

**UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS**

LUCIANA DIAS DE MACEDO

**LEXICAL BUNDLES ACROSS SECTIONS
OF APPLIED LINGUISTICS RESEARCH ARTICLES**

**BELO HORIZONTE
FACULDADE DE LETRAS DA UFMG
2018**

Luciana Dias de Macedo

**LEXICAL BUNDLES ACROSS SECTIONS
OF APPLIED LINGUISTICS RESEARCH ARTICLES**

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Faculdade de Letras da Universidade Federal de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Linguística Aplicada.

Área de Concentração: Linguística Aplicada
Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras (3A)

Orientadora: Profa. Dra. Deise Prina Dutra

Belo Horizonte
Faculdade de Letras da UFMG
2018

Ficha catalográfica elaborada pelos Bibliotecários da Biblioteca FALE/UFMG

M1411 Macedo, Luciana Dias de.
Lexical bundles across sections of applied linguistics research articles [manuscrito] / Luciana Dias de Macedo. – 2018.
102 f., enc.: il. grafs (p&b)
Orientadora: Deise Prina Dutra.
Área de concentração: Linguística Aplicada.
Linha de Pesquisa: Ensino/Aprendizagem de Línguas Estrangeiras.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.
Bibliografia: f. 92-96.
Apêndices: f. 97-103.

1. Língua inglesa – Estudo e ensino – Falantes estrangeiros – Teses. 2. Aquisição de segunda linguagem – Teses. 3. Linguística de corpus – Teses. 4. Redação acadêmica – Teses. 5. Língua inglesa – Lexicologia – Teses. I. Dutra, Deise Prina. I. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD: 420.7



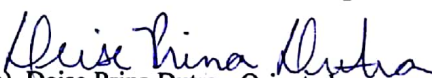
FOLHA DE APROVAÇÃO


LEXICAL BUNDLES ACROSS SECTIONS OF APPLIED LINGUISTICS RESEARCH ARTICLES

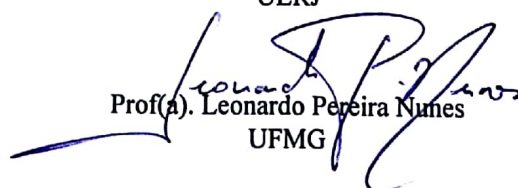
LUCIANA DIAS DE MACEDO

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGÜÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGÜÍSTICOS, área de concentração LINGÜÍSTICA APLICADA, linha de pesquisa Ensino/Aprendizagem de Línguas Estrangeiras.

Aprovada em 23 de fevereiro de 2018, pela banca constituída pelos membros:


Prof(a). Deise Prina Dutra - Orientadora
UFMG


Prof(a). Patricia Pereira Bértoli
UERJ


Prof(a). Leonardo Pereira Nunes
UFMG

Belo Horizonte, 23 de fevereiro de 2018.

To my parents, sisters and Guilherme.

AGRADECIMENTOS

Primeiramente, gostaria de expressar minha profunda gratidão aos meus queridos pais, principalmente à minha mãe, que sempre se dispôs a nos ajudar, a nos guiar para que pudéssemos ter oportunidades melhores, apesar de todas as dificuldades. Reconheço que nossas conquistas são como se fossem as suas, e isto certamente sempre nos motivou a crescer.

Gostaria de agradecer às minhas irmãs, Fernanda e Fabiana, especialmente à Fa, por ter se disposto a me ajudar, sem pestanejar, com a compilação do corpus para esta pesquisa. Agradeço às minhas amigas, especialmente à Karol, que me auxiliou com a revisão do texto.

Agradeço também aos meus amigos e colegas do Poslin, em especial, à Janaina, Danilo, Barbara e Bruna que sempre me ajudaram com dicas preciosas e discussões sempre interessantes e frutíferas. Aos meus alunos e coordenadores da Rizvi, pelo apoio e cooperação quando eu tive que me ausentar para participar de eventos acadêmicos. Muito obrigada!

Sou imensamente grata à minha orientadora, Professora Deise, que me guiou e ajudou a trilhar esse caminho novo e desafiador que é trabalhar com Linguística de Corpus. Muito obrigada por seu profissionalismo, competência, flexibilidade, e principalmente por sempre nos motivar a ser pesquisadores melhores.

Finalmente, serei eternamente grata ao meu querido Guilherme, que sempre tenta me animar nos momentos difíceis e celebra comigo as pequenas conquistas. Você é minha inspiração e minha força. Te amo.

ACKNOWLEDGMENTS

Firstly, I would like to express my deepest gratitude to my dear parents, especially to my mother, who was always there to help and guide us so we could have better opportunities despite the adversities. I know she always recognizes our achievements as her own, and this certainly has greatly encouraged us to keep on evolving.

I would like to thank my dear sisters, Fernanda and Fabiana, especially Fa for having promptly offered and supported me with the compilation of the corpus for this study. I'm thankful to my friends, especially Karol, who helped me with the text revision.

I would also like to thank my friends and colleagues from Poslin, particularly Janaina, Danilo, Barbara, and Bruna who were always there to help with precious tips, insightful and fruitful discussions. I really appreciate all the encouragement I got from my students and also Rizvi's and its coordinators' support and cooperation at the times when I had to be absent from work in order to take part in academic events. Thank you!

I am really thankful to my advisor, Professor Deise, who has guided me throughout this new and challenging path of Corpus Linguistics. Thank you for being professional, competent, flexible, and mostly for encouraging us to become better researchers.

Finally, I am eternally grateful to my dearest Guilherme, who has cheered me up in every moment of discouragement and who has celebrated all the small achievements with me. You are my inspiration and strength. I love you.

*“The most damaging phrase in the language is:
‘It’s always been done that way.’”
(Grace Hopper)*

ABSTRACT

The increasing use of computer-held text corpora for the analysis of lexicogrammatical patterns, once unavailable to observers, has allowed researchers to get a better and more precise understanding of language. Corpus Linguistics has also shown to be capable of revealing valuable and detailed features of expressions that play a crucial role in discourse with linguistic investigations that comprehend not only single words but also idiomatic or non-idiomatic expressions in a given context. The aim of the current research is to investigate 4-7 word lexical bundles in sections of Applied Linguistics Research Articles, published in English. Lexical bundles, fundamentally defined by frequency (BIBER et al., 1999; CORTES, 2013), are generated with no pre-defined linguistic categories. Despite the frequency with which they occur, lexical bundles are “not idiomatic in meaning and not perceptually salient” (BIBER; BARBIERI, 2007, p. 269). Their use, however, has been thoroughly investigated due to the role those devices play as building blocks of discourse. In order to offer details of in-text linguistic variation, a corpus was compiled from 180 articles and split into four subcorpora: Introduction, Methods, Results, and Discussion (IMRD), with more than 1 million words in total. The lexical bundles were individually classified into structural and functional categories, including their subtypes, according to previous and established taxonomies (BIBER et al., 1999, 2004; CORTES, 2013; SIMPSON-VLACH; ELLIS, 2010). Two different statistical treatments, confidence intervals and the null-hypothesis significance z-test, were adopted as complementary in order to check whether differences across subcorpora were significant. Results show that sections of Applied Linguistics research articles should be treated as separate texts for they display strong distinctions, and grammatical structures may play singular functional roles. The main distinctions between sections entail frequency of devices and a preference for particular lexicogrammatical elements in relation to the pragmatic nature of the section, e.g. fragments with exclusive roles can be related to passive and non-passive voice forms containing specific verbs. Additionally, the proportion of some structures and functional categories that emerged in the current research challenge generalizations made in previous studies about academic register. Finally, we argue for more research on sub-registers with representative and specialized corpora so as to give more insight into academic writing of research articles sections.

Key words: lexical bundles, corpus linguistics, academic writing, research article sections

RESUMO

O crescente uso de corpora textuais em computadores para a análise de padrões léxico-gramaticais, antes indisponível para observadores, tem permitido aos pesquisadores uma compreensão melhor e mais precisa sobre aspectos linguísticos. A Linguística de Corpus tem demonstrado capacidade de revelar características valiosas e detalhadas de expressões que desempenham papel crucial no discurso, por meio de pesquisas linguísticas que abrangem desde palavras soltas a expressões idiomáticas e não-idiomáticas em um dado contexto. O objetivo da atual pesquisa é investigar pacotes lexicais com sequências de 4-7 palavras em seções de artigos acadêmicos pertencentes à Linguística Aplicada, publicados em inglês, e entender seu papel pragmático dentro de cada seção. Pacotes lexicais, fundamentalmente definidos pela frequência (BIBER et al., 1999; CORTES, 2013) são gerados sem quaisquer categorias linguísticas pré-definidas. Apesar de sua frequência, os pacotes lexicais não são considerados idiomáticos em definição ou perceptualmente salientes (BIBER; BARBIERI, 2007, p. 269). Seu uso, no entanto, tem sido amplamente investigado devido ao papel que esses dispositivos desempenham como blocos construtores de discurso. Com o intuito de oferecer mais detalhes da variação linguística interna de artigos, um corpus foi compilado de 180 artigos e dividido em quatro subcorpora: Introdução, Métodos, Resultados e Discussão (IMRD), com mais de 1 milhão de palavras no total. Os pacotes lexicais foram classificados individualmente em categorias estruturais e funcionais, incluindo seus subtipos, de acordo com taxonomias existentes e estabelecidas na literatura (BIBER et al., 1999, 2004; CORTES, 2013; SIMPSON-VLACH; ELLIS, 2010). Dois tratamentos estatísticos, intervalos de confiança e z-teste de hipótese nula, foram empregados como complementares para checar se as diferenças entre os subcorpora são relevantes. Os resultados mostram que seções de artigos da Linguística Aplicada deveriam ser abordadas como textos distintos dadas as grandes diferenças identificadas, assim como estruturas gramaticais podem desempenhar papéis funcionais singulares. As principais distinções entre as seções envolvem a frequência de pacotes lexicais e uma preferência por elementos lexicogramaticais particulares em relação à natureza pragmática da seção, por exemplo, fragmentos com papéis exclusivos podem ser relacionados às formas em voz passiva ou não-passiva contendo verbos específicos. Além disso, a proporção de algumas estruturas e categorias funcionais que emergiram na atual pesquisa desafiam generalizações feitas em estudos anteriores sobre o registro acadêmico. Finalmente, sugerimos mais pesquisas de sub-registros com corpora representativos e especializados para que haja mais esclarecimentos em relação à escrita acadêmica de seções de artigos científicos.

Palavras-chave: pacotes lexicais, linguística de corpus, escrita acadêmica, seções de artigos acadêmicos

LIST OF ABBREVIATIONS

AFL	Academic Formulas List
AL	Applied Linguistics
CL	Corpus Linguistics
DA	Discourse Analysis
EAP	English for Academic Purposes
ESP	English for Specific Purposes
IMRD	Introduction, Methods, Results, and Discussion
MI	Mutual information
RA	Research Articles
TESOL	Teaching English to Speakers of Other Languages

LIST OF GRAPHS

Graph 1: Bundle types and bundle tokens across the IMRD subcorpora	45
Graph 2: Normalized frequency and confidence intervals of “PP fragments”	48
Graph 3: Normalized frequency and confidence intervals of noun phrases with of-phrase fragments	50
Graph 4: Normalized frequency and confidence intervals of non-passive voice lexical bundles	52
Graph 5: Normalized frequency and confidence intervals of lexical bundles containing verb phrases with passive voice structures	54
Graph 6: Normalized frequency and confidence intervals of “ <i>anticipatory it + adjective</i> ” lexical bundles	57
Graph 7: Normalized frequency and confidence intervals of specification of attributes: Intangible framing attributes (Referential expressions)	62
Graph 8: Normalized frequency and confidence intervals of specification of attributes: quantity specification (Referential expressions)	64
Graph 9: Normalized frequency and confidence intervals of contrast and comparison (Referential expressions)	66
Graph 10: Normalized frequency and confidence intervals of deictics and locatives (Referential expressions)	67
Graph 11: Normalized frequency and confidence intervals of metadiscourse textual reference (Discourse organizing functions)	70
Graph 12: Normalized frequency and confidence intervals of topic elaboration: cause and effect (Discourse organizing functions)	73
Graph 13: Normalized frequency and confidence intervals of discourse markers (Discourse organizing functions)	75
Graph 14: Normalized frequency and confidence intervals of evaluation devices (Stance expressions)	77
Graph 15: Normalized frequency and confidence intervals of ability and possibility devices (Stance expressions)	79
Graph 16: Normalized frequency and confidence intervals of hedging devices (Stance expressions)	80

LIST OF TABLES

Table 1: Structural types of lexical bundles adapted from Biber et al. (1999, 2004) and Cortes (2013)	24
Table 2: Functional categories and subtypes of lexical bundles adapted from Biber et al. (1999, 2004) and Simpson-Vlach & Ellis (2010)	26
Table 3: List of high-impact journals	33
Table 4: IMRD Corpus size.....	34
Table 5: Raw frequency of lexical bundles generated from each IMRD subcorpus by AntConc and corpus size	45
Table 6: Major categories: structural types of lexical bundles	48
Table 7: Lexical bundles subtypes that incorporate noun phrase or prepositional phrase fragments.....	48
Table 8: Prepositional phrase fragments occurrences	49
Table 9: Noun phrases with of-fragment occurrences	52
Table 10: Lexical bundles subtypes that incorporate verb phrase fragments..	52
Table 11: Lexical bundles containing non-passive structures occurrences	54
Table 12: Lexical bundles containing passive structures occurrences	56
Table 13: Lexical bundles subtypes that incorporate dependent clause fragments..	57
Table 14: Anticipatory it + adjective fragments occurrences	58
Table 15: Major categories: functional types of lexical bundles (percentage)	59
Table 16: Lexical bundles subtypes: Referential expressions (frequency per 1000 words)....	62
Table 17: Intangible framing attributes occurrences	64
Table 18: Quantity specification occurrences	65
Table 19: Contrast and comparison occurrences	67
Table 20: Deictics and locatives occurrences	68
Table 21: Lexical bundles subtypes: Discourse organizing functions	70
Table 22: Metadiscourse and textual reference occurrences	72
Table 23: Topic elaboration: cause and effect occurrences	74
Table 24: Lexical bundles subtypes: Stance expressions	77
Table 25: Evaluation occurrences	78

Table 26: Hedges occurrences	81
Table 27: Lexical bundles subtypes: Applied Linguistics RA subtypes.....	82
Table 28: Structural categories and subtypes of lexical bundles and a pairwise comparison analysis of subcorpora (p-value).....	84
Table 29: Functional categories and subtypes of lexical bundles and a pairwise comparison analysis of subcorpora (p-value).....	85

LIST OF FIGURES

FIGURE 1: Lexical bundles from the Methods subcorpus generated by AntConc 3.4.4	38
FIGURE 2: Lexical bundles from the Methods subcorpus displayed in a Google spreadsheet for classification	41
FIGURE 3: Concordance lines of the lexical bundle “the end of the”	41

TABLE OF CONTENTS

1.INTRODUCTION.....	17
1.1. The motivations for this research	17
1.2. Research aims	18
1.2.1 Research goals	18
1.2.2. Research objectives	18
1.3. The research questions	19
1.4. Relevance of this study	19
1.5. The following chapters	19
2. LITERATURE REVIEW.....	21
2.1 Lexical Bundles	22
2.1.1 Structural types	23
2.1.2 Functional taxonomy	25
2.1.2.1 Referential expressions	26
2.1.2.2 Discourse organizing functions	28
2.1.2.3 Stance expressions	29
2.2 Research article sections	30
2.2.1 Research article sections: lexical bundles	31
3. METHODOLOGY.....	33
3.1 The corpus.....	34
3.2 The procedure for data analysis.....	35
3.2.1 Generating the lexical bundles	36
3.2.2 Structural and functional types of lexical bundles.....	38
3.2.3 Statistical treatment.....	41
4. RESULTS AND DISCUSSION.....	44
4.1 Lexical bundle types across the IMRD subcorpora.....	46
4.1.1 Structural types of lexical bundles.....	46
4.1.1.1 Prepositional phrase, noun phrase fragments, and other nouns.....	47
4.1.1.1.1 Prepositional phrase fragments.....	47
4.1.1.1.2 Noun phrases with of-phrase fragments.....	49
4.1.1.2 Verb phrase fragments.....	51
4.1.1.2.1 Non-passive voice lexical bundles.....	52
4.1.1.2.2 Passive voice lexical bundles.....	53
4.1.1.3 Dependent clause fragments.....	56
4.1.1.3.1 Anticipatory it + adjective fragments.....	57
4.1.2 Functional types of lexical bundles.....	59
4.1.2.1 Referential expressions.....	60
4.1.2.1.1 Specification of attributes: Intangible framing attributes.....	61

4.1.2.1.2 Specification of attributes: quantity specification.....	63
4.1.2.1.3 Contrast and comparison.....	65
4.1.2.1.4 Deictics and locatives.....	66
4.1.2.1.5 Identification and focus/ Tangible framing attributes.....	68
4.1.2.2 Discourse organizing function.....	69
4.1.2.2.1 Metadiscourse and textual reference.....	70
4.1.2.2.2 Topic elaboration: cause and effect.....	73
4.1.2.2.3 Discourse markers.....	74
4.1.2.3 Stance expressions.....	75
4.1.2.3.1 Evaluation.....	76
4.1.2.3.2 Ability and possibility.....	78
4.1.2.3.3 Hedges.....	79
4.1.2.4 Applied Linguistics RAs special devices.....	81
4.2 Secondary statistical treatment, the null-hypothesis test.....	82
5. CONCLUSION.....	85
REFERENCE.....	92
APPENDIX A.....	97
APPENDIX B.....	99
APPENDIX C.....	102
APPENDIX D.....	103

1. INTRODUCTION

1.1. The motivations for this research

Among various reasons for having a Research Article (RA) turned down by an international journal, one that is consistently distressing is the non-compliance with editors' expectations of language and style, which include levels of consistency and quality of texts. These are usually unwritten rules which are mostly related to the register inherent to a certain discourse community. Such rules, though often mastered by members of the community, are likely to cause frustration to novice writers and/or native or non-native speakers of that language. Hence, fully understanding the discursive features of scientific papers from specific fields of study may open the way for more international publication of interesting pieces of research that are curbed by the unenlightenment of language specifications in the academic community.

I believe that offering useful and applicable information to those who want to be part of an international academic community is paramount. Researchers who focus on the understanding of academic texts may side with what is called genre-related analysis (SWALES, 1990), which is called *register* studies by more recent researchers (BIBER et al, 2004, 2009, and so on; HYLAND, 2007, 2008, 2012, etc). Despite referring to the same topic, *register* has been more commonly used due to the myriad of definitions, sometimes blurred, given to *genre* in the literature.

Having in mind the constraints that an understudied register might impose, this study aims at bringing more validity to the findings by providing both quantitative and qualitative analyses rather than merely relying on intuition about the key features of RA sections. This view is shared by a number of contemporary linguists, among whom Sampson (2002), who maintains that data that come from intuition are “hopelessly unreliable” (p.2). Combining a strong foundation based on accepted and peer-reviewed RAs, written in English, from high-impact journals and statistical significance tests, this study may assist not only those who are willing to break the barriers of publication in international journals, but also enthusiasts who are keen on improving their academic writing skills and style. Additionally, this might also provide language instructors with more insights into academic writing features.

In order to do so, our objective is to investigate lexical bundles within each of the major sections of RAs, namely Introduction, Methods, Results, and Discussion (IMRD).

Lexical bundles are important building blocks of written and spoken academic discourse, hence the appropriate use of these structures may offer texts more readability and a sense of compatibility with other texts belonging to the same register. Although considerable research has been done on the use of lexical bundles in academic discourse (BIBER, 2010; BIBER et al., 1999, 2004; BIBER; BARBIERI, 2007; BYRD; COXHEAD, 2010; CORTES, 2008, 2004, 2013; HYLAND, 2008; SIMPSON-VLACH; ELLIS, 2010, to name a few), very little is known as to how these structures are employed in each section of an RA.

Not only does the present research investigate the frequency of those devices, according to previous classifications by renowned researchers, Biber et al. (1999) and (2004), Cortes (2013), and Simpson-Vlach and Ellis (2010), but it also offers a qualitative analysis of each category and subtype considering their role as building blocks of discourse (BIBER et al., 2004) in each RA section.

1.2. Research aims

Many studies have been carried out in order to understand the linguistic features of academic discourse. Nonetheless, very few have detailed salient features from sections of Research Articles. Academic discourse is considered a register, and it is widely known that RA sections also display distinctive discursive and grammatical structures. However, it is not always clear what exactly makes RA sections different. Moreover, there is still a need to understand what specific linguistic characteristics comprise those elements of differentiation. Below we present the general aims (goals) and specific aims (objectives) of this research.

1.2.1 Research goals

The goal of the present research is to describe the use of lexical bundles in academic research article sections from Applied Linguistics high-impact journals, whose language of publication is English. Our main purpose is to understand how those linguistic devices are used in the Introduction, Methods, Results, Discussion sections by sorting their structures and functions according to previous taxonomies. A fundamental assumption of this study is to contribute to the teaching of writing skills to learners in EAP courses by learning more about the writing of RA sections by experienced authors.

1.2.2. Research objectives

This research has three main objectives, all of which relate to the writing composition of scientific articles. They are as follows:

1. after the structural and pragmatic classifications of lexical bundles from each IMRD subcorpus, compare and contrast the major and minor categories;
2. investigate the relationship between structure, function and the role of devices in each RA section and present the emerging findings;
3. provide a list of the main lexical bundles used in each section and sorted into the subtypes.

The steps above involve both quantitative and qualitative analyses.

1.3. The research questions

The research objectives led to the following three sets of research questions:

- A. Which subcorpora (IMRD) present the highest proportion of lexical bundles?
What is the ratio bundle token/type of each subcorpus?
- B. What are the most frequently used lexical bundles sorted into the structural types in each subcorpora? What do these structures tell us about the pragmatic function of each Applied Linguistics RA section?
- C. Which functional types are most commonly employed in each subcorpora?
What types of pragmatic features are revealed by their proportion and elements in each RA section?

1.4. Relevance of this study

Features of academic discourse might be too broad for those who are willing to improve their writing skills of RAs. Therefore, breaking down academic discourse into smaller pieces could help learners/researchers have access to more objective and specific rules. The decision to study the sections of RAs came from a need to provide more details on those sub-registers. Additionally, it is important to bear in mind that different fields of study have different discourse communities. In conclusion, this research aims at providing insights to those who want to improve their writing skills of RA sections within the Applied Linguistics discourse community.

1.5. The following chapters

The next chapters are organized into Literature Review, Methodology, Results and Discussion, and Conclusion. In the Literature Review, we present an overview of corpus linguistics and discourse analysis and previous studies on research article sections. The Methodology describes the steps taken to compile the IMRD corpus, the measures adopted when classifying the lexical bundles and their scrutiny, and covers the two statistical

procedures used to test significance. In the Results and Discussion chapter, the most relevant distinctions regarding the structural and functional types of devices across the subcorpora are presented and discussed. Finally, the Conclusion chapter covers the main findings, approaches the limitations of the study, and addresses pedagogical relevance.

2. LITERATURE REVIEW

Research into language use has seen a considerable growth in the employment of Corpus Linguistics (CL) in the last years. Its increasing popularity is due to the possibility of researching authentic material and the relatively current ease to access a great deal of data from thousands or millions of words. It can thus offer much more representativeness to the data analyzed so as to avoid cherry-picking or focusing on atypical language aspects. Similarly, CL has proven to be a resourceful tool not only due to the features mentioned above, but also because CL analyses do not rely on intuition.

In addition to “bringing together linguistic theory and data” (FLOWERDEW, 2011, p. 81), CL is also related to discourse analysis (DA). As noted by Biber et al. (2007), “corpus linguistic studies are generally considered to be a type of DA because they describe the use of linguistic forms in context”. Additionally, CL and discourse analysis both 1) take selected examples of naturally occurring discourse as their starting point; 2) identify recurring patterns in those examples; and 3) relate their findings to the social, intellectual or ideological contexts in which discourse plays a role (CHARLES; HUNSTON; PECORARI, 2011). However, their priorities tend to diverge in that discourse analysis focuses on entire texts and their cultural context, while CL sometimes applies techniques that disregard individual texts and prioritizes recurrent patterns of small scale items, i.e. words and phrases. Yet, one is likely to regard these two approaches as complementary methodologies.

The combination of CL and DA have yielded studies of written corpora that range from register-based approach (including those based on the Swalesian (SWALES, 2004) notion of genre) to linguistic devices with discourse functions, such as lexical bundles (FLOWERDEW, 2011). Other corpus studies have focused on the use of personal pronouns, passive/active voice and the identity in students’ academic writing (HYLAND, 2002), the pragmatic function of word sequences across registers (BIBER 2010; BIBER et al., 1999, 2004; BYRD & COXHEAD, 2010; HYLAND, 2008; SIMPSON-VLAC; ELLIS, 2010; etc), among many others.

In this chapter, we present an overview of the main studies on lexical bundles, the taxonomies of structural and functional types adopted for the current analysis, and finally a brief presentation of past studies attributed to the discursive role of RA sections.

2.1 Lexical Bundles

Firth (1935), cited by Stubbs (1993), claims that the complete meaning of a word is always contextual. In other words, studies of meaning must take context into consideration in order to be taken seriously. Later Firth (1957) also states that “you shall know a word by the company it keeps” (p. 11). Since then, there has been an increasing interest in investigating groups of words that occur together. For instance, Sinclair (1987) demonstrates that words co-occur in specific patterns with specific meanings. This led him to formulate the “idiom principle”, which means that words are governed by a co-selection rather than by an item-by-item criteria.

The term *lexical bundle* coined by Biber et al. (1999) is commonly referred in a corpus linguistic perspective as *recurrent word combinations* (ALTENBERG, 1998; De COCK, 1998), *n-grams* (BANERJEE; PEDERSON, 2003), *prefabricated patterns* (GRANGER, 1998), *formulas* (GRANGER; MEUNIER, 2008; SINCLAIR, 1991), *clusters* (HYLAND, 2008; SCHMITT; GRANDAGE; ADOLPHS, 2004), *sentence stems* (PAWLEY; SYDER, 1983), *formulaic sequences* (SCHMITT; CARTER, 2004), among many others.

Frequency is a fundamental characteristic that defines lexical bundles (BIBER et al., 1999; CORTES, 2013). The extraction of these multi-word sequences consists in the use of a sizable corpora, and it should disregard any pre-defined linguistic categories¹. Despite the frequency with which they occur, lexical bundles are “not idiomatic in meaning and not perceptually salient” (BIBER; BARBIERI, 2007, p. 269). In other words, they do not usually coincide with traditional grammatical units and may be phrase or clause fragments, such as “it is important to”, “it should be noted”, “in order to be”. Cortes (2013), however, claims that some lexical bundles that are made up of more than six words can represent complete units. She postulates that, although lexical bundles might not represent complete structural units, they are still seen as “important building blocks in discourse” and are able to convey complete semantic units (p. 270).

In order to offer more pedagogical insights into this field, researchers, namely Biber et al. (2004), Cortes (2004, 2013), Hyland (2008), and Simpson-Vlach & Ellis (2010), have created frequency-based lists of bundles regarding their context of realization and classified them into structural and functional types encompassing other subcategories presented below.

¹This is the so called corpus-driven approach which is closely related to the analysis of lexical bundles or any other features that are not recognized by traditional linguistic theories (refer to Biber, 2010).

2.1.1 Structural types

Biber et al. (2004) created a list of bundles according to their functions and structures in conversation and academic prose. Lexical bundles can be structurally sorted into three main groups that incorporate 1) noun phrases or prepositional phrase fragments; 2) verb phrase fragments; and 3) dependent clause fragments.

In academic prose, over 60 per cent of lexical bundles carry noun phrases or prepositional phrase fragments, such as *the objective of this paper, in the next section, one of the most important* (BIBER et al., 1999). In addition, according to Biber et al. (1999), prepositional phrase expressions are quite frequent in the academic discourse, especially bundles beginning with the preposition *in*. This preposition is found in expressions that communicate places or parts of the text, such as *in the United States* and *in the next section*. The most frequently used prepositional phrase expressions found by the authors are *at the same time* and *on the other hand*. They both carry an idiomatic meaning and work as linking adverbials. Note that the prepositional phrase fragments do not include noun phrases with of-phrase fragments.

There are a large number of lexical bundles in academic prose consisting of a noun phrase followed by a post-modifying of-phrase, e.g. *the end of the, the beginning of the, the base of the, the position of the, the shape of the, the size of the*. The authors maintain that this subtype covers a wide range of meanings, but some are especially important, for instance, a considerable number of these lexical bundles are used for physical description, including identification of place, size, and amount (*other parts of the; the shape of the; the total number of*). Another function is to mark simple existence or presence (*the existence of; the presence of*) or a variety of abstract qualities (*the nature of the*). Finally, the last group of bundles in this subcategory describes processes or events lasting over a period of time (*the course of the*).

The second category regards the lexical bundles that incorporate verb phrase fragments, in expressions such as *little is known about, is related to the, it has been shown that, it is necessary to, it has been suggested that*. This category encompasses two subtypes: verb phrase with non-passive verb; and verb phrase with passive verb. Biber et al. (1999) state that, in academic prose, just a few lexical bundles are built around a verb, and most of them incorporate a passive voice verb followed by a prepositional phrase, which marks a locative or logical relation, such as *are shown in table 3.7, and is shown in figure 6.20*. The authors

maintain that two expressions are moderately common in this category: one that identifies tabular/graphic and another that identifies some finding or assertion, e.g. *is based on the*.

The third category includes lexical bundles that incorporate dependent clause fragments. In this study we will consider three types: “to-clause fragment”; “that-clause fragments”; and “anticipatory it + adjective fragments”. Biber et al. (1999, 2004) demonstrated that dependent clause fragments are more commonly used in the academic discourse than verb phrase fragments. See Table 1.

Table 1: Structural types of lexical bundles adapted from Biber et al. (1999, 2004) and Cortes (2013)

Structural types of lexical bundles
1. Lexical bundles that incorporate noun phrase and prepositional phrase fragments
1a. Prepositional phrase expressions
1b. (connector +) Noun phrase with of-phrase fragment
1c. Noun phrase with other post-modifier fragment or Other noun phrase expressions
2. Lexical bundles that incorporate verb phrase fragments
2a. Verb phrase with non-passive verb
2b. Verb phrase with passive verb
3. Lexical bundles that incorporate dependent clause fragments
3a. to-clause fragment
3b. That-clause fragments
3c. Anticipatory it + adjective fragments
3b. WH-clauses
4. Lexical bundles that include NP and VP, fragments or whole phrases or clauses

Cortes (2013), however, after identifying that longer bundles are in some cases complete structures, complete clauses, and sometimes even sentences, argues that there should be a fourth group: lexical bundles that include noun phrases and verb phrases, fragments or whole phrases or clauses, in bundles such as *the rest of the paper is organized as follows*, and *the objective of this study was to evaluate*.

2.1.2 Functional taxonomy

The AFL (Academic Formulas List) developed by Simpson-Vlach and Ellis (2010) provides a collection of the most commonly used phrases employed in academic oral discourse and in writing compositions, such as the expressions used to indicate quantity, to show the stance of the speaker or writer, or to attribute an idea to a specific source of information. The AFL's main objective was to create a pedagogically useful list of formulaic sequences for academic speech and writing. The authors were inspired by Biber et al. (2004), but they did not simply attain to frequency based list of formulaic sentences. They also took into consideration a statistical measure of cohesiveness, mutual information (MI), insights from experienced professionals into which formulas are perceived to be the important for teaching. The authors kept the three great functional groups: Referential expressions, Discourse organizing functions, and Stance expressions from Biber et al. (2004), yet proposed other subtypes.

Hyland (2008) similarly proposed three functional categories of bundles, also based on research articles and dissertations: research-oriented bundles, text-oriented bundles, and participant-oriented bundles, which could loosely correspond respectively to Referential, discourse-orienting, and Stance expressions.

The current study investigates the following categories and subcategories across the sections:

Table 2: Functional categories and subtypes of lexical bundles adapted from Biber et al. (1999, 2004) and Simpson-Vlach & Ellis (2010)

Functional types

Referential expressions

1. Specification of attributes:
 - 1.a. Intangible framing attributes
 - 1.b. Tangible framing attributes
 - 1.c. Quantity specification
 2. Topic introduction and focus
 3. Contrast and comparison
 4. Deictics and locatives
-

Discourse organizing functions

1. Metadiscourse and textual reference
 2. Topic elaboration: cause and effect
 3. Discourse markers
 4. Topic introduction and focus
-

Stance expressions

1. Evaluation
 2. Expressions of ability and possibility
 3. Hedges
 4. Intention/volition, prediction
-

2.1.2.1 Referential expressions

Referential expressions “make direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as especially important” (p. 384). According to Hyland (2012), Referential expressions play an important rhetorical role, because they “frame, scaffold, and present arguments as a coherently managed and organized arrangement” (p. 160). The use of these expressions reflects the writer’s awareness of discursive conventions in that particular community or register. In the AFL, Referential expressions comprehend the largest group of the pragmatic functional taxonomy (SIMPSON-VLACH & ELLIS, 2010). Additionally,

Biber et al. (1999, 2004) and Dutra & Berber Sardinha (2013) concluded that Referential expressions are the most recurring in comparison to Stance expressions and Discourse organizing functions in corpora of written academic discourse. While the former investigated academic textbooks and research articles, the latter focused on essay learner corpora.

Simpson-Vlach & Ellis (2010) analyzed five Referential expression subcategories in the spoken and written registers: *specification of attributes*, *identification and focus*, *contrast and comparison*, *deictics and locatives*, and *vagueness markers*. Since this study focuses on the written register, we disregard the last subcategory, *vagueness markers*, due to the lack or extremely low frequency of these instances in the academic written register (see Table 2 for the subcategories analyzed).

What follows is a list of four subcategories from Referential expressions:

1) *Specification of attributes* encompass three distinctive attributes: 1.a) *intangible*, 1.b) *tangible*, and 1.c) *quantity specification*. They convey different and essential discursive meanings in the academic register.

1.a) The *Intangible framing attributes* category includes phrases that frame both concrete entities (A.1) and abstract concepts (A.2) or categories (SIMPSON-VLACH; ELLIS, 2010, p. 504), such as in:

(A.1) ... based on the total volume passing through each cost center

(A.2) so even with the notion of eminent domain and fair market value...

The 1.b) *Tangible framing attributes* category refers to physical or measurable attributes to the coming information or noun. Some examples from the AFL are *over a period of*, *the frequency of*, and *the amount of*.

The 1.c) *Quantity specification* category is closely related to the category of tangible framing attributes. Elements belonging to this category enumerate or specify amounts of the following nouns (cataphoric), such as *[a/large/the] number of*, *there are three*, or refer to a prior noun phrase (anaphoric), e.g. *both of these*, *of these two*.

The second most frequently employed subcategory of Referential expression in the AFL is 2) *identification and focus*. This includes typical expository phrases such as *as an example*, *such as the*, *referred to as*. Simpson-Vlach and Ellis (2010) also conclude that it is not surprising that this category is so common in academic discourse for “exemplification and identification are basic pragmatic functions” in this register (p.504). Biber et al. (2004) also

argue that these expressions “can be used to introduce a discussion by stating the main point first, and then giving the details”.

3) *Contrast and comparison* was a category coined by Simpson-Vlach & Ellis (2010). As the name suggests, contrast and comparison expressions encompass lexical bundles with the words *same, different, between, more, etc.*

4) *Deictics and locatives* expressions, the fourth and last to be analyzed from the Referential expressions. They refer to “physical locations in the environment or to temporal or spatial reference points in the discourse”, e.g. *in the United States, in the classroom, etc.* Biber et al. (2004) add that these types of expression are mostly prevailing in the written registers.

2.1.2.2 Discourse organizing functions

The other pragmatic function that is a concern in this study is Discourse organizing functions, which sets “relationships between prior and coming discourse” (BIBER et al., 2004, p.384). According to Biber (2004) and Simpson-Vlach and Ellis (2010) cause–effect subcategory is somewhat common in academic written discourse, while discourse markers are rare. In the AFL, these functions fall into four main subcategories: 1) *metadiscourse*, 2) *topic introduction*, 3) *topic elaboration* (3.a *non-causal* and 3.b *cause and effect*), and 4) *discourse markers*.

1) *Metadiscourse and textual reference*, in the AFL, is the most common subcategory within the Discourse organizing functions. Another term coined by Simpson-Vlach and Ellis, metadiscourse and textual reference expressions include *in the next section, in this chapter, this study was to, etc.* The researchers found a clear difference between the metadiscourse formulas in the spoken and written discourses, concluding that these expressions tend to be genre-specific

2) *Topic introduction and focus* consists of phrases that often frame an entire clause or upcoming segment of discourse, the AFL basically presents two instances: *For example [if/in/the]* and *what are the*.

3) *Topic elaboration* includes two groups 3.a) *non-causal* and 3.b) *cause and effect* expressions. Their function is to signal further explanation of a previously introduced topic.

3.a) *Non-causal topic elaboration* expressions are “used to mark elaboration without any explicit causal relationship implied” (SIMPSON-VLACH; ELLIS, 2010, p. 507). Consequently, this category presents phrases that summarize or rephrase, such as *it turns out*

that and *what happens is*. In this category, it is more likely to find lexical bundles which are present in the spoken discourse.

3.b) *Cause and effect topic elaboration* is an important group in the academic discourse, according to Simpson-Vlach and Ellis. These expressions signal a reason, effect, or causal relationship, for example, *in order to*, *as a result the*.

4) Finally, *discourse markers* include connectives, such as *as well as the*, *at the same time*, *in other words*. These devices are used to connect and signal transitions between clauses or constituents. Biber et al. (2004) claim that the bundles *as well as the* and *on the other hand* are used for explicit comparison and contrast and are considerably more common in written than in the spoken discourse.

2.1.2.3 Stance expressions

Stance expressions refer to the speaker's knowledge of or attitude toward the information in the proposition to be stated. Hyland (1999, p. 101) maintains that this is how "writers project themselves into their texts to communicate their integrity, credibility, involvement, and a relationship to their subject matter and their readers". This research looked at the following categories: 1) *evaluation*, 2) *expression of ability and possibility*, 3) *hedges*, and 4) *intention/volition and prediction*.

1) *Evaluation* was also developed and separately presented in the AFL. This category includes bundles such as *it is important to*, *it is necessary to*. It also includes *the importance of*, *is consistent with*. Simpson-Vlach and Ellis argue that most of the evaluative devices presented in the AFL were found in the written corpus.

2) *Expression of ability and possibility* "frame or introduce some possible or actual action or proposition". (SIMPSON-VLACH; ELLIS, 2010, p. 506)

3) *Hedges*, within the Stance expressions category, are among the most commonly used devices employed in the AFL. These formulas express some degree of qualification, mitigation, or tentativeness (Hyland, 1998). For instance, *there may be*, *to some extent*.

4) *Intention/volition and prediction* formulas occur mostly in the spoken register, according to Simpson-Vlach and Ellis. These expressions communicate the speaker's intention to do something in the future (BIBER et al., 2004).

2.2 Research article sections

As noted by Swales (1990), not only do texts pertaining to a certain register exhibit purpose, but also offer various patterns of similarity regarding structure, style, content and intended audience. It is necessary, however, to understand what structures are salient in specific registers, more importantly, what structures are salient within each RA from a given register. In this section, we present a brief overview of past studies regarding the IMRD RA sections.

According to Swales (1990), many academic writers claim that it is more difficult to get started on a composition than to work on the rest of the text. He also claims that Introductions are read first, so this might require an eye-catching composition. Also they are usually shorter and simpler than the other main sections. Additionally, the first paragraphs of a text might represent a challenge due to the countless possibilities of beginning it; such as what and how much background should be included, what kind of appeal will be made for the audience, how direct the approach could be. Perhaps for this reason, there has been a considerable number of studies involving this RA section.

Swales (1990) claims that unlike the Introduction and Discussion section, the Method paragraphs might be characterized as *broken linear*. In other words, “the sentences are like islands in a string” (p. 168) which could be read as if they were mere topics by those with specialist knowledge. His claims are based on botany, agriculture, engineering, biochemistry, medicine and zoology RAs, where many research methodologies are well established or protocolized (*apud* Swales 1990: WEISSBERG, 1984; GILBERT & MULKAY BRUCE, 1984). Nonetheless, “softer”, emerging or interdisciplinary fields tend to deal with given and new information more cohesively. Information in the Methods in this field is carefully presented with step-by-step descriptions and supported by anaphoric reference and lexical repetition (SWALES, 1990). However, in order to support this claim, Swales analyzes only one paragraph from the methods section of an Applied Linguistics paper from *TESOL Quarterly*.

As noted by Swales (1990), Result sections present a great deal of repetitive regularity in paragraph organization, in grammatical structure and in lexical choice. He also states that this is done deliberately in order to avoid “associative contamination with commentary or observation” (p.171).

The Discussion section is presented as a mirror image of the Introduction in Hill, Soppelsa, and West (1982). This means that they are expected to proceed from particular to general, “from the specific information reported in the Method and the Results sections to a more general view of how the findings should be interpreted” (WEISSBERG; BUKER, 1990, p. 161). Discussion sections have been a subject of register-based investigation carried out by various researchers. Hopkins (1988) investigated the discussion section of RAs and dissertations. Holmes (1997) analyzed the discussion sections of 30 RAs, 10 from each of the following fields: History, Political Science, and Sociology.

2.2.1 Research article sections: lexical bundles

Cortes (2013) briefly presents the findings regarding the use of lexical bundles and their functional and structural taxonomies from RA introduction sections. This study proved to be very robust and detailed, for it gathered a one-million word corpus of RA introductions from various academic disciplines. The results are in line with previous studies (BIBER et al., 1999; BIBER; CONRAD, 1999; BIBER et al., 2003, 2004), thus revealing that the longer the device the least frequent it is found in the corpus. Cortes also created a fourth structural category: “**Lexical bundles that include noun phrases and verb phrases** (fragments or whole phrases or clauses)”, for example, *the rest of the paper is organized as follows*, and *the objective of this study was to evaluate* (p. 38).

Le & Harrington (2015) compiled a corpus of 124 Discussion texts from leading applied linguistics journals. They focused their analysis on the relationship between 3-word bundles with the rhetorical move *commenting on results*, proposed by Ruiying and Allison (2003), in the Discussion section of quantitative research articles. The authors also found that writers, when giving interpretations and making comparisons between findings from other studies, usually rely on the present simple. They also highlight that the use of hedging devices: modal verbs (may, might) and modal adjectives (possible) are very commonly used to comment on results, which had already been stated by Hyland (1998).

Although there has been a considerable increase in the number of studies investigating academic register, one rarely finds studies on lexical bundles, structures and functions within each of the RA section. It is widely recognized, however, that RA sections are distinctive texts, so they should be treated as different sub-registers². Yet, most of the studies hitherto

²On sub-register: “register should be considered a continuous construct with particular register being defined at various levels of generality” (BIBER; FINEGAN, 1994, p. 221)

have focused on the cross-comparison comprehending an entire field of study, e.g. Engineering, Applied Linguistics, or even an entire register as in classroom language, textbooks, spoken or academic texts (BIBER et al., 1999, 2004; BIBER; BARBIERI, 2007; CORTES, 2013; HYLAND, 2008, 2012; ELLIS et al., 2008; SIMPSON-VLACH; ELLIS, 2010). Their contribution are undoubtedly precious to the field of EAP or ESP, but there is a need to better understand how these devices are employed in each section. Therefore, the present study aims at connecting lexical bundles from each of the sections analyzed (IMRD) to the structural and functional taxonomies of Biber et al. (1999; 2004) and Simpson-Vlach and Ellis (2010). In the next chapter, we present more details on the methodology adopted in the current study.

3. METHODOLOGY

In order to investigate the structures of lexical bundles and their role as building blocks of discourse, we compiled a specialized corpus restricted to RAs³ published in Applied Linguistics academic journals and/or whose authors or co-authors are members of Applied Linguistics departments. We collected 180 RAs from high-impact journals⁴ which necessarily went through peer review, see Table 3:

Table 3: List of high-impact journals

High-Impact Journals - Applied Linguistics	JIF (2016)	Number of RAs
International Journal of Intercultural Relations	1.183	13
Journal of English for Academic Purposes	1.414	12
Journal of Memory and Language	3.065	7
Journal of Second Language Writing	1.591	31
Language Learning	2.079	22
Linguistics and Education	0.833	35
System	1.400	3
TESOL Quarterly	2.056	19
Modern Language Journal	1.745	14
Patient Education and Counseling	2.429	2
Language Learning and Technology	2.293	22
TOTAL	-	180

All the RAs collected cover the following sub-areas: language acquisition and language processing; language teaching and learning; language in the professions; language in society; and analysis of spoken and written discourse⁵. As mentioned above, the RAs were peer reviewed and published in well-known journals. These journals very often impose strict

³All RAs were downloaded from Portal CAPES (<http://www.periodicos.capes.gov.br/>).

⁴Journal impact factors in Applied Linguistics range from 3.593 to 0.048 (2016), source: Journal Citation Reports (JCR).

⁵RAs which only belonged to Medical, Psychology or Law schools or other faculties were all discarded.

requirements upon their authors' composition⁶. Therefore, not only do we assume there is an acceptable use of the English language, but also presume that the authors were able to comply with the style and register inherent to RAs and, especially, conform to the community's discourse rules. The decision of not limiting this research to RAs written by native English speakers came as a reflection of the vast and increasing number of nationalities resorting to this language as *lingua franca*⁷ in the scientific community.

In this chapter, we describe the paths and experimental procedures adopted to carry out this analysis. Firstly, we describe and detail the Corpus of RAs sections especially compiled for this study. Secondly, we explain how the lexical bundles were generated and structurally and functionally classified according to existing taxonomies. Finally, we discuss the statistical procedures adopted in this research.

3.1 The corpus

The decision to investigate experimental papers was based on the premise that there is a "low chance of little-known people being "invited" (p.208) to contribute with review articles, review essays, general articles, state of the art surveys, etc. Therefore, since this study is meant to assist novice writers or academic new-comers, we eliminated the theoretical ones, despite their increasingly frequent phenomenon (SWALES, 2004).

For the present research, we compiled 180 RAs to make 4 subcorpora of the following sections: Introduction, Methods, Results and Discussion (IMRD). Each of which presents a distinctive sum of words.

Table 4: IMRD Corpus size

Subcorpora	Subcorpora size
Introduction	133,499
Methods	295,887
Results	337,295
Discussion	240,786
Total	1,007,467

⁶See https://www.elsevier.com/data/assets/pdf_file/0003/91173/Brochure_UPP_April2015.pdf as an illustration

⁷"a contact language between speakers or speaker groups when at least one of them uses it as a second language" (MAURANEN, 2017, p. 7)

Some experimental RAs did not present the traditional IMRD organization structure. They either did not have a clear methodology section or the discussion part was found along with the conclusion section. Our criterion involved discarding those articles. It is widely known that results, discussion and conclusion sections might come under distinctive labels or along with other topics in the same subdivision. In order to avoid jeopardizing our analysis, we either read the section to ensure it could be part of our corpus or decided not to include any articles which could cast doubt over its category. Another rule regarded the headings of certain sections: if they were encountered under the label “Results and Discussion”, or “Discussion and Conclusion”, i.e. “hybrid headings” (LIN; EVANS, 2012), their corresponding RA was also excluded, given potential differences in the communicative functions of these sections.

Having all articles sorted, the next step entailed the compilation of every section in .txt extension. This work demanded arduous hours of: 1) selecting and copying texts only, i.e. skipping excerpts of dialogue, tables, figures, and other elements which did not belong to the prose of the sections so we could avoid inflation of data; and 2) pasting texts into text documents. After all text files were ready, another clean-up process and adjustment were applied with Python 3.0. This measure was taken because, when the texts are pasted into the text files, lines are broken causing a difference in number of generated lexical bundles due to the separation of elements. Texts previous to the Python treatment presented a different number of lexical bundles generated in AntConc 3.4.4 (ANTHONY, 2016). Since sentences and clauses were also broken, AntConc did not recognize the broken lines as such. Therefore, a Python script was employed to adjust this issue turning texts with broken clauses and sentences into single paragraphs, see Appendix C.

We also searched for any duplicated files but found none.

3.2 The procedure for data analysis

With very few pre-conceived ideas, the purpose of this study was to analyze the bundles that emerged from our corpus. As Biber et al. (2004, p. 176) claimed: “frequency data identifies patterns that must be explained” rather than regarding frequency as explanatory. The relevance of a frequency-based study comes as a necessity to investigate patterns which are likely to go unanalyzed due to their high frequency.

Our data analysis comprehended the following steps:

1. generating the list of lexical bundles of 4, 5, 6, and 7 words from each subcorpus, with a minimum frequency of 4, occurring in at least 2 different texts, as a way to avoid idiosyncrasy;
2. calculation of the ratio bundle type/bundle token in order to estimate which sections rely on more word sequences;
3. deletion of overlapping bundles, or bundles that contained elements duplicated were clustered together as one;
4. understanding and manually selecting lexical bundles according to their structural and functional-semantic purposes, using taxonomies previously developed in the literature (BIBER et al., 1999; BIBER et al., 2003, 2004; CORTES, 2013; SIMPSON-VLACH; ELLIS, 2010);
5. labeling and classifying new lexical bundles (not found in previous taxonomies);
6. Cross-comparison of the subcorpora regarding the structural and functional types of bundles and the linguistic features of RA sections; and
7. conducting statistical treatment.

What follows is a thorough description of the procedures adopted in this study. The following sections further explain the steps presented above.

3.2.1 Generating the lexical bundles

The frequency cut-off point used to identify lexical bundles is somewhat arbitrary, so this leads to varying practices in the literature. Some studies of written corpus data have employed cut-offs of 25 times per million words (e.g. CHEN & BAKER, 2010), and others have used 20 times per million words (e.g. CORTES, 2004). The dispersion criterion is also arbitrary, a criterion of three to five texts is often used for four-word bundles (e.g. BIBER & BARBIERI, 2007; BIBER et al., 2004; CHEN & BAKER, 2010; CORTES, 2004), but percentages are also sometimes used (HYLAND, 2008).

For this study, we used the application AntConc 3.4.4 (ANTHONY, 2016) to generate the bundles and defined the length of bundles as four, five, six and seven-word sequences that occurred at least four times in two different papers. AntConc generates the bundles, presents how many different bundles there are and also the total number of bundles, the so-called bundle types and bundle tokens respectively. These details allow us to calculate 1) the frequency of bundles utilized in each subcorpus, obviously considering a normalized frequency due to the different sizes of the subcorpora; 2) the variability of the lexical bundles employed by calculating the ratio *bundle type/lexical bundle token*.

Fig 1: Lexical bundles from the Methods subcorpus generated by AntConc 3.4.4

Rank	Freq	Range	N-gram
1	47	33	at the end of
2	46	28	participants were asked to
3	45	34	the end of the
4	41	19	appendix s in the
5	41	26	in the united states
6	41	19	s in the supporting
7	40	18	appendix s in the supporting
8	40	18	in the supporting information
9	39	32	at the time of
10	39	18	in the supporting information online
11	39	18	the supporting information online
12	38	18	s in the supporting information
13	38	22	the total number of
14	37	17	appendix s in the supporting information
15	37	18	s in the supporting information online
16	36	17	appendix s in the supporting information online
17	33	28	as well as the
18	33	24	the beginning of the
19	32	24	at the beginning of

The second process of data analysis entailed the generation of long lists of word sequences from each of the four subcorpora. As it is already known, AntConc 3.4.4 is not designed to ignore overlapping bundles, for example in “at the time of data” (8 occurrences) and “at the time of data collection” (8 occurrences) are considered two different bundles by the application. This, then, required a clean-up process which could not rely on manual selection of bundles since the lists had approximately 1,000 bundles each.

Having the lists of bundles generated for each subcorpus, the next step was to discard overlapping bundles so as to guard against inflated results. This was done by running an *R* script (see Appendix B) with the use of three packages, namely *tidyverse*, *stringr*, and *readxl*. It is noteworthy that a previous methodology has adopted a similar process (BOHORQUEZ, 2015). The main idea was to delete the shortest bundles which occurred as frequently as their counterpart, see an illustration below with the raw frequency between brackets, also known as complete overlap (CHEN & BAKER, 2010):

randomly assigned to one (7)
 randomly assigned to one of (7)

by the participants in (4)
 by the participants in this (4)
 by the participants in this study (4)

Hence, the highlighted bundles remained in the list for they are more complete versions of the others. Other occurrences of overlapping bundles were preserved, they are called complete subsumption (CHEN; BAKER, 2010), such as in:

were asked to read (9)
 participants were asked to read (5)

This allowed us to join the overlapping bundles with lower frequency with the most frequent one. This resulted in bundles like “(participants) were asked to read” (9).

In occurrences of almost identical bundles, as shown in the chart below, bundles were combined by adding the missing element to one bundle and clustering them together as one. Such combinations can also be found in the AFL.

the result of the (5)
 the results of the (40)
 the result(s) of the (45)

If bundles offered interesting combination to be preserved, the most frequent one was deducted by the frequency of the others that remained. As an illustration, *there was a significant* was kept in the list, but the overlapping occurrences of this bundle had to have their frequency deducted from *there was a significant*, hence $43 - 29 = 14$ occurrences.

43 - 13 - 12 - 4 = 14 | there was a significant

13	there was a significant difference (between (the)/ in (the)
12	there was a significant effect (of)
4	there was a significant interaction

Finally, results were normalized per 1,000 words to make our data comparable, following a recommendation found in Biber, Conrad & Reppen (1998, p.264) where it is stated that “frequency counts should be normed to the typical text length in a corpus”, due to the fact that we are dealing with a specialized corpus of RA sections⁸.

3.2.2 Structural and functional types of lexical bundles

Our main analysis consisted in sorting the list of lexical bundles generated per subcorpus, taking into consideration their structural and functional types. This action meant to help us understand if bundles presented any special features in regard to different structures,

⁸On average, each section compiled contain approximately 1,400 words.

such as passive voice, dependent clause, prepositional phrases, etc; or functional ones, i.e. stance markers, discourse organizers and Referential expressions (see Table 1 in the Literature Review chapter).

The selection of the structural types to be analyzed herein was based on the fact that Biber et al. (2004) created the nomenclature of types that occur in both spoken and written discourse. “Yes-no question fragments”, for instance, are not expected to be found in the academic written discourse. The same applies for “WH-question fragments”. These instances are inherent to the oral discourse, thus their scrutiny is deemed unnecessary for the present study. It is worth pointing out that the passive voice structures analyzed included bare passive, e.g. “as shown in figure”, and the non-passive structures encompass all “VP bundles” with no passive voice forms, such as verb BE and modal verbs.

The table of functions was established by Biber et al. (2004), based on previous theoretical investigations on discourse functions, i.e. Hymes (1974), Halliday (1978), Brown and Fraser (1979), Biber (1988, 1995), as a preliminary functional taxonomy of the bundles in university classroom teaching and textbooks. It was then further developed by Simpson-Vlach and Ellis (2010), and it was also employed in this analysis (see Table 2 in the Literature Review chapter).

Most of the bundles generated were sorted individually into the categories and subcategories listed above. The functional taxonomy introduced by Biber et al. (2004) is composed of lexical bundles found in classroom teaching and textbooks, with a corpus size of over 2 million words. This means that the bundles generated and classified did not always fit the bundles generated, due to the size of corpora or the specificity of vocabulary encountered. Therefore, it was necessary to create new subcategories taking into consideration discursive role in the text.

Fig 2: Lexical bundles from the Methods subcorpus displayed in a Google spreadsheet for classification

	A	B	C	D	E
1	47	at the end of (the semester)	PP	referential	Deictics and locatives
2	46	participants were asked to (complete/ indicate/ read/ select/ write)	passive - past	stance	obligation and directive
3	44	in order to address/ avoid/ examine/ facilitate/ gain/ make/ obtain/ provide	PP	Discourse organizing functi	Topic elaboration: cause and effect
4	41	in the United States	PP	referential	Deictics and locatives
5	41	(appendix) s in the supporting information online	other NP		
6	39	at the time of	PP	referential	Deictics and locatives
7	38	the total number of (participants/ words)	Noun phrase with of-phrase fragment	referential	Quantity specification
8	33	the beginning of the (semester/ study)	Noun phrase with of-phrase fragment	referential	deictic
9	33	as well as the	PP	Discourse organizing functi	Discourse markers
10	32	at the beginning of	PP	referential	deictic
11	31	in the present study	PP	Discourse organizing functi	Metadiscourse and textual reference
12	28	in the case of (the)	PP	referential	Identification and focus
13	28	for each of the	PP	referential	Quantity specification
14	26	at the beginning of the (semester/ study)	PP	referential	deictic
15	26	on the basis of (the/their)	PP	referential	Intangible framing attributes
16	26	for the purpose(s) of (the/ this)	PP	Discourse organizing functi	Topic elaboration: cause and effect
17	24	they were asked to	passive - past		
18	23	the analysis of the	Noun phrase with of-phrase fragment	referential	Intangible framing attributes
19	23	in the form of (a)	PP	referential	Intangible framing attributes
20	23	over the course of	PP	referential	Intangible framing attributes
21	23	in the current study	PP	Discourse organizing functi	Metadiscourse and textual reference

Another inherent challenge of this type of analysis lies on the fact that certain bundles can be classified into more than one category. For instance, “the end of the” may belong to Deictics and locatives from Referential expression or Metadiscourse and textual reference from the Discourse organizing functions. For this matter, it was also necessary to qualitatively check the concordance lines and understand to which category those determined bundles belong.

Fig 3: Concordance lines of the lexical bundle “the end of the”

AntConc 3.4.4w (Windows) 2014

File Global Settings Tool Preferences Help

Corpus Files: EAPMS1.txt, EAPMS2.txt, EAPMS3.txt, EAPMS4.txt, EAPMS5.txt, EAPMS9.txt, EAPMS12.txt, EAPMS17.txt, EAPMS18.txt, EAPMS19.txt, EAPMS20.txt, EAPMS21.txt, LEMS3.txt, LEMS5.txt, LEMS6.txt, LEMS10.txt, LEMS11.txt, LEMS12.txt, LEMS14.txt, LEMS15.txt, LEMS18.txt, LEMS20.txt, LEMS21.txt, LEMS22.txt, LEMS24.txt, LEMS26.txt, LEMS28.txt, LEMS29.txt, LEMS30.txt, LEMS33.txt, LEMS34.txt, LEMS35.txt, LEMS36.txt, LEMS37.txt

Total No. 180
Files Processed

Concordance Concordance Plot File View Clusters/N-Grams Collocates Word List Keyword List

Concordance Hits 45
Hit KWIC

1 evaluation survey was distributed to students by the end of the course. It covered the
2 in the last twenty years (Chan, 2014). Towards the end of the last century the widespread
3) one afternoon after normal class hours towards the end of the semester. The participants were
4 the Major League Baseball Players Association. At the end of the conversation, Robert concluded that
5 teams. The present study took place at the end of the project's third and
6), Ms. Youssef's reflection log completed at the end of the unit, and the written
7 reflective groups were required to answer at the end of the four treatment sessions. Furthermore
8 to enhance the noticeability of recasts. At the end of the session, the learners practiced
9 to be having difficulties in understanding. After the end of the CELTA course, Phiona then
10 were administered at the beginning and at the end of the study (immediately after the 7
11 . Lastly, I included a writing prompt at the end of the short story that read:
12 . The writing prompt that I added to the end of the text stated: "Imagine that
13 with the less forthcoming young rappers at the end of the sessions. From practitioner research
14 word was associated with each picture at the end of the session. This allowed us
15 access explicit knowledge. A bell sound indicated the end of the response time, and the
16 etalinguistic knowledge test, was administered at the end of the experiment, in order not
17 . Then the visual display was shown until the end of the trial. The first nonword
18 CF, depending on the treatment condition. At the end of the first classroom session, learners
19 University's Certificate of French Studies at the end of the program, the international students
20 the program. Finally, the third was toward the end of the academic year when the

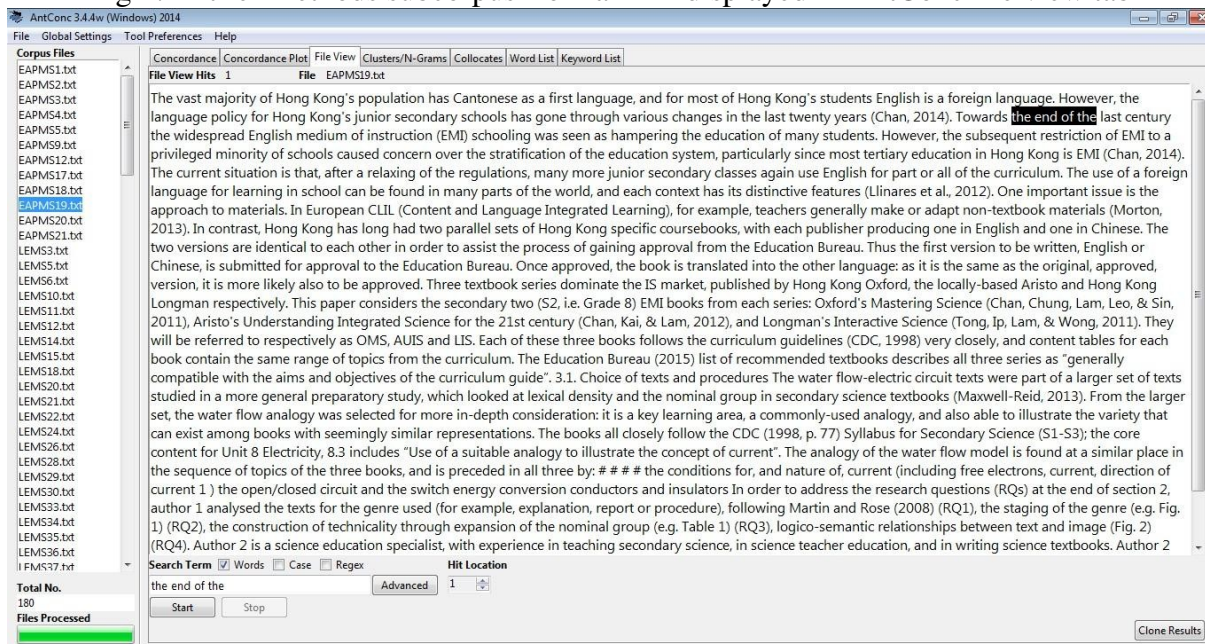
File: EAPMS9.txt, EAPMS19.txt, EAPMS21.txt, LEMS12.txt, LEMS21.txt, LEMS21.txt, LEMS24.txt, LEMS24.txt, LEMS29.txt, LEMS34.txt, LEMS37.txt, LEMS37.txt, LEMS51.txt, LEMS51.txt, LEMS8.txt, LEMS19.txt, LEMS19.txt, LEMS19.txt, LEMS20.txt, LEMS20.txt, LLTMS15.txt, LSMS20.txt, LSMS20.txt

Search Term Words Case Regex Search Window Size 50

the end of the

Kwic Sort Level 1 1R Level 2 2R Level 3 3R

Fig 4: Entire Methods subcorpus from an RA displayed in AntConc file view tab



3.2.3 Statistical treatment

For this research, we relied on two different types of statistical treatment, namely confidence interval and a null-hypothesis significance test, the z-test. In this section, we detail the reasons why these two different statistical concepts were chosen and how they were applied in our analysis.

The subcorpora containing 180 RA sections from this research is a representative sample of the population of sections of Applied Linguistics RAs. In this case, one usually relies on confidence intervals in order to estimate the range of confidence that can be considered when comparing findings. In other words, confidence intervals can bring more reliable generalizations regarding the differences between each of the subcorpora analyzed. They are rigorous statistical testing, provide appropriate effect sizes, and are “an alternative approach to significance testing proper” (GRIES, 2006, p. 199). Confidence interval also allows us to choose how confident we can be about the true value of a population parameter (SHESKIN, 2003). As Gries (2013, p. 133) states:

The so-called confidence interval, which is useful to provide with your mean, is the interval of values around the sample mean around which we will assume there is no significant difference with the sample mean. From the expression “significant difference”, it follows that a confidence interval is typically defined as 1-significance level, i.e., typically as $1 - 0.05 = 0.95$.

The confidence intervals for this study were defined as 0.99, i.e. we can then assume with 99% of confidence that our samples are significantly different or not by looking at the error bars. If the error bars overlap, differences across sections are not considered statistically significant. This type of estimation has not been commonly used by Applied Linguistics researchers, even though it is highly recommended (GRIES, 2006), mainly when dealing with small findings from big samples⁹. As an illustration, the subtypes of lexical bundles generated in this study represented very small quantities, the highest was 14.43 (see Table 6), the lowest, when higher than 0, was 0.04 (Table 24) out of subcorpora of 250,000 words, on average. Relying on null-hypothesis, in this case, might result in disregarding relevant results, according to Gries (2006).

The estimation was run on Google Sheets, where all the data of this research were stored and managed. The function below illustrates how the standard errors for our estimates were calculated.

$$se = \sqrt{p*(1-p) / n}$$

In this equation, p means the normalized frequency of a expression or word and n is the size of our sample. For the standard error, we calculate the confidence interval with:

$$CI = [p - 2.54 * se; p + 2.54 * se]$$

The value of 2.54, in a normal distribution with 0 mean and 1 variance, represents 99% of the whole mass probability.

In spite of the reservations about the use of null-hypothesis testing (KILGARRIFF, 2005, GRIES, 2006) in corpus studies, we decided to test the null-hypothesis of each pair of subcorpus by running a test designed to deal with larger samples, the z-test. This decision was based on the premise that there is no consensus on what type of statistical treatment to be employed when testing the significance of results, and also because both statistical measures should be reported in scientific articles (GRIES, 2005; du PREL et al., 2009, p.338) for “they provide complementary types of information”.

A script was created in order to run the z-test. In the “script editor” of Google Sheets, we entered a function, see Appendix D.

⁹This is a problem that often arises in corpus-linguistic studies due to small but significant findings whose relevance may be limited by the comparison or presence of high frequency events, such as corpus size (KILGARRIFF, 2005).

After that, in Google Sheets, we ran the following:

```
=NORMDIST(z_test($B4,$D4,133499,337295), 0, 1, FALSE)
```

Where, **NORMDIST** means normal distribution, **B4** corresponds to the raw frequency of a finding, **D4** is another finding, here we tested the pair Introduction vs Methods. The values between the parentheses are the corpus size of each subcorpus. The values “0” and “1” mean that the normal distribution must have a mean of zero and variance of 1. **FALSE** is a required function argument by Google Sheets. This argument might vary in other spreadsheet programs.

This chapter covered the criteria and steps followed to compile the four subcorpora; the procedure for data analysis, including the generation of lexical bundles, their analysis; and finally their statistical treatment. In the following chapter, we present the results and discussion of the most statistically significant findings.

4. RESULTS AND DISCUSSION

For the present study, we generated 3,976 bundle types and 25,361 bundle tokens from the Introduction, Methods, Results and Discussion subcorpora taken from Applied Linguistics RAs, as detailed in the Methodology chapter.

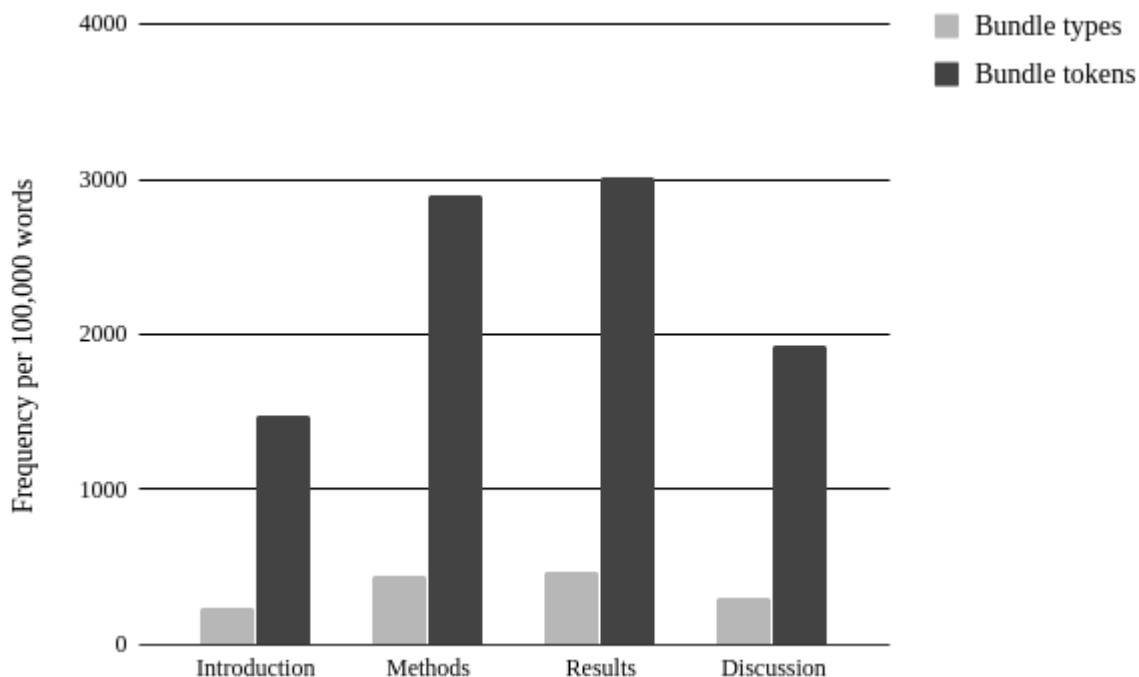
Table 5: Raw frequency of lexical bundles generated from each IMRD subcorpus by AntConc and corpus size

	Introduction	Methods	Results	Discussion	Total
Bundle types	316	1320	1597	743	3976
Bundle tokens	1978	8573	10,153	4657	25,361
Corpus size	133,499	295,887	337,295	240,786	1,007,467

The following process entailed the normalization of frequency of each subcorpus. From the findings in Graph 1, we can infer that the subcorpora of Methods and Results present a higher frequency of lexical bundles than the other subcorpora, Introduction and Discussion. It is worth pointing out that the Introduction subcorpus contains just as half as the sum of bundles found in the Methods and Results. This pattern recurs throughout the analysis herein presented, i.e. the Introduction subcorpus usually shows the lowest occurrence of the investigated categories, even after the frequency of findings is normalized.

Nonetheless, when calculating the ratio bundle type/ bundle token, all subcorpora present similar ratio. Bundle types are the same as different lexical bundles, and bundle tokens represent the overall frequency of lexical bundles across the subcorpora. In the Introduction, Results and Discussion subcorpora, the ratio is 0.16, in Methods, 0.15. Therefore, we can assume that the entire IMRD corpus shows no distinction regarding the range of different lexical bundles, in general.

Graph 1: Bundle types and bundle tokens across the IMRD subcorpora



After the clean-up process, described in the Methodology chapter, the bundle types were individually analyzed and sorted into the structural and functional categories based on Biber et al. (1999, 2004) and Simpson-Vlach & Ellis's (2010) taxonomies. Due to corpus size and to the specificity of the sub-register herein analyzed, it was necessary to create and present a list of new subtypes regarding their pragmatic function.

In order to understand the proportion of these devices in each subcorpus, we present a separate table with normalized frequency (per 1000 words) of each set of categories and subtypes sorted into structural and functional taxonomies. Subsequently, the normalized frequency (per 1000 words) with confidence interval error bars are found in the Graphs as a way to cross compare the amount of those devices in each subcorpus. We also display a list of the 10-15 most frequently instances with some of the categories/subcategories herein investigated so that we could illustrate and present a brief qualitative analysis of the most notable distinctions in the subcorpora.

4.1 Lexical bundle types across the IMRD subcorpora

4.1.1 Structural types of lexical bundles

We searched and sorted bundles based on their types as in Biber et al. (2004), namely lexical bundles that incorporate verb phrase fragments, lexical bundles that incorporate noun phrases or prepositional phrase fragments, and lexical bundles that incorporate clause fragments. In addition, one subtype *anticipatory it + adjective fragments*, clausal bundles starting with an anticipatory “it”, which is not separately presented by Biber et al. (2004), but it is in Biber et al. (1999), was also included in the category of lexical bundles that incorporate dependent clause fragments, considering its pedagogical relevance (HYLAND, 2012) and frequency in the list of bundles generated.

Below, the normalized frequency of all bundles from Introduction, Methods, Results and Discussion subcorpora sorted into three major structural categories of lexical bundles is presented. From the bundles analyzed, the great majority falls into noun phrases or prepositional phrase fragments. These results corroborate with Biber et al. (1999, 2004) and Biber (2010) who claim that “most lexical bundles in academic prose are building blocks for extended noun phrases or prepositional phrases” (BIBER et al., 1999, p. 992).

Table 6: Major categories: structural types of lexical bundles (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
NP or PP frag.	7.29	14.43	12.12	10.18
VP frag.	1.04	4.96	3.85	3.03
Dep. clause frag.	0.52	0.91	1.24	1.54

In spite of the fact that the proportion of “NP or PP fragments” corroborates with previous research, the same cannot be stated about the amount of “VP” and dependent clause fragments. Dependent clause fragments, in the current study, constitute the smallest share of the structural types. This finding challenges the claim (BIBER et al., 1999, 2004; BIBER, 2010)¹⁰ that academic discourse presents a considerable frequency of dependent clause fragments, higher than “VP fragments”. This distinction, therefore, might be a particularity from the sub-register subject to the present study, Applied Linguistics RAs.

¹⁰Their corpus comprises academic research articles (2.7 million words) and advanced academic books (2.6 million words), see Biber et al. (1999, p. 32-34)

4.1.1.1 Prepositional phrase, noun phrase fragments, and other nouns

Within each of the structural types of bundles, we analyzed prepositional phrase fragments, noun phrase fragments as well as other nouns. The table below illustrates that prepositional phrase fragments, such as *in the United States*, *at the same time*, *as well as the*, *to the field of*, represent the greatest share in all subcorpora. Other noun phrases are found, for example, in *teachers in this study*, and *the second research question*. See Tables 8 and 9 for further occurrences of lexical bundles within these subcategories.

Table 7: Lexical bundles subtypes that incorporate noun phrase or prepositional phrase fragments (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
PP frag	4.46	7.84	4.97	5.96
NP with of-phrase frag.	1.18	3.64	3.26	2.26
Other NP	1.64	2.96	3.89	1.96

In the following sections, we present graphs with confidence interval testing for the prepositional phrase and noun phrase with of-phrase fragments in order to better illustrate the difference in frequency (per 1000 words) and error bars which define the range of confidence one can have when comparing the populations or subcorpora.

4.1.1.1.1 Prepositional phrase fragments

The image below shows that the normalized frequency of prepositional phrase fragments in the Introduction and Results sections do not differ significantly. Nevertheless, Methods and Discussion present a high and significant difference in the use of these structures. The error bars, which represent the confidence interval, do not overlap, thus revealing the significant difference of the findings.

Graph 2: Normalized frequency and confidence intervals of prepositional phrase fragments

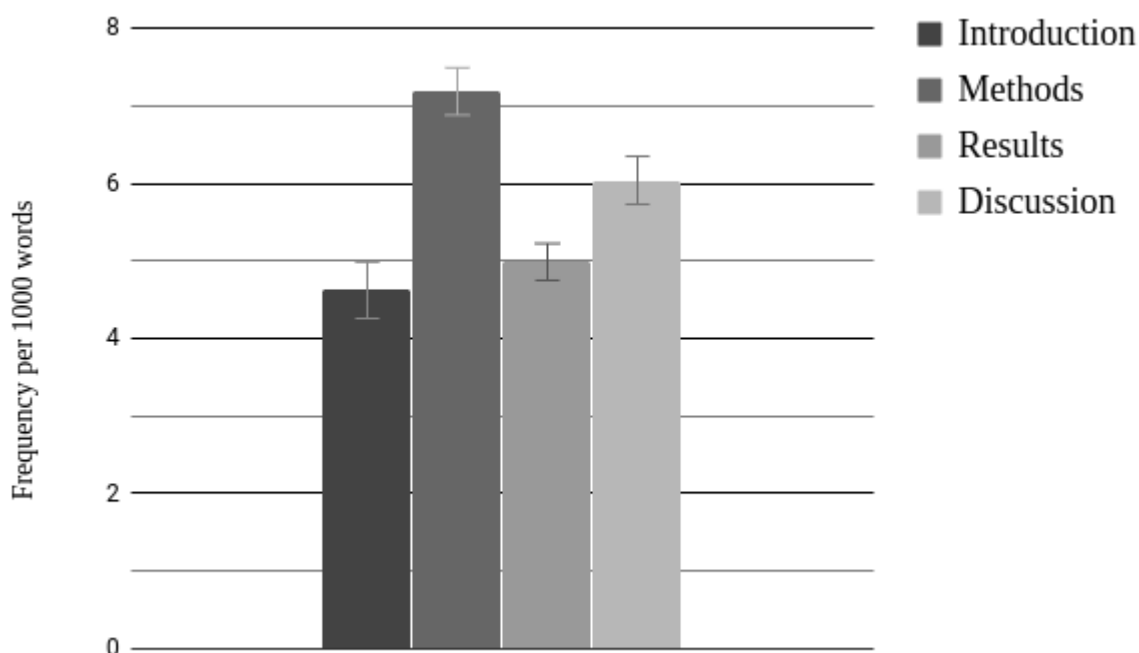


Table 8 displays the most frequently used prepositional phrase expressions in the subcorpora. As Biber et al. (1999, p. 1019) claim, lexical bundles beginning with the preposition *in* are commonly used in the academic prose, they are employed to identify a particular location, e.g. *in the United States*, or “to specify a particular discourse context”, such as *in the current study* and *in the case of the*. Other very frequently used prepositional phrase bundles are *at the same time*, *on the other hand*, *as well as the*. These bundles have relatively idiomatic meanings (BIBER et al., 1999) and are used as linking adverbials to compare or contrast two propositions or events. The bold typed and italic bundles are found in more than one subcorpus. It is interesting to note that Table 8 does not reveal much exclusivity of bundles containing PP phrase fragments. This might mean that IMRD subregisters do not reveal to be substantially different regarding the use of this subtype of lexical bundle.

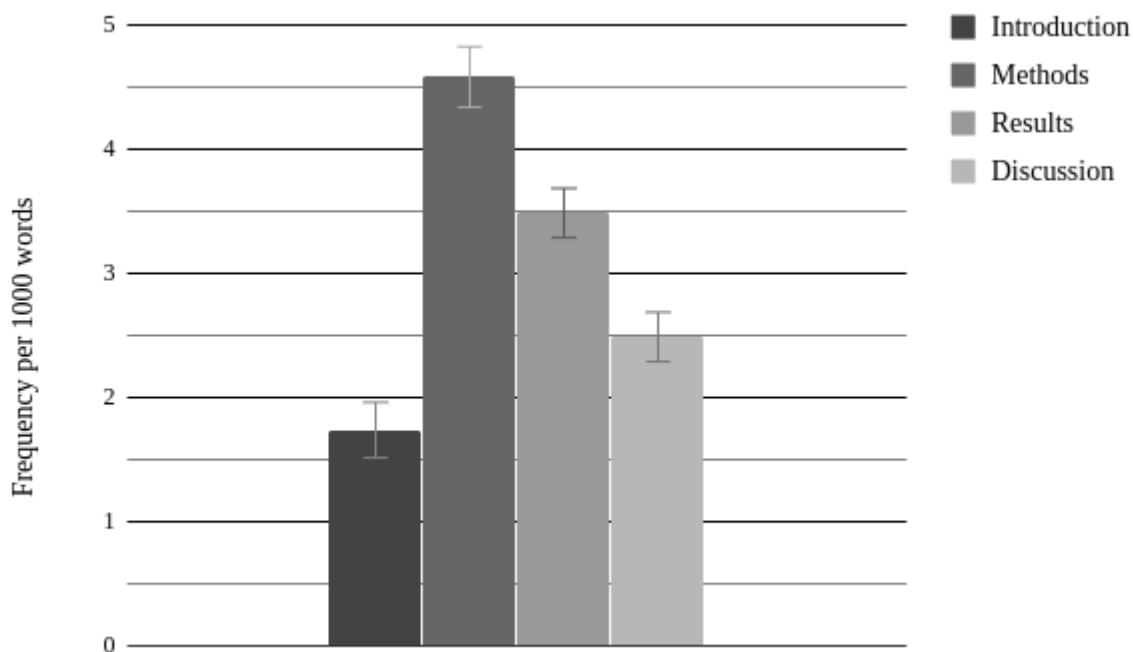
Table 8: Prepositional phrase fragments occurrences

Introduction	Methods	Results	Discussion
<i>in the United States</i>	<i>at the end of</i> (the semester)	<i>(and) at the same time</i>	<i>in the present study</i>
<i>at the same time</i>	in order to address/ avoid/ examine/ facilitate/ gain/ make/ obtain/ provide	<i>on the other hand</i> <i>(the)</i>	<i>in the current study</i>
<i>on the basis of (the)</i>	<i>in the United States</i>	<i>in the context of</i>	<i>on the other hand</i> <i>(the)</i>
<i>on the other hand</i>	at the time of	<i>in the case of (the)</i>	<i>at the same time</i> <i>(the)</i>
<i>in the present study</i>	<i>as well as the</i>	<i>in terms of the</i>	<i>in the context of</i> <i>(the)</i>
<i>as well as the</i>	<i>in the present study</i>	<i>as well as the</i>	of the present study
<i>in the case of</i>	for each of the	<i>at the beginning of</i> <i>(the)</i>	<i>as well as the</i>
in the target language	<i>in the case of (the)</i>	between the two groups	in line with the
of english as a	<i>at the beginning of</i> <i>the</i> (semester/ study)	<i>at the end of (the)</i>	<i>in the case of (the)</i>
of the present study (is)	for the purpose(s) of (the/ this)	<i>in the United States</i>	<i>in the United States</i>
<i>over the course of</i> <i>(the year/ a semester)</i>	<i>on the basis of</i> <i>(the/their)</i>	in the use of	in this study the
to the development of	<i>in the current study</i>	of the variance in	<i>in terms of the</i>

4.1.1.1.2 Noun phrases with of-phrase fragments

Graph 3 shows that noun phrases with of-phrase fragments are employed in all sections with a significant difference in frequency across all of them. This is revealed by the error bars, which do not overlap, representing the confidence interval.

Graph 3: Normalized frequency and confidence intervals of noun phrases with of-phrase fragments



Noun phrases with of-fragments are quite diverse throughout the subcorpora, but most of them convey Intangible framing attributes, e.g. *the meaning of the*, *the use of the*, and Quantity specification¹¹, such as *a wide range of*, *the total number of*, and *a large number of*. Quantity specification (see 4.1.2.1.2) devices are very often used in Methods as well as in Results RA sections. The relationship between these two subtypes might explain why the Methods and Results subcorpora make a considerable use of “NP with of-phrase fragments”. See Table 9. The bold typed and italic bundles are found in more than one subcorpus. Bundles from a given subcorpus, which are not found in the other subcorpora, might be exclusive to the type of subregister of RA sections.

¹¹Subcategories of Referential expressions later discussed in this chapter.

Table 9: Noun phrases with of-fragment occurrences

Introduction	Methods	Results	Discussion
a number of studies (have)	the total number of (participants/ words)	(a) significant main effect of	<i>the results of the</i> (study/ present study)
<i>a wide range of</i>	the analysis of the	<i>the results of the</i>	the use of the
a large number of (extensive/ growing/ considerable) body of research on	<i>the meaning of the</i>	a significant effect of	the results of this (study)
our understanding of the	<i>the use of the</i>	<i>the use of the</i>	the findings of this study
the aim of the	the majority of the (participants)	the course of the	<i>a wide range of</i>
the process of learning	one of the two	the main effect of	the findings of the (study/ present study)
the purpose of the	native speakers of English	the majority of the	the context of the
the use of a	part of a larger	the mean number of	a great deal of
and the development of	the center of the	the rest of the	the nature of the
patterns of interaction in	<i>the content of the</i>	the total number of	the quality of the
the role of language	the course of the	<i>the content of the</i>	a greater number of
	the purpose of the (study)	<i>the meaning of the</i>	beyond the scope of (the)

4.1.1.2 Verb phrase fragments

The following step was to investigate the verb phrase fragments. Phrases with non-passive verbs constitute the largest portion of all lexical bundles that incorporate verb phrase fragments in the Introduction, Results and Discussion subcorpora (see Table 10). The most distinctive feature of the subtypes of lexical bundles that incorporate verb phrase fragments in Methods might be the dramatic difference between the frequency of passive¹² and non-passive constructions. This finding corroborates with Swales (1990) in which he claims that “the past

¹²We also included bare passive voice, e.g. *as shown in Fig X*

passive is consistently chosen and the identity of the underlying agent is consistently that of the experimenters” (p.167). Other verb phrases are those which could not be identified as containing passive or non-passive structures because the elements after the linking verbs are not revealed, such as *the present study is* or *the target language is*.

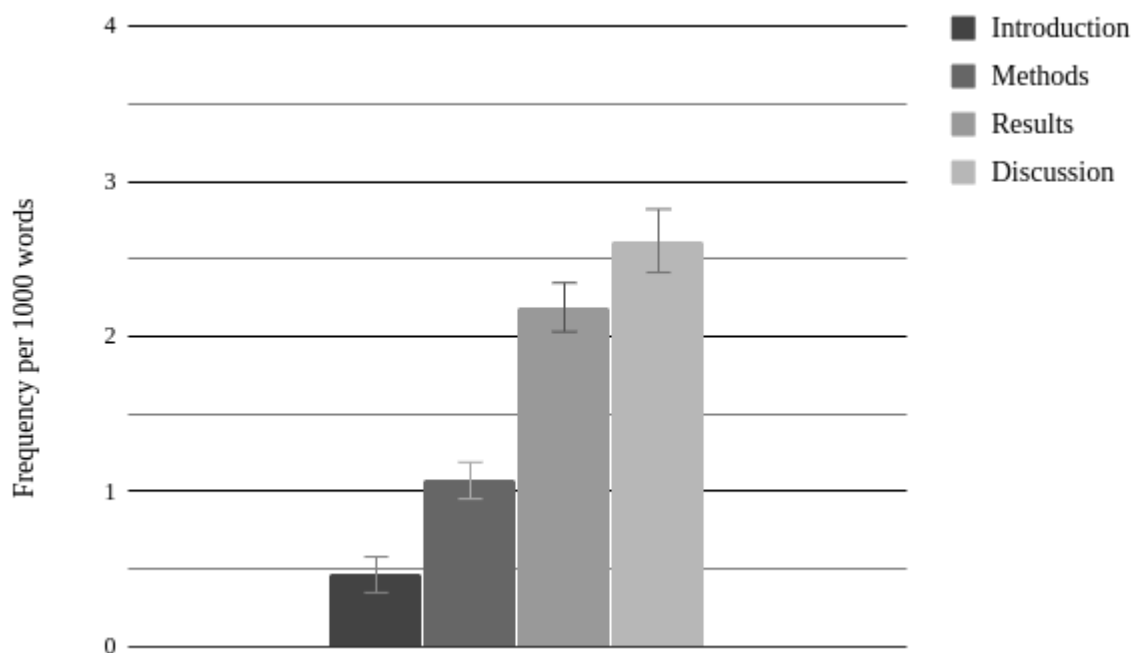
Table 10: Lexical bundles subtypes that incorporate verb phrase fragments (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
Non-passive	0.52	0.99	1.90	2.10
Passive	0.25	3.61	1.70	0.78
Other VP	0.27	0.36	0.26	0.15

4.1.1.2.1 Non-passive voice lexical bundles

The normalized frequency related to corpus size reveals that the Discussion and Results subcorpora present a dramatic difference in the use of non-passive voice bundles in comparison to the other subcorpora.

Graph 4: Normalized frequency and confidence intervals of non-passive voice lexical bundles



The subcorpora display a diverse frequency of lexical bundles with non-passive voice structures. The findings revealed that the non-passive structures in the Methods subcorpus

communicate the active role that participants and the tests play by employing the verbs BE, PARTICIPATE, TAKE PLACE, RATE and HAVE, always in the past simple. On the other hand, the Results subcorpus demonstrates a massive use of the existential pronoun THERE + copula BE as the main verb, and verbs such as SHOW, DIFFER, REVEAL, and DEMONSTRATE, predominantly in the past simple. Conversely, the Discussion subcorpus presents a greater use of bundles with modal verbs or present simple verbs, most of which conveying stance. This section is where authors need to negotiate with their peers, hence the use of modalized sentences.

Table 11: Lexical bundles containing non-passive structures occurrences

Introduction	Methods	Results	Discussion
play(s) an important role	participated in the study	there was no significant (group)	(is/are) in line with (the)
(to) shed light on the	were native speakers of	there was a significant difference (between (the)/ in (the)	(may) be due to the
there has been a	ranged in age from X to Y	there was a significant effect (of)	be the case that
is more likely to	that appeared in the	did not differ significantly (from)	are more likely to
plays a role in	the study took place	test showed that there	was not the case
this study was to	appeared on the screen	showed a significant main effect (of)	it may be that
et al found that	data collection took place	was not statistically significant	this suggests that the
investigated the effects of	participated in this study	table shows the descriptive (statistics)	(our/the) results show that (the)
studies have investigated the	rated on a point	did not have a	studies have shown that
this study is to	had more than one	model demonstrated that the	these findings suggest that

4.1.1.2.2 Passive voice lexical bundles

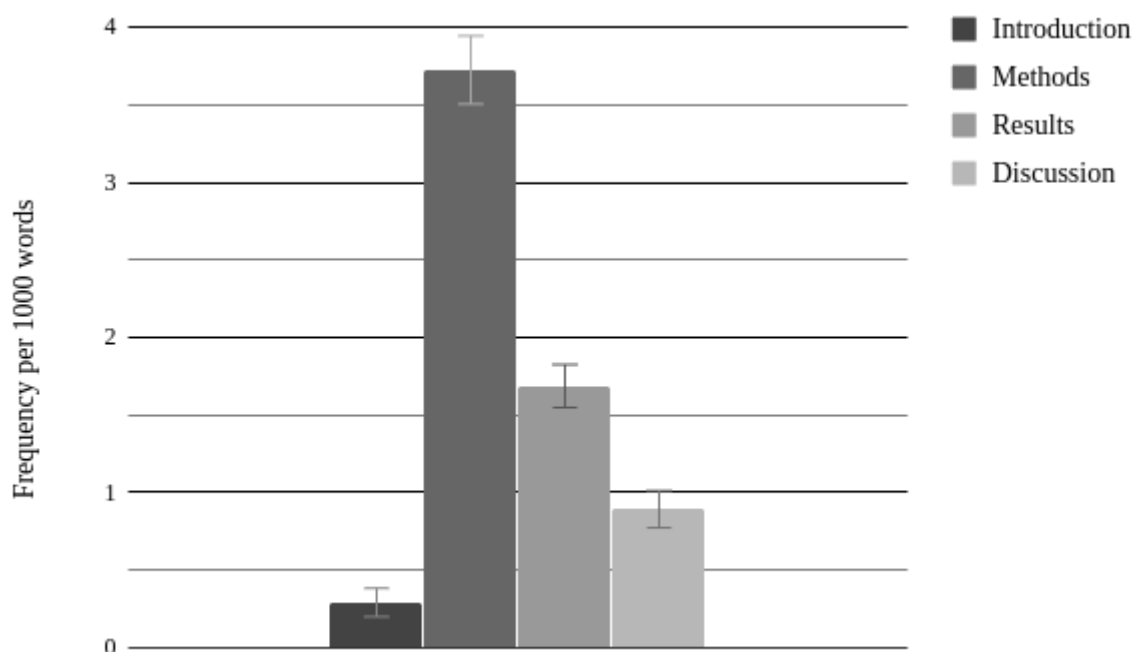
The Methods subcorpus presents a considerably high frequency of lexical bundles with passive voice constructions, because agentives are realized by the method rather than by the protagonists (SWALES, 1990), such as the participants and procedures, as an illustration, *participants were instructed to, data were collected from, used in this study*. The Results subcorpus also displays a high use of passive constructions. These constructions were thoroughly analyzed, and they revealed an interesting pattern. In addition to corroborating

with Biber et al. (1999) who identified that tabular/graphic displays of data are marked by passive voice verbs and a preposition, we also observed that no present simple passive voice bundles carries out any other function but the reference to tables, figures or appendices in the Methods and Results subcorpora, e.g. *are shown in table (the)* and *are shown in the appendix*.

It is worth pointing out, however, that the bundles with passive voice construction in the present simple in the Introduction and Discussion subcorpora play other general roles, such as indicating a gap in the literature (Introduction), e.g. *little is known about*, which is also communicated by present perfect devices, e.g. *attention has been paid*, and commenting on results and recommending further research in Discussion, e.g. *students are expected to*, *is supported by the*, and *research is needed to (investigate)*¹³.

While the Discussion subcorpus generated three instances of passive voice constructions in the present simple, the most frequently used passive devices are those containing a modal verb, such as *it should also be (noted)*, *it could be argued that*, and *can be attributed to*, all of which clearly belong to the Stance expression category to be discussed in the following section.

Graph 5: Normalized frequency and confidence intervals of lexical bundles containing verb phrases with passive voice structures



¹³Refer to rhetorical moves and steps in Swales (1990, 2004), Cortes (2013), and Ruiying and Allison (2003) for further information.

Again, the Introduction section presents the lowest frequency of use within the category of verb phrases. The image above reveals an extremely modest frequency in the use of lexical bundles with passive voice in Introduction. This finding challenges previous studies which overgeneralize the use of non-passive and passive voice in the academic discourse, e.g. Tarone et al., 1998, Hyland, 2002, and Biber et al., 1999. For instance, Biber et al. (1999, p. 1020) claim that “[o]nly a few lexical bundles in academic prose are built around a verb phrase, and the majority of these lexical bundles incorporate a passive voice verb”.

In similar fashion, Hyland (2002) goes further to claim that humanities and social science papers present a higher frequency of non-passive structures since these areas tend to hold a stronger identity by not downplaying their personal role when highlighting issues under study.

At the same time, another study¹⁴ (HESLOT, 1982), in line with the findings of the present research, has postulated that the Introduction has a low use of passive voice, but Methods has a high frequency, while Results and Discussion present variable levels of this grammatical structure in 16 RAs from the journal *Phytopathology*. We conclude therefore that different frequency levels in the use of passive and non-passive voice should be specifically related to sections of RAs, not to an entire register or fields of study¹⁵. The bold typed and italic bundles are found in more than one subcorpus:

¹⁴Biber and Finegan (1994) by analyzing 20 medical RAs also conclude that Methods sections present a high frequency of passive constructions, while the other sections do not present a considerable variability.

¹⁵All non-passive and passive lexical bundles from the whole corpus total 42 and 38 per cent.

Table 12: Lexical bundles containing passive structures occurrences

Introduction	Methods	Results	Discussion
(has/have) been shown to be	participants were asked to (complete/ indicate/ read/ select/ write)	as shown in table (the)	it should be noted (that)
can be used to	they were asked to	as can be seen (from the/ in the)	used in this study
little is known about	participants were instructed to	are presented in table	more research is needed (to investigate)
attention is drawn to	the data were collected	are summarized in table (the)	was found to be
attention has been paid	participants were presented with (a)	can be found in	as indicated by the
	and were asked to	(no) significant differences were found (between)	be argued that the
	were used in the	as shown in figure/fig	be attributed to the
	has been shown to (be)	it should (also) be noted (that the)	can be seen as
	were randomly assigned to	were used in the	found in this study
	students were asked to	be explained by the	be related to the
	were included in the		can be used to

It is worth noting that Tables 8 and 9, containing lexical bundles with PP and NP fragments, show more similarities of devices used across the subcorpora than Tables 11 and 12, which present the occurrences containing VP fragments (see boldtyped devices). These tables reveal that there seem to be more lexical diversity with VP than with NP.

4.1.1.3 Dependent clause fragments

The last subtype of structural category concerns dependent clause fragments, which encompasses “that-clauses”, “anticipatory it + adjective”, ‘to-clause’ and “WH-fragments”¹⁶. From those occurrences, we decided to scrutinize the structure “anticipatory it + adjective fragments”.

¹⁶However, a table of findings regarding the other subtypes and statistical tests can be found in Appendix A and C.

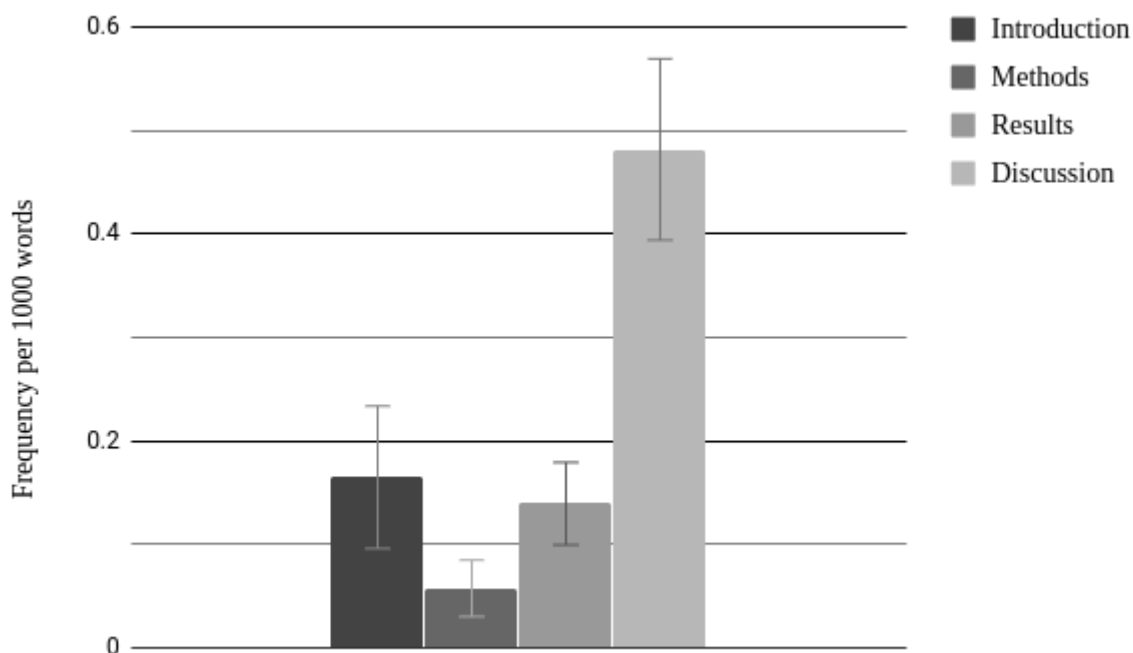
Table 13: Lexical bundles subtypes that incorporate dependent clause fragments (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
that clauses	0.10	0.18	0.61	0.48
It (be) + adj.	0.16	0.06	0.14	0.48
to-clauses	0.19	0.58	0.40	0.53
WH-frag.	0.07	0.08	0.09	0.05

4.1.1.3.1 Anticipatory it + adjective fragments

The subcorpora of Introduction, Methods and Results do not display a significant difference in the use of “anticipatory it + adjective fragments”, as revealed below by the error bars, but the Discussion subcorpus has the highest number of occurrence of this structure. “Anticipatory it + adjective” reports “the stance of the writer; for example, possibility/likelihood, importance, necessity” (BIBER et al., 1999, p. 1018). Not surprisingly, the Discussion section is expected to present a greater use of this structure. It is in the Discussion section that authors interpret the significance of their findings and explain new understanding or insights, hence resorting to possibility, importance, and necessity stances.

Graph 6: Normalized frequency and confidence intervals of “*anticipatory it + adjective*” lexical bundles



The table below presents all the “anticipatory it + adjective” bundles. *It is important to* is found in all the subcorpora. In addition to possibility/likelihood, importance, necessity claimed by Biber et al. (1999), “anticipatory it + adjective” bundles across the Applied Linguistics RAs sections convey other stances, such as in *it is true that*, *it is clear that*, *it is difficult to*, and *it is reasonable to*. The bold typed and italic bundles are found in more than one subcorpus.

Table 14: Anticipatory it + adjective fragments occurrences

Introduction	Methods	Results	Discussion
<i>it is important to</i>	<i>it is important to</i>	<i>it is important to (note that)</i>	<i>it is possible that (the)</i>
it is also important	it was possible to	it is interesting to note that	<i>it is important to (note/ bear in mind that)</i>
it is difficult to	it was not possible to	<i>it is likely that</i>	<i>it is clear that</i>
<i>it is possible that</i>		<i>it is difficult to</i>	it is possible to
		<i>it is possible that</i>	(it) is not possible to
		it is true that	<i>it is difficult to</i>
		<i>it is clear that</i>	<i>it is worth noting (that)</i>
		<i>it is worth noting</i>	it is also possible (that)
			<i>it is likely that</i>
			it is necessary to
			it is not clear
			it is reasonable to

In this section, we presented the main findings regarding the use of lexical bundles sorted into structural types. Firstly, there is an outstanding difference in frequency of lexical bundles generated, Methods and Results present almost twice as much in comparison with the other subcorpora. Secondly, unlike previous studies (BIBER et al., 1999, 2004; BIBER, 2010) the current research revealed that dependent clauses are used less frequently than verb phrase fragments. Thirdly, the Methods subcorpus revealed a close relationship between noun phrase with of-fragments with quantity specification (a functional subtype). Additionally, the “anticipatory *it* + adjective” is closely related to Stance expressions and it is very commonly used in the Discussion section. Finally, the use of lexical bundles with passive and non-passive verbs are considerably unbalanced, especially regarding the Methods and Introduction

subcorpora. We could also identify that specific verbs are used exclusively in passive or non-passive constructions according to the role they play in the sections.

Given all these distinctions in structural types across the IMRD corpus, it is also expected to encounter bundles with distinctive discourse functions. In the following section, we present the investigation of functional categories and subtypes.

4.1.2 Functional types of lexical bundles

The functional categories of lexical bundles comprehend the so-called Referential expressions, Stance expressions, and Discourse organizing functions. From the lexical bundle tokens classified, the greatest share is made up of Referential expressions. The considerable amount of this category is a feature of academic prose (BIBER et al., 2004; SIMPSON-VLACH; ELLIS, 2010; DUTRA & BERBER SARDINHA et al., 2013).

In the Methods subcorpus, Stance expressions are the least occurring subcategory of functional types in the Introduction and Methods subcorpora, 14 and 15 per cent, while Referential expressions represent a great deal of them. Discourse organizing functions are very little used as well, 19 per cent. The functional types in the Results subcorpus display an interesting pattern: although most of the types are constituted by Referential expressions, Stance expressions also represent a considerable amount of functional types, 32 per cent. Finally, from the Discussion subcorpus, the findings below show that there is a somewhat balance across the subcategories in the Discussion subcorpus.

Table 15: Major categories: functional types of lexical bundles (percentage)

	Introduction	Methods	Results	Discussion
Referential expressions	62%	66%	49%	40%
Discourse organizing functions	24%	19%	19%	29%
Stance expressions	14%	15%	32%	31%
Total	100%	100%	100%	100%

What follows is an investigation of the subtypes belonging to the major categories described above.

4.1.2.1 Referential Expressions

A closer look at the Referential expressions in all subcorpora, except for Methods, reveals that Intangible framing attributes constitute the majority of this category. Intangible framing devices are important academic phrases (SIMPSON-VLACH; ELLIS, 2010, p.17), and include “phrases that frame both concrete entities and abstract concepts or categories”, see Table 14. The Intangible Framing subcategory is also claimed to be the largest pragmatic category within the specification of attributes in the Referential Expressions, which also includes Tangible Framing Attributes and Quantity Specification (SIMPSON-VLACH; ELLIS, 2010).

The Methods subcorpus shows that Quantity Specification devices represent the highest frequency of the analyzed bundles. Similarly, Deictic and Locative devices are very frequently used in Methods. In the Results subcorpus, the subcategories of Referential Expressions also show a pattern worth pointing out. Similar to the other subcorpora, intangible framing attributes and Quantity Specification bundles make up most of the bundles from this subcategory. Nevertheless, the frequency of Contrast and Comparison bundles (e.g. *between the two groups, in contrast to the*) is considerably high. It is expected, therefore, to find this feature in Results subcorpus, since it is where authors contrast and compare with results found or results from previous studies. See extracts below from two different texts from the Results subcorpus:

[1 LRSR7] ***On the other hand**, respondents who were born in the US or immigrated to the US before or at the age of 5 (Group 2) rated substantially higher in their English proficiency and English use frequency than those who immigrated after the age of 5 (Group 1).*

[2 LLRS9] ***At the same time**, there was a wide range of text length within most course types and grade levels, as can be seen in standard deviations listed in the table.*

Finally, despite not representing a great proportion within the Referential expressions bundles, Quantity Specification and Contrast and Comparison also play a role in the Discussion section.

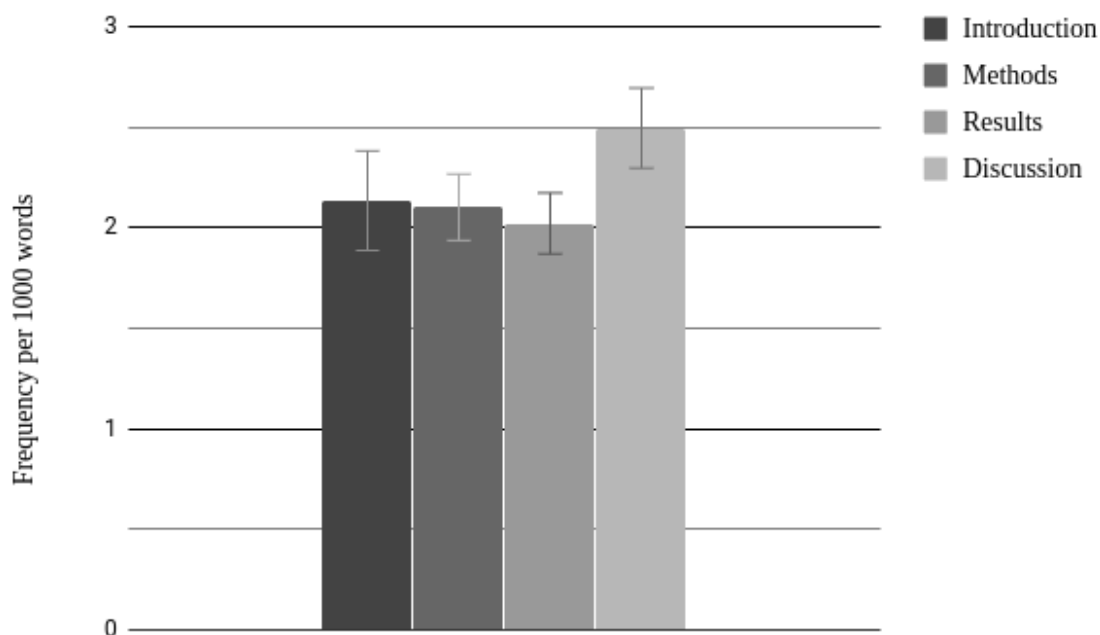
Table 16: Lexical bundles subtypes: Referential Expressions (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
Intangible framing attributes	2.13	2.10	2.02	2.50
Quantity specification	1.30	2.97	1.70	0.93
Deictics and locatives	0.50	2.05	0.68	0.31
Identification and focus	0.24	0.36	0.27	0.33
Contrast and comparison	0.23	0.45	1.43	0.81
Tangible framing attributes	0.20	0.28	0.14	0.13

4.1.2.1.1 Specification of attributes: Intangible framing attributes

As already mentioned, the Introduction subcorpus, unlike the others, does not present any significant difference regarding a greater use of structural or functional types. On the contrary, it mostly displays the lowest frequency of categories and subtypes of the investigated devices, or no statistically significant difference at all. The analysis that presented something a little peculiar was the frequency of Specification of attributes: Intangible framing attributes (Graph 7). This subtype belongs to the Referential expressions set and includes phrases that frame both concrete entities and abstract concepts or categories (SIMPSON-VLACH; ELLIS, 2010). The Introduction subcorpus shows a slight higher frequency of it in relation to Methods and Results, but not to Discussion. However, with the normalized frequency and the confidence interval applied, see Graph 7, the difference across the subcorpora is not considered statistically significant.

Graph 7: Normalized frequency and confidence intervals of specification of attributes:
Intangible framing attributes (Referential expressions)



As can be seen in Table 17, from the most frequent instances, a substantial amount of identical lexical bundles is used across the sections. A relevant amount is composed of “a/the N of”, or “NP with of-fragments”, as in Simpson-Vlach and Ellis (2010) list, and most of them frame an attribute of a following noun phrase. Taking into account the frequency (Graph 7) and the bundle types (Table 15), we can assume that Intangible framing attributes do not present a significant distinction neither in frequency nor in use across the subcorpora.

Table 17: Intangible framing attributes occurrences

Introduction	Methods	Results	Discussion
<i>the extent to which (the)</i>	<i>on the basis of (the/their)</i>	<i>in the context of</i>	<i>in the context of (the)</i>
<i>on the basis of (the)</i>	the analysis of the	<i>in the case of (the)</i>	<i>in the case of (the)</i>
<i>the ways in which</i>	<i>in the form of (a)</i>	<i>in terms of the</i>	<i>the use of the</i>
<i>in the case of</i>	<i>over the course of (the)</i>	<i>(in) the use of the</i>	<i>the extent to which</i>
as a function of (the)	<i>in terms of the</i>	<i>over the course of (the)</i>	<i>in terms of the</i>
to the development of	<i>the meaning of the</i>	with regard to the	to the fact that (the)
<i>in the context of</i>	<i>the use of the</i>	<i>in relation to the</i>	<i>on the basis of (the)</i>
in the process of	<i>the extent to which (a)</i>	<i>on the basis of (the)</i>	<i>with respect to the</i>
<i>in the use of</i>	<i>in the context of (a/the)</i>	the course of the	<i>the way(s) in which</i>
on the development of (the)	<i>the content of the</i>	<i>in the form of (a)</i>	<i>in relation to the</i>
<i>in the form of</i>	in terms of their	<i>the extent to which (the)</i>	<i>over the course of</i>
with a focus on	the rest of the	as a function of	in the absence of
the degree to which	over a period of	<i>the content of the</i>	<i>in the form of</i>
<i>with respect to the</i>	an overview of the	<i>the meaning of the</i>	the nature of the

4.1.2.1.2 Specification of Attributes: Quantity Specification

In the subcategory of bundles conveying Quantity and Specification, as illustrated in the graph below, there is significant difference in frequency across all the subcorpora. The Methods subcorpus, as expected, presents a much higher use of this type of bundle. This distinction might be due to the fact that it is in the Methods subcorpus, where one can find information on the quantity of participants, questions, tests, length of experiments, etc.

Graph 8: Normalized frequency and confidence intervals of specification of attributes: quantity specification (Referential expressions)

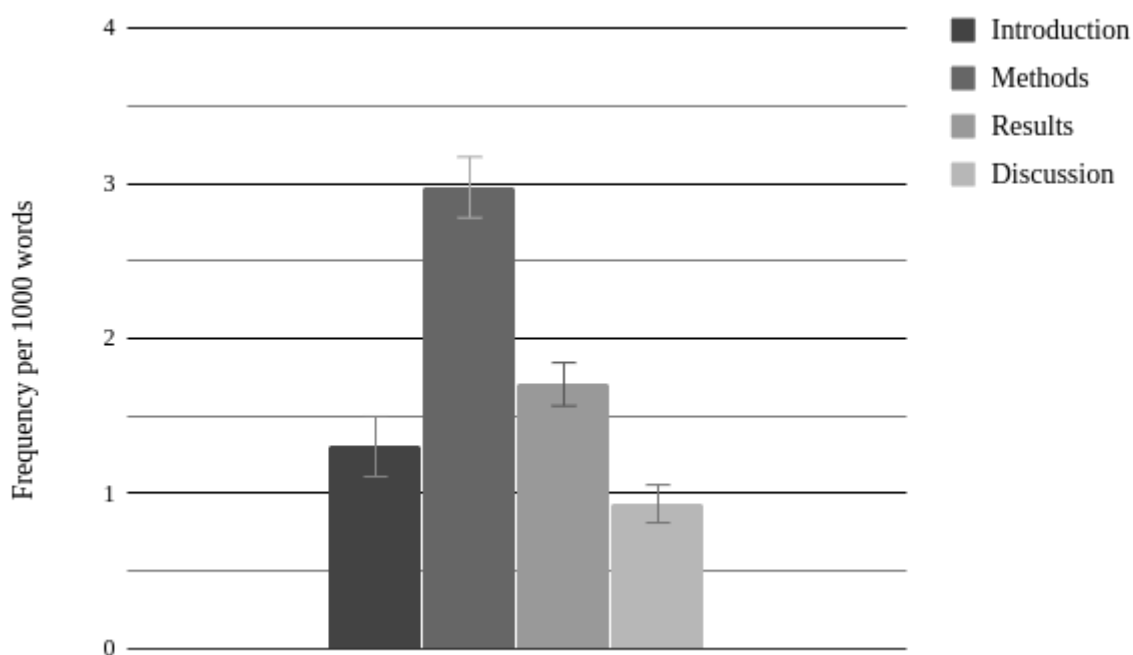


Table 18 reveals a variability in the structures of bundles conveying Quantity specification across the sections. Just a few instances are found within the most frequently employed lexical bundles, such as *a wide range of*, *a small number of*, *each of the three*. As already mentioned, a great deal of lexical bundles of Quantity specification belong to the structural type “NP with of-fragments”. The bold typed and italic bundles are found in more than one subcorpus.

Table 18: Quantity specification occurrences

Introduction	Methods	Results	Discussion
a number of studies (have)	(by) the total number of (participants/ words)	<i>each of the three</i>	the second research question
<i>a wide range of</i>	for each of the	a large effect size	the first research question
is one of the (most)	by the first author	the majority of the	<i>a wide range of</i>
a large number of	<i>each of the three</i>	the mean number of	a great deal of
(extensive/ growing/ considerable) body of research on	the majority of the (participants)	the total number of	a greater number of
little is known about	<i>a wide range of</i>	the second research question	<i>a small number of</i>
of a number of	one of the two	the first research question	in their first year (of)
both teachers and students	(and) the number of words	the mean percentage of	in the number of
little research has been	the second research question	of the two groups	as one of the
the full range of	one of the three	<i>a small number of</i>	to a greater extent than
in the two languages	the mean length of	<i>a wide range of</i>	that many of the
of a range of	in each of the	with a large effect size	a higher degree of

4.1.2.1.3 Contrast and Comparison

The frequency of contrast and comparison lexical bundles are much higher in the Results subcorpus than in the other ones. As already stated in this study, the Results subcorpus usually offers contrast and comparison between findings of the study itself or previous studies. Some of the most frequently employed bundles are *on the other hand (the)*, *significant difference between(s) the (two)*, *in contrast to the*, and *in the same way*.

Table 19 shows that, while the Introduction subcorpus displays a modest quantity of contrast and comparison bundles, from which all devices work as linking adverbials, the other subcorpora present lexical bundles whose general attribution is to participants, tests, or results specifically. The bold typed and italic bundles are found in more than one subcorpus.

Graph 9: Normalized frequency and confidence intervals of contrast and comparison (Referential expressions)

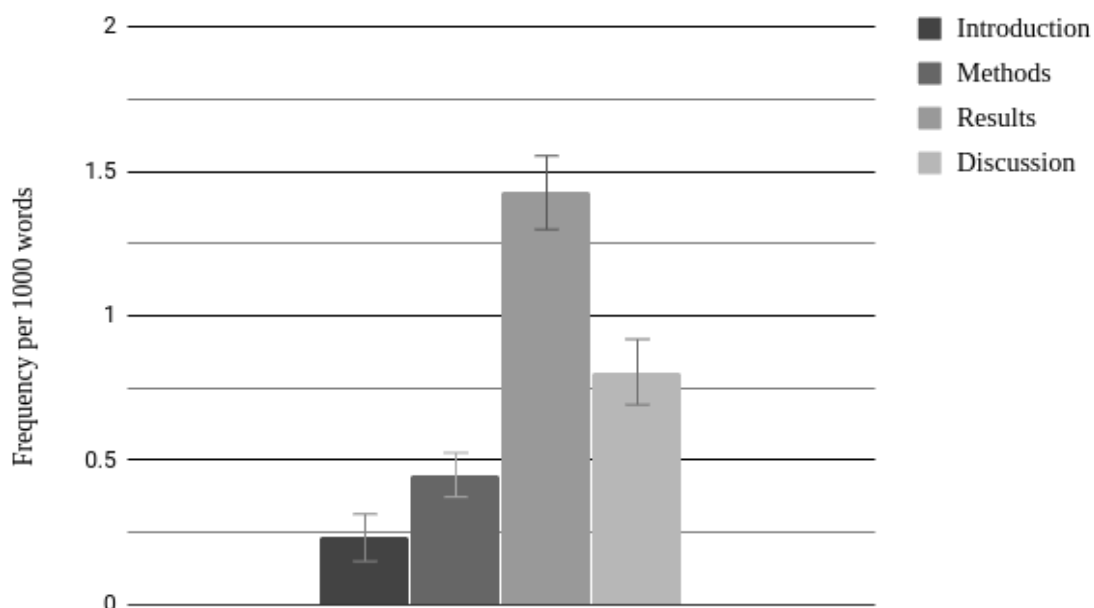


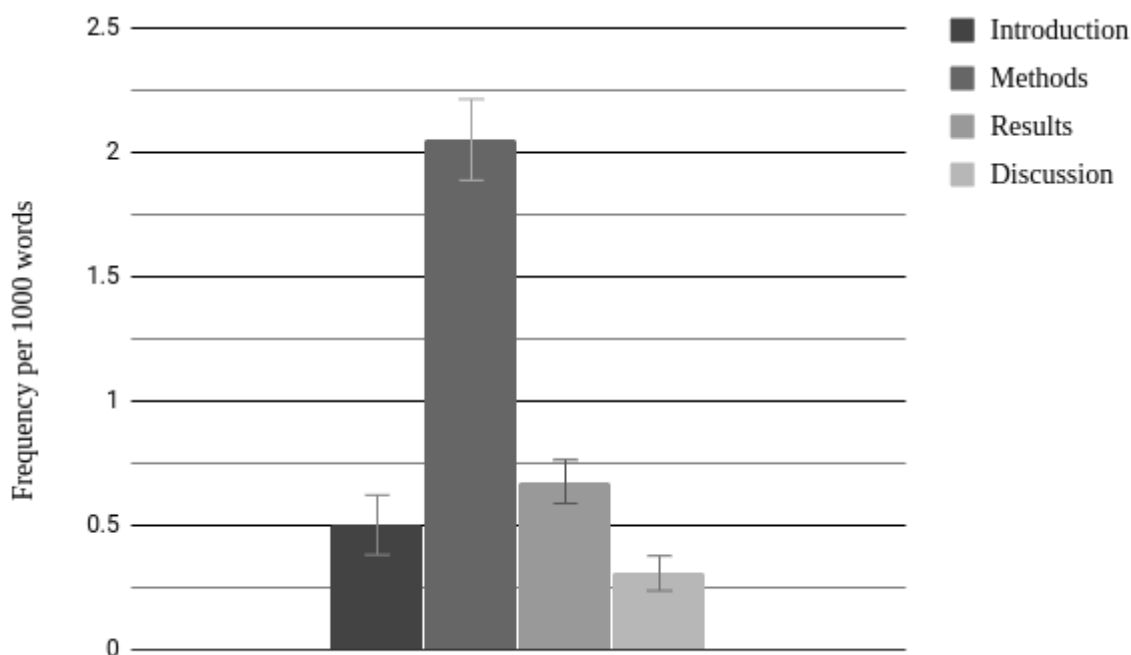
Table 19: Contrast and comparison occurrences

Introduction	Methods	Results	Discussion
<i>on the other hand</i>	difference(s) between the two	<i>on the other hand (the)</i>	<i>on the other hand (the)</i>
<i>on the one hand</i>	the relationship between the	between the two groups	<i>on the one hand (the)</i>
<i>in contrast to the</i>	other writing change functions	<i>in contrast to the</i>	the difference between the (two)
in the same way	<i>in contrast to the</i>	in the same way	be attributed to the
	to be related to	between the two variables	there were no significant (differences)
	the same as the	significant difference between(s) the (two)	be related to the
	languages other than english	the difference between the (groups)	there were differences in
	associated with the target	interaction between the two	<i>in contrast to the</i>
	the same procedure was	other writing change functions	the link between the
	differed significantly from the	relationship between the two	as opposed to the

4.1.2.1.4 Deictics and locatives

Following the same pattern, Deictics and Locatives are considerably more common in the Methods sections than in the others. Because they are devices used to express details about the experiment procedures, such as in *at a university in, an English speaking country, in the original text, at the end of (the semester)*. These types of bundles, therefore, are considerably prevailing in the Methods subcorpus.

Graph 10: Normalized frequency and confidence intervals of deictics and locatives (Referential expressions)



As can be seen in Table 20, lexical bundles conveying Deictics and Locatives do not differ in structure across the sections. Their core, nouns between prepositions and articles, are quite the same, e.g. *end, beginning, United States, time*, etc. The bold typed and italic bundles are found in more than one subcorpus.

Table 20: Deictics and locatives occurrences

Introduction	Methods	Results	Discussion
<i>the end of the</i>	<i>at the beginning of</i>	<i>(at) the end of the (semester)</i>	<i>in the United States</i>
<i>in the US</i>	<i>the end of the semester</i>	<i>(at) the beginning of the (semester/ study/ year)</i>	at the end of (the/ their)
(arrived) <i>in the United States</i>	at a university in	<i>in the United States</i>	<i>the end of the semester</i>
where English is the	the start of the	<i>in the US</i>	<i>in the US</i>
	the onset of the	at the time of (the)	(at) the beginning of the
	the middle of the	at the start of (the)	the location of a
	in the source text	written at the beginning	at the time of
	the position of the	by the end of (the)	and outside the classroom
	from the source texts	time of the study	in the classroom and
	at a US university	parts of the world	
	in a quiet room	the end of this	

4.1.2.1.5 Identification and focus/ Tangible framing attributes

With the normalized frequency and confidence interval (see Appendix A), we found that the use of identification and focus lexical bundles throughout the IMRD subcorpora is balanced with no significant difference.

Examples of Identification and focus devices:

- **Introduction:** *that there is a, as a type of, and there has been a,*
- **Methods:** *in the case of (the), that the use of, and of different types of,*
- **Results:** *that none of the, an example of the, and test showed that there were,*
- **Discussion:** *as a resource for, as a tool for, and as indicated by the.*

It is interesting to highlight that in the AFL (SIMPSON-VLACH; ELLIS, 2010), Identification and focus devices are more frequent than Contrast and comparison and Deictics and locatives, on the grounds that exemplification and identification (identification and focus) are basic pragmatic functions in academic writing. This pattern however does not replicate in this study. As can be seen in Table 14, other subcategories are much more frequent than Identification and focus, such as Intangible framing attributes, Quantity and specification, and Contrast and comparison (the latter in the Results and Discussion subcorpora only). This

might be explained by the fact that when sections of Research Articles are analyzed, clear differences appear. The communicative function of each RA section is different. Therefore, the bundles are different as well.

Another subtype that belongs to specification of attributes is tangible framing attributes. This subtype “refer[s] to physical or measurable attributes” (SIMPSON-VLACH; ELLIS, 2010, p. 18), as an illustration, our corpus generated the following:

Tangible framing attributes:

- **Introduction:** *over the course of (the year/ a semester),*
- **Methods:** *as part of the,*
- **Results:** *at the level of,*
- **Discussion:** *the frequency of the.*

The Methods subcorpus displays a slight difference in comparison to the Results and Discussion sections considering the confidence interval of this subtype (see Appendix A).

4.1.2.2 Discourse organizing function

Discourse organizing function devices play an important role in the text. They reflect the connection between prior and coming discourse (Biber et al., 2004). This category comprehends the subtypes: Metadiscourse and textual reference, Topic introduction and focus, Topic elaboration: non-causal, Topic elaboration: cause and effect, and Discourse markers.

Discourse organizing functions bundles in the whole corpus are mostly represented by Metadiscourse and textual references. The substantial amount of Metadiscourse and textual reference bundles corroborates with Simpson-Vlach and Ellis (2010), who also found that this is the most occurring subcategory from Discourse organizing functions in academic prose.

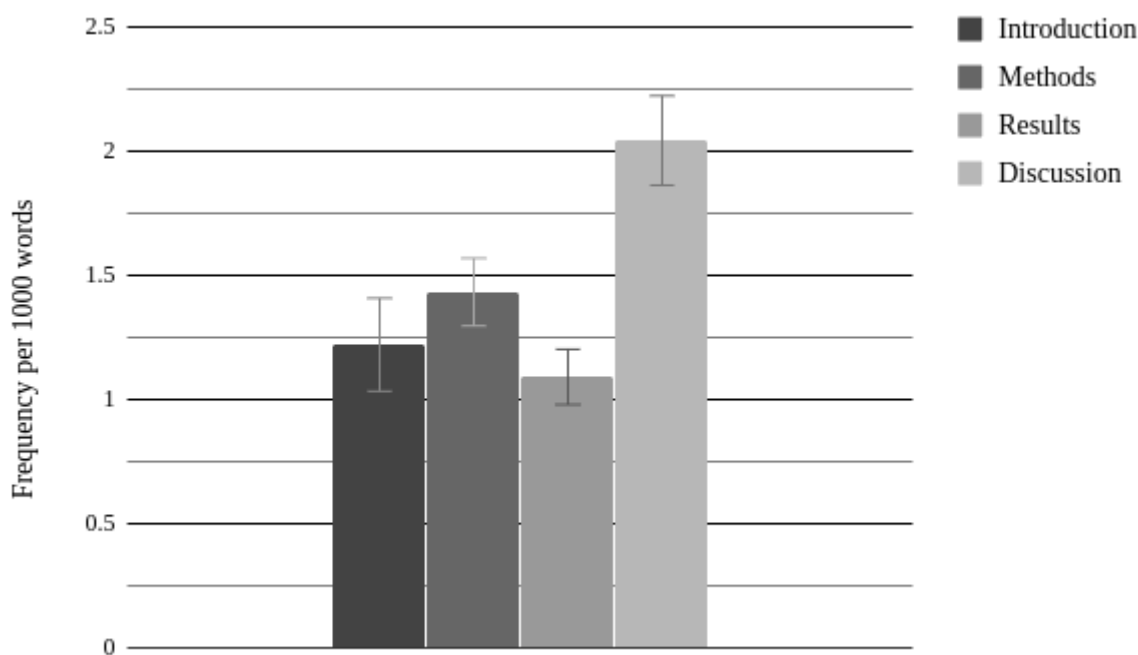
Table 21: Lexical bundles subtypes: Discourse organizing functions
(frequency per 1000 words)

	Introduction	Methods	Results	Discussion
Metadiscourse and text. ref.	1.22	1.43	1.09	2.04
Cause and effect	0.31	0.54	0.95	1.27
Discourse markers	0.28	0.32	0.33	0.39
Topic introduction and focus	0.00	0.06	0.09	0.00

4.1.2.2.1 Metadiscourse and textual reference

The graph below displays the normalized frequency of Metadiscourse and textual reference with confidence interval represented by the error bars. The Discussion subcorpus is the one with the highest use of this subtype. The other subcorpora do not present a statistically significant difference.

Graph 11: Normalized frequency and confidence intervals of metadiscourse textual reference (Discourse organizing functions)



The table below shows a pattern in the use of bundles belonging to the subcategory of Metadiscourse. Most of them contain the word *study*. This might not be surprising, considering we are dealing with academic discourse. Moreover, this subcategory does not present a variety in use within each subcorpus, i.e. the ratio: bundle types/ bundle tokens is somewhat the same, Introduction 0,14, Methods 0,14, Results 0,13 and Discussion 0,12. If the Discussion shows a significantly higher frequency of this subcategory, we can assume that certain bundles are preferred over others.

We found that the bundles *in the present study* and *in the current study* occur 54 and 47 times in the Discussion subcorpora. These bundles are by far the most employed ones, and are found in all the subcorpora. In the Discussion, the high use of these bundles perhaps signals a preference for these constructions when presenting the discussion while referring back to present findings or relating to previous research.

Another interesting feature of this subcategory is its frequency in the Methods subcorpus. As mentioned in the Literature review chapter, Method paragraphs might be characterized as *broken linear* or containing sentences as if they were islands in a string (SWALES, 1990). Methods sections like this are usually found in the physical and life sciences. They are claimed to be “enigmatic, swift, presumptive of background knowledge, not designed for easy replication, and with little statement of rationale or discussion of the choices made” (SWALES, 1990, p. 170).

On the other hand, “softer”, emerging or interdisciplinary fields tend to deal with given and new information more cohesively, supported by anaphoric reference and lexical repetition (SWALES, 1990). The combination of Corpus Linguistics and the analysis of lexical bundles can contribute by revealing that this information cohesiveness is also supported by Metadiscourse and textual reference devices similarly across the Introduction, Methods and Results subcorpora, with different lexical choice but rather similar frequency of use in Applied Linguistics RAs. The bold typed and italic bundles are found in more than one subcorpus.

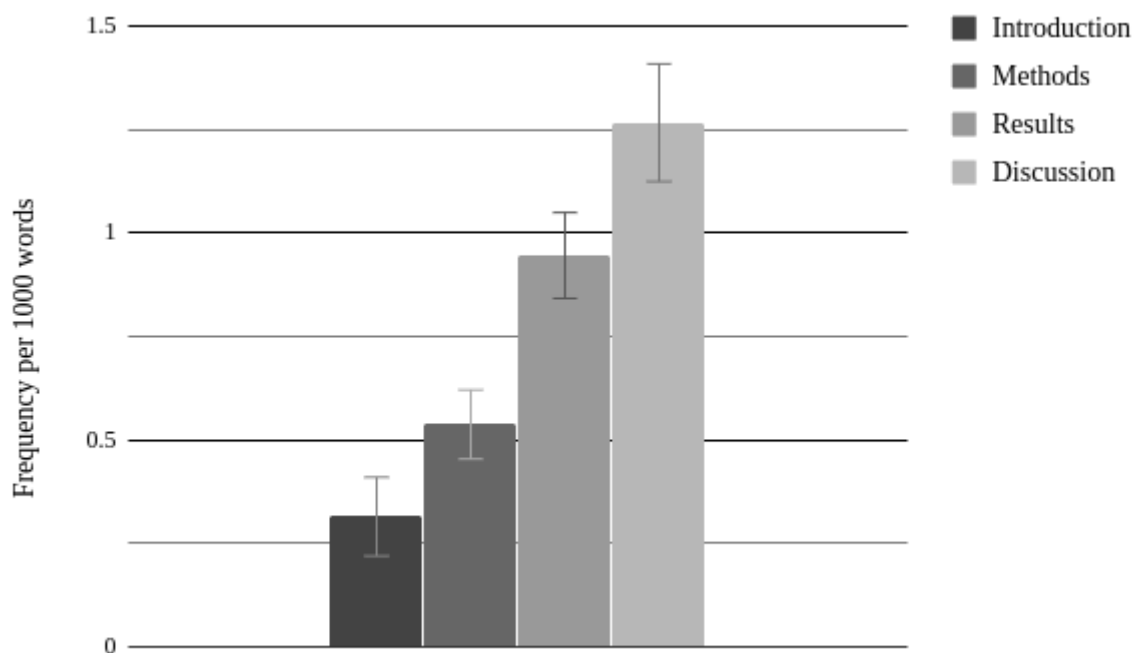
Table 22: Metadiscourse and textual reference occurrences

Introduction	Methods	Results	Discussion
the present study is	<i>in the present study</i>	as shown in table (the)	<i>in the present study</i>
within the field of	<i>in the current study</i>	are shown in table (the)	<i>in the current study</i>
in the next section (we)	as shown in table/ figure	are presented in table	of the present study
<i>In the present study</i> we	in this study were	in the supporting information online	in this study the
the current study is	for the present study	as shown in figure	of the current study
in this study we	in this study is	as shown in fig	participants in this study (were)
the present study was	in this study the	in the post test	in this study were
in this article we	the following research questions	in the test set	of this study was (to)
(of) this study was to	for the current study	are summarized in table (the)	in this study we
In this article I	in this study was	<i>in the present study</i>	from the current study
in this paper we	of the current study	in the source text	this study did not
the following research questions	of the present study	as indicated by the	findings of the study
this study is to	in the study were	in the following section(s)	the third research question
of the source text	included in this study	in the next turn	found in this study
of the study was to	for this study the	in the following excerpt	in the current study we

4.1.2.2.2 Topic elaboration: cause and effect

Cause and effect bundles express reason, effect, or causal relationship (SIMPSON-VLACH & ELLIS, 2010). The trend shown in the figure below seems to be quite interesting. There is a gradual increase in the use of Cause and effect lexical bundles across the IMRD subcorpora. The Discussion subcorpus reveals a great number of bundles from this subcategory. There is a modest difference from one subcorpus to another, but it is all statistically significant, as the error bars signal.

Graph 12: Normalized frequency and confidence intervals of topic elaboration: Cause and effect (Discourse organizing functions)



As can be seen in the Table 23, bundles containing the element *result(s)* are the most frequently encountered expressions in the subcategory of Topic elaboration: Cause and effect. However, it is also important to note that expressions, such as *the purpose of the*, *in order to address/avoid/examine*, *the findings of the* are also considerably employed across the sections. The bold typed and italic bundles are found in more than one subcorpus.

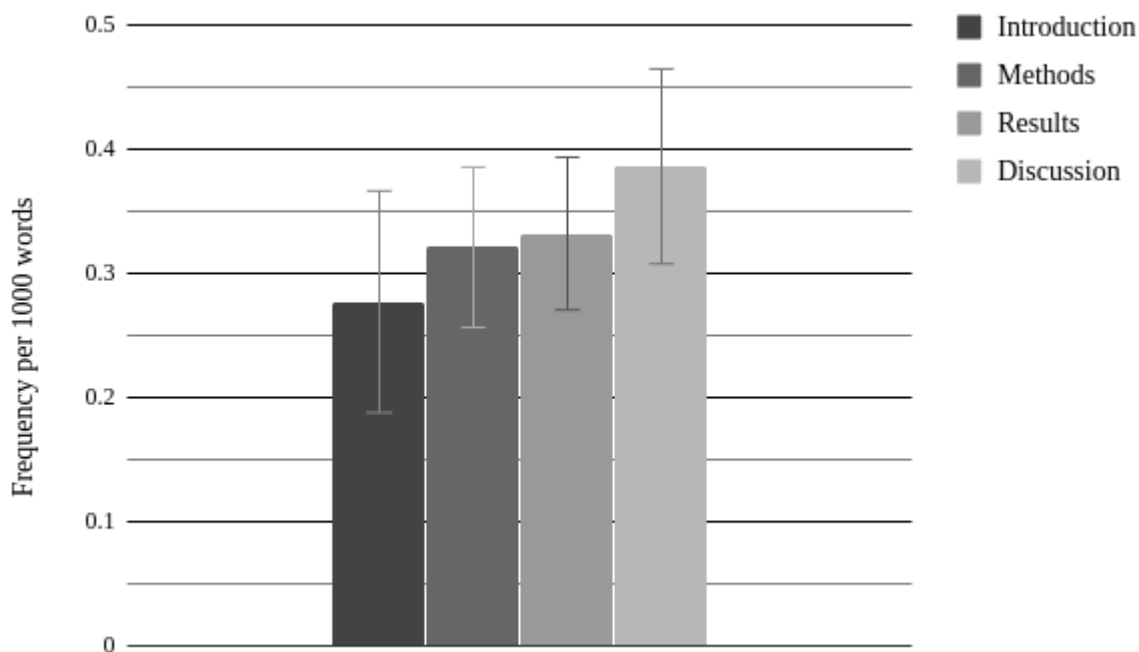
Table 23: Topic elaboration: cause and effect occurrences

Introduction	Methods	Results	Discussion
the aim of the	in order to address/ avoid/ examine/ facilitate/ gain/ make/ obtain/ provide	<i>the results of the</i>	<i>the results of the</i> (study/ present study)
the purpose of the	for the purpose(s) of (the/ this)	the main effect of	the fact that the
as a result of	the purpose of the (study)	accounted for of the variance (in the)	as a result of
investigated the effects of	the purpose of this (study)	the results for the	the results of this (study)
the goal of this	the results of the	the effect(s) of the	the findings of this study
the results of a	as a result of	as a result of (the)	the findings of the (study/ present study)
	the goal of the	the results from the (test set model)	be due to the
	to determine whether the	in order to test	at the expense of
	whether or not the	the results showed that	for the use of
	for this reason we	the results show that	for the development of
	due to the time	a main effect of	the results showed that (the)
	the results of a	to the fact that (they)	due to the fact that (the)
	to examine whether the	in this way the	as a result the

4.1.2.2.3 Discourse markers

In the academic written register, *discourse markers* work as connectives that signal transitions between clauses and constituents (SIMPSON-VLACH & ELLIS, 2010), for example, *at the same time* - Introduction subcorpus, *as well as the* in Methods. The image below shows no significant differences of this use across the subcorpora.

Graph 13: Normalized frequency and confidence intervals of discourse markers
(Discourse organizing functions)



4.1.2.3 Stance expressions

Stance expressions “express attitudes or assessment of certainty that frame some other proposition” (Biber et al., 2004, p. 384) and are used to present argumentation by expressing judgment and opinions (DUTRA; ORFANÓ; BERBER SARDINHA, 2014). The current study analyzed the frequency of four subtypes, namely evaluation, expressions of ability and possibility, hedges, and intention/ volition/ prediction. The most expressive finding regarding the use of Stance expressions is the frequency of evaluation and hedges in the Results subcorpus. It is interesting to note that a vast majority of the evaluation bundles, in Results, are to communicate the significance or non-relevance of statistical findings. This is sustained by Swales’s (1990, p. 171) proposition that the style and structure of the Results sections “seem to be designed to deny on the author’s part any associative contamination with commentary or observation”, so evaluation devices are overwhelmingly related to what statistical tests have shown (see Table 23).

On the other hand, from the bundles analyzed in the Discussion subcorpus, the most occurring subtype are hedges. According to Hyland (1999, p. 433), “hedges allow writers to anticipate possible opposition to claims by expressing statements with precision, caution, and

diplomatic deference to the views of colleagues”. Therefore, finding a high frequency of this type of bundle in the Discussion subcorpus should not be a surprise, given the fact that authors use this section to comment on results, which expresses the main communicative purpose of the Discussion section (YANG & ALLISON, 2003) and make new knowledge claims (BASTURKMEN, 2009).

Table 24: Lexical bundles subtypes: Stance expressions (frequency per 1000 words)

	Introduction	Methods	Results	Discussion
Evaluation	0.41	0.10	2.16	0.80
Exp. of ability and possibility	0.16	0.37	0.48	0.56
Hedges	0.15	0.04	0.85	1.51
Intention/ volition, prediction	0.00	0.03	0.04	0.00

The proportion of Stance expression subtypes across the corpus reveals that the main subtype in both the Introduction and Results subcorpora is evaluation; in Methods, expressions of ability and possibility constitute highest frequency; and in Discussion, hedges are more frequently utilized. Expressions of ability and possibility are represented in the Introduction subcorpus by *to be able to*, *can be used to*, and others. Two bundle types are identified as hedges in the Introduction subcorpus: *is more likely to* and *were more likely to*¹⁷. In this section, we present the most distinctive features of Stance expressions that emerged in the present analysis.

4.1.2.3.1 Evaluation

The most frequently employed bundles expressing evaluation were *it is important to*; *play(s) an important role*; *it is also important*.

[1 LLINTRO3] *In the interest of developing effective L2 pedagogy, **it is important to** examine whether the input might be structured in ways that might help learners to notice distributional cues to grammatical categories while considering other known variables that impact L2 learning trajectories.*

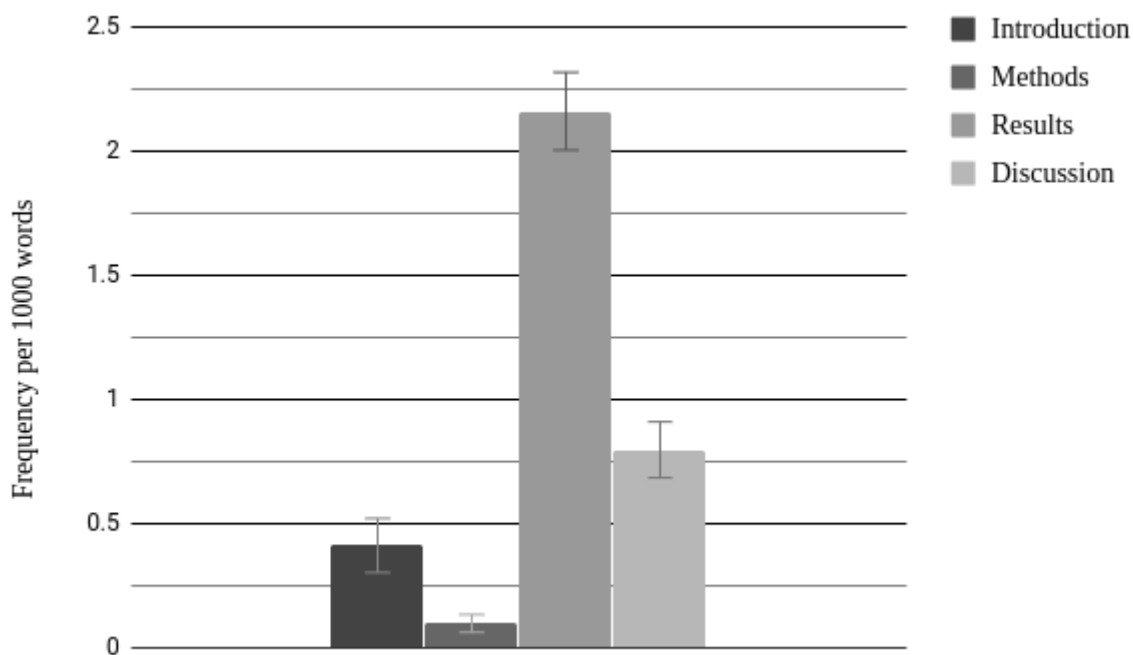
[2 TSINTRO12] *Whereas teachers’ beliefs have been shown to strongly influence the development of teachers’ practice and views about teacher preparation (e.g., Borg,*

¹⁷In a corpus study, Hyland (1999) reveals that hedges in Applied Linguistics RAs are the most commonly used subtype of stance. It is important to have in mind that the author investigated specifically selected vocabulary and different categories from those adopted for the current research. See Hyland, K. (1999). Disciplinary discourses: Writer stance in research articles.

2003b; Fang, 1996; Kagan, 1992; Peacock, 2001), learners' beliefs have also been observed to ***play an important role*** in second language (L2) learning.

The normalized cross-comparison of the subcorpora reveals that the Results subcorpus is, as expected, the subcorpus that presents the greatest amount of evaluation devices.

Graph 14: Normalized frequency and confidence intervals of evaluation devices (Stance expressions)



The table below shows the bundles conveying Evaluation in the Introduction and Methods subcorpora. On the other hand, the Results and Discussion subcorpora present a great amount of this subcategory, as shown in Graph 14. While the Results subcorpus displays a much more repetitive pattern with the use of *significant/significantly* evaluative devices, the Discussion subcorpus presents a more diverse use of elements, such as the adjectives *clear*, *consistent*, *difficult*, *surprising*, etc.

Table 25: Evaluation occurrences

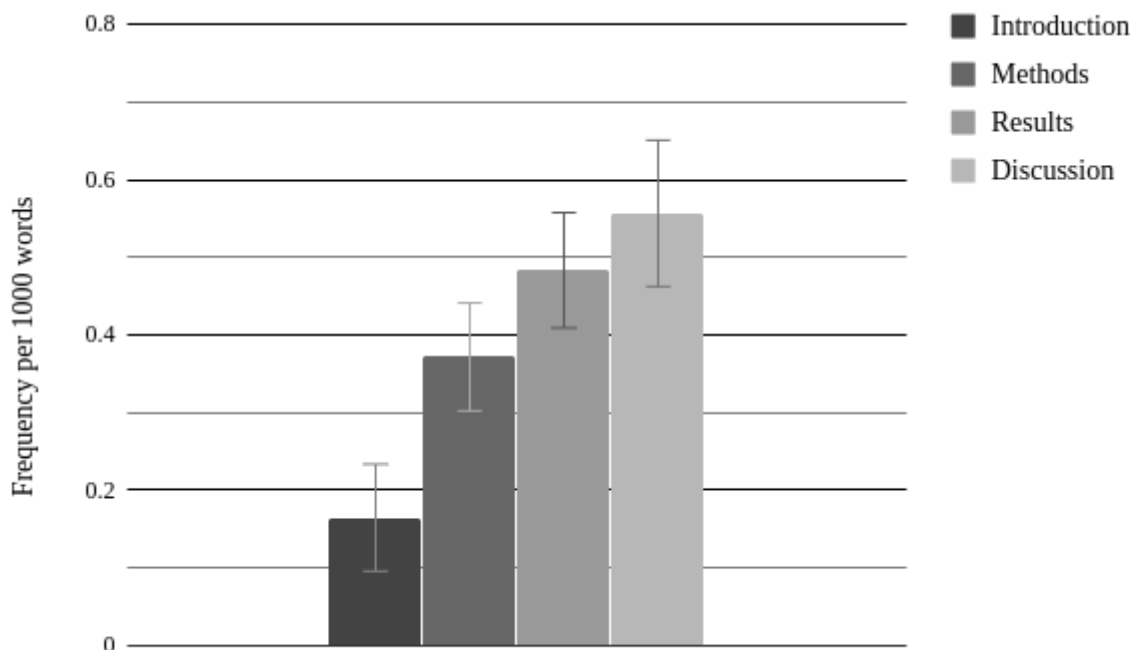
Introduction	Methods	Results	Discussion
<i>it is important to</i>	<i>it is important to</i> (note that)	(a) significant main effect of	<i>it is important to</i> (note/ bear in mind that)
play(s) an important role in	successful completion of the	(there was) a significant effect of	it is clear that
it is also important	how well the indices	there was no significant (group)	this is consistent with the
to be sensitive to	strongly disagree to strongly agree	no significant difference(s) between (the)	this finding is consistent with
to better understand the	a better understanding of	significant main effect(s) for	it is difficult to
it is difficult to		there was a significant difference (between (the)/ in (the))	it is worth noting (that)
a critical period for		<i>it is important to</i> (note that)	a better understanding of
sensitive to cumulative frequency		no statistically significant difference(s) between	it is necessary to
a better understanding of		no significant effect of	it is not clear
		was not statistically significant	this is not surprising
		(no) significant differences were found (between)	the overall quality of
		an important role in	it is reasonable to
		found to correlate significantly with	findings are consistent with
		it is interesting to note that	has the potential to

4.1.2.3.2 Ability and possibility

As noted by Simpson-Vlach and Ellis (2010), the ability and possibility expressions establish or introduce some possible or actual action or proposition. As the graph reveals, the

Introduction subcorpus is the only one that presents a significant low level of ability and possibility devices.

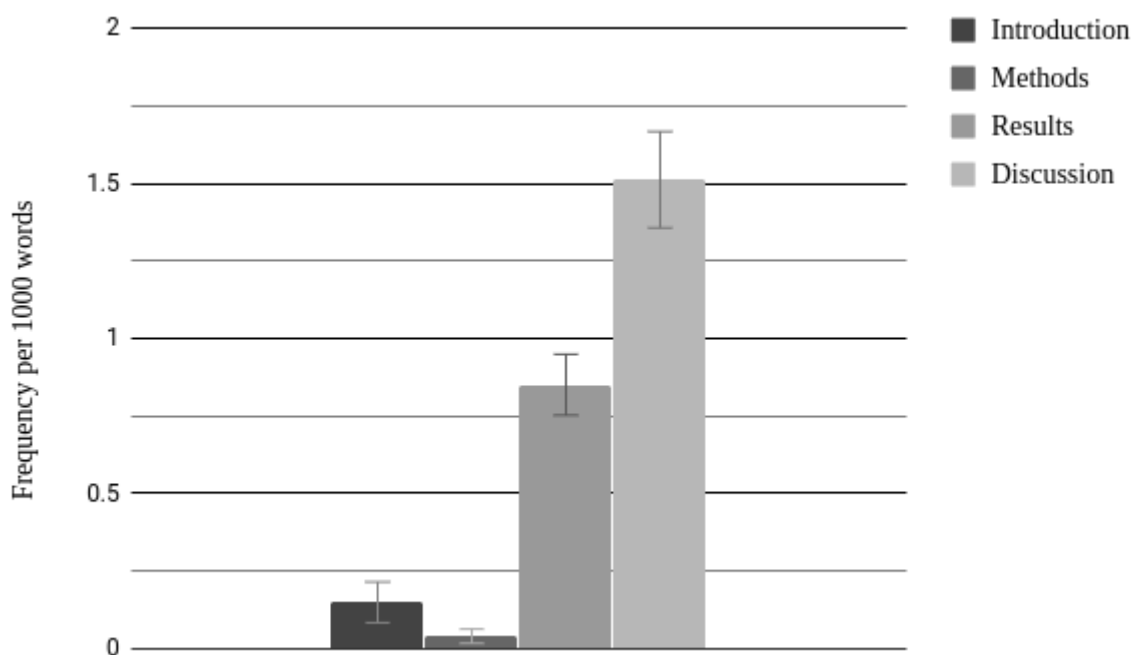
Graph 15: Normalized frequency and confidence intervals of ability and possibility devices (Stance expressions)



4.1.2.3.3 Hedges

Hedges are used to “gain acceptance for their work by balancing conviction with caution, and by conveying an appropriate disciplinary persona of modesty and assertiveness” (HYLAND, 2000, p. 179). This research shows that these devices are much more employed in the Results and Discussion sections. According to Hyland (1999), the use of hedge in the academic discourse “reflects the critical importance of distinguishing fact from opinion” as well as the “the need for writers to evaluate their assertions in ways that are likely to be persuasive to their peers” (p.106). Therefore, this finding perhaps coincides with our intuitions that authors tend to employ more hedges in Results and significantly more in Discussion, since it is the section used to convince the audience with appropriate caution and deference.

Graph 16: Normalized frequency and confidence intervals of hedging devices
(Stance expressions)



Swales (1990, p. 175) postulates that “as we move towards the diffuse end of the continuum the more necessary it becomes for authors to engage in acts of persuasion that will encourage the readerships to share particular visions of the research world”. This is clearly observed in Graph 16 with the normalized frequency of hedges across the subcorpora.

As can be seen in the table below, the most occurring instances of hedges in the Results and Discussion subcorpora are somewhat different. The devices used in the Results are less modalized than in the Discussion, i.e. in the former there is a high frequency of the reporting verbs *SHOW*, *INDICATE*, and *REVEAL*, while in the latter, modal verbs, such as *MAY*, *COULD*, and other less assertive reporting verbs are more commonly employed, such as *SUGGEST* and *APPEAR*.

Table 26: Hedges occurrences

Introduction	Methods	Results	Discussion
is more likely to	convey the meaning of	test showed that there	more likely to be
were more likely to	and at least one of	showed a significant main effect (of)	may be due to (the)
our understanding of the		table shows the descriptive (statistics)	it may be that
on the assumption that		more likely to use	this suggests that the
		we found a significant	the results showed that (the)
		the results indicate that	it could be argued that
		post hoc tests showed that	studies have shown that
		it is likely that	is likely to be
		did not appear to	may be the case
		tests revealed that the	these findings suggest that
		we found that the	was found to be
		these results suggest that	did not appear to
		it was found that	it appears that the
		analysis revealed that the	it is likely that
		it seems that the	there appears to be

4.1.2.4 Applied Linguistics RAs special devices

From the remaining lexical bundles, for not having fit previous taxonomies (BIBER et al., 1999, 2004 and SIMPSON-VLACH & ELLIS, 2010), we created seven new subtypes whose references are to: 1) languages; 2) participants; 3) processes or interactions; 4) theories or claims; 5) procedures or task details; 6) tools or appendix; and 7) tests or results. See table below. Between brackets is the subcorpus (IMRD) from where the examples were extracted.

Table 27: Lexical bundles subtypes: Applied Linguistics RA subtypes
(frequency per 1000 words)

	Introduction	Methods	Results	Discussion	
Reference to languages	0.87	0.36	0.00	0.14	<i>(English/ German/ Spanish) as a foreign language (I)</i>
Reference to participants	0.03	0.65	0.21	0.27	<i>male and female students (M)</i>
Reference to processes or interactions	0.52	0.08	0.00	0.22	<i>(in the) face to face interaction(s) (I)/ web based collaborative writing (D)</i>
Reference to theories or claims	0.14	0.00	0.00	0.00	<i>(of) the critical period hypothesis (I)</i>
Reference to procedures or task details	0.00	0.32	2.97	0.31	<i>used in this study (M)/ in task negotiation and (D)</i>
Reference to tools or appendix	0.00	0.31	0.07	0.00	<i>corpus of contemporary american english coca (M)</i>
Reference to tests or results	0.00	0.76	2.22	0.11	<i>means and standard deviations (for) (R)</i>

4.2 Secondary statistical treatment, the null-hypothesis test

In addition to confidence interval, we tested the null hypothesis of the 4 sets of categories above, namely major structural categories (3), major functional categories (3), structural subtypes (7), and functional subtypes (14). We compared the proportion of lexical bundle subtypes for each pair of subcorpus, thereby performing 204 comparisons, and tested the significance of the difference in proportion using a standard z-test with significance level 0.001. 116 of the 204 comparisons resulted in significant differences (see Table 28). They all confirmed the significance difference revealed by the confidence intervals.

For each pair, the null hypothesis assumes that all subcorpora are equal. In other words, there can not be a difference between the frequency of two subtypes. In each test, we

checked the p-value, i.e. the probability of obtaining a value greater or equal to 0 for the difference, given the null hypothesis is correct. We assume a threshold of 0.001 for rejecting the null hypothesis. If $p > 0.001$, we reject it.

In grey are the pairs with no statistically significant difference.

Table 28: Structural categories and subtypes of lexical bundles and a pairwise comparison analysis of subcorpora (p-value)

	Introd. / Methods	Introd. / Results	Introd. / Disc.	Methods / Res.	Methods / Disc.	Results / Disc.
Total: NP / PP	0	0	0	0	0	0
PP	0	0.0272	0.0000	0	0	0.1544
NP + of	0	0	0	0	0	0
Other NP	0	0	0.0346	0.0000	0	0
Total: VP	0	0	0	0	0	0
Non-passive	0.0000	0	0	0	0	0.3486
Passive	0	0	0	0	0	0
Other VP	0.1184	0.3892	0.0259	0.0312	0.0000	0.0006
Total: Dep. Clauses	0.0000	0	0	0.0001	0	0.1750
that clauses	0.0326	0	0	0	0.0000	0.0020
It (be) + adj.	0.0070	0.3277	0.0000	0.0016	0	0
to-clauses	0	0.0001	0.0000	0.0018	0.2848	0.1871
WH-frag./ if clauses	0.3317	0.3180	0.3211	0.3981	0.1179	0.0462

Interpretation: $p > 0.001$ the null hypothesis could not be rejected; $p < 0.001$ null hypothesis was rejected.

Table 29: Functional categories and subtypes of lexical bundles and a pairwise comparison analysis of subcorpora (p-value).

	Introd. / Methods	Introd. / Results	Introd. / Disc.	Methods / Res.	Methods / Disc.	Results / Disc.
Referential expressions	0	0	0.0973	0	0	0
Intangible framing attributes	0.3898	0.2984	0.0334	0.3119	0.0047	0.1456
Quantity specification	0	0.0017	0.0027	0	0	0
Deictics and locatives	0	0.0287	0.0093	0	0	0
Identification and focus	0.0385	0.3234	0.1035	0.0652	0.3501	0.3633
Contrast and comparison	0.0003	0	0	0	0.0000	0
Tangible framing attributes	0.1150	0.1429	0.1246	0.0003	0.0002	0.2928
Discourse organizing functions	0.0004	0.0000	0	0.2699	0	0.0000
Metadiscourse and textual reference	0.0798	0.2015	0.0000	0.0003	0.000	0
Topic elaboration: cause and effect	0.0010	0	0	0.0000	0	0.0540
Discourse markers	0.2940	0.2438	0.0791	0.3875	0.1811	0.3948
Topic introduction and focus	0.0001	0.0000	0	0.1351	0.0000	0.0000
Stance expressions	0	0	0	0	0	0.0012
Evaluation	0.0000	0	0.00001	0	0	0
Expressions of ability and possibility	0.0000	0.0000	0	0.0395	0.003	0.3974
Hedges	0.0035	0	0	0	0	0.0000
Intention/volition, prediction	0.0073	0.0006	0	0.2892	0.0073	0.0006

This chapter presented the overall proportion of lexical bundles in Applied Linguistics RA sections, the analysis of lexical bundles according to their structural and functional types, and pointed out distinctive patterns across the IMRD corpus. In the following chapter, we will discuss the most relevant findings of this study and the implications of this type of analysis for the field of English for Academic Purposes.

5. CONCLUSION

In the current study, we analyzed lexical bundles in subcorpora of Applied Linguistics Research Article (RA) sections, their structural and functional types considering previous and consolidated taxonomies in the literature. A long list of lexical bundles was generated by utilizing a corpus especially designed for this research. The corpus is composed of 180 Applied Linguistics RAs from high-impact journals and split into four subcorpora of Introduction, Methods, Results and Discussion sections, with more than 1 million words in total. In the Literature Review chapter, we presented some background on the research of lexical bundles and the importance of investigating these devices in academic discourse. We also covered each category and subcategory to be analyzed and how bundles were sorted. The results were cross-compared by looking at the normalized frequency within the structural and functional category. Finally, we adopted two statistical treatments, confidence intervals and the null-hypothesis significance z-test in order to check whether differences across subcorpora were significant.

The findings of this study show that, although the IMRD subcorpora share certain features concerning the structural and functional analysis of lexical bundles, Applied Linguistics RA sections should be treated as separate texts for they display strong distinctions, and some grammatical structures may play singular functional roles. Firstly, this research revealed the proportion of 4-7 word lexical bundles found in the subcorpora is strikingly different. There are twice as many lexical bundles in the Methods and Results subcorpora as in the other subcorpora. Secondly, our analysis entailed the calculation of the ratio bundle token/type in order to estimate how much lexical bundles undergo repetitive use. The ratios showed that the subcorpora vary their devices in similar ways. In other words, the repetition of lexical bundles is somewhat the same across the subcorpora. These findings answer our first research question about the proportion of lexical bundles and their ratio bundle token/type in each subcorpus.

The following process allowed us to answer the second research question which regards the frequency and patterns of structural types across the IMRD subcorpora. The first set involved the frequency of the main structural categories: noun phrase and prepositional phrase fragments, verb phrase fragments and dependent clause fragments. Unlike previous studies on academic prose (BIBER et al., 1999, 2004; BIBER, 2010), none of the subcorpora

of the present research generated more dependent clause fragments than verb phrase fragments. We assume, therefore, this divergence may be caused by particularities of Applied Linguistics RAs, still unknown in the present study, but they should be further considered and addressed so that we can cross-compare with the findings from Biber et al.'s academic register corpora.

We also analyzed each subtype from the structural categories, and found an interesting correlation between “NP with of-phrase fragments”, for instance, *the use of the, the meaning of the, a wide range of, the majority of the*, and two functional subtypes: Intangible framing attributes and Quantity specification. Almost half of the “NP with of-phrase fragments” are also Quantity specification devices. Therefore, this should explain why the Methods and Results subcorpora present the highest use of that structure.

The proportion of non-passive and passive voice structures across the subcorpora is also noteworthy. Some studies generalize the use of non-passive or passive across registers, but while the frequency of these structures is fairly even throughout the entire corpus, 42 and 38 per cent, this study has shown that Methods and Results present a statistically significant difference in the use of passive voice, with the former presenting twice as many passive structures as the other subcorpora. Although there has already been studies (HESLOT, 1982; SWALES, 1990) proving that the methods sections are expected to display a great deal of passive structures, none of them has been able to determine its proportion or whether results sections would follow suit.

The findings also suggest that no present simple passive voice bundles carry out any other function but the reference to tables, figures or appendices in the Methods and Results subcorpora, e.g. *are shown in table* and *are shown in the appendix*. Conversely, in the Introduction subcorpus, lexical bundles with passive voice structures in the present tense (simple or perfect aspects) indicate a gap or review items of previous literature (CORTES, 2013; SWALES, 2004, 1990). In the Discussion subcorpus, the most frequently used bundles in the passive voice are those containing a modal verb, such as *it should also be (noted)*, *it could be argued that*, and *can be attributed to*.

The non-passive structures in the Methods subcorpus communicate the active role that participants and the tests play by employing specific verbs, namely PARTICIPATE, TAKE PLACE, RATE and HAVE, always in the past simple. The Results section demonstrates a massive use of other particular structures, such as THERE BE, and reporting verbs such as SHOW, DIFFER, REVEAL,

and DEMONSTRATE, predominantly in the past simple. We conclude, therefore, that the proportions of passive and non-passive voice should be specifically and more carefully related to sections of RAs, not to entire registers or fields of specialization, given their considerable in-text variation.

While the Introduction, Methods, and Results subcorpora display a similar frequency of the *it* + adjective lexical bundle, the Discussion subcorpus presents a considerably higher use of this device. This finding can be related to Biber et al.'s (1999, p. 1018) proposition that this structure is employed to express “the stance of the writer; for example, possibility/likelihood, importance, necessity”. It is in the discussion section that writers tend to express their opinion, attitudes or assessment of certainty.

The final analysis helped us answer the third research question relating to the most commonly employed pragmatic functions in each subcorpora, and the features revealed by their proportion and elements. The current study revealed some more remarkable features. From the lexical bundles classified from all subcorpora, the greatest share is made up of Referential expressions. The considerable amount of this category is a feature of the academic prose (BIBER et al., 2004; SIMPSON-VLACH & ELLIS, 2010; DUTRA & BERBER SARDINHA, 2013). In the Methods subcorpus, Stance expressions are not surprisingly the least occurring subcategory of functional types. However, the functional types in the Results subcorpus display an interesting pattern: although most of the types are constituted by Referential expressions (49 per cent), Stance expressions also represent a considerable amount, 32 per cent. Finally, in the Discussion section subcorpus, the proportion of three categories are somewhat balanced.

Notably, when considering the analysis of each subcategory separately, bundles conveying Quantity and specification are used significantly differently across the subcorpora. The Methods subcorpus presents a much higher use of this type of bundle. This distinction may be due to the fact that it is in the methods section that one can find information on the quantity of participants, questions, tests, length of experiments, etc. The present research also revealed that quantity specification bundles are often framed by ‘the/a N of’, or “NP with of-fragments”. This is reflected in the high frequency of “NP with of-fragments” in the Methods and Results sections, as discussed above.

The amount of Contrast and comparison lexical bundles is much higher in the Results subcorpus than in the other ones, because findings of the study itself or previous studies are

compared and contrasted in that section. However, an interesting pattern emerges in the Introduction, all Contrast and comparison devices are linking adverbials, such as *on the other hand*, *on the one hand*, *in contrast to the*, *in the same way*. In the other subcorpora, Contrast and comparison bundles are mostly related to participants, tests, or results, e.g. *between the two groups*, *did not differ significantly (from)*, *the link between the*, etc. The most recurrent bundle is *in contrast to the*, this device is found among the most common bundles belonging to the contrast and comparison subtype in the four subcorpora analyzed in this study.

Deictics and locatives are considerably more common in the Methods subcorpora than in the others. By looking at the entire sections of the Methods subcorpus, Deictics and Locatives bundles are expressions mostly used to give details about the experiment procedures, such as in *at a university in*, *an English speaking country*, *in the original text*, *at the end of (the semester)*. These types of bundles therefore are considerably prevailing in the Methods subcorpus. Additionally, the instances across the investigated subcorpora do not present much variability, i.e. prepositions and articles frame the same elements: *end*, *beginning*, *United States*, *time*, etc.

Discourse organizing functions reflect the connection between prior and coming discourse (BIBER et al., 2004). In the current study, we have analyzed the subtypes: Metadiscourse and textual reference, Topic introduction and focus, Topic elaboration: non-causal, Topic elaboration: cause and effect, and Discourse markers. Discourse organizing function bundles in every subcorpus are mostly represented by metadiscourse and textual references. The substantial amount of metadiscourse and textual reference bundles corroborates with Simpson-Vlach and Ellis (2010), who also found that this is the most occurring subcategory from Discourse organizing functions in academic prose.

Our findings show that metadiscourse and textual reference devices are greatly represented by the word *study*. This might not be surprising, considering we are dealing with academic discourse. Moreover, this subcategory does not present variability in use within each subcorpus, i.e. they all have somewhat the same bundle types/ bundle tokens ratio. If the Discussion shows a significantly higher frequency of this subcategory, we can assume that certain bundles are preferred over others. We found that the bundles *in the present study* and *in the current study* are the most commonly resorted Metadiscourse and textual reference devices, especially in the Discussion subcorpus.

The research has also shown that the use of Metadiscourse and textual reference devices in Methods is somewhat the same across the other subcorpora. Swales (1990) suggests that “softer”, emerging or interdisciplinary fields tend to deal with given and new information more cohesively. He also says that information in the Methods in this field is carefully presented with step-by-step description and supported by anaphoric reference and lexical repetition. We have not investigated the presence of anaphoric reference or lexical repetition in order to test Swales’ claim, but by showing that the use of Metadiscourse and textual reference devices in the Methods subcorpus is balanced with the other subcorpora, we can also add that the cohesiveness feature in methods sections of Applied Linguistics may also be played by those lexical bundles.

The most expressive finding regarding the use of Stance expressions is the frequency of evaluation, 61 percent, and hedges, 24 per cent, in the Results subcorpus. It is interesting to note that a vast majority of the evaluation bundles, in the Results subcorpus, are to communicate the significance or non-relevance of statistical findings. Swales (1990) states that the style and structure of results sections “seem to be designed to deny on the author’s part any associative contamination with commentary or observation” (p.171). This claim can be supported by the fact that evaluation devices, in the Results sections subcorpus, are overwhelmingly related to what statistical tests have shown rather than the authors’ proposition.

Finally, the Discussion subcorpus is composed of 53 percent of hedges. Hedges are employed in order to help writers avoid or diminish the possibility of opposition (HYLAND, 1999). Finding a high frequency of this type of lexical bundle in subcorpus does not come as a surprise given the fact that authors use the RA section to comment on results, which expresses the main communicative purpose of the Discussion section (RUIYING & ALLISON, 2003) and to make new knowledge claims (BASTURKMEN, 2009). In addition, our qualitative analysis indicates that hedge devices used in the Results are less modalized than in the Discussion, i.e. in the former there is a high frequency of the reporting verbs *SHOW*, *INDICATE* and *REVEAL*, while in the latter, modal verbs, *MAY*, *COULD*, and other less assertive reporting verbs are more commonly employed, such as *SUGGEST* and *APPEAR*.

Given the specificity of the corpus investigated in this study, it was expected to encounter devices that would not be suitable to the taxonomies of Biber et al., 1999, 2004 and Simpson-Vlach & Ellis, 2010. Consequently, taking into account the remaining devices, we

created seven new subtypes whose references are to: 1) languages; 2) participants; 3) processes or interactions; 4) theories or claims; 5) procedures or task details; 6) tools or appendix; and 7) tests or results.

The current study was an attempt to offer a type of analysis which has not yet been provided by the literature, namely the investigation of each subtype from the major structural and functional categories of lexical bundles in an Applied Linguistics RA corpus considering their discursive role. However, further investigation is needed in order to fully understand some distinctions in frequency which could not be interpreted due to lack of resources or simply because it was not part of our objectives. Some of our main suggestions are to investigate what makes introductions so different from the other sections. For that purpose, not only should lexical bundles be studied but also other elements which could be “replacing” lexical bundles given their low frequency. It would also be interesting to better understand the relationship between lexical bundles and the rhetorical moves from each section, and provide the identification of devices exclusively linked to each move or step. Incorporating the Abstract section into the corpus could also yield interesting findings. Additionally, further research should be undertaken to investigate the findings that are not in line with previous studies, namely lower frequency of dependent clause fragments than verbal phrase fragments in the whole IMRD corpus, and the proportion of Referential, Discourse organizing functions, and Stance expressions in the Results and Discussion sections of Applied Linguistics RAs. By replicating the methodology of the present study, a cross-comparison of other RA sections from different fields would also be very useful.

The scope of this study was limited in terms of corpus size. It is widely recommended that lexical bundles be generated from larger corpora. Although the sum of the section corpora of the current study complies with that criteria, the devices were generated from each subcorpus separately, the sums varied from 200,000 to 300,000 words. Another limitation is the type of classification of bundles. Classifying lexical bundles into both structural and functional categories usually poses challenges due to the subjective nature of this analysis. Additionally, the confidence interval test applied in our analysis, recommended by Gries (2013), was meant to test if the difference in frequency is statistically significant. However, according to Gries (2013) and Crawley (2014) very small differences, despite statistically significant, should be disregarded. This explains why we covered the main identified distinctions in the subcorpora.

Notwithstanding these limitations, the study suggests that the structures and discursive functions that emerged in this analysis be used in workshops or be related to what is taught in the classrooms of EAP. The corpus especially compiled for the current study can also be an aid when designing lessons. EAP instructors may use it as a tool to relate, illustrate, or even better understand the myriad of possibilities in terms of writing features of RA sections from Applied Linguistics.

REFERENCE

- ALTENBERG, B. Recurrent verb-complement constructions in the London-Lund Corpus. *English language corpora: Design, analysis and exploitation*, 1993, p. 227-245.
- ANTHONY, L. Antconc (version 3.4. 4) [software]. Waseda University, 2016.
- BANERJEE, S.; PEDERSEN, T. The design, implementation, and use of the ngram statistics package. In: *CICLing*, v.. 2588, 2003, p. 370-381.
- BASTURKMEN, H. Commenting on results in published research articles and masters dissertations in language teaching. *Journal of English for Academic Purposes*, v. 8, n. 4, 2009. p. 241-251.
- BIBER, D.; FINEGAN, E. Adverbial stance types in English. *Discourse processes*, v. 11, n. 1, 1988, p. 1-34.
- BIBER, D. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.
- BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus linguistics: Investigating language structure and use*. Cambridge University Press, 1998.
- BIBER, D. Corpus-Based and Corpus-Driven Analyses of Languages: Variation and Use. In: HEINE, B.; NARROG, E. (Ed) *The Oxford handbook of linguistic analysis*. Oxford Handbooks in Linguistic, 2010, p. 159-191.
- BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. *English for specific purposes*, v. 26, n. 3, 2007, p. 263-286.
- BIBER, D.; CONNOR, U.; UPTON, T. A. *Discourse on the move: Using corpus analysis to describe discourse structure*, v. 28, John Benjamins Publishing, 2007.
- BIBER, D.; CONRAD, S. Lexical bundles in conversation and academic prose. *Language and Computers*, v. 26, 1999, p. 181-190.
- BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E.; QUIRK, R. *Longman grammar of spoken and written English*. v. 2. Cambridge, MA: MIT Press, 1999.
- BIBER, D.; CONRAD, S.; CORTES, V. *Lexical bundles in speech and writing: An initial taxonomy*. na, 2003.
- BIBER, D.; CONRAD, S.; CORTES, V.. If you look at...: Lexical bundles in university teaching and textbooks. *Applied linguistics*, v. 25, n. 3, 2004. p. 371-405.
- BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. *English for specific purposes*, v. 26, n. 3, 2007. p. 263-286.

BIBER, D.; FINEGAN, E. Intra-textual variation within medical research articles. In: OOSTDIJK, N.; HAAN, P. de (Ed.). *Corpus-based research into language: in honour of Jan Aarts*, v. 12, Rodopi, 1994. p. 201-222.

BOHORQUEZ, C. Eliminação de pacotes lexicais relacionados ao tópico e de pacotes lexicais em contexto de sobreposição: uma proposta metodológica para os estudos da linguística de corpus, 103 f. Dissertação (Mestrado em Linguística Aplicada) - Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte. 2015.

BROWN, P.; FRASER, C. Speech as a marker of situation. In: *Social markers in speech*. Cambridge University Press, 1979. p. 33-62.

BYRD, P.; COXHEAD, A. On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL*, v. 5, n. 5, 2010, p. 31-64.

CHARLES, M.; HUNSTON, S.; PECORARI, D. (Ed.) *Academic writing: At the interface of corpus and discourse*. A&C Black, 2011.

CHEN, Y.-H.; BAKER, P. Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, v. 14, n.2, 2010. p. 30-49.

de COCK, S. A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-Native Speakers of English. *International Journal of Corpus Linguistics*, v. 3, n. 1, 1998. p. 59-80.

CORTES, V. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for Specific Purposes*, v. 23, n. 4, 2004. p. 397-423.

CORTES, V. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, v. 3, n. 1, 2008, p. 43-57.

CORTES, V. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. *Journal of English for academic purposes*, v. 12, n.1, 2013. p. 33-43.

DUTRA, D. P.; ORFANO, B. M. ; BERBER SARDINHA, A. P. STANCE BUNDLES IN LEARNER CORPORA. In: ALUISIO, S. M.; TAGNIN, S. (Org.). *New Language Technologies and Linguistic Research: A Two-Way Road*. Newcastle upon Tyne: Cambridge Scholars Publishing, 2014, p. 02-15.

DUTRA, D.P.; BERBER SARDINHA, T. Referential expressions in English learner argumentative writing. In: S. Granger, G. Gilquin & F. Meunier (Ed.) *Twenty Years of Learner Corpus Research: Looking back, Moving ahead*. Corpora and Language in Use – Proceedings 1, Louvain-la-Neuve: Presses universitaires de Louvain, 2013. p. 117-127.

ELLIS, N.; SIMPSON-VLACH, R.; MAYNARD, C. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *Tesol Quarterly*, v. 42, n. 3, 2008, p. 375-396.

FIRTH, J. R. A synopsis of linguistic theory, 1930–1955. *Studies in linguistic analysis*. Oxford: Basil Blackwell, 1957. *apud* STUBBS, M. British traditions in text analysis. *Text and technology: in honour of John Sinclair*, 1993, p.1-33.

FLOWERDEW, L. *Corpora and language education*. Springer, 2011.

GILBERT, G. N.; MULKAY BRUCE, M. Opening Pandora's box: a sociological analysis of scientific discourse. Cambridge: Cambridge University Press, 1984. *apud* SWALES, J. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

GRANGER, S. Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. *Phraseology: Theory, analysis and applications*, 1998. p. 145-160.

GRANGER, S.; MEUNIER, F. (Ed.). *Phraseology: An interdisciplinary perspective*. John Benjamins Publishing, 2008.

GRIES, S.Th. Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff." *Corpus linguistics and linguistic theory*, v. 1, n. 2, 2005, p. 277-294.

GRIES, S. Th. Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik*, v. 54, n. 2, 2006, p. 191-202.

GRIES, S. Th. *Statistics for linguistics with R: A practical introduction*. Walter de Gruyter, 2013.

HALLIDAY, M. A. K. *Language as Social Semiotic: The social interpretation of language and meaning*. London: Edward Arnold. 1978.

HESLOT, J. Tense and other indexical markers in the typology of scientific texts in English. *Pragmatics and LSP*, 1982. p. 83-103.

HILL, S.S.; SOPPELSA, B. F.; WEST, G. K. Teaching ESL Students to Read and Write Experimental-Research Papers. *TESOL quarterly*, v. 16, n. 3, 1982, p. 333-347.

HOLMES, R. Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for specific Purposes*, v. 16, n. 4, 1997. p. 321-337.

HOPKINS, A.; DUDLEY-EVANS, T. A genre-based investigation of the discussion sections in articles and dissertations. *English for specific purposes*, v. 7, n. 2, 1988, p. 113-121.

HYLAND, K. *Hedging in scientific research articles*. John Benjamins Publishing, 1998.

HYLAND, K. Disciplinary discourses: Writer stance in research articles. *Writing: Texts, processes and practices* 99121, 1999.

HYLAND, K. Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts.

Language Awareness, v. 9, n. 4, 2000. p. 179-197.

HYLAND, K. Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, v. 34, n. 8, 2002. p. 1091-1112.

HYLAND, K. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, v. 27, n. 1, 2008. p. 4-21.

HYLAND, K. Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, v. 18, n. 1, 2008. p. 41-62.

HYLAND, K.. Bundles in academic discourse. *Annual Review of Applied Linguistics*, v. 32, 2012. p. 150-169.

HYMES, D. *Foundations in Sociolinguistics: An Ethnographic Approach*. Philadelphia, PA: University of Pennsylvania Press, 1974. p. 145-178.

KILGARRIFF, A. Language is never, ever, ever, random. *Corpus linguistics and linguistic theory* 1, n. 2, 2005, p.: 263-276.

LE, T.N.P.; HARRINGTON, M. Phraseology used to comment on results in the Discussion section of applied linguistics quantitative research articles. *English for Specific Purposes*, v. 39, 2015, p. 45-61.

LIN, L.; EVANS, S. Structural patterns in empirical research articles: A cross-disciplinary study. *English for Specific Purposes*, v. 31, n. 3, 2012. p. 150-160.

MAURANEN, A. Conceptualising ELF in: JENKINS, J.; BAKER, W.; DEWEY, M. (Ed) *The Routledge handbook of English as a lingua franca*. Routledge, 2017. p. 7- 24.

PAWLEY, A.; SYDER, F. H. Two puzzles for linguistic theory: Native-like selection and native-like fluency. *Language and communication*, v. 191, 1983. p. 225.

du PREL, J., HOMMEL, G.; RÖHRIG, B.; BLETTNER, M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, v. 106, n. 19, 2009. p.335-339.

RUIYING, Y; ALLISON, D. Research articles in applied linguistics: Moving from results to conclusions. *English for specific purposes*, v. 22, n.4, 2003. p. 365-385.

SAMPSON, G. Briefly noted-English for the computer: the SUSANNE corpus and analytic scheme. *Computational Linguistics*, v. 28, n. 1, 2002, p. 102-103.

SCHMITT, N.; CARTER, R. Formulaic sequences in action. *Formulaic sequences: Acquisition, processing and use*, 2004, p. 1-22.

SCHMITT, N.; GRANDAGE, S.; ADOLPHS, S. Are corpus-derived recurrent clusters psycholinguistically valid. *Formulaic sequences: Acquisition, processing and use*, 2004. p.

127-51.

SHEKIN, D. *Handbook of parametric and nonparametric statistical procedures*. crc Press, 2003.

SIMPSON-VLACH, R.; ELLIS, N. An academic formulas list: New methods in phraseology research. *Applied Linguistics*, v. 31, n.4, 2010. p. 487-512.

SINCLAIR, J. *Collins COBUILD English language dictionary*. Harper Collins Publishers, 1987.

SINCLAIR, J. *Corpus, concordance, collocation*. Oxford University Press, 1991.

SWALES, J. and NAJJAR, H. The writing of research article introductions. *Written communication*, v. 4, n. 2, 1987. p.175-191.

SWALES, J. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

SWALES, J. *Research genres: Explorations and applications*. Ernst Klett Sprachen, 2004.

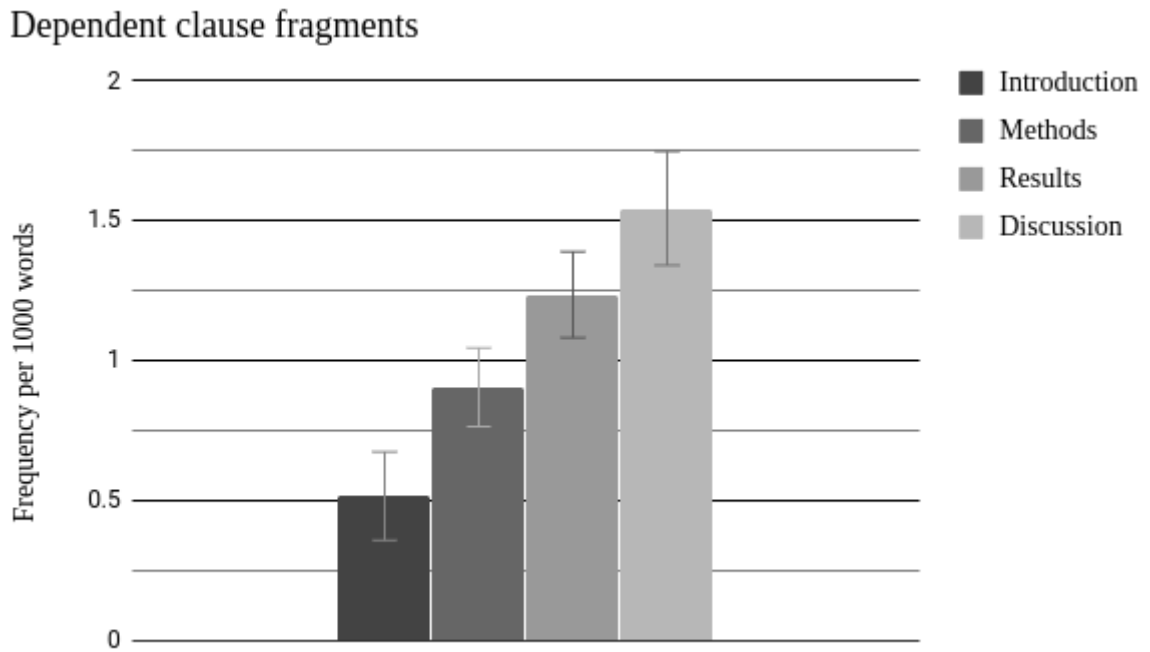
TARONE, E.; DWYER, S.; GILLETTE, S.; ICKE, V. On the use of the passive and active voice in astrophysics journal papers: With extensions to other languages and other fields. *English for specific purposes*, v. 17, n. 1, 1998. p. 113-132.

WEISSBERG, R.C. Given and new: Paragraph development models from scientific English. *Tesol Quarterly*, v. 18, n. 3, 1984, p. 485-500 *apud* SWALES, J. *Genre analysis: English in academic and research settings*. Cambridge University Press, 1990.

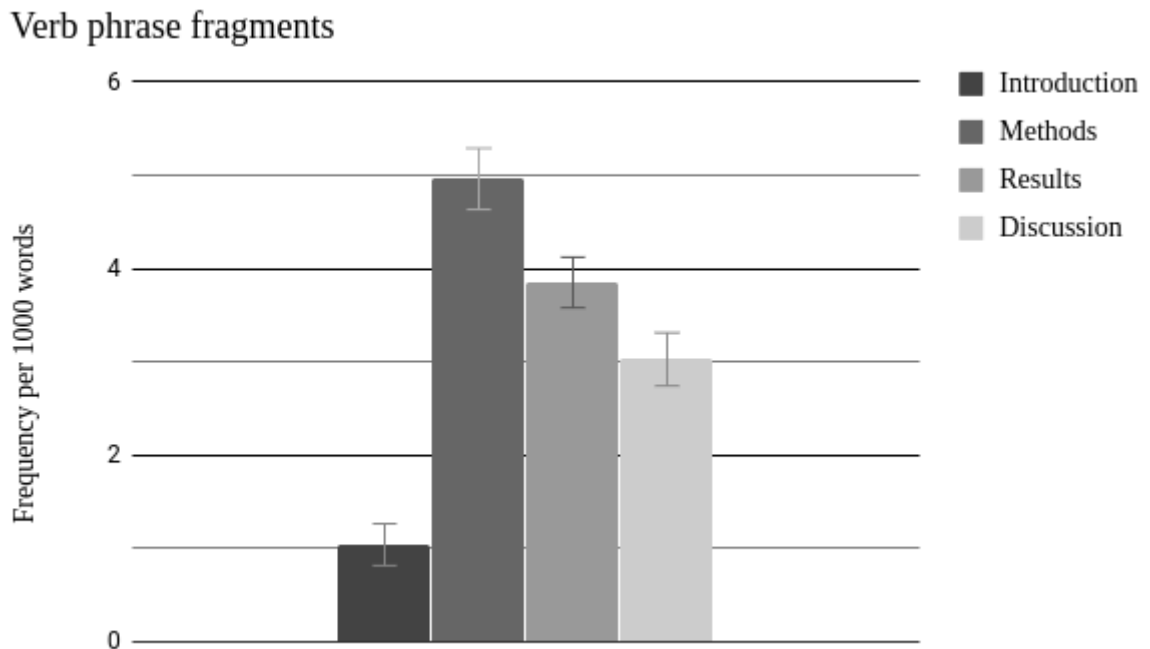
WEISSBERG, R.; BUKER, S. *Writing up research*. Englewood Cliffs, NJ: Prentice Hall, 1990.

APPENDIX A - GRAPHS OF REMAINING ANALYSES

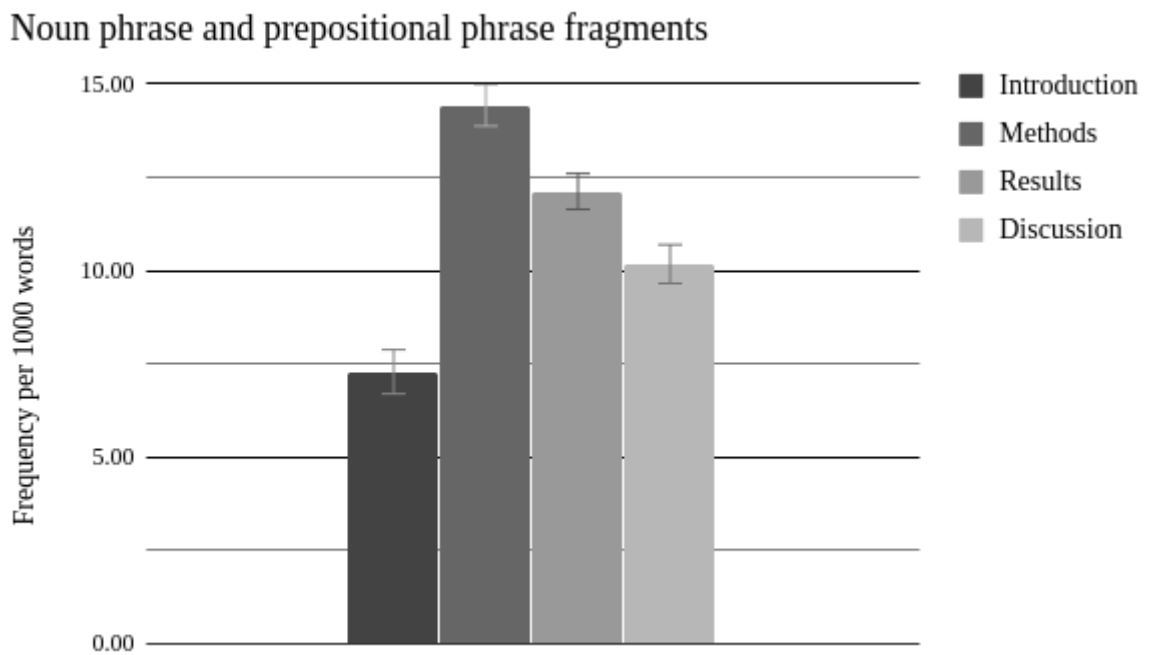
Graph 1a: Normalized frequency and confidence intervals of dependent clause fragments



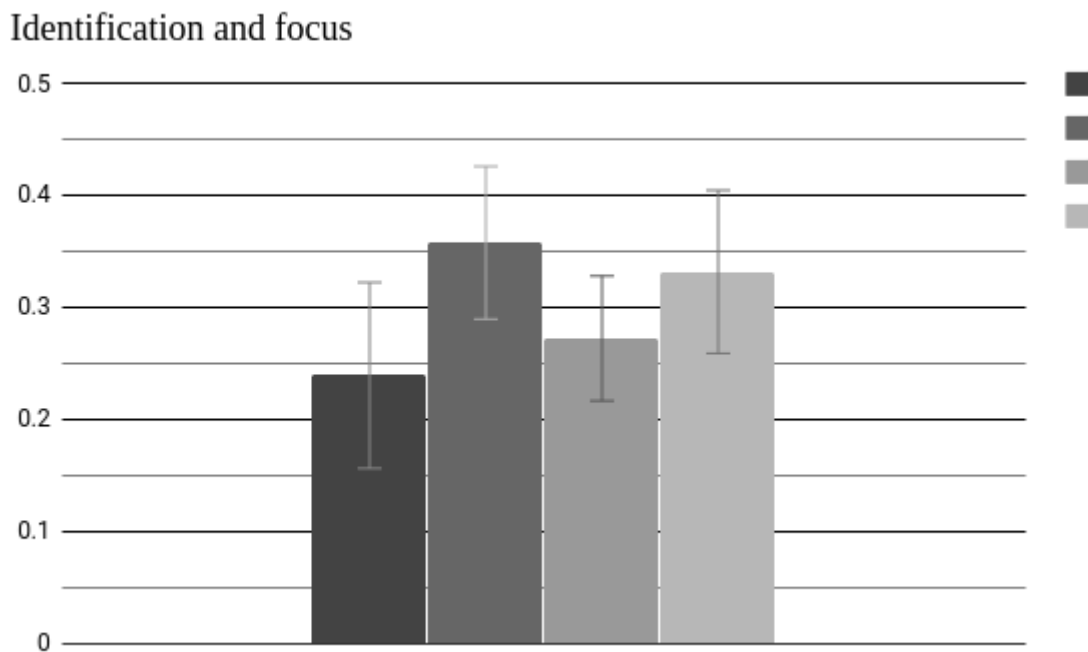
Graph 2a: Normalized frequency and confidence intervals of verb phrase fragments



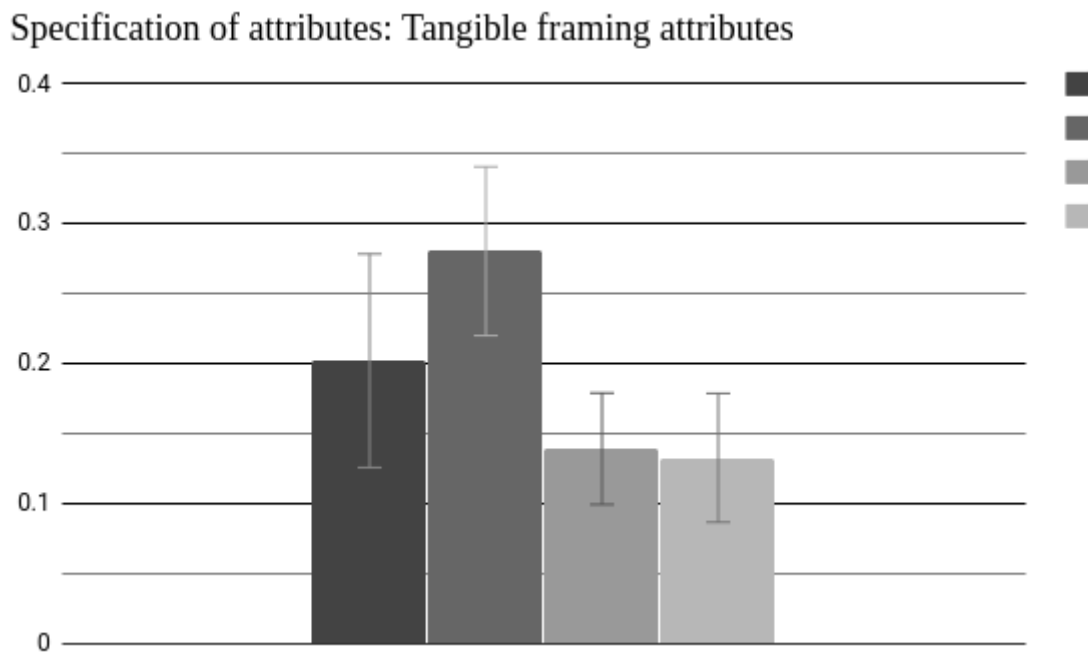
Graph 3a: Normalized frequency and confidence intervals of noun phrase and prepositional phrase fragments



Graph 4a: Normalized frequency and confidence intervals of identification and focus lexical bundles (Referential expressions)



Graph 5a: Normalized frequency and confidence intervals of specification of attributes: tangible framing attributes lexical bundles (Referential expressions)



APPENDIX B - R script

1b. R script used to delete the shortest bundles which occurred as frequently as their counterpart:

```

library(tidyverse)
library(stringr)
library(readxl)

dados <- read_excel('~Downloads/RESULTS - bundles Dissertação.xlsx', col_names = F)
dados <- dados %>%
  set_names(c("numero", "bundle"))#, "X3", "X4", "X5")

dados_2 <- expand_grid(dados$bundle, dados$bundle)

dados_2 <- filter(dados_2, !(Var1 == Var2))

dados_2$Var1 <- as.character(dados_2$Var1)
dados_2$Var2 <- as.character(dados_2$Var2)
dados_2$contido <- NA
for (i in 6036:nrow(dados_2)) {
  print(i)
  try(dados_2$contido[i] <- str_detect(dados_2$Var2[i], dados_2$Var1[i]), T)
}
dados_2 <- filter(dados_2, contido)

joined <- left_join(dados_2, dados, by=c("Var1" = "bundle"))
joined$numero_1 <- joined$numero
joined <- select(joined, -numero)

joined <- left_join(joined, dados, by=c("Var2" = "bundle"))
joined$numero_2 <- joined$numero

joined <- joined %>%
  select(-numero) %>%
  mutate(numero_char = str_count(Var2)) %>%
  arrange(desc(numero_char))

joined2 <- joined %>%
  filter(numero_1 == numero_2)

joined <- joined %>%
  filter(numero_1 != numero_2)

bundles_removal <- joined2 %>%
  distinct(Var1)

```

```
write.csv(bundles_remover, "~/bundles_remover_RESULTS.csv", row.names=F)
dados <- filter(dados, !(bundle %in% pull(bundles_remover)) )
joined <- joined %>%
  mutate(numero_char1 = str_count(Var1)) %>%
  arrange(desc(numero_char1)) %>%
  arrange(desc(numero_char))

write.csv(dados, "~/DISCUSSION_dados_filtrados.csv", row.names=F)
```

APPENDIX C - Python script

```
“import os
import glob
import re
os.chdir('/home/mydocuments/corpusfiles/')
os.system('find . | grep .txt > lista_arquivos')
with open('lista_arquivos', 'r') as f:
    lista_arquivos = f.read()
lista_arquivos = lista_arquivos.split('\n')[:-1]
for i in lista_arquivos:
    print(i)
    conteudo = ""
    print("Lendo")
    with open(i,'rb') as file:
        conteudo = file.read()
    print("Escrevendo")
    with open(i,'wb') as file:
        file.write(re.sub(b'\n', b' ', conteudo))”
```

APPENDIX D - Script created to run the z-test in Google Sheets

```
function z_test(mean1, mean2, n1, n2) {  
  var p1 = mean1/n1;  
  var p2 = mean2/n2;  
  var var1 = (p1*(1-p1));  
  var var2 = (p2*(1-p2));  
  var valor1 = var1/n1;  
  var valor2 = var2/n2;  
  var soma = valor1 + valor2;  
  var raiz = Math.sqrt(soma);  
  z = (p1 - p2)/raiz;  
  return z;  
}
```