of
11	pain	herbal tea	the prevalence of
12	anxiety	respiratory problems	to determine the
13	case	data collection	in order to
14	tooth	common way	in this study
15	information	risk factor	of oral cancer
16	dentistry	case report	the relationship between
17	evidence	tooth surface	the results of
18	results	case reports	as well as
19	population	use cessation	in a population
20	use	tobacco use	more likely to
21	clinical	comprehension questions	risk factor for
22	based	relationship between	the control group
23	subjects	caused by	years of age
24	used	periodontal disease	cross sectional study
25	groups	association between	non experimental research
26	data	at least	based on the
27	caries	increased risk	how would you
28	abstract	over time	in other words
29	disease	based on	of the tooth
30	age	oral cancer	the development of
31	care	oral cavity	this type of
32	report	other words	a cross sectional
33	studies	more likely	conventional black tea
34	factors	oral hygiene	risk factor for periodontitis
35	related	smoking status	the proportion of
36	survey	university students	the purpose of
37	dentists	control group	there is a
38	mean	cohort studies	tobacco use cessation
39	objective	risk factors	common way to say
40	control	associated with	in the dental
41	dentist	research project	is associated with
42	researchers	non-experimental research	the aim of
43	cancer	oral health	the majority of
44	common	health care	tooth surface loss
45	medical	dental anxiety	type of research
46	periodontal	compared with	dental anxiety in
47	design	dental fear	of the following
48	examination	research design	one of the
49	practice	cross-sectional study	oral health care
50	risk	cohort study	prevalence of dental

Appendix 3. Frequency and position of adversative Linking Adverbials by category and subcategory, in both LOCNESS and MUC (adapted from Liu, 2008:514)

	NS (LOCNESS): 149,574w			**NNS (MUC):** 148,068w		
LAs	**Total hits**	**Initial**	**Medial**	**Total hits**	**Initial**	**Medial**
Proper adversative/Concessive						
at the same time	10	4	6	12	3	9
however	175	106	69	178	144	34
nevertheless	3	2	1	54	50	4
nonetheless	0	0	0	20	19	1
of course	25	13	12	26	14	13
once again	13	4	9	13	6	7
tough	35	10	25	21	4	14
even though*	31	17	14	33	18	15
although*	51	45	6	73	49	24
yet	52	26	26	58	20	21
Subtotal	**395**	**227**	**168**	**492**	**327**	**142**
Contrastive						
actually	38	3	35	17	2	15
as a matter of fact	0	0	0	6	6	0
conversely	1	1	0	0	0	0
in comparison	3	0	3	9	4	5
in contrast	4	3	1	28	15	13
in fact	35	20	15	38	29	9
in reality	3	0	3	7	5	2
on the other hand	22	13	9	24	15	9
Subtotal	**106**	**40**	**66**	**129**	**76**	**53**
Correction						
instead	62	23	39	40	10	30
on the contrary	2	2	0	3	2	1
rather	56	6	50	44	1	43
Subtotal	**120**	**31**	**89**	**87**	**13**	**74**
Dismissal						
admittedly	2	2	0	3	2	1
after all	8	8	0	4	2	2
all the same	0	0	0	1	0	1
anyhow	0	0	0	0	0	0
anyway	9	0	9	1	0	1
at any rate	0	0	0	0	0	0
despite	6	3	3	28	18	10
in any case	0	0	0	2	0	2
in spite of	0	0	0	2	2	0
still	9	0	9	11	8	3
Subtotal	**34**	**13**	**21**	**52**	**32**	**20**
TOTAL	**655**	**311**	**344**	**737**	**448**	**289**

	NS (LOCNESS): 149,574w			**NNS (MUC):** 148,068w		
LAs	**Total hits**	**Initial**	**Medial**	**Total hits**	**Initial**	**Medial**
Normed 1000w	**4,38**			**4,84**		
Normed 1000s	**82,16**			**107,65**		

*subordinating conjunction

Appendix 4. Frequency and position of adversative LAs by category and subcategory, in the NNS corpus (MUC) according to students' L1

	DUTCH			**GERMAN**			**FRENCH**			**ITALIAN**			**SPANISH**		
LAs	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**
Proper Adversative/ Concessive															
at the same time	0	0	0	4	2	2	2	0	2	2	1	1	4	0	4
however	37	27	10	59	46	13	30	22	8	22	20	2	30	29	1
nevertheless	11	9	2	13	12	1	11	10	1	7	7	0	12	12	0
nonetheless	1	1	0	7	7	0	5	5	0	1	1	0	6	5	1
of course	1	1	0	1	1	0	21	9	12	0	0	0	4	3	1
once again	1	1	0	1	0	1	2	0	2	3	1	2	6	4	2
though	6	2	4	4	1	3	3	1	2	2	0	2	3	0	3
even though*	3	1	2	5	3	2	6	4	2	9	5	4	10	5	5
although*	10	8	2	25	18	7	16	10	6	10	5	5	12	8	4
yet	2	1	1	4	1	3	14	5	9	18	11	7	3	2	1
Subtotal	**72**	**51**	**21**	**123**	**91**	**32**	**110**	**66**	**44**	**74**	**51**	**23**	**90**	**68**	**22**
Contrastive															
actually	6	1	5	2	0	2	4	1	3	3	0	3	2	0	2
as a matter of fact	0	0	0	1	1	0	0	0	0	4	4	0	1	1	0
conversely	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
in comparison	0	0	0	6	3	3	1	0	1	0	0	0	2	1	1
in contrast	15	10	5	7	3	4	2	1	1	1	1	0	3	0	3
in fact	6	5	1	2	1	1	15	9	6	10	10	0	5	4	1
in reality	2	2	0	0	0	0	2	2		2	1	1	1	0	1
on the other hand	2	2	0	5	3	2	5	3	2	1	1	0	11	6	5
Subtotal	**31**	**20**	**11**	**23**	**11**	**12**	**29**	**16**	**13**	**21**	**17**	**4**	**25**	**12**	**13**
Correction															

	DUTCH			**GERMAN**			**FRENCH**			**ITALIAN**			**SPANISH**		
LAs	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**	**Total hits**	**initial**	**medial**
instead	7	3	4	13	5	8	12	1	11	6	1	5	2	0	2
on the contrary	0	0	0	1	1	0	1	0	1	0	0	0	1	1	0
rather	10	0	10	18	0	18	3	0	3	5	0	5	8	1	7
Subtotal	**17**	**3**	**14**	**32**	**6**	**26**	**16**	**1**	**15**	**11**	**1**	**10**	**11**	**2**	**9**
Dismissal															
admittedly	2	1	1	1	1	0	0	0	0	0	0	0	0	0	0
after all	1	1	0	2	1	1	0	0	0	0	0	0	1	0	1
all the same	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
anyhow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
anyway	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
at any rate	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
despite	2	2	0	9	5	4	9	7	2	1	1	0	7	3	4
in any case	1	0	1	0	0	0	1	0	1	0	0	0	0	0	0
in spite of	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0
still	4	4	0	1	1	0	3	1	2	3	2	1	0	0	0
Subtotal	**10**	**8**	**2**	**14**	**8**	**6**	**16**	**10**	**6**	**4**	**3**	**1**	**8**	**3**	**5**
TOTAL	**130**	**82**	**48**	**192**	**116**	**76**	**171**	**93**	**78**	**110**	**72**	**38**	**134**	**85**	**49**
Normed 1k W	**4,32**			**5,42**			**5,45**			**4,22**			**5,34**		
Normed 1k S	**85,75**			**118,52**			**112,06**			**102,23**			**120,94**		

*subordinating conjunction

Appendix 5. List of academic journals used to compile the expert corpus

MED journals
BMJ Quality & Safety
European Journal of Clinical Investigation
Journal of international medical research
Journal of investigative medicine
Journal of the Canadian Association of Emergency Physicians
Lancet Neurol
Nursing Older People
Regenerative Medicine
The new England Journal of Medicine
Tissue Engineering

LIN journals
Applied linguistics
Computer Learner Corpora, Second Language Acquisition, and Foreign Language Teaching
Corpora and Language Teaching
English for Specific Purposes
Journal of Second Language Writing
Language Teaching Research
Lingua
Linguistics and the human sciences
TESOL Quarterly
Text: Interdisciplinary Journal for the Study of Discourse

Appendix 6. Global results for metadiscourse categories in the learner and the expert corpus (raw and normed results per 1000 words)

Reflexive metadiscourse		**BDs**				**RAS**			
Discipline		**LIN**	**LIN norm**	**MED norm**	**MED**	**LIN**	**LIN norm**	**MED norm**	**MED**
Tokens		**65,180**			**38,791**	**177,041**			**81,182**
Types		**5,537**			**4,656**	**9,853**			**7,553**
Metatext	**Tags**								
Reference to the text									
Full text	_MD_MT_RT_FT	115	1.76	1.39	54	321	1.81	1.22	99
Part of the text	_MD_MT_RT_PT	128	1.96	1.16	45	194	1.10	0.60	49
Semiotic modes	_MD_MT_RT_SM	79	1.21	0.80	31	475	2.68	3.07	249
TOTAL RT		**322**	**4.94**	**3.35**	**130**	**990**	5.59	4.89	**397**
Endophoric markers									
Anaphoric	_MD_MT_EN_AN	90	1.38	0.46	18	159	0.90	0.59	48
Cataphoric	_MD_MT_EN_CA	53	0.81	0.67	26	210	1.19	0.64	52
Deictic	_MD_MT_EN_DE	13	0.20	0.00	0	132	0.75	0.02	2
TOTAL EN		**156**	**2.39**	**1.13**	**44**	**501**	**2.83**	**1.26**	**102**
Code Glosses									
Reformulators	_MD_MT_CG_RE	138	2.12	1.01	39	442	2.50	1.34	109
Exemplifiers	_MD_MT_CG_EX	155	2.38	0.80	31	696	3.93	1.64	133
Parentheticals	_MD_MT_CG_PA	266	4.08	7.53	292	618	3.49	5.57	452
Dashes (-)	_MD_MT_CG_DA	22	0.34	0.00	0	40	0.23	0.16	13
Colons (:)	_MD_MT_CG_CL	192	2.95	2.55	99	277	1.56	0.65	53
Semicolons (;)	_MD_MT_CG_SC	67	1.03	0.62	24	216	1.22	1.68	136
TOTAL CG		**840**	**12.89**	**12.50**	**485**	**2289**	**12.93**	**11.04**	**896**
Linking Devices									
Adding	_MD_MT_LD_AD	158	2.42	2.60	101	282	1.59	1.88	153
Constrasting	_MD_MT_LD_CN	304	4.66	2.81	109	785	4.43	2.82	229
Consecutive	_MD_MT_LD_CO	110	1.69	1.50	58	347	1.96	1.13	92
Organizers	_MD_MT_LD_OR	152	2.33	1.16	45	388	2.19	1.39	113
Topicalizers	_MD_MT_LD_TO	44	0.68	0.36	14	184	1.04	0.18	15
TOTAL LD		**768**	**11.78**	**8.43**	**327**	**1986**	**11.22**	**7.42**	**602**
TOTAL METATEXT		**2086**	**32.00**	**25.42**	**986**	**5766**	**32.57**	**24.60**	**1997**
Interpersonal	**Tags**								
Writer oriented									
Self-mention	_MD_IP_WO_SF	119	1.83	5.13	199	725	4.10	3.82	310
Reader oriented									
Rethorical Questions	_MD_IP_RO_RQ	3	0.05	0.00	0	12	0.07	0.00	0
Directives	_MD_IP_RO_DI	28	0.43	0.90	35	200	1.13	0.17	14
TOTAL RO		**31**	**0.48**	**0.90**	**35**	**212**	**1.20**	**0.17**	**14**
Participant oriented									
Inclusive we	_MD_IP_PO_IW	58	0.89	0.34	13	173	0.98	0.04	3
TOTAL INTERPERSONAL		**208**	**3.19**	**6.37**	**247**	**1110**	**6.27**	**4.03**	**327**
TOTAL METADISCOURSE		**2294**	**35.19**	**31.79**	**1233**	**6876**	**38.84**	**28.63**	**2324**
TOTAL MD PERCENTAGE			**3.52%**	**3.18%**			**3.88%**	**2.86%**	

Appendix 7. Top-3 textual and interpersonal markers in each corpus[21]

	LIN		MED	
TEXTUAL	**BDs**	**RAs**	**BDs**	**RAs**
Reference to the text				
Full text	(this) paper (64)	(current, this) study (213)	(this) study (43)	(current, present) study (77)
	(this) study (21)	(this) paper (75)	(this) project (4)	(this) paper (8)
	(this, final) project (10)	(this) article (23)	(this) document (2)	(our) trial (5)
Part of the text	(in this) section (45)	(in this) section (89)	(see) appendix (15)	(in) appendix (19)
	(in) appendix (20)	(see) appendix (17)	(see) annex (10)	(in) sections (14)
	(in the) introduction (6)	(in the) discussion (14)		(see) annex (2)
Semiotic modes	(in) table (27)	(in) table (149)	table (11)	figure (83)
	figure (17)	(in) figure (63)	figure (7)	table (48)
	in (x) (15)	in (x) (39)	diagram (4)	image (4)
Endophoric markers				
Anaphoric	(explained, stated) above (14)	(noted, listed) above (75)	above (8)	(described) previously (16)
	(the) latter (11)	(the) latter (21)	(as) mentioned (3)	(described) above (9)
	(the) previous (7)	(as) mentioned (12)		(as) mentioned (5)
Cataphoric	(in the) following (22)	(are the, in the) following (86)	(the) following (14)	(the) following (17)
	(as) follows (8)	(discussed) below (64)	(as) follows (7)	(as) follows (10)
	(described) below (3)	next (section) (12)		
Deictic	here (we) (9)	(adopted, used) here (94)	N.A.*	here (1)
	now (we) (3)	(let us) now (26)		
		so far (8)		
Code Glosses				
Reformulators	i.e. (21)	i.e. (114)	especially (12)	specifically (32)
	(defined, known, referred to) as (19)	especially (55)	defined as (8)	defined as (26)
	that is, (17)	particularly (50)	specifically (6)	especially (14)
Exemplifiers	such as (38)	e.g. (243)	such as (13)	such as (66)
	(for) instance (22)	such as (138)	e.g. (7)	for example (36)
	e.g. (20)	for example (135)	for instance (4)	e.g. (22)
Parentheticals	refer to sections (7)	list examples (13)	refer to sections (23)	refer to semiotic modes (68)
	list examples (4)	refer to sections (5)	specify type of variable (7)	
	refer to semiotic modes (4)	cataphoric markers (3)		
Dashes (-)[22]	- also known as (2)	- and (4)	N.A	- and (5)
	- e.g. (2)	- thus (2)		- for example (2)
	- i.e. (2)	- that is, (2)		

[21] The function *Cluster* in the text analysis software AntConc has been used to identify the top-3 markers in each category; a minimum range of two was set (i.e. markers had to be present in at least two different texts to be included in the top-3 list).

[22] In the case of dashes, colons, and semicolons, we provide the words that followed or preceded these marks more frequently. As for parentheticals, we indicate three of the most frequent functions they perform in all texts -i.e. contain lists of examples, refer to semiotic modes, or to parts of the text.

Colons (:)	for example: (8)	the following: (6)	are: (3)	as follows: (4)
	are: (6)	categories: (5)	for example: (2)	
	as follows: (4)	research question: (3)		
Semicolons (;)	; the (13)	; and (23)	; however (3)	; and (16)
	: and (7)	; however (12)		; however (10)
	; in (6)	; see (7)		; therefore (4)
Linking Devices				
Adding	moreover (32)	in addition (57)	moreover (16)	in addition (29)
	furthermore (23)	moreover (30)	furthermore (15)	additionally (29)
	another (12)	another (27)	in addition (11)	furthermore (18)
Contrasting	however (76)	however (247)	however (54)	however (89)
	whereas (49)	although (109)	although (5)	although (46)
	although (30)	while (75)	nonetheless (6)	while (20)
Consecutive	thus (35)	thus (184)	therefore (29)	therefore (42)
	therefore (30)	therefore (103)	thus (12)	thus (26)
	hence (16)	hence (24)	consequently (5)	As a result (10)
Organizers	finally (24)	(the) second (127)	respectively (7)	respectively (25)
	on the one hand (13)	finally (44)	(the) second (5)	finally (15)
	first (11)	third (42)	then (5)	then (14)
Topicalizers	in the (case, context) of (13)	in (terms, the case, the context) of (63)	in terms of (6)	with respect to (8)
	regarding (9)	with (respect, regard) to (45)	regarding (4)	in the context of (5)
	as far as (x) is concerned (2)	regarding (21)	as for (2)	with regard to (2)

	LIN		MED	
INTERPERSONAL	**BDs**	**RAs**	**BDs**	**RAs**
Writer oriented				
Self-mention	we (44) (have, can, found)	we (410) (will, have, examined)	we (will, have, expect) (133)	we (194) (found, used, examined)
	I (40) (would like to)	our (188) (study, data, investigation)	our (study, results) (55)	our (112) (study, findings, knowledge)
	our (20) (findings, analysis)	I (69) (have, will, would)	(allows) us (6)	
Reader oriented				
Directives	see (21)	see (118)	see (32)	see (15)
	consider (1)	cf. (30)		
		consider (12)		
Rhetorical Qs.	N.A	N.A	N.A	N.A*
Participant oriented				
Inclusive we	we (can see, have seen) (45)	we (can see, need) (129)	we (can, need) (13)	we (should) (3)
	(let) us (6)	(gives, helps, let) us (22)		

*Non-Applicable

Appendix 8. Complete list of reflexive metadiscourse markers found in the corpora

Textual Metadiscourse

Full text	Part of the text	Semiotic modes	Anaphoric	Cataphoric	Deictic	Reformulators	Exemplifiers
my paper	analytical framework	diagram	above	as follows	here	also known as	an example
our investigation	annex	extract	abovementioned	below	now	at the same time	and so on
our research	annexes	fig.	aforementioned	follows with	so far	defined as	and so on and so forth
our study	appendices	figure	Again	further on	up to this point	especially	as in
our work	appendix	fragment	as it has been mentioned	in the following		generally speaking	be it
the current article	in the analysis	graph	as mentioned	in the next		i.e.	e.g.
the current study	in the results	in (	as noted above/earlier	later on		in other words	for example
the present paper	section		as seen in	subsequent		in particular	for instance
the present research	sections		the first of	the following		known as	like
the present study	subsection		the former			more accurately	such as
the study	the conclusion		the latter			more specifically	
this article	the discussion		this first			namely	
this essay	the introduction		this second			particularly	
this investigation	the methodology					put differently	**Parentheticals**
this paper	the theoretical					put it	()
this project						put it simply	
this study						referred to as	**Colons**
this trial						simply put	:
this work						so-called	
						specifically	**Semicolons**
						that is to say	;
						that is,	
						to be more precise	**Dashes**
							–

					Interpersonal metadiscourse		
Adding	**Contrasting**	**Consecutive**	**Organizers**	**Topicalisers**	**Self-mention**	**Directives**	**inclusive we**
additionally	alternatively	as a consequence	(a) (b) (c)	as far as	I	cf.	brings us
Also	although	as a result	(i) (ii) (iii)	as for	me	consider	if we
And	and yet	consequently	1) 2) 3)	as regards	our	if you look at	let us
Another	But	hence	Afterwards	concerning	the authors'	note that	we can
apart from	but still	So	all in all	in regard to	us	one can	we find
as well as	by contrast	therefore	eventually	in terms of	we	one could	we may
Besides,	contrarily	thus	finally	in that light		one could	we might
furthermore	contrastively		first of all	in the case of		one might	we see
in addition	conversely		first,	in the context of		see	we should
moreover	despite		firstly	regarding		view	we will
Other	even though		in the first place	turning to			
Similarly	however		Last	with regard to			
	in any case		lastly	with respect to			
	in contrast		next,				
	in spite of		on the one hand				
	in turn		overall				
	instead		pn the other hand				
	nevertheless		respectively				
	nonetheless		Second,				
	notwithstanding		secondly				
	on the contrary		The first				
	on the other hand		The first of				
	otherwise		the last				
	Rather		the second				
	Still		Then				
	though		then,				
	unlike		third				
	whereas		thirdly				
	While		to begin with				
	whilst		To conclude				
	Yet		to sum up				

Appendix 9. Lexical bundles found in the learner and the expert corpus according to sections and disciplines (sorted by frequency)

LIN BD introduction	
in order to	14
the aim of	8
of this paper (is to)	8
the analysis of (the)	7
the use of	6
as well as the	5
the fact that	4
(one) of the most	4
in this paper	4
the study of	4
it has been	4
(one) of the main	4
due to the (fact that)	4
to the study (of the)	4
paper aims to	3
attention to the	3
related to the	3
to do so	3
followed by the	3
there is a	3
aim of this paper is	3
this paper aims (to)	3
this paper will (focus on)	3

LIN BD conclusion	
the use of	11
in order to	8
the fact that	5
of this paper	4
as well as	4
most of the	4
it has been	4
this study has	3
one of the	3
analysis of the	3

MED BD introduction	
in order to	7
as well as	7
such as the	5
according to the	5
the rest of	5
the prevalence of	5
the result of	4
of the most	4
the use of	4
the risk of	4
the development of	4
there is no	4
is one of the	4
the conclusion that	3
as a result	3
of the population	3
lack of a	3
the most prevalent	3
is the most	3
it is a	3
it is the	3
is not a	3
recent studies have	3
have been proposed	3
although there is (no)	3

MED BD conclusion	
of this study	6
in order to	5
the possibility of	5
due to the (fact that)	5
the results of	3
impact of the	3
one of the	3
will not be	3

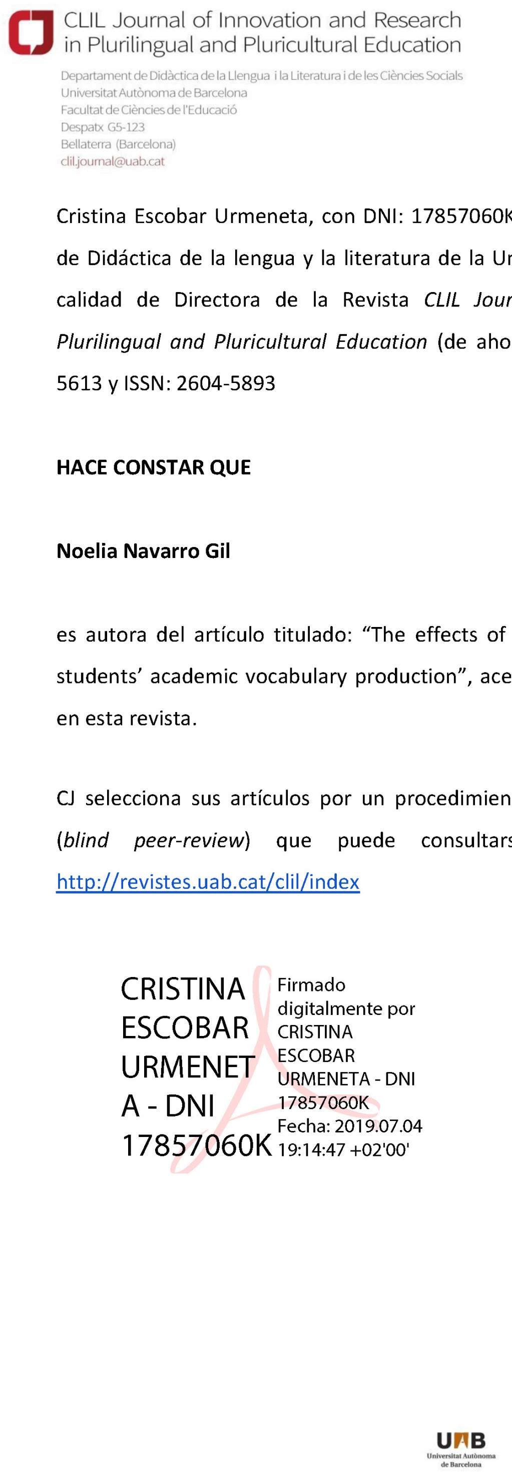
LIN RA introduction	
the use of	17
in order to	14
(used) to refer to	10
in terms of	9
refer to the	8
of the most	8
the effects of	7
one of the	7
the basis of	6
as well as	6
some of the	6
different types of	6
of the same	5
that they are	5
the current study	5
the present study	5
there is a	5
based on the	5
in the field	5
the nature of	5
are likely to	5
the comparison of	4
between the two	4
interest in the	4
in this study	4
in this paper	4
the focus of	4
the results of	4
the area of	4
the context of	4
the fact that	4
the range of	4
the role of	4
the ways that	4
can be used	4
the field of	4
in the current	4
the notion of	4
the study of	4
that it is	4
it has been	4
argue that the	4
in the context of	4
a (wide) range of	4
to contribute to (the)	4
differences in the	3
to find out	3
the importance of	3
in the study	3
focusing on the	3
as a result	3
in relation to	3
the number of	3
a number of	3
as part of	3
in a number of	3

MED RA introduction	
the use of	10
the risk of	7
as well as	6
a number of	5
of this study	4
in order to	4
the effect/s of	4
the presence of	4
been shown to	4
to be the	4
it is not	3
there is a	3
the prevalence of	3
changes in the	3
the ability to	3
be able to	3
as a result (of)	3

MED RA conclusion	
the use of	11
the current study	9
as well as	7
is associated with	7
in this study	6
was associated with a/an	6
a number of	6
the proportion of	6
the presence of	6
the present study	5
the results of	5
consistent with the	5
can be used (to)	5
in addition to	4
there was no	4
the prevalence of	4
in our study	3
because of the	3
the application of	3
the field of	3
we did not	3
are needed to	3
there is a	3

				to develop a	3
				analysis of the	3
				in what ways	3
				is used to	3
				understanding of the	3
				the form of	3
				body of research	3
				the potential to	3
				contribute to the	3
				be argued that	3
				is the use of	3
				in the form of	3
				a growing interest in the	3
				on the basis of the	3
				it can be argued that	3
				of the use (of)	3
				is likely to (be)	3
				has the potential (to)	3
LIN RA conclusion					
the use of	26	but it is	3		
the present study	10	that there are	3		
in this study	10	it is important	3		
in order to	10	study has shown	3		
the fact that	10	study has been	3		
in this paper	8	this study is	3		
there is a	8	is that the	3		
as well as	8	the part of	3		
seems to be	8	reference to the	3		
the case of	7	in this way	3		
in the use of	7	the following three	3		
in relation to	6	a variety of	3		
in terms of	6	some of the	3		
the lack of	6	the majority of	3		
differences in the	6	the number of	3		
of the most	6	be used to	3		
in the present study	6	can be used to	3		
the importance of	5	the construction of	3		
the current study	5	the level of	3		
based on the	5	the process of	3		
with respect to	5	the role of	3		
should not be	5	the beginning of	3		
in the case of	5	for the present	3		
this study has	4	found in the	3		
this paper has	4	the complexity of	3		
such as the	4	understanding of the	3		
for future research	4	on their own	3		
due to the	4	to be the	3		
has shown that	4	it should be	3		
greater use of	4	there is no	3		
the analysis of	4	on the other hand	3		
the quality of	4	avenues for future research	3		
in the literature	4	on the part of	3		
that there is	4	it is true that	3		
needs to be	4	still in its infancy	3		
our understanding of	4	play an important role in	3		
it would be (interesting to)	4				

Appendix 10. Letters of acceptance from editors



CLIL Journal of Innovation and Research
in Plurilingual and Pluricultural Education

Departament de Didàctica de la Llengua i la Literatura i de les Ciències Socials
Universitat Autònoma de Barcelona
Facultat de Ciències de l'Educació
Despatx G5-123
Bellaterra (Barcelona)
clil.journal@uab.cat

Cristina Escobar Urmeneta, con DNI: 17857060K, Profesora Titular del Departamento de Didáctica de la lengua y la literatura de la Universitat Autònoma de Barcelona, en calidad de Directora de la Revista *CLIL Journal of Innovation and Research in Plurilingual and Pluricultural Education* (de ahora en adelante, CJ) con eISSN: 2604-5613 y ISSN: 2604-5893

HACE CONSTAR QUE

Noelia Navarro Gil

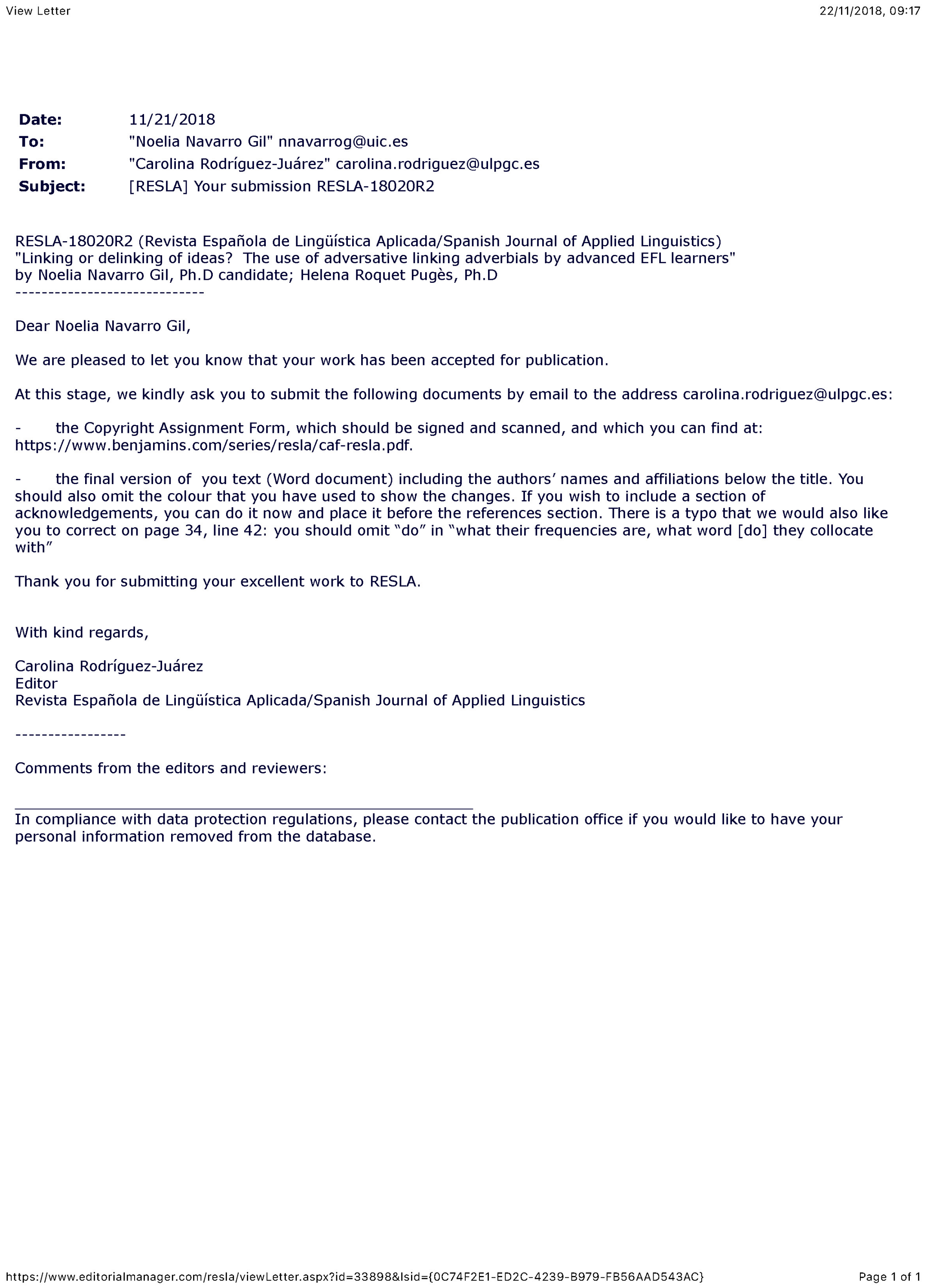
es autora del artículo titulado: "The effects of a content-based language course on students' academic vocabulary production", aceptado para su publicación en el 2019 en esta revista.

CJ selecciona sus artículos por un procedimiento de evaluación anónima por pares (*blind peer-review*) que puede consultarse en la web de la revista: http://revistes.uab.cat/clil/index

CRISTINA
ESCOBAR
URMENET
A - DNI
17857060K

Firmado
digitalmente por
CRISTINA
ESCOBAR
URMENETA - DNI
17857060K
Fecha: 2019.07.04
19:14:47 +02'00'

UAB
Universitat Autònoma
de Barcelona



View Letter 22/11/2018, 09:17

Date: 11/21/2018
To: "Noelia Navarro Gil" nnavarrog@uic.es
From: "Carolina Rodríguez-Juárez" carolina.rodriguez@ulpgc.es
Subject: [RESLA] Your submission RESLA-18020R2

RESLA-18020R2 (Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics)
"Linking or delinking of ideas? The use of adversative linking adverbials by advanced EFL learners"
by Noelia Navarro Gil, Ph.D candidate; Helena Roquet Pugès, Ph.D

Dear Noelia Navarro Gil,

We are pleased to let you know that your work has been accepted for publication.

At this stage, we kindly ask you to submit the following documents by email to the address carolina.rodriguez@ulpgc.es:

- the Copyright Assignment Form, which should be signed and scanned, and which you can find at: https://www.benjamins.com/series/resla/caf-resla.pdf.

- the final version of you text (Word document) including the authors' names and affiliations below the title. You should also omit the colour that you have used to show the changes. If you wish to include a section of acknowledgements, you can do it now and place it before the references section. There is a typo that we would also like you to correct on page 34, line 42: you should omit "do" in "what their frequencies are, what word [do] they collocate with"

Thank you for submitting your excellent work to RESLA.

With kind regards,

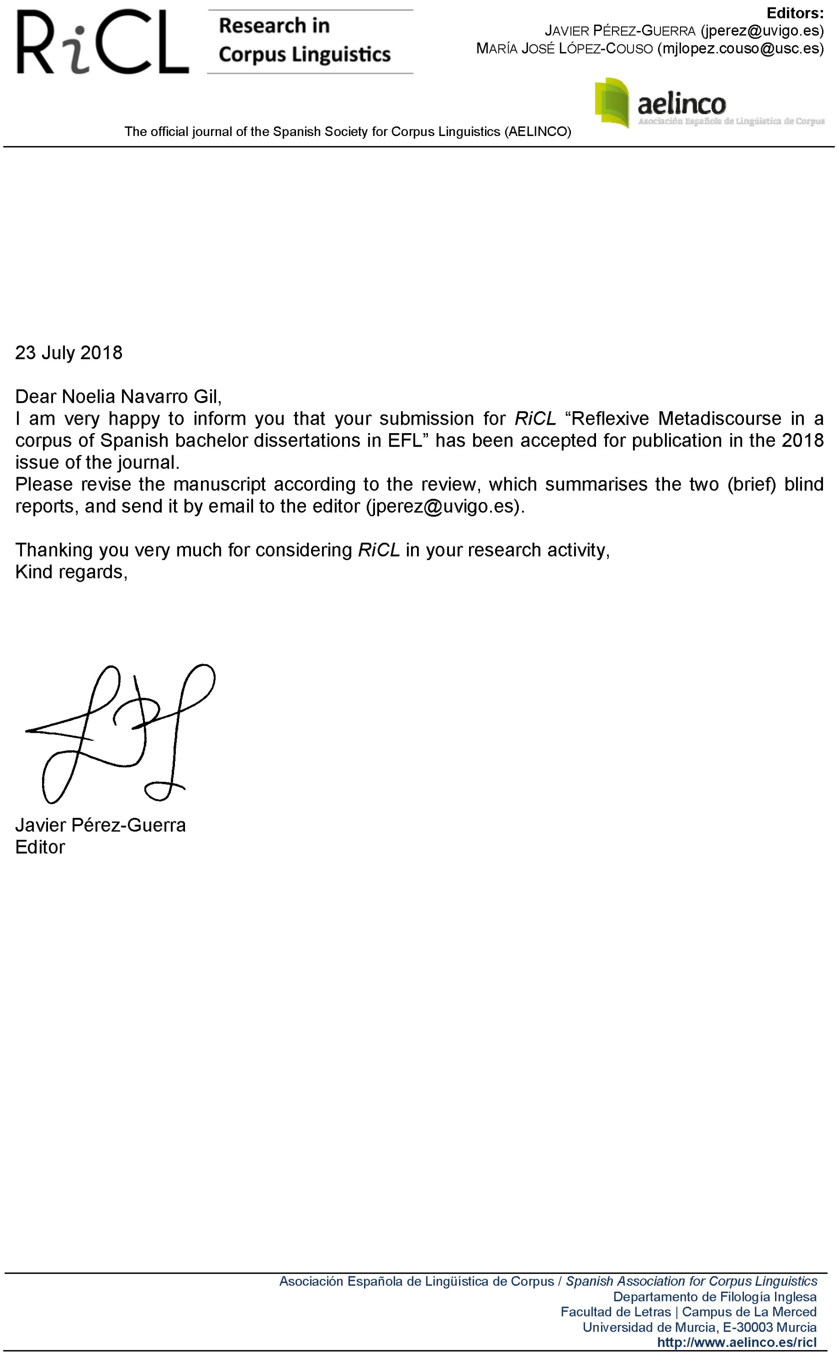
Carolina Rodríguez-Juárez
Editor
Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics

----

Comments from the editors and reviewers:

In compliance with data protection regulations, please contact the publication office if you would like to have your personal information removed from the database.

https://www.editorialmanager.com/resla/viewLetter.aspx?id=33898&lsid={0C74F2E1-ED2C-4239-B979-FB56AAD543AC} Page 1 of 1



RiCL Research in Corpus Linguistics

Editors:
JAVIER PÉREZ-GUERRA (jperez@uvigo.es)
MARÍA JOSÉ LÓPEZ-COUSO (mjlopez.couso@usc.es)

aelinco
Asociación Española de Lingüística de Corpus

The official journal of the Spanish Society for Corpus Linguistics (AELINCO)

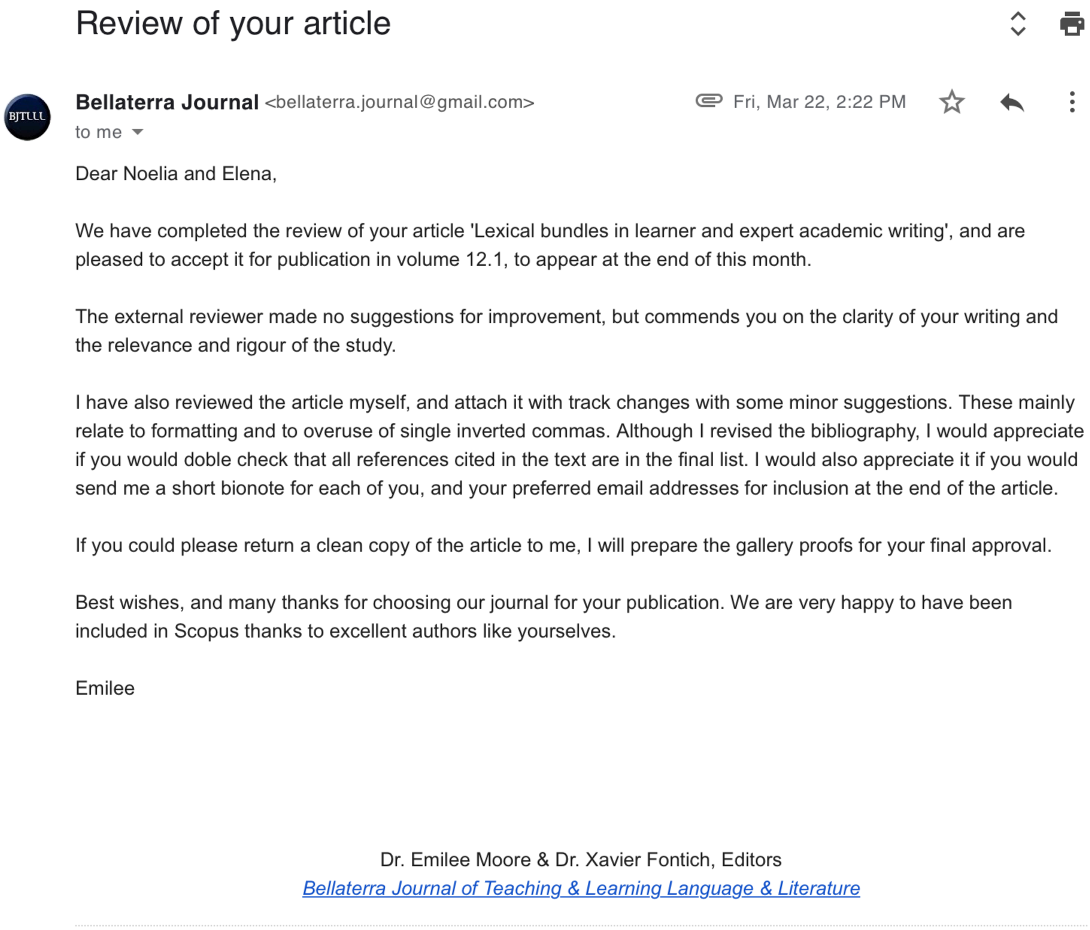
23 July 2018

Dear Noelia Navarro Gil,
I am very happy to inform you that your submission for *RiCL* "Reflexive Metadiscourse in a corpus of Spanish bachelor dissertations in EFL" has been accepted for publication in the 2018 issue of the journal.
Please revise the manuscript according to the review, which summarises the two (brief) blind reports, and send it by email to the editor (jperez@uvigo.es).

Thanking you very much for considering *RiCL* in your research activity,
Kind regards,

Javier Pérez-Guerra
Editor

Asociación Española de Lingüística de Corpus / *Spanish Association for Corpus Linguistics*
Departamento de Filología Inglesa
Facultad de Letras | Campus de La Merced
Universidad de Murcia, E-30003 Murcia
http://www.aelinco.es/ricl



Review of your article

Bellaterra Journal <bellaterra.journal@gmail.com>

to me

Fri, Mar 22, 2:22 PM

Dear Noelia and Elena,

We have completed the review of your article 'Lexical bundles in learner and expert academic writing', and are pleased to accept it for publication in volume 12.1, to appear at the end of this month.

The external reviewer made no suggestions for improvement, but commends you on the clarity of your writing and the relevance and rigour of the study.

I have also reviewed the article myself, and attach it with track changes with some minor suggestions. These mainly relate to formatting and to overuse of single inverted commas. Although I revised the bibliography, I would appreciate if you would doble check that all references cited in the text are in the final list. I would also appreciate it if you would send me a short bionote for each of you, and your preferred email addresses for inclusion at the end of the article.

If you could please return a clean copy of the article to me, I will prepare the gallery proofs for your final approval.

Best wishes, and many thanks for choosing our journal for your publication. We are very happy to have been included in Scopus thanks to excellent authors like yourselves.

Emilee

Dr. Emilee Moore & Dr. Xavier Fontich, Editors
Bellaterra Journal of Teaching & Learning Language & Literature

Summary of the thesis in Spanish

Introducción

En las últimas décadas ha crecido un interés científico en describir cómo se construye el discurso académico en diferentes disciplinas y géneros. La globalización y la aparición del inglés como *lingua franca*, y también como el lenguaje de la ciencia y de la investigación, han hecho del discurso académico en inglés un requisito para la publicación y, por lo tanto, una habilidad básica para investigadores noveles y estudiantes universitarios. Esto ha tenido un impacto considerable en una mayoría de instituciones europeas de educación superior, en las que el número de programas y asignaturas que se ofrecen en inglés está (y aún sigue) experimentando un aumento constante. A los estudiantes universitarios se les exige habitualmente escuchar (ej. clases magistrales, conferencias), hablar (ej. hacer presentaciones), leer (ej. la literatura pertinente) y escribir (ej. redactar trabajos y proyectos), a diferentes niveles de inmersión, en inglés. En el contexto español, Bolonia y el requisito de obtener un certificado de B1 – B2 en inglés (u otra lengua extranjera) que la mayoría de las universidades han establecido para que los estudiantes puedan graduarse, ha desencadenado, en parte, el aumento de programas que se imparten en inglés como medio de instrucción (EMI por sus siglas en inglés), y/o en inglés con fines académicos o específicos (EAP, ESP, respectivamente por sus siglas en inglés) (Pérez-Vidal et al., 2018). Esto representa un desafío para hablantes no nativos de la lengua inglesa, especialmente si el estatus del inglés es de ‘idioma extranjero’ y no de ‘segunda lengua’, es decir, no es un idioma oficial en el país. Los estudiantes para los que el inglés es una lengua extranjera (EFL por sus siglas en inglés) no solo tienen que aprender y producir 'inglés general' para poder obtener los certificados B1 o B2 requeridos en la mayoría de los grados universitarios, sino que también necesitan adquirir el ‘discurso especializado’ de sus disciplinas, para poder tener éxito en el ámbito académico.

Diferentes líneas de investigación que analizan las implicaciones de la enseñanza de contenido a través del inglés (por ejemplo, EMI) o la enseñanza del discurso académico en inglés (por ejemplo, EAP) en la educación superior, desde diferentes perspectivas (por ejemplo, desde el punto de vista de los alumnos, de los instructores, o del tipo de materiales utilizados) está ganando terreno. Con el desarrollo de la lingüística de corpus, es decir, el estudio del lenguaje auténtico en forma de textos electrónicos (que pueden provenir de eventos hablados o ya escritos), la forma en que se perciben las lenguas, en términos de gramática, vocabulario, estructuras, funciones y patrones, ha cambiado drásticamente, y ha pasado de basarse en percepciones que eran principalmente intuitivas a centrarse en interpretaciones basadas en la evidencia. Los resultados que emanan de las investigaciones basada en corpus han permitido que la lengua no se contemple ya como una entidad invariable, compuesta de palabras o tipos individuales que forman estructuras gramaticales, sino más bien como un concepto orgánico, que se adapta, cambia y evoluciona según el modo (ej. hablado o escrito), el registro (ej. formal o informal), los géneros (ej. académico, ficción), y las disciplinas (ej. historia, biología). Esto inevitablemente ha cambiado, y sigue cambiando, la forma en que se enseñan y se aprenden las lenguas.

Con el tiempo, se han desarrollado diferentes herramientas y métodos de corpus en un intento por explicar y capturar esta variabilidad del lenguaje. De hecho, los métodos de corpus se utilizan cada vez más en otros campos y con diferentes propósitos (como por ejemplo, en los campos de adquisición del lenguaje, pragmática, sociología, etc.), principalmente porque permiten realizar exploraciones del lenguaje de forma contextualizada. De hecho, una de las principales contribuciones que ha realizado la investigación de corpus en los campos relacionados con el lenguaje es la identificación de combinaciones recurrentes de palabras (también llamado 'formulaicidad', 'fraseología', o 'patrones de lenguaje' [Hunston, 2002; Meunier y Granger, 2008; Wray, 2002]) que actúan como ‘bloques de construcción’ (Biber y Barbieri, 2007) y que son fundamentales para la construcción del mismo. Como se ha demostrado, el aprendizaje y la producción de lenguaje a través de estas ‘fórmulas’ es sin duda más efectivo que el aprendizaje de palabras aisladas, fuera de contexto; estas combinaciones de palabras recurrentes, al mismo tiempo, han demostrado reducir el tiempo de procesamiento y de producción para usuarios e interlocutores. Este hallazgo ha unido dos paradigmas previamente separados, como son el léxico y la gramática, y ha cambiado desde entonces el estudio de la lexicografía y la fraseología. Otra contribución importante de las investigaciones de corpus que a menudo se menciona en la literatura es la confirmación de la variabilidad en los géneros: académico, de ficción, revistas, noticias, etc., todos los géneros tienden a mostrar convenciones lingüísticas particulares; A veces, algunas de estas prácticas lingüísticas son afines entre géneros. Estas características compartidas proporcionan un enfoque ‘general’ de la lengua entre los diversos géneros. Otras, en cambio, son solo características de algunos géneros en particular. Sin embargo, si bien es cierto que se pueden definir prácticas por género (por ejemplo, el uso de "dispositivos de señalización" en el género académico), éstos siguen siendo todavía demasiado amplios para formar un concepto generalizable; por ejemplo, los ensayos, los artículos, las rúbricas de evaluación y los correos electrónicos formales son subgéneros que forman parte del "género académico", pero éstos pueden diferir enormemente en el vocabulario que usan, las estructuras que contienen y los objetivos que persiguen. Numerosos autores han defendido la especificidad del lenguaje y la necesidad de contextualizar las exploraciones lingüísticas, especialmente si se quieren extraer implicaciones pedagógicas de los resultados de investigación.

En el caso particular de la escritura académica, existe un consenso generalizado que defiende la especificidad como elemento clave, tanto para la enseñanza como para el aprendizaje del discurso académico, y también que los métodos de corpus pueden ayudar a proporcionar una imagen más precisa y específica del lenguaje en uso. Desafortunadamente, el grado en que los descubrimientos y las implicaciones pedagógicas que surgen de los estudios de corpus se aplican más tarde en el aula o son utilizados en los materiales de enseñanza es todavía relativamente bajo (Gilquin et al., 2007; Paltridge, 2002; Römer, 2011; Springer, 2012). En la presente tesis doctoral se realizan análisis cuantitativos y cualitativos de la redacción académica producida por estudiantes de EFL en la universidad, utilizando herramientas y metodologías de la lingüística de corpus. El objetivo principal es identificar posibles características de la escritura académica en una segunda lengua (lengua *extranjera* o *segunda* lengua se utilizan indistintamente aquí para referirse a escritores no nativos de la lengua inglesa) para poder sugerir implicaciones pedagógicas que puedan ser útiles para estudiantes e instructores de la escritura académica. Tres tipos de textos diferentes que representan tareas habituales a las que los estudiantes se enfrentan en algún momento durante sus estudios de grado, a saber, una actividad de escritura en el aula, el trabajo final de una asignatura y la tesis final de grado, se han recopilado y convertido en diferentes corpus para su análisis. Así mismo, se ha realizado una exploración de cuatro fenómenos lingüísticos descritos en la literatura como componentes importantes para el desarrollo de la escritura académica: (1) terminología académica general y terminología específica de la disciplina (cf. Coxhead, 2017; Durrant, 2016; Granger, 2017a), (2) conjunciones (adversativas en particular) (cf. Granger y Tyson, 1996; Liu, 2008; Rica-Peromingo, 2012) (3) metadiscurso (reflexivo en particular) (cf. Ädel, 2006, 2016 ; Hyland, 2010; Mauranen, 2010) y (4) paquetes léxicos (cf. Biber et al., 1999; 2004; Biber y Barbieri, 2007; Hyland, 2008a).

El primer fenómeno, la terminología o el vocabulario académico, es un aspecto particularmente importante para la escritura académica. Existen numerosas palabras, colocaciones y sintagmas que pueden clasificarse como "académicas". Algunas de estas se pueden encontrar en todas las disciplinas (ej. *hipótesis*, *sin embargo, resultados preliminares*, etc.); hay otras palabras y expresiones más técnicas, que, por otro lado, solo se pueden encontrar en algunas disciplinas específicas (ej. *realizar una extracción, motivo de consulta, interlengua,* etc.). El conjunto anterior se conoce como vocabulario académico "general", mientras que el último se conoce como vocabulario "específico de la disciplina" o "técnico". La efectividad de enseñar y aprender escritura académica centrándose en uno u otro tipo de vocabulario es todavía una cuestión de debate. Se han realizado numerosos esfuerzos para unificar y describir el vocabulario académico 'general' o 'interdisciplinar' de manera que pueda ser útil para escritores noveles y no nativos que escriben en diferentes áreas disciplinarias. Por ejemplo, recientemente se han desarrollado listas basadas en corpus que contienen palabras, combinaciones de palabras y fórmulas que se encuentran en una amplia gama de géneros académicos (ej., Ackermann y Chen, 2013; Gardner y Davies, 2014; Simpson-Vlach y Ellis, 2010). Por otro lado, algunos estudios han agregado más peso a la especificidad disciplinaria de los géneros académicos y afirman que la presencia de estos ítems "generales" en disciplinas específicas es relativamente baja (Granger, 2017a; Hyland, 2008; Hyland y Tse, 2007); éstos estudios también indican que un enfoque pedagógico basado exclusivamente en terminología general sería, por lo tanto, menos efectivo que la enseñanza y el aprendizaje de la terminología académica más específica, contextualizada en una disciplina particular. Independientemente del enfoque, la producción de vocabulario académico general y específico puede ser un desafío para los estudiantes de EFL que escriben en sus disciplinas académicas. Además, relativamente pocos estudios han analizado el uso real y el desarrollo de la terminología académica (tanto general como específica) en un corpus de estudiantes. En el primer estudio presentado en esta tesis doctoral se analizan ambos tipos de vocabulario académico en un corpus de estudiantes para ver si éstos mejoran su sofisticación léxica y fraseológica después de un curso que proporciona instrucción sobre ambos tipos de vocabulario.

El segundo área de estudio trata sobre el uso de las conjunciones adversativas (ej. *sin embargo, alternativamente, por otra parte*). Estos elementos juegan un papel importante para conseguir coherencia y cohesión en el discurso académico. Sin embargo, a pesar de que la mayoría de los cursos de inglés tratan con estos dispositivos desde etapas muy tempranas, los estudiantes no nativos a menudo tienen dificultades para usarlos adecuadamente. Numerosos estudios que exploran el uso de conectores en la redacción académica han encontrado que, en comparación con escritores nativos o expertos, los estudiantes pueden utilizar estos conectores de manera completamente diferente en términos de ubicación, categoría y frecuencia (Biber et al., 1999, 2004; Granger y Tyson, 1996; Lei, 2012; Rica-Peromingo, 2012; Swales, 2002). De hecho, se ha encontrado que las conjunciones que pertenecen a la categoría adversativa (ej. *a pesar de que, sin embargo, por el contrario*) representan el mayor desafío para los estudiantes; éstos son, además, uno de los tipos más comunes de conectores en la escritura argumentativa. Los ítems en la categoría adversativa pueden poseer diferentes grados de contraste (ej., concesivos *–aún*, correctivos *–más bien*, contrastivos *–de hecho*). Éste y el hecho de que puedan tomar diferentes posiciones dentro de una misma oración también supone un desafío para escritores no experimentados o no nativos. Los textos escritos por estudiantes que no son conscientes de estas particularidades a menudo muestran un uso incorrecto de las conjunciones adversativas. El segundo estudio trata sobre el uso de estas conjunciones en textos argumentativos producidos por alumnos no nativos y desvela ciertas características que merecen atención pedagógica.

El tercer área de estudio es el metadiscurso. El metadiscurso en la escritura académica se refiere a aquellos elementos o marcadores lingüísticos que ayudan a los escritores a referirse a dos entidades principales: (1) el texto que se desarrolla, y (2) el lector y/o el autor del texto. El metadiscurso difiere del contenido ideacional o proposicional de un texto en que no agrega información nueva, pero que es, al mismo tiempo, un componente vital para que se entienda este contenido. Existe una amplia gama de ítems lingüísticos que se pueden calificar como metadiscursivos, pero generalmente se agrupan en dos macro categorías: marcadores que se refieren al texto, es decir, metadiscurso textual (ej., *en la figura 1, en segundo lugar, como se mencionó anteriormente*), y marcadores que se refieren a el escritor o el lector del texto, es decir, metadiscurso interpersonal (ej., *nuestro propósito, vea el apéndice 1, el lector se puede preguntar si*). Los textos que contienen un uso equilibrado de ambos tipos de marcadores de metadiscurso suelen ser más comprensibles y fáciles de leer. Dado que una estructuración clara de la información, un enmarcado cuidadoso de los argumentos y una guía constante del lector son prácticas comunes en textos académicos escritos en inglés, el uso y la comprensión de estos marcadores es de gran importancia, especialmente cuando se escriben textos académicos largos (ej. trabajos de final de grado). Una dificultad añadida es el hecho de que estas prácticas metadiscursivas suelen ser altamente específicas, lo que significa que los tipos y la medida en la que aparecen están determinadas por la disciplina y por el género específico del texto (Hyland, 2000, 2005, 2012). Los estudiantes de EFL no siempre son conscientes de estas particularidades y, a menudo, no utilizan los dispositivos metadiscursivos de manera adecuada. El tercer estudio explora el metadiscurso reflexivo en textos académicos escritos por estudiantes de EFL en su último año de carrera en dos disciplinas diferentes (medicina y lingüística), y lo compara con el uso del metadiscurso en un corpus de expertos compuesto por artículos de investigación en las mismas disciplinas. Las diferencias que se reportan entre géneros y disciplinas pueden ser de utilidad pedagógica tanto para escritores no nativos como para instructores de inglés académico.

Los paquetes léxicos son el cuarto dispositivo lingüístico explorado. Un paquete léxico es una combinación de palabras recurrentes que puede tener diferentes longitudes (ej. de tres, cuatro o cinco palabras), diferentes estructuras (ej. parte de un sintagma nominal, verbal, preposicional, etc.) y realizar diferentes funciones (ej. *por otro lado* es un paquete léxico de tres palabras que realiza una función de 'transición' en el texto): su cantidad y diversidad abundan en el lenguaje. Los paquetes léxicos se han analizado en la literatura desde diferentes perspectivas en términos de disciplinas, modos y registros (Ädel y Erman, 2012; Biber y Barbieri, 2007). Diversas investigaciones han demostrado que, desafortunadamente, no parece haber un conjunto único de paquetes léxicos que se puedan emplear de forma general, sino todo lo contrario: cada modo, registro y disciplina tiende a usar, con mayor o menos frecuencia, un grupo de paquetes léxicos para sus propósitos particulares. Si bien es cierto que algunos de estos paquetes se pueden encontrar en todos los modos, registros y disciplinas, otros paquetes son más específicos y, para demostrar pertenencia a una comunidad de expertos determinada, hay que estar seguros de cuáles, cómo y cuándo usar estos paquetes. Los aprendices de inglés como lengua extranjera generalmente no adquieren estos paquetes léxicos a través de la mera exposición y se requiere de una enseñanza más explícita. Esto hace que habitualmente se encuentren problemas como la infrautilización, el uso excesivo y/o el uso indebido de estos ítems cuando se comparan con textos escritos por nativos o por autores expertos (Ädel y Erman 2012; Chen y Baker, 2010; Liu, 2012; Meunier y Granger, 2008). El cuarto estudio se centra en el uso de estos paquetes en la escritura académica de alumnos y de expertos y describe los paquetes más frecuentes encontrados en cada subcorpus, los que tienen en común y las estructuras y las funciones que caracterizan cada tipo de escritura.

En la presente tesis doctoral, estas cuatro áreas problemáticas han sido analizadas en varios textos académicos escritos por estudiantes en la universidad. Estos alumnos son escritores de EFL, con diferentes L1 (principalmente europeas) en los dos primeros estudios; en los dos últimos estudios, por otro lado, los estudiantes fueron específicamente escritores con español como L1. Esta exploración de corpus que provienen de estudiantes con diferentes perfiles lingüísticos ha ayudado a la autora a analizar las características de los estudiantes desde perspectivas diferentes, es decir, la escritura académica producida en un contexto de aula internacional por un lado, y la escritura académica producida por una población con una L1 específica por el otro. Además, los participantes representan estudiantes universitarios en su primer año de estudios (en los estudios uno y dos), y en su último año de estudios (en los estudios tres y cuatro). Dado que la complejidad de la escritura aumenta de un texto a otro en los cuatro estudios realizados (por ejemplo, de una breve tarea de escritura realizada en el aula, a un trabajo de final de grado), las posibilidades de analizar fenómenos lingüísticos más complejos aumentaron en consecuencia (ej. en el estudio uno se analiza el vocabulario académico mientras que en el estudio tres se estudian las diferentes expresiones metadiscursivas utilizadas en los textos). Además, los textos utilizados se produjeron en cuatro grados diferentes, a saber, odontología, estudios europeos, medicina y lingüística. Trabajar con textos de diferentes longitudes, disciplinas y propósitos ha llevado a la autora a explorar diferentes paradigmas lingüísticos, y a utilizar diferentes métodos de corpus: los estudios incluidos en la presente tesis reflejan metodologías basadas en el corpus (ej. el uso de listas predefinidas, uso de taxonomías validadas) y metodologías impulsadas por el corpus (i.e. exploración de elementos realmente presentes en los textos, sin ideas preconcebidas). Finalmente, análisis cuantitativos (ej. comparaciones de frecuencia de uso) y cualitativos (ej. clasificación según estructura gramatical y funcional) de los diferentes corpus han sido realizados en todos los estudios con la intención de capturar de forma más contextualizada el uso de los diferentes fenómenos lingüísticos estudiados.

Como hemos visto, la literatura aboga por la necesidad de enseñar y aprender la escritura académica desde perspectivas basadas en corpus y contextualizadas en disciplinas específicas. Por lo tanto, el objetivo de esta tesis es doble: explorar textos académicos escritos en inglés por estudiantes no nativos con respecto al uso de cuatro fenómenos lingüísticos para ver cómo los utilizan y poder identificar prácticas que caracterizan a este tipo de escritura y, en segundo lugar y tras considerar la necesidad de investigar de forma empírica la escritura académica en una segunda lengua y remarcar posibles implicaciones pedagógicas, esta tesis tiene como objetivo proporcionar asesoramiento pedagógico sobre el uso de estos dispositivos, mediante la utilización de diferentes corpus de referencia. En este sentido, los corpus de referencia utilizados en los estudios provienen de corpus generales o académicos de grandes dimensiones y ampliamente conocidos, así como también de corpus auto compilados y más especializados. Estos textos representan escritores nativos (textos de estudiantes universitarios) o expertos (artículos de investigación publicados). Tras realizar comparaciones entre los corpus de aprendices y los de referencia se extraen implicaciones pedagógicas de cada estudio, con el objetivo de ayudar a escritores no nativos a mejorar sus habilidades de escritura académica en relación con el uso del vocabulario académico, las conjunciones adversativas, el uso de metadiscurso y los paquetes léxicos. Se espera que los hallazgos que emergen de los estudios presentados en esta tesis puedan ser de interés para instructores de escritura académica y diseñadores de material pedagógico, ya que proporcionan un análisis de la producción de la escritura académica realizada por estudiantes en la universidad que está contextualizado en disciplinas concretas.

Objetivos y preguntas de investigación

La pregunta general que guía los cuatro estudios llevados a cabo para la presente tesis es: ¿Cómo pueden los métodos de lingüística de corpus contribuir a identificar las características de escritura académica producida en inglés como lengua extranjera por estudiantes universitarios? Esta pregunta general está motivada por la necesidad de encontrar patrones de uso que caracterizan la escritura de estudiantes y que difieren en textos comparables escritos por nativos o expertos. El objetivo principal de esta tesis doctoral es, por lo tanto, contribuir al campo de investigación que estudia la escritura en una segunda lengua y la alfabetización disciplinaria a través de la lingüística de corpus, y a la vez, servir como un recurso pedagógico útil tanto para instructores de L2 como para alumnos que necesitan desarrollar sus habilidades de escritura académica en inglés en la universidad.

En el primer estudio se analiza el uso del vocabulario académico general y técnico antes y después de un curso de lengua basado en contenido (CBI por sus siglas en inglés). La principal pregunta de investigación que guía este estudio es: ¿tuvo algún efecto el curso de CBI en la producción de vocabulario académico de los estudiantes? La hipótesis es que habría una mayor producción de vocabulario académico en los textos escritos después del curso (T2) como un efecto positivo resultante de la instrucción recibida. Además, se espera una mayor producción de vocabulario académico por parte de uno de los grupos (que estudian en la modalidad EMI), en comparación con el otro grupo (en la modalidad L1), debido posiblemente a una mayor exposición a la lengua inglesa en un contexto académico.

En el segundo estudio se explora el uso de las conjunciones adverbiales de tipo adversativo en textos argumentativos escritos por estudiantes de EFL con diferentes L1s, y se compara con textos argumentativos escritos por estudiantes cuya L1 es el inglés. Hay dos preguntas principales de investigación: 1) ¿Cómo utilizan los escritores no nativos las conjunciones adversativas en términos de frecuencia, ubicación y categorías en comparación con escritores nativos? Siguiendo estudios previos sobre el uso de los conectores, se espera que los estudiantes no nativos utilicen más conectores que sus homólogos nativos. Esto se debe al hecho de que estos conectores se enseñan con frecuencia en los cursos de inglés desde niveles tempranos y de que además estos conectores se suelen presentan en largas listas (a menudo descontextualizadas), y se premia con calificaciones más altas a los estudiantes que los usan en sus textos (Granger y Tyson, 1996; Granger, 2004; Lei, 2012; Rica-Peromingo, 2012; Wray, 2002). La segunda pregunta formulada para este estudio es: 2) ¿Cómo utilizan las conjunciones adverbiales estudiantes de dos familias lingüísticas diferentes cuando se comparan entre sí? Se espera que el grupo de estudiantes con L1s romances (es decir, francés, italiano y español) y el grupo con L1s germánicas (es decir, holandés y alemán) muestren frecuencias y patrones de uso similares cuando se comparan entre sí en sus grupos, debido a una posible influencia de su lengua materna.

El tercer estudio investiga la presencia del metadiscurso reflexivo (MD) en trabajos finales de grado (TFGs) escritos en inglés por los estudiantes de L1 español, y lo compara con el uso de estos dispositivos por escritores expertos de artículos de investigación (AIs) en las mismas disciplinas (lingüística y medicina). Las principales preguntas formuladas para este estudio son: 1) ¿Cómo usan los marcadores de MD en inglés académico los estudiantes de español en comparación con un corpus de expertos? Y 2) ¿Hay alguna diferencia en el uso de MD entre disciplinas? Se estudian la frecuencia general y dos categorías principales de marcadores de MD (es decir, textuales e interpersonales), para luego analizar la variación interdisciplinaria (lingüística vs. medicina) y variación según el perfil del escritor (aprendiz vs. experto). La hipótesis que se plantea es que los estudiantes usarán algunos marcadores de MD con mayor frecuencia (ej. marcadores textuales para proporcionar la estructura de sus textos) y con menor frecuencia (ej. marcadores interpersonales para dirigirse a los lectores) que los expertos, y que esto posiblemente denote características de la escritura del alumno (como transferencia de prácticas que vienen de su L1) o de un género en particular (TFGs); también se espera que algunos marcadores de MD estén presentes solo en textos de lingüística o solo en textos de medicina, como una posible consecuencia de la variación interdisciplinaria.

Finalmente, el cuarto estudio analiza el uso de paquetes léxicos (LBs por sus siglas en inglés) en TFGs y AIs. Las preguntas principales que guían este estudio son: 1) ¿Cómo usan los LBs los estudiantes con español como L1 en las secciones de introducción y conclusión cuando escriben en inglés académico, en comparación con un corpus de expertos? Y 2) ¿Cómo se usan estos LBs en términos de estructura gramatical y función retórica? Se espera que los estudiantes usen LBs similares en sus TFGs, independientemente de su disciplina, particularmente debido a las rigurosas convenciones de género (es decir, las directrices y requisitos específicos de los TFGs), y también a la estructura canónica y los objetivos retóricos de las secciones de introducción y conclusión en este género académico. También se espera que el uso de LBs difiera de los AIs publicados en la misma disciplina debido, principalmente a la experiencia de los autores. La última hipótesis es que el uso de LBs en textos de lingüística y de medicina también difiera en términos de frecuencia, estructuras y funciones, posiblemente debido a diferentes convenciones disciplinarias.

Discusión y conclusión

La presente tesis doctoral ha tenido por objetivo investigar cómo diferentes enfoques de corpus pueden servir para identificar las características de la escritura académica de aprendices en una segunda lengua. Con esta intención, se han realizaron cuatro estudios. Primero, se ha analizado el uso de la terminología general y específica de la disciplina en una tarea de escritura en inglés realizada por estudiantes de primer año con diferentes L1s antes y después de un curso de CBI. En el segundo estudio se ha examinado el uso de conjunciones adversativas en textos argumentativos de mayor longitud escritos también por estudiantes de primer año y con diferentes L1s, y se ha comparado con el uso de estos dispositivos por estudiantes nativos de inglés. En el tercer artículo se analizan las prácticas metadiscursivas en textos aún más largos, es decir, en trabajos de final de grado producidos por los estudiantes de último año de medicina y de lingüística con español como L1; éstas se han comparado con un corpus de expertos que consistía en artículos de investigación académicos publicados en las mismas disciplinas. Finalmente, en el cuarto estudio, se han extraído los paquetes léxicos presentes en las secciones de introducción y conclusión de estos textos (TFGs y AIs), para, tras analizar su frecuencia y clasificar su estructura y su función, comparar su uso en los textos escritos en inglés por estudiantes y en artículos escritos por expertos.

Los diferentes enfoques de corpus que se han empleado han servido para identificar características importantes de la escritura de los alumnos a partir de las cuales se extraen algunas implicaciones pedagógicas: en primer lugar, se encontró que los alumnos producían más vocabulario académico (tanto general como específico de la disciplina) después de un curso de CBI, lo que muestra un posible beneficio a corto plazo (un semestre) de este tipo de instrucción en la universidad. Sin embargo, la instrucción CBI parece tener poco o ningún efecto sobre el uso de fórmulas académicas y colocaciones, las cuales permanecieron iguales tras el curso. Este resultado hace hincapié en la necesidad de prestar una mayor atención pedagógica a las fórmulas y colocaciones en la escritura académica. El análisis de los textos según la modalidad de estudios (es decir, EMI o L1) también mostró que el uso de la terminología académica por parte del grupo EMI después del curso fue significativamente mayor, en comparación con el grupo L1, los cuales no mostraron tal mejora. Una mayor exposición a la lengua inglesa en un contexto académico que experimentó el grupo EMI podría explicar las diferencias encontradas.

En segundo lugar, se encontró que el uso de conjunciones adversativas en textos argumentativos escritos por escritores nativos y no nativos era comparable en términos de frecuencia. Sin embargo, un análisis más cualitativo de los tipos y categorías de estos ítems mostró que los escritores no siempre estuvieron de acuerdo con la elección de las conjunciones, especialmente en relación con la ubicación de estos elementos dentro de oraciones y párrafos. Diferentes casos que mostraban un uso excesivo (ej. *sin embargo*) infrautilización (ej. *en realidad*) y uso indebido (ej. *en contraste*) fueron encontrados y explorados en detalle. El hecho de que el corpus de referencia pertenecía a estudiantes universitarios y de que estos textos no siempre modelan buenas prácticas de redacción, planteó la necesidad de realizar comparaciones adicionales con un corpus expertos (como el subcorpus académico incluido en BNC y COCA) para comparar frecuencias y usos de ciertos ítems que parecían problemáticos de forma que se pudiera determinar de forma más adecuada si los casos de sobreuso e infrautilización requerían atención pedagógica. Además, la exploración del uso de conjunciones adversativas de acuerdo a dos familias lingüísticas diferentes (lenguas romance y lenguas germánicas) también produjo resultados interesantes: todos los subcorpus compartían preferencias generales con respecto a las categorías utilizadas y la ubicación de conjunciones adversativas en las oraciones. El hecho de que estos estudiantes, a pesar de tener diferentes L1, recibieron el mismo tipo de instrucción y revisaron los textos en clase podría haber aliviado cualquier posible diferencia. Las prácticas de revisión y la metodología basada en proyectos aplicada en el curso al que estuvieron expuestos parecen haber tenido un efecto positivo en la reducción de posibles problemas de transferencia de L1 en cuanto al uso de conjunciones adversativas en textos académicos.

En el tercer estudio se analizaron los marcadores textuales e interpersonales de metadiscurso reflexivo en TFGs escritos por alumnos con español como L1, y AIs escritos por expertos. Este análisis reveló diferencias no solo por parte de los dos tipos de escritores (aprendiz frente a experto) y/o género (ej. como la infrautilización de marcadores interpersonales en los TFGs en lingüística o el predominio de marcadores aditivos en todos los TFGs en general); estas diferencias también fueron indicativas de diferentes convenciones disciplinarias, como fue el predominio general de la marcadores textuales en textos de lingüística, en comparación con los textos médicos. Estos resultados demuestran que la enseñanza y aprendizaje de marcadores metadiscursivos requieren, por lo tanto, de una contextualización en la disciplina estudiada.

Finalmente, en el cuarto artículo se realizó la extracción de paquetes léxicos (LBs) de diferentes longitudes en las secciones de introducción y conclusión de estos textos (TFGs y AIs); los LBs obtenidos se clasificaron manualmente según su estructura gramatical y sus funciones retóricas principales. Se identificaron preferencias que podrían denotar la inmadurez de los escritores (ej. menor diversidad de estructuras y funciones de los LBs utilizados por los escritores de TFGs), y que también destacan las prácticas de diferentes comunidades académicas (ej. la alta frecuencia de LBs con funciones de ‘describir procesos de investigación’ en textos de medicina frente al predominio de LBs orientados a ‘describir procesos textuales’ en los textos de lingüística). Los hallazgos presentados en estudio doctoral refuerzan la utilidad de los métodos cuantitativos y cualitativos de corpus aplicados a textos escritos en L2 y la comparación de los resultados con un corpus de expertos para identificar ciertas características de la escritura académica en L2 que requieren atención pedagógica.

Se espera que los hallazgos obtenidos sobre el discurso académico y los enfoques de corpus aplicados a la escritura de estudiantes que se han presentado en esta tesis doctoral puedan ser útiles para futuros estudiantes de escritura académica, así como también para instructores y diseñadores de materiales pedagógicos, contribuyendo así a los campos que investigan la escritura en una segunda lengua y la lingüística de corpus. Para acabar, esta tesis refuerza la necesidad y el beneficio de usar materiales basados en corpus en el aula y apoya la instrucción de la escritura académica desde una perspectiva de género y de disciplina.