v. 25, n. 1, e50007, 2025
Corpus Linguistics: Studies and Applications
ISSN: 1984-6398 · DOI: 10.1590/1984-6398202550007

Revista Brasileira de Linguística Aplicada

Thematic Dossier

De acordo com a...: Lexical Bundles and Disciplinary Variation in Academic Brazilian Portuguese

De acordo com a...: *pacotes lexicais e variação disciplinar no português acadêmico brasileiro*

Larissa Goulart*
*Montclair State University (MSU)
Montclair | NJ | USA
goulartl@montclair.edu
https://orcid.org/0000-0002-4190-4589

Abstract: Research on formulaic language has shown that few lexical bundles overlap across disciplines in expert academic writing, with disciplinary writing being characterized by specific bundles. This was proven to be the case in both English (Hyland, 2008) and Spanish (Cortes, 2008). This paper explores the use of lexical bundles in Portuguese academic writing in two disciplinary groups, *life sciences* and *humanities*. Using a corpus of research articles written by Brazilian scientists, four-word bundles were extracted and compared for their function (Biber; Conrad; Cortes, 2004) and structure (Biber; Johansson; Leech; Conrad; Finegan, 1999). The results of this study indicate that writing for *life sciences* contains more formulaic language, with referential bundles being used more frequently to refer to research participants and places; and *humanities* writing contains fewer bundles overall and more diversity in function types than *life sciences*. The results of this paper can help us understand the use of lexical bundles in Portuguese and inform the development of materials for teaching Portuguese as both an L1 and L2.

Keywords: lexical bundles; academic Portuguese; formulaic language; disciplinary writing.

Resumo: Pesquisas sobre linguagem formulaica mostram que poucos pacotes lexicais ocorrem em diferentes disciplinas acadêmicas, sendo a escrita disciplinar caracterizada por estruturas específicas. Isso foi comprovado tanto em inglês (Hyland, 2008) quanto em espanhol (Cortes, 2008). O presente artigo

This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.

explora o uso de pacotes lexicais na escrita acadêmica em português em dois grupos disciplinares: ciências biológicas e humanidades. Utilizando um *corpus* de artigos de pesquisa escritos por cientistas brasileiros, pacotes lexicais de quatro palavras foram extraídos e comparadas em relação à sua função semântica (Biber; Conrad; Cortes, 2004) e estrutura gramatical (Biber; Johansson; Leech; Conrad; Finegan, 1999). Os resultados deste estudo indicam que a escrita para ciências biológicas contém mais linguagem formulaica, com estruturas referenciais sendo usadas mais frequentemente para se referir a participantes da pesquisa e locais; e a escrita nas humanidades contém menos estruturas no geral e mais diversidade nos tipos de funções do que nas ciências biológicas. Os resultados deste artigo podem ajudar a entender o uso de estruturas léxicas em português e informar o desenvolvimento de materiais para o ensino de português como L1 e L2.

Palavras-chave: pacotes lexicais; português acadêmico; sequências formulaicas; escrita disciplinar.

1 Introduction

Over the past twenty years there has been a sharp increase in the number of students attending higher education in Brazil. According to the Higher Education Census, the number of students enrolled in public universities in Brazil has more than doubled since 2010 (INEP, 2019). This means that there are now more students in the university classroom who need to learn how to use Portuguese as an academic language. In addition, new initiatives to promote the internationalization of Portuguese encourage international students to attend university in Brazil. One such case is the Languages without Borders program which offers classes of Portuguese as a Foreign Language (PFL) at the university level. As a result of this increase in the number of students, Nunes and Lopes-Perna (2015) suggest that there is a need for more research describing academic Portuguese. Therefore, the goal of this paper is to examine the patterns of lexical bundles that are common in Brazilian research articles from the fields of life sciences and humanities.

Biber, Johansson, Leech, Conrad and Finegan (1999) define lexical bundles as words that usually go together in natural discourse, regardless of their idiomaticity and structural status. Different from other formulaic sequences, lexical bundles are empirically identified and allow for structural and functional comparisons between corpora. Biber and Conrad (1999) mention that, even though lexical bundles do not have specific grammatical functions, they have strong structural correlations. By using this correlation, researchers can categorize

bundles into prepositional, phrasal, and clausal bundles and use these categories to conduct systematic comparisons between bundles in two or more corpora. In addition to this structural classification, Biber, Conrad and Cortes (2004) introduced a functional taxonomy used to classify bundles according to their discourse functions. These functions – stance expression, discourse organizers, referential expressions, and special conversational functions – have been used in several studies to classify bundles (e.g., Cortes, 2004; Biber; Barbieri, 2007) in spoken and written registers. Other researchers have developed new classification schemes to match the discourse domains they were investigating. Hyland (2008), for instance, proposes new categories – research-oriented, text-oriented and participant-oriented – for the discourse functions associated with bundles in research articles.

In English, most research on lexical bundles has focused on academic writing, with the majority having investigated register variation (e.g., Biber; Barbieri, 2007), L1 and L2 production of lexical bundles in academic writing (e.g. Chen; Baker, 2010), student versus expert writing (e.g. Cortes, 2002, 2004), and disciplinary differences (e.g. Hyland, 2008; Cortes, 2004). The results of these studies have revealed patterns of variation, such as the use of more stance bundles in the life sciences compared to the humanities (Cortes, 2004), and the limited overlap of bundles across disciplines (Hyland, 2008).

Few studies have attempted to describe the use of bundles in languages other than English; these studies described Spanish (Cortes, 2008; Tracy-Ventura; Cortes; Biber, 2007), Korean (Kim, 2009), French (Granger, 2014) and Portuguese (Berber Sardinha; Teixeira; Ferreira, 2014). Their results reveal the need for more research describing academic language patterns in languages other than English, as this type of research allows for comparisons across languages and helps refine the method used for bundle extraction (see Granger, 2014).

Previous work looking at the use of lexical bundles in Portuguese did not provide an in-depth description of academic writing. Berber Sardinha, Teixeira and Ferreira (2014), for instance, explored the use of lexical bundles in 48 different registers in Portuguese. Their results revealed that research articles contain fewer bundles than political speeches and many other routinized registers (e.g., liturgy, recipes, medicine labels, etc.), but more bundles than spoken registers, such as soap operas, TV news and interviews. This study, however, focused on comparing the 48 registers in the corpus, lacking a detailed description of academic language taking into account bundle functions and structures, or comparisons between disciplines in academic Portuguese.

In sum, previous research has shown that there is disciplinary variation in the use of lexical bundles in research articles in English (Hyland, 2008) and Spanish (Cortes, 2004). The goal of this paper is to describe how two broad disciplinary groups (humanities and life sciences) use lexical bundles in research articles published in Portuguese. The research questions that guided this study were:

(1) What differences and similarities exist between the overall number of bundles extracted in humanities and life sciences?

(2) What differences exist in the structural and functional types of bundles between humanities and life sciences?

(3) What is the relationship between the structural and functional classification of bundles in humanities and life sciences?

To answer these research questions, a corpus of academic writing in Brazilian Portuguese was used, the Corpus de Português Escrito em Periódicos (CoPEP). Additionally, bundles were extracted and classified for function and structure. The following section (Methods) introduces the corpus and the method used for bundle extraction. Next, the results are presented and discussed in light of previous studies.

2 Methods

2.1 Corpus de Português Escrito em Periódicos (CoPEP)

The Corpus de Português Escrito em Periódicos (CoPEP), referred to in English as the Written Corpus of Portuguese Published in Academic Journals, was compiled by Kuhn and Ferreira (2020) in order to develop the dictionary of Portuguese for university students (Dicionário Online de Português para Universitários – DOPU) (Kuhn, 2017). This corpus contains research articles published in peer-reviewed Brazilian and Portuguese journals. According to Kuhn (2017, p. 103), these are "academic written texts portraying exemplary language". When CoPEP was developed the researchers' goal was to represent both Brazilian and European varieties of Portuguese. Therefore, 50% of the words in the corpus come from Brazilian journals and 50% come from Portuguese journals. In order to guarantee the quality of these research articles, Kuhn and Ferreira (2020) extracted the texts from a research database (SciELO) that only includes peer-reviewed papers, published in high impact journals in their fields. CoPEP includes research articles published from 1992 to 2018. For this study, a subcorpus of CoPEP was used, including only texts written by Brazilian scholars in the fields of life sciences and humanities.

Only Brazilian texts were included because lexical bundles reveal lexico-grammatical patterns, which are different in European (EU) Portuguese and Brazilian Portuguese. In EU Portuguese, for instance, the progressive aspect is expressed using the verb *to be* followed by a definite article (*a*) plus the infinitive form of the main verb (e.g., *estou a comprar* – I am shopping, *está a estudar* – He is studying), while in Brazilian Portuguese the progressive aspect is expressed using the verb *to be* plus the gerund form of the main verb (e.g. *estou comprando* – I am shopping, *está estudando* – He is studying). These grammatical differences would affect the results of the study if both European and Brazilian varieties were analyzed together, and contrasting the bundles used in European and Brazilian Portuguese is beyond the scope of this research. Another exclusion criterion was data of publication. Only papers published after 2000 were included in this study, as they represent more recent research practices. Table 1 details the subcorpora used in this study.

Table 1 – CoPEP - Brazilian Portuguese

	Disciplines Represented	Tokens	Texts
Humanities	Sociology, Psychology, Education, History, Geography, Philosophy, and Anthropology	15,294,433	3,100
Life Sciences	Physical Education, Nursing, Pharmacy, Medicine, Odontology, and Collective Health	7,875,989	1,596
Total		**23,170,422**	**4,696**

Source: the author.

2.2 Lexical Bundles Identification and Classification

Tracy-Ventura, Cortes and Biber (2007, p. 219) emphasize that lexical bundles are identified empirically based on frequency and range. This section addresses the criteria of frequency and range used to extract lexical bundles in both subcorpora. Although Granger (2014) suggests that in romance languages, it would be best to extract bundles of different sizes, the bundle size selected for this study was four-word sequences. There are three reasons for this: first, extracting bundles of different sizes would result in an unmanageable number of lexical bundles to classify and interpret; second, all three-word bundles in this corpus would be part of the four-word bundles; third, previous research of bundles in Portuguese (Berber Sardinha; Teixeira; Ferreira, 2014) and academic Spanish (Cortes, 2008) have adopted the four-word approach, thus making the results of the current study comparable to these previous studies.

Considering frequency, previous research on lexical bundles has adopted cut-off points varying from ten to forty times per million. After conducting some exploratory analysis, the cut-off point of twenty times per million was adopted in this study. This cut-off allowed a manageable number of bundles to be extracted from both subcorpora. In addition, for range, bundles had to occur in at least 2% of the texts in the corpus to guarantee that the bundles were not a representation of only one individual.

In order to answer research question two, the bundles were also classified functionally and structurally. Bundle structure was classified using a classification framework similar to the one adopted in previous studies (e.g. Cortes, 2008; Pan; Reppen; Biber, 2016). That is, the first word in the bundle was used as the base for classification, followed by the other phrases or clauses in the bundle. After adapting the classification to meet the specificities of Portuguese, the data was classified using the following criteria:

1. Noun Phrase

   a. Noun phrase: *consentimento livre e esclarecido* (known and free consent)[1]

   b. Noun phrase + prepositional phrase: *os profissionais da saúde* (health professionals)

[1] All bundles given as an example were translated by the author following the most approximate English translation; in some cases the translation is not grammatical, or not the most natural construction.

c. Noun phrase + prepositional phrase + prepositional phrase: *curso de graduação em* (undergraduate course in)

d. Noun phrase + preposition + dependent clause: *o fato de que* (the fact that)

e. Noun phrase + dependent clause: *uma vez que a* (once the)

f. Noun phrase + verb phrase: *os dados foram coletados* (the data was collected)

2. Prepositional phrase

a. Prepositional phrase + noun phrase: *de sólidos solúveis totais* (of total soluble solids)

b. Prepositional phrase + noun phrase + prepositional phrase: *ao mesmo tempo em* (at the same time as)

c. Prepositional phrase + prepositional phrase: *de serviços de saúde* (of health services)

d. Prepositional phrase + preposition + dependent clause: *na medida em que* (to the extent that)

e. Preposition + dependent clause + verb phrase + verb phrase: *no que se refere* (in what concerns)

3. Verb phrase

a. Verb phrase + prepositional phrase: *foi aprovado pelo comitê* (was approved by the committee)

b. Verb phrase + prepositional phrase + prepositional phrase: *aprovado pelo comitê de* (approved by the committee of)

c. Verb phrase + preposition + dependent clause: *tendo em vista que* (considering that)

d. Dependent clause + verb phrase + prepositional phrase: *que se refere à* (that refers to)

e. Dependent clause + noun phrase + prepositional phrase: *que a maioria dos* (that most of the)

It is worth noting some modifications to previous classifications; the first one was the exclusion of fragments. Cortes (2008) added a fragment marking (e.g. PP*f*) whenever a bundle ended in a fragment of a new phrase. While these fragments occur in the data analyzed here, they were categorized as complete phrases. *Ao mesmo tempo em* (at the same time as), for example, ends in a fragment, but it was classified as prepositional phrase + prepositional phrase. The reason for doing this is that the fragment category does not add significant information to the bundle classification, but it would mean adding at least two new categories to the data. In addition, the distinction between *of*-phrases and other prepositional phrases was not marked, since most of the prepositional phrases are controlled by "*of*". Furthermore, dependent clauses and verb phrases are grouped together since all bundles with a dependent clause included a verb phrase or are followed by one.

For the functional classification, bundles were initially classified following Biber, Conrad and Cortes (2004)'s classification, based on their function in context. In order to do

so, concordance lines for all the lexical bundles were extracted and then analyzed. Bundles were classified following their main discourse function. Subcategories were added to identify the specific meaning of the bundles. For the study conducted here, instead of following a previously existing classification, the subcategories were based on the bundles extracted. The category of stance, however, had few examples in the corpus, therefore no subcategory of stance was developed for this study. The final functional categories are the following:

1. Discourse organizer: have the function of structuring the text. They are usually recurrent in specific sections of a paper.

    a. Framing: *de acordo com o* (according to)

    b. Resultative: *com a finalidade de* (in order to)

    c. Explanatory: *tendo em vista que* (considering that)

    d. Contrast: *na medida em que* (to the extent that)

2. Referential: have the function of referencing physical or abstract entities.

    a. Entity: *da universidade federal de* (of the federal university of)

    b. Time: *segunda metade do século* (second half of the century)

    c. Place: *Rio Grande do Sul*

    d. Topic: *na área de saúde* (in the health field)

    e. Method: *a coleta de dados* (data collection)

    f. Subject: *os profissionais da saúde* (health professionals)

3. Stance: help writers express attitude towards themselves or others.

    o fato de que (the fact that)

Previous studies (e.g., Cortes, 2004; Chen; Baker, 2010) have excluded names of places or other entities from their bundle list. Nevertheless, this was not done in this study, since names of places and institutions reveal disciplinary practices in life sciences and humanities. Log-likelihood was calculated comparing disciplines for each subcategory.

Finally, following the criteria established above, 156 bundles were extracted, classified and analyzed in both disciplines. The next section discusses the results of the bundle analysis, divided by overall comparison, structural and functional comparison across disciplines.

3 Results and Discussion

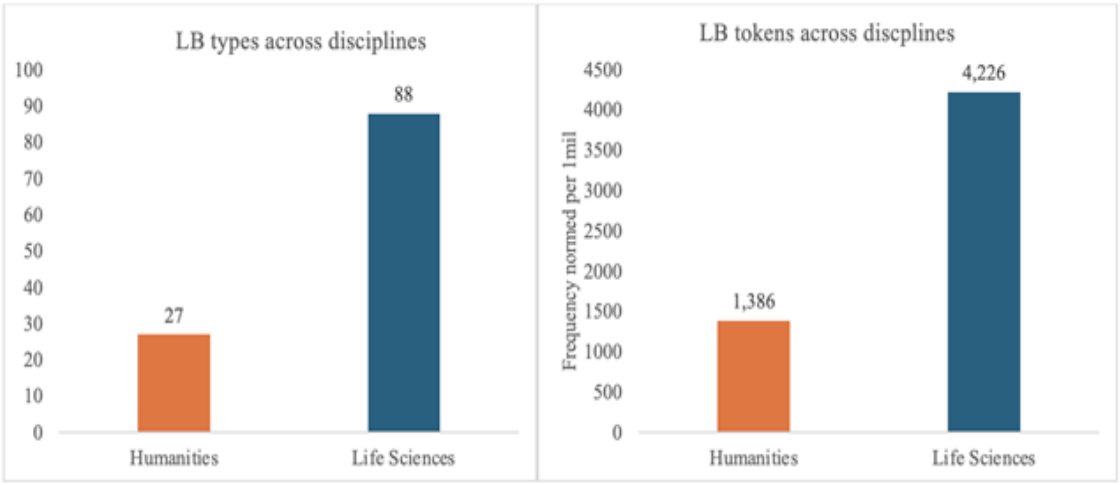
The results and discussion sections are organized in four different analyses. First, the overall comparison between both disciplinary subcorpora, focusing on the commonalities found across disciplines. Next, the comparison of structures taking into account types and tokens.

The functional types are compared across disciplines. The last part focuses on the comparison between functions identified across disciplines.

3.1 Lexical Bundles in Humanities and Life Sciences

Adopting the thresholds for extraction described above, there were 121 lexical bundles in the life sciences subcorpus and thirty-five in the humanities subcorpus. This considerable difference between the life sciences and humanities subcorpora is a result of linguistic variation in the use of gender markers and number markers in bundles with the same stem (e.g., *os pesquisadores* vs. *as pesquisadoras* – **the researchers**). These constructions were especially relevant in the life sciences subcorpus. Cortes (2008) reported encountering the same issue when comparing lexical bundles in Spanish and English. Cortes (2008) grouped bundles with the same stem (e.g. *en la que se*, *en las que se* and *en los que se*) in order to focus on bundle variation, rather than having the same bundle being analyzed multiple times. The same procedure was adopted with the bundles extracted in the Portuguese corpus (e.g *de acordo com o* and *de acordo com a* – both being '**according to**' in English). In addition, bundles with the same stem, but controlled by different prepositions, were grouped together (e.g. *dos/nos/do/os/o serviço de saúde* – **of/in/the health services**). This resulted in twenty-seven bundles in humanities and eighty-eight bundles in life sciences. Figure 1 shows the number of bundle types and tokens in each discipline. For the structural analysis, however, the original bundles were analyzed given that the structure information of the grouped bundles combined different grammatical categories.

Figure 1 – Lexical bundle types and tokens across disciplines[2]



Source: the author.

[2] For this figure and the following ones, the scale in the graph referring to the number of types is different from the graph referring to the number of tokens, because tokens are more frequent than types. In addition, the number of types was not normed, but the number of tokens was.

Figure 1 shows that the life sciences subcorpus has considerably more bundles than the humanities subcorpus for both types and tokens, even after grouping the bundles with the same stem. Furthermore, the results yielded thirteen bundles in common for both disciplines (48% of bundles in humanities and 15% in life sciences). Table 2 presents the lexical bundles shared by both disciplinary subcorpora. All the bundles extracted from both corpora can be found in Appendix A.

Table 2 – Shared bundles across disciplines

		Humanities		Life Sciences	
Bundles	Discourse Function	Normed per 1mil	Range N = 3,100	Normed per 1mil	Range N = 1,596
ao mesmo tempo em (at the same time as)	Referential	34.59	225	22.22	122
com o objetivo de (with the goal of)	Discourse organizer	47.21	266	71.10	362
de acordo com o/a/os/as (according to)	Discourse organizer	140.97	345	208.48	396
do ponto de vista (from the point of view)	Stance	114.75	353	41.90	202
estado de São Paulo (state of São Paulo)	Referential	34.33	103	57.39	155
mesmo tempo em que (at the same time as)	Referential	34.06	217	22.09	121
na medida em que (at the extent that)	Discourse organizer	128.48	304	51.68	185
no que diz respeito (in what concerns)	Discourse organizer	67.41	281	44.57	200
no que se refere (in what concerns)	Discourse organizer	59.83	224	75.55	161
o fato de que (the fact that)	Stance	72.05	410	25.39	147
que diz respeito à (what concerns the)	Discourse organizer	29.88	205	22.47	120
rio grande do sul	Referential	35.44	92	41.26	123
uma vez que a/o (once)	Discourse organizer	48.91	247	66.15	213

Source: the author.

Table 2 reveals that most of the shared bundles are discourse organizers, with *de acordo com o* being the most frequent bundle in both disciplines. Four referential bundles are shared in both disciplines, two of them (*Rio Grande do Sul* and *estado de São Paulo*) refer to specific places. These bundles are usually used to situate the data collection reported in the study, as excerpt 1 illustrates.

Excerpt 1 – doc#2392 - LS

Desse total, 945 foram coletados no estado do ***Rio Grande do Sul****; 602 no Maranhão e 248 no* ***estado de São Paulo****.*

From this total, 945 were collected in the state of **Rio Grande do Sul**; 602 in Maranhão and 248 in the **state of São Paulo**.[3]

The other two referential bundles found in common in both subcorpora are time references. These bundles could be nested together to form the expression *ao mesmo tempo em que* (at the same time as). As we can see in excerpt 2, in the humanities subcorpus this bundle is used to report events or actions that occur (or should occur) together. Excerpt 3 shows that in the life sciences subcorpus the same bundle is used to report on methodological steps in an experiment, treatment or diagnosis.

Excerpt 2 – doc#2330 – HU

Em outras palavras, ***ao mesmo tempo em que*** *reconhece que cabe a ele a gestão do seu desenvolvimento, o jovem percebe a competitividade profissional.*
In other words, **at the same time as** she/he recognizes that it is up to him/her to manage his/her development, the young person notices the professional competitivity.

Excerpt 3 – doc#757 – LS

... presença de palidez nos lobos medianos e lateral esquerdo, ***ao mesmo tempo em que*** *os lobos inferiores são sujeitos a isquemia.*
... presence of paleness in the center and left lobes, **at same time as** the inferior lobes are subject to ischemia.

The use of this bundle in both corpora indicates that although these are shared bundles, they can be used in different contexts across disciplines. Considering the stance bundles shared by both subcorpora, we can see that *o fato de que* (the fact that) is used by writers in both disciplines in contexts where the authors present new information and emphasize their discovery or point of view.

Excerpt 4 – doc#315- LS

...críticos estão causando grande prejuízo para os produtores de laranja. Aliado a isso está ***o fato de que*** *os europeus, acostumados a consumir o limão de casca amarela, estão aprendendo a consumir o Tahiti.*
...critics are causing great loss to orange producers. In combination is **the fact that** Europeans, who are used to consuming lemon are learning to consume lime.

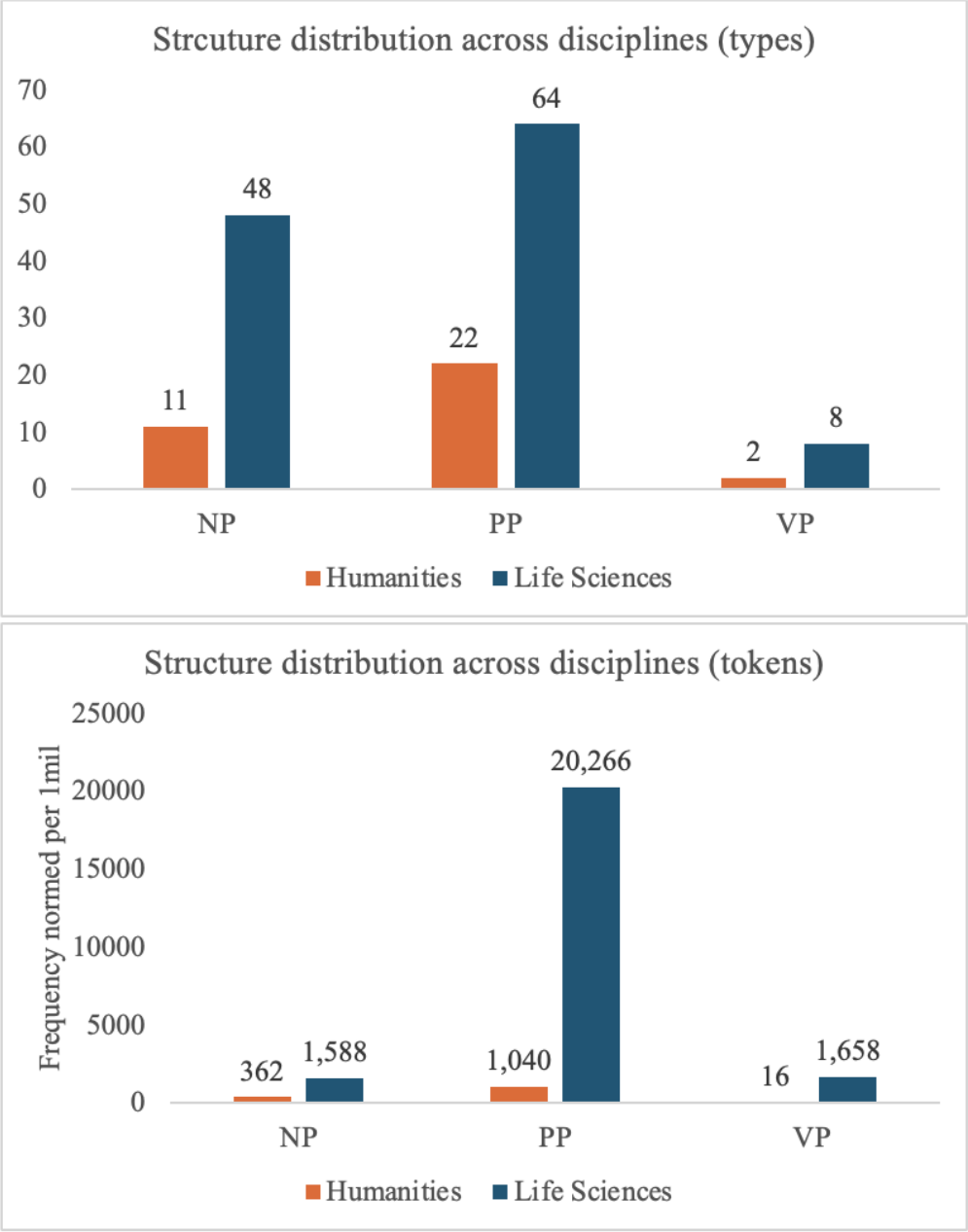
In academic English, Hyland (2008) also finds five bundles that are shared across several disciplines in academic writing. The limited number of shared lexical bundles across the disciplines highlights the need to take into account discipline variation when investigating and teaching academic Portuguese. In addition, some of the shared bundles, as *ao mesmo tempo em que* (at the same time as) are used in different contexts in these disciplines, suggesting that even shared bundles might not have the same meaning across disciplines. The following sections focus on comparing the structural and functional classification in the humanities and life sciences subcorpora.

[3] All excerpts were translated by the author following the most approximate English translation; in some cases the translation is not grammatical, or not the most natural construction.

3.2 Comparison of the Structural Types of Lexical Bundles

Using the structural classification detailed in the methods section, the bundles from the two disciplines were classified according to their structure. The main structure classification was prepositional phrases (PP), noun phrases (NP) and verb phrases (VP), followed by the subcategories described in the methods section. Figure 2 presents the distribution of the three main categories in both disciplinary subcorpora.

Figure 2 – Structure distribution across disciplines



Source: the author.

As we can see from Figure 2, although the number of types between noun phrase (forty-eight) and prepositional phrase (sixty-four) bundles are somewhat similar for the life sciences subcorpus, the number of tokens is significantly different. This indicates that the different types of noun phrases are not as frequent as the prepositional bundles. Table 3 presents the distribution of these main categories in the disciplines, considering their percentages in the text.

Table 3 – Distribution of main structural categories

	Types		Tokens	
	Humanities	Life Sciences	Humanities	Life Sciences
NP	31.43% (11)	40% (48)	25.54% (362)	36.34% (1,588)
PP	62.86% (22)	53.33% (64)	73.33% (1,040)	58.85% (20,266)
VP	5.71% (2)	6.67% (8)	1.13% (16)	4.81% (1,658)

Source: the author.

Table 3 shows that 25% of the bundles in humanities are noun phrase based, while 73% are prepositional phrase based. Table 3 indicates that the humanities subcorpus follows the same trend as the life sciences subcorpus, with prepositional bundles being used more frequently. This trend was also found in studies of academic writing in other languages, such as Spanish (Cortes, 2008) and English (Hyland, 2008). As pointed out by Cortes (2008), this happens in Romance languages, such as Spanish and Portuguese, because there are fewer forms of noun modification. English, for instance, allows noun-noun sequences; in Romance languages, these structures are translated as nouns followed by prepositional phrases.

Table 4 presents the raw number of types and tokens for each structure across disciplines. Numbers with an asterisk mark a statistically significant difference according to the log-likelihood test. This test was used to compare the raw frequency of tokens for the structural categories of both disciplines. The results show which structures have a significant difference and indicate which discipline uses that structure more frequently.

Table 4 – Distribution of structural subcategories

Categ	Structural subcategory	Types		Tokens	
		HU	LS	HU	LS
NP		11	48	5,540	12,514*
	NP	-	2	-	729
	NP + PP	4	33	1,705	8,603*
	NP + PP + PP	1	8	398	1,921*
	NP + PP + VP	-	1	-	198
	NP + prep + that-cl	4	2	2,689*	374
	NP + that-cl	2	2	748	521*
	NP + VP	-	1	-	168
PP		22	64	15,907	20,266*
	PP + NP	1	1	422	694*
	PP + NP + PP	2	2	1,251*	463
	PP + PP	15	57	9,999	17,756*
	PP + prep + that-cl	2	1	2,289*	407
	prep + that-cl + VP	2	2	1,941	946

VP		2	8	246	1,658*
	that-cl + NP + VP		1	-	168
	that-cl + VP + PP		3		613
	VP + PP		2		411
	VP + PP + PP		1		283
	VP + prep + that-cl	2	1	246	183*

* = significant at p <.001, * also indicates overuse by that discipline

Source: the author.

The results of the log-likelihood test indicate that bundles containing dependent clauses (NP + prep + that-cl and PP + prep + that-cl) are more common in the humanities subcorpus. Excerpt 5 illustrates the use of this construction in the humanities subcorpus.

Excerpt 5 – doc#2417 – HU

...nas pequenas porcentagens de grupos domésticos de famílias extensas, ***ao contrário do que*** *se poderia esperar.*
...in the small percentages of domestic groups with extended families, **contrary to** what might be expected.

The frequency of bundles with dependent clauses suggests that the humanities subcorpus contains more elaboration than the life sciences subcorpus, while the life sciences subcorpus contains more compression represented in the frequency of the prepositional bundles. Nevertheless, prepositional phrases modifying the noun phrase are the most common form of elaboration in both corpora, as excerpts 6 and 7 exemplify.

Excerpt 6 – doc#2285 - HU

...se uniram ***com o objetivo de*** *trabalhar em conjunto para buscar melhores negociações...*
...(workers) united **in order to** work together to seek better negotiations...

Excerpt 7 – doc#287 - LS

...tem grande importância ***do ponto de vista*** *da conservação biológica e da aplicação no melhoramento genético.*
...have great significance **from the point of view of** biological conservation and of applications in genetic improvement.

Excerpt 7 highlights the use of several prepositional phrases modifying noun phrases, a finding that has previously been discussed by Cortes (2008) when analyzing Spanish bundles. Appendix B contains all the bundles classified according to their subcategories, and Table 5 presents the most representative bundles for each subcategory in the data.

Table 5 – Examples of structural subcategories

Base	Subcategory	Humanities	Life Sciences
NP	NP		consentimento livre e esclarecido (known and free consent) material e métodos o (materials and methods the)
	NP + PP	Rio Grande do Sul estado de São Paulo (state of São Paulo) segunda metade do século (second half of the century) a existência de uma (the existence of a)	Rio Grande do Sul sistema único de saúde (national health system) a coleta de dados (the data collection)
	NP + PP + PP	ponto de vista da (point of view of the)	comitê de ética em (ethics committee in) curso de graduação em (undergraduate course in) escola de enfermagem da (nursing school of) ética em pesquisa da (ethics in research of)
	NP + PP + VP		coleta de dados foi (data collection was)
	NP + prep + that-cl	o fato de que (the fact that) mesmo tempo em que (same time that) a idéia de que (the idea that)	o fato de que (the fact that) mesmo tempo em que (same time that)
	NP + that-cl	uma vez que a uma vez que o (once)	uma vez que a uma vez que o (once)
	NP + VP		os dados foram coletados (the data was collected)
PP	PP + NP	com os Estados Unidos (with the US)	de consentimento livre e (of free consent) de sólidos solúveis totais (of total soluble solids)
	PP + NP + PP	com o objetivo de (with the objective of) ao mesmo tempo em (at the same time in)	com a finalidade de (with the objective of) ao mesmo tempo em (at the same time as)
	PP + PP	do ponto de vista (from the point of view) do rio de janeiro (of Rio de Janeiro) no mercado de trabalho (in the job market) de acordo com a (according to)	pelo comitê de ética (for the ethics committee) do departamento de enfermagem (of the nursing department) de coleta de dados (of the data collection) da universidade federal de (of the federal university of)
	PP + prep + that-cl	na medida em que (at the extent that) ao contrário do que (instead of what)	na medida em que (at the extent that)
	prep + that-cl + VP	no que diz respeito no que se refere (in what concerns)	no que diz respeito no que se refere (in what concerns)

VP	that-cl + NP + VP		que a maioria dos (that the majority of)
	that-cl + VP + PP	que diz respeito à que se refere à (what concerns the)	que se refere à que se refere ao que diz respeito à (what concerns the)
	VP + PP		foi aprovado pelo comitê (was approved by the committee) trata-se de um estudo (it is a study)
	VP + PP + PP		aprovado pelo comitê de (approved by the committee of)
	VP + prep + that-cl		tendo em vista que (considering that)

Source: the author.

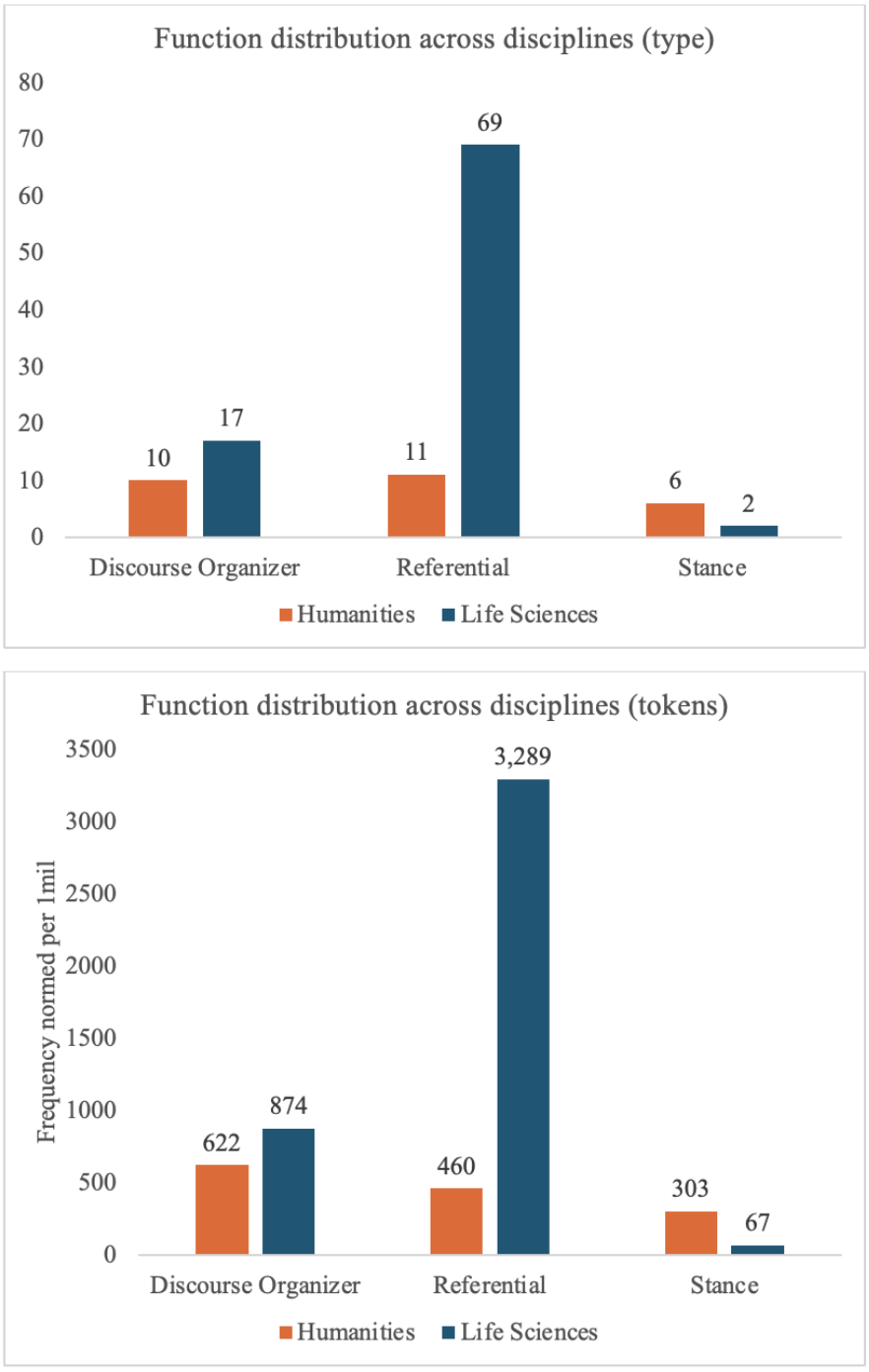
Table 5 shows that the life sciences subcorpus has more variety in bundle structure than the humanities subcorpus, with seven subcategories not occurring in the humanities bundles and all occurring in the life sciences bundles. We can also see the use of synonymous bundles across the data in both disciplines, as for example, *que se refere à* and *que diz respeito à* (what concerns the). Another interesting finding is that most of the prepositional bundles in life sciences are controlled by *de* (of); while this preposition is frequent in humanities, other prepositions, such as *em* (in), are also present in humanities.

In summary, the structure comparison revealed that prepositional bundles are the most frequent form in both the life sciences and the humanities subcorpora. Nevertheless, clausal bundles appear more frequently in the humanities subcorpus than in the life sciences one. The results also showed patterns of academic writing similar to Spanish, with prepositional phrases always appearing as noun modifiers. Finally, the life sciences subcorpus seems to be more prolific in relation to different types of structures when compared to the humanities subcorpus. Next section will present the results of the functional classification in both disciplines.

3.3 Comparison of the Functional Types of Lexical Bundles

The lexical bundles extracted in the life sciences and the humanities subcorpora were compared in relation to their discourse functions. This section discusses the results of this analysis. The bundles were first categorized for their main function (discourse organizers, referential or stance), and then analyzed within these categories in order to determine their subcategories. Figure 3 presents the functional distribution by types and tokens in both disciplines.

Figure 3 – Functional distribution across disciplines



Source: the author.

Figure 3 shows that the life sciences subcorpus has significantly more referential bundles in both tokens and types than the humanities subcorpus. The figure also indicates that stance bundles are more common in the humanities, which is the opposite trend than what was found in L1 English writing (Cortes, 2004). This contrast should be taken with caution, though, because stance bundles are not that frequent in the corpus. Table 6 indicates that stance represents 22% of the total types in the humanities subcorpus, while in the life

sciences subcorpus it represents only 2% of the total number of types. Table 6 also highlights the difference in the number of referential bundles across disciplines, with 77.1% of the tokens in life sciences coming from this category and only 33.23% in humanities.

Table 6 – Distribution of main functional categories

	Types		Tokens	
	Humanities	Life Sciences	Humanities	Life Sciences
Discourse Organizer	37.04% (10)	19.32% (17)	44.87% (622)	20.70% (874)
Referential	40.74% (11)	78.41% (69)	33.23% (460)	77.71% (3,289)
Stance	22.22% (6)	2.27% (2)	21.90% (303)	1.59% (67)

Source: the author.

Referential bundles might be more common in the life sciences subcorpus because these writers use more bundles to refer to the institutions connected to the research and to the participants in the study reported, as can be seen in excerpts 8 and 9.

Excerpt 8 – doc#208 – LS

Todos os abacaxizeiros obtidos foram plantados na área experimental ***da Universidade Federal de*** *Goiás.*
All the pineapple plants were planted in the experimental area of the **Federal University of** Goiás.

Excerpt 9 – doc#1115 - LS

Neste estudo, 78,13% dos membros ***da equipe de enfermagem*** não verificou o VRG antes de alterar a dieta.
In this study, 78.13% of the **nursing team** members did not check the GRV before changing the diet.

Similar to the structure analysis, a Log-likelihood was computed for the number of tokens in each subcategory across disciplines. Table 7 presents the number of types and tokens for each subcategory of lexical bundles identified in both disciplinary subcorpora.

Table 7 – Distribution of functional subcategories

Category	Subcategory	Types		Tokens	
		HU	LS	HU	LS
Discourse Organizers		10	17	9,518	6,889*
	Framing	3	3	3,291	2,029*
	Contrast	1	1	1,965*	407
	Explanatory	5	10	3,514	3,319*
	Resultative	1	3	748	1,131*
Referential		11	69	7,048	25,869*
	Entity	1	21	1,650	6,943*
	Methods	-	10	-	4,179
	Place	4	10	2,616	2,811*
	Subjects	-	6	-	3,392
	Time	5	3	2,468*	613
	Topic	1	19	314	7,931*
Stance		6	2	4,645*	530

* = significant at p <.001, * also indicates overuse by that discipline

Source: the author.

The results of the log-likelihood test reveal that most of the patterns are more common in the life sciences subcorpus than in the humanities subcorpus, with referential bundles in the subcategories of entities depicting the greatest difference between token frequency. As it was shown in excerpt 8, bundles referring to entities are used in life science research articles to account for an institution where the research took place. Excerpt 10 illustrates a bundle that mentions an institution that supported the research.

Excerpt 10 – doc#937 – LS

Houve predomínio de fornecimento gratuito de medicamento através do ***sistema único de saúde****.*
Most of the free medication supplies came from the **national health system**.

Table 7 also shows that there are two subcategories that are not found in the humanities subcorpus, "subjects" and "methods". Bundles associated with methods relate to the steps taken for data collection, such as ethical considerations and specific procedures, as excerpt 11 shows.

Excerpt 11 – doc#980 – LS

O instrumento ***de coleta de dados*** *foi um questionário com perguntas estruturadas relativas às questões norteadoras...*
The **data collection** instrument was a questionnaire with structured questions concerning the guiding questions...

The log-likelihood also showed that three subcategories are more frequent in the humanities subcorpus than in the life sciences one. Not surprisingly, stance is one of them,

followed by time referential bundles and contrast bundles. Time references are more common in humanities as they relate to the topics being investigated, as can be seen in excerpt 12.

Excerpt 12 – doc#2341 – HU

...foi o modelo utilizado em muitas nações industrializadas a partir da ***segunda metade do*** século passado.
...it was the model used in many industrialized nations in the **second half of** last century.

Another interesting finding is that both the life sciences and humanities subcorpora have only one type of contrast bundle, *na medida em que*. Nevertheless, this bundle is significantly more frequent in the humanities subcorpus. In this subcorpus, this bundle is used to introduce facts and the work of others in research articles. In addition, this bundle occurs in almost every text in the humanities subcorpus. Excerpt 13 shows an example of fact organization, while excerpt 14 presents an example of integrating the work of others.

Excerpt 13 – doc#2300 – HU

Na medida em que *existam, além das sobras, outros indicadores para representar a reputação do conselho junto ao cooperado...*
As long as there are, aside from the surplus, other indicators to represent the council reputation to the partners...

Excerpt 14 – doc#2304 – HU

...para o alcance dos resultados pela organização, ***na medida em que****, segundo Matias-Pereira (2009) "o planejamento possibilita a coordenação...*
... to achieve the results of the organization, **to the extent that**, according to Matias-Pereira (2009) "the planning allows the coordination...

We can see that in both excerpts *na medida em que* is being used to introduce a different side of an argument, but this new side can be introduced either by new facts or citations. Appendix C contains all the bundles with their functional classification. Table 8 has examples of the most salient bundles in each subcategory.

Table 8 – Examples of functional subcategories

	Humanities	Life Sciences
Discourse organizer		
Framing	de acordo com o/a/os/as (according to) a partir de um/uma (from a) por meio de uma (through a)	de acordo com o/a/os/as (according to) a análise dos dados (the data analysis) trata-se de um estudo (it is a study)
Contrast	na medida em que (at the extent that)	na medida em que (at the extent that)
Explanatory	no que diz respeito no que se refere (in what concerns) com o objetivo de (with the goal of) que diz respeito à (in what concerns)	no que se refere (in what concerns) com o objetivo de (with the objective of) que se refere à/ao (what concerns the) no que diz respeito (in what concerns)
Resultative	uma vez que a/o (once)	uma vez que a/o (once) para o desenvolvimento da/de (to the development of) que a maioria dos (that most of)
Referential		
Entity	do/no mercado de trabalho (of the job market)	do/o/pelo ministério da saúde (of the health ministry) da universidade federal de/da (of the federal university of) sistema único de saúde (national health system)
Methods		de/a coleta de dados (of the data collection) de ética em pesquisa (of ethics in research)
Place	do/no rio de janeiro (of Rio de Janeiro) Rio Grande do Sul estado de São Paulo (state of São Paulo)	estado de São Paulo (state of São Paulo) do/no estado de São (of the state of São) do Rio de Janeiro (of Rio de Janeiro)
Subjects		dos/os profissionais de saúde (of health professionals) da/a equipe de enfermagem (of the nursing team)
Time	ao longo do tempo (throughout time) ao mesmo tempo em (at the same time as) mesmo tempo em que (same time that)	durante o período de (during the period of) ao mesmo tempo em (at the same time as) mesmo tempo em que (same time that)
Topic	a existência de uma (the existence of a)	dos/nos/os/de/ao serviços de saúde (of health services) na/da/de/a qualidade de vida (of quality of life) de/a promoção de/da saúde (of health promotion)
Stance		
	do ponto de vista (of the point of view) fato de que (the fact that) a idéia de que (the idea that)	do ponto de vista (of the point of view) fato de que (the fact that)

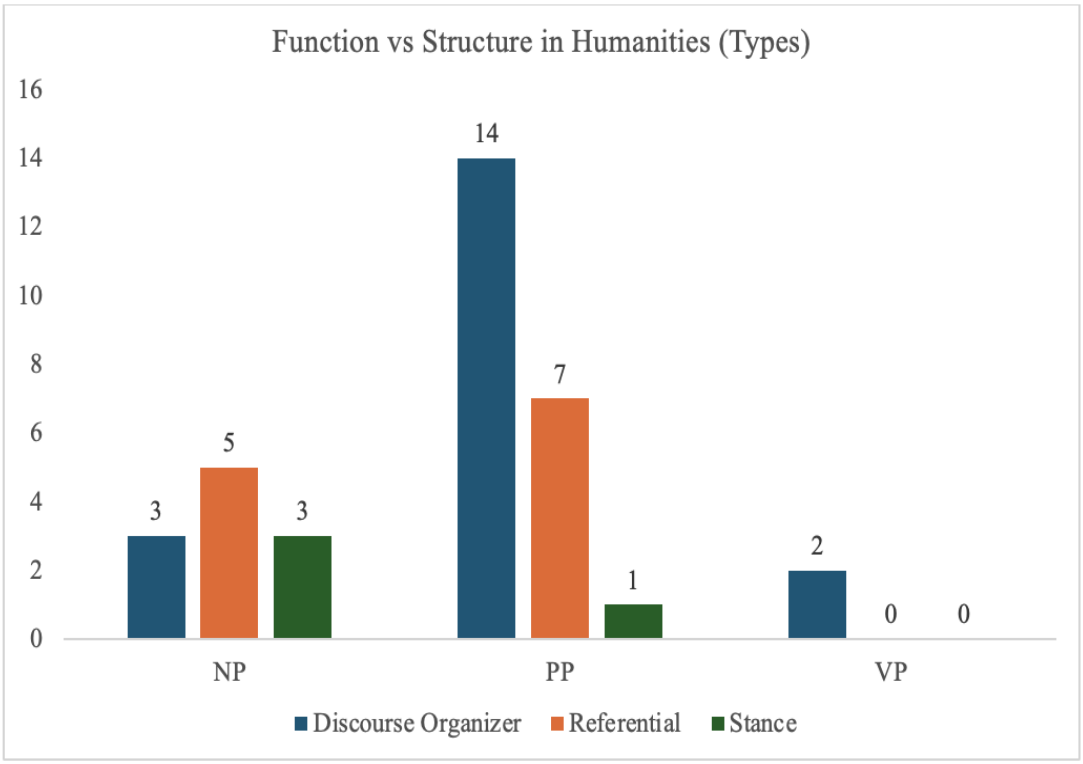
Source: the author.

The functional comparison revealed that research articles in life sciences tend to have more referential bundles, because they report on specific research subjects, place of data collection and methodological procedures, while the same does not occur in humanities. Research articles in the humanities are characterized by bundles referring to time of research, and contrast in order to show opposite sides of an argument and expressions of stance. The following section analyses the relationship between form and function in both disciplines.

3.4 The Relationship Between Form and Function Across Disciplines

In English, Biber, Conrad and Cortes (2004) have found a relationship between form and function in lexical bundles, with stance bundles usually having a clausal structure. Since this paper seeks to provide a description of lexical bundles in academic Portuguese, the same relation was explored for the main categories of structure and function across disciplines. Figures 4 and 5 present the results of this comparison for the number of types in each disciplinary subcorpora. Figure 4 shows that for humanities, noun phrase bundles usually have referential function and prepositional based bundles usually occur as discourse organizers.

Figure 4 – Function and structure in humanities



Source: the author.

The use of noun phrases as referential bundles in the humanities subcorpus relates to mentions of place and time markers. The five bundle types identified as noun phrases are part of the following subcategories: place – *estado de São Paulo* (excerpt 1), *Rio Grande do Sul* (excerpt 1); time – *segunda metade do século* (excerpt 12), *mesmo tempo em que* (excerpt 2); and

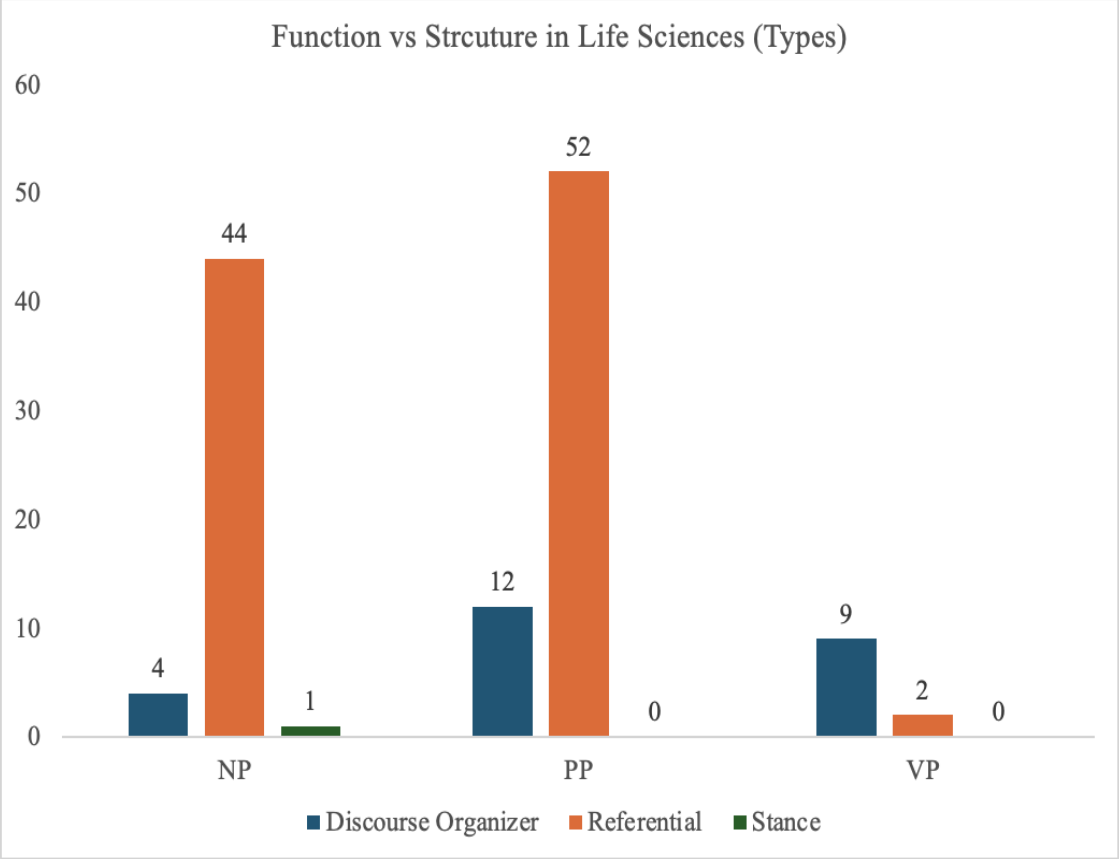
topic – *a existência de uma*. The prepositional phrase bundles are usually discourse organizers, as excerpt 15 illustrates. One of the reasons for this are these fixed structures such as *de acordo com o/a/os/as* and *na medida em que*.

Excerpt 15 – doc#2335 – HU

***De acordo com o** objetivo proposto neste estudo, busca-se encontrar um modelo de análise do ambiente...*
According to the goals of this study, we seek to find a model of environmental analysis...

As we can see, the bundle in excerpt 15 is used to organize the different parts of the text and to connect these sections. Figure 5 shows that for the life sciences subcorpus, both noun and prepositional phrases bundles are used with referential meaning, reinforcing the findings of section 3.3 that reference bundles are prevalent in life sciences. In addition, verb phrase bundles tend to be discourse organizers.

Figure 5 – Function and structure in life sciences



Source: the author.

As was illustrated in excerpts 8 and 9, prepositional bundles in the life sciences subcorpus usually have the referential function of mentioning research participants, or institutions involved in the data collection. Excerpt 16 illustrates the use of noun phrases as referential bundles in life sciences. In this case we have two bundles, referring to the same term, namely *termo de consentimento livre e esclarecido*, which relates to the methods used for data collection.

Excerpt 16 – doc#2335 – HU

Os voluntários assinaram ***o termo de consentimento livre e esclarecido****.*
Volunteers signed the **free and known consent form**.

The comparison between form and function shows once more that there are disciplinary differences in the use of lexical bundles in the life sciences and the humanities subcorpora. Figures 4 and 5 suggest that humanities bundles have a pattern distinguishing between form and function. In humanities, prepositional bundles tend to be organizers and noun-phrase bundles tend to be referential. In life sciences, however, referential bundles are the preferred semantic category for both noun and prepositional bundles.

The goal of this section was to present the results and discuss the comparisons between the functional and structural categorization in humanities and life sciences. The next section will summarize the results and discuss possible implications for academic Portuguese teaching.

4 Conclusion

The study reported here explored the use of lexical bundles in research articles written in Portuguese across two disciplines, humanities and life sciences. Four-word bundles were extracted from both corpora, using as a threshold the frequency of twenty times per million. The bundles that matched the criteria were then classified for their functional and structural categories. Finally, the bundles extracted from each discipline were compared to account for variation in humanities and life sciences writing.

The results suggest that writing for life sciences in Portuguese is more formulaic than for humanities. The same was reported for English writing in biology in Cortes (2004). A possible explanation for this is that research articles in this discipline tend to use the same structure, even using the same sequences of words, especially in the methods section, where the authors report ethical approval, the institution where the study was conducted, and the funding agencies.

In relation to structure and function categories, the analysis revealed that life sciences has more details as to the places of data collection, entities involved in the study and subjects that took part in the projects. As a result, this discipline had more prepositional phrase bundles and more referential bundles as well. Humanities seems to have more stance bundles and to use more bundles with dependent clauses when compared to life sciences. Furthermore, the fact that humanities contained fewer bundles might be a result of topic variation in this discipline.

For teachers of Portuguese as an academic language, this study emphasizes the need to address discipline specific writing in the classroom, since the results suggest that there are differences across disciplines. Furthermore, it points out the fact that students in life sciences should be exposed to referential bundles if they want to participate in the disciplinary discourse of their fields. The results also indicate that students writing for humanity classes should learn how to use discourse organizer bundles, as these are more relevant for this discipline than referential bundles. The list in Appendix A presents all the bundles extracted

for this study with their normed frequency and range. Teachers can use it to introduce the most frequent bundles in their discipline, or use the lists in appendices B and C to create activities with lexical bundles by function or structure.

Finally, this study is not without its limitations. First, exploring bundles of larger size in Portuguese might show different disciplinary patterns. For example, some of the bundles extracted were combinations of the same sequences as *ao mesmo tempo em que*. In addition, it might be worth considering less stringent measures of frequency and range when extracting bundles, as this resulted in a limited number of bundle types that could be addressed in the classroom. Finally, it might be interesting to investigate *p*-frames in Portuguese, as this would give students of Portuguese sequences that they can adapt in order to fit their research topic.

References

BERBER SARDINHA, T. B.; TEIXEIRA, R.; FERREIRA, R. Lexical Bundles in Brazilian Portuguese. *In*: BERBER SARDINHA, T.; FERREIRA, T. (ed). *Working with Portuguese Corpora*. London: Bloomsbury, 2014. p. 33-68.

BIBER, D.; BARBIERI, F. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, v. 26, n. 3, p. 263-286, 2007.

BIBER, D.; CONRAD, S. Lexical Bundles in Conversation and Academic Prose. *Language and Computers*, v. 26, p. 181-190, 1999.

BIBER, D.; CONRAD, S.; CORTES, V. If You Look at...: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, v. 25, n. 3, p. 371-405, 2004.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *Longman Grammar of Spoken and Written English*. London: Longman, 1999. p. 1204.

CHEN, Y. H.; BAKER, P. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology*, v. 14, n. 2, p. 30-49, 2010.

CORTES, V. A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish. *Corpora*, v. 3, n. 1, p. 43-57, 2008.

CORTES, V. Lexical Bundles in Freshman Composition. *In*: REPPEN, R.; FITZMAURICE, S.; BIBER, D. (ed.). *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins, 2002. p. 131-145.

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

GRANGER, S. A Lexical Bundle Approach to Comparing Languages: Stems in English and French. *Languages in Contrast*, v. 14, n. 1, p. 58-72, 2014.

HYLAND, K. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes*, v. 27, n. 1, p. 4-28, 2008.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. Sinopse estatística da educação Superior. Brasília: Inep, 2019. Available at: http://inep.gov.br/sinopses-estatisticas-da-educacao-basica. Accessed on: Jan. 2024.

KIM, Y. Korean Lexical Bundles in Conversation and Academic Texts. *Corpora*, v. 2, n. 4, p. 135-165, 2009.

KUHN, T. *A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students*. 2017. Dissertation (PhD in Linguistics) – University of Lisbon, Lisbon, 2017.

KUHN, T. Z.; FERREIRA, J. P. O Corpus de Português Escrito em Periódicos – CoPEP. *DELTA*: Documentação de Estudos em Lingüística Teórica e Aplicada, v. 2, n. 36, p. 1-42, 2020.

NUNES, S.; LOPES-PERNA, C. Uma análise dos substantivos como marcadores de posicionamento em artigos acadêmicos em Língua Portuguesa. *Brazilian English Language Teaching Journal*, v. 6, n. 2, p. 172-195, 2015.

PAN, F.; REPPEN, R.; BIBER, D. Comparing Patterns of L1 Versus L2 English Academic Professionals: Lexical Bundles in Telecommunications Research Journals. *Journal of English for Academic Purposes*, v. 21, p. 60-71, 2016.

TRACY-VENTURA, N.; CORTES, V.; BIBER, D. Lexical Bundles in Spanish Speech and Writing. *In*: PARODI, G. (ed.). *Working with Spanish Corpora*. London: Bloomsbury, 2007. p. 217-231.

History

Date of submission: 13/11/2024.
Date of approval: 16/06/2025.

Responsible Editor

Andréa Machado de Almeida Mattos, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais/MG, Brasil. Lattes: http://lattes.cnpq.br/7749222257907067, ORCID: https://orcid.org/0000-0003-3190-7329, e-mail: andreamattos@ufmg.br.

Declaration of Conflict of Interest

The author declares that there is no conflict of interest.

Data Availability Statement

Data is available upon request.

Use of AI

During the preparation of this work, the author did not use AI.

Reviews

As part of the commitment made by the Brazilian Journal of Applied Linguistics to Open Science, the journal publishes the reviews issued regarding its published works, when authorized by all parties involved.

Review 1

Round 1

O manuscrito apresenta um estudo bem fundamentado sobre o uso de pacotes lexicais na escrita acadêmica brasileira, explorando as particularidades das áreas de ciências biológicas e humanidades. A pesquisa se destaca pela abordagem cuidadosa na coleta e análise de dados, além do embasamento teórico consistente, que dialoga com estudos anteriores na área. De forma geral, o manuscrito demonstra qualidade e potencial impacto na área, sendo uma excelente contribuição para os estudos sobre escrita acadêmica em português.

Nesse sentido, recomendo a aprovação do manuscrito para publicação, com pequenas sugestões de melhoria, a saber: i) A inclusão de um parágrafo de transição entre a introdução e a seção de metodologia ajudaria a conectar as ideias de forma mais fluida. ii) Revisão das descrições de quadros e tabelas para assegurar conformidade com as normas mais recentes da ABNT.

Reviewed on: 21/01/2025.
Recommendation: Accept Submission.

Round 2

Após examinar a última versão do manuscrito submetido a este periódico/evento, verifico que as solicitações de revisão apontadas anteriormente foram satisfatoriamente contempladas pelos autores. O trabalho encontra-se, a meu ver, em condições adequadas para publicação. Desta forma, recomendo o aceite do artigo para publicação.

Reviewed on: 28/04/2025.
Recommendation: Accept Submission.

Reviewer: Leonardo Neves Correa.

Appendix A – Lexical bundles in humanities and life sciences

Humanities			Life Sciences			Life Sciences		
Lexical Bundle	Norm Freq	Range	Lexical Bundle	Norm Freq	Range	Lexical Bundle	Norm Freq	Range
de acordo com o/a/os/as	140.97	345	de acordo com o/a/os/as	208.48	396	na maioria das vezes	37.46	189
na medida em que	128.48	304	dos/nos/os/de/ao serviços de saúde	177.63	176	com a finalidade de	36.57	195
do ponto de vista	114.75	353	dos/os profissionais de saúde	109.32	181	aprovado pelo comitê de	35.93	281
do/no mercado de trabalho	107.88	91	da/a equipe de enfermagem	108.81	121	o termo de consentimento	35.55	266
do/no rio de janeiro	73.69	102	de/a coleta de dados	101.70	288	do curso de enfermagem	34.03	71
o fato de que	72.05	410	na/da/de/a qualidade de vida	93.07	155	com o intuito de	33.77	186
no que diz respeito	67.41	281	do/o/pelo ministério da saúde	88.37	178	durante o período de	33.52	163
no que se refere	59.83	224	da universidade federal de/da	82.28	195	do conselho nacional de	32.88	245
a partir de um/uma	51.26	234	da/a assistência de enfermagem	76.69	104	do sistema único de	32.76	166
uma vez que a/o	48.91	247	no que se refere	75.55	161	ética em pesquisa da	32.00	252
com o objetivo de	47.21	266	de/a promoção de/da saúde	75.42	97	conselho nacional de saúde	31.11	228
ao longo do tempo	46.95	213	com o objetivo de	71.10	362	do rio grande do	30.85	106
a idéia de que	46.68	289	do/o cuidado de enfermagem	69.83	91	material e métodos o	29.71	234
rio grande do sul	35.44	92	da/a equipe de saúde	69.58	141	de sólidos solúveis totais	29.08	51
ao mesmo tempo em	34.59	225	da/na área de saúde	68.44	178	teor de sólidos solúveis	28.57	54
estado de São Paulo	34.33	103	de ética em pesquisa	67.80	482	foi aprovado pelo comitê	27.68	216
mesmo tempo em que	34.06	217	de graduação em enfermagem	67.17	104	do curso de graduação	27.55	84
que diz respeito à	29.88	205	do/o processo de trabalho	66.40	83	do sistema de saúde	27.30	109
com os estados unidos	27.59	57	uma vez que a/o	66.15	213	do processo de enfermagem	27.04	40
ponto de vista da	26.02	192	de saúde da família	64.63	75	no mercado de trabalho	26.41	90
que se refere à	25.43	189	consentimento livre e esclarecido	62.85	444	o fato de que	25.39	147
a partir dos anos	24.58	158	de/a educação em saúde	60.69	44	unidade de terapia intensiva	25.39	73
fato de que a	23.01	234	termo de consentimento livre	59.80	422	a análise dos dados	24.63	166

por meio de uma	22.95	212	comitê de ética em	59.17	448	trata-se de um estudo	24.50	191
segunda metade do século	21.18	121	de consentimento livre e	59.04	419	associação brasileira de enfermagem	24.25	72
ao contrário do que	21.18	208	sistema único de saúde	57.64	224	serviços de saúde e	24.00	125
a existência de uma	20.53	190	da escola de enfermagem	57.52	145	tendo em vista que	23.24	117
			pelo comitê de ética	57.52	443	qualidade de vida e	23.11	98
			estado de são paulo	57.39	155	os critérios de inclusão	22.98	122
			curso/s de graduação em	57.01	85	enfermagem da universidade federal	22.47	88
			para o desenvolvimento da/de	56.12	170	que diz respeito à	22.47	120
			que se refere à/ao	55.36	161	ao mesmo tempo em	22.22	122
			escola de enfermagem da/de	54.85	87	mesmo tempo em que	22.09	121
			na medida em que	51.68	185	universidade federal do rio	21.58	99
			do/no estado de são	50.66	105	da faculdade de medicina	21.46	90
			de atenção à saúde	49.26	113	estratégia saúde da família	21.46	56
			do rio de janeiro	46.34	96	os dados foram coletados	21.33	150
			dos/os profissionais de enfermagem	45.07	85	que a maioria dos	21.33	125
			no que diz respeito	44.57	200	profissionais de saúde e	21.20	109
			do departamento de enfermagem	42.79	130	unidades básicas de saúde	21.08	69
			do ponto de vista	41.90	202	curso de enfermagem da	20.57	43
			rio grande do sul	41.26	123	de saúde da população	20.57	86
			universidade de são paulo	40.88	128	atenção primária à saúde	20.31	36
			de enfermagem da universidade	40.63	133	fatores de risco para	20.31	67

Source: the author.

Appendix B – Structural classification of bundles

Base	Sub-category	Humanities	Life Sciences
NP	NP		consentimento livre e esclarecido; material e métodos o
	NP + PP	rio grande do sul; estado de são paulo; segunda metade do século; a existência de uma	rio grande do sul; sistema único de saúde; a coleta de dados; termo de consentimento livre; os profissionais de saúde; a equipe de enfermagem; estado de são paulo; a qualidade de vida; universidade de são paulo; o termo de consentimento; a promoção da saúde; o cuidado de enfermagem; conselho nacional de saúde; a educação em saúde; o processo de trabalho; a equipe de saúde; os serviços de saúde; teor de sólidos solúveis; o ministério da saúde; unidade de terapia intensiva; a análise dos dados; associação brasileira de enfermagem; serviços de saúde e; a assistência de enfermagem; qualidade de vida e; os critérios de inclusão; enfermagem da universidade federal; universidade federal do rio; estratégia saúde da família; profissionais de saúde e; unidades básicas de saúde; atenção primária à saúde; os profissionais de enfermagem
	NP + PP + PP	ponto de vista da	comitê de ética em; curso de graduação em; escola de enfermagem da; ética em pesquisa da; cursos de graduação em; escola de enfermagem de; curso de enfermagem da; fatores de risco para
	NP + PP + VP		coleta de dados foi
	NP + prep + that-cl	o fato de que; mesmo tempo em que; a idéia de que; fato de que a	o fato de que; mesmo tempo em que
	NP + that-cl	uma vez que a; uma vez que o	uma vez que a; uma vez que o
	NP + VP		os dados foram coletados
PP	PP + NP	com os estados unidos	de consentimento livre e; de sólidos solúveis totais
	PP + NP + PP	com o objetivo de; ao mesmo tempo em	com a finalidade de; ao mesmo tempo em

	PP + PP	do ponto de vista; do rio de janeiro; no mercado de trabalho; de acordo com a; no rio de janeiro; ao longo do tempo; de acordo com o; do mercado de trabalho; a partir de uma; a partir dos anos; por meio de uma; da américa do sul; de acordo com os; a partir de um; de acordo com as	pelo comitê de ética; do departamento de enfermagem; de coleta de dados; da universidade federal de; de ética em pesquisa; da escola de enfermagem; do rio de janeiro; da assistência de enfermagem; dos profissionais de saúde; com o objetivo de; de acordo com o; da equipe de enfermagem; de saúde da família; de atenção à saúde; de acordo com os; de acordo com as; da qualidade de vida; de acordo com a; do ponto de vista; dos serviços de saúde; na área da saúde; da equipe de saúde; de enfermagem da universidade; na maioria das vezes; do ministério da saúde; do processo de trabalho; do cuidado de enfermagem; nos serviços de saúde; da universidade de são; do curso de enfermagem; com o intuito de; durante o período de; para o desenvolvimento de; do conselho nacional de; do sistema único de; da universidade federal do; do rio grande do; de educação em saúde; do estado de são; da área da saúde; do curso de graduação; do sistema de saúde; do processo de enfermagem; no mercado de trabalho; de serviços de saúde; pelo ministério da saúde; dos profissionais de enfermagem; na qualidade de vida; para o desenvolvimento da; da faculdade de medicina; aos serviços de saúde; de qualidade de vida; de promoção da saúde; de saúde da população; no estado de são; da criança e do; de graduação em enfermagem
	PP + prep + that-cl	na medida em que; ao contrário do que	na medida em que
	prep + that-cl + VP	no que diz respeito; no que se refere	no que diz respeito; no que se refere
VP	that-cl + NP + VP		que a maioria dos
	that-cl + VP + PP	que diz respeito à; que se refere à	que se refere à; que se refere ao; que diz respeito à
	VP + PP		foi aprovado pelo comitê; trata-se de um estudo
	VP + PP + PP		aprovado pelo comitê de
	VP + prep + that-cl		tendo em vista que

Source: the author.

Appendix C – Functional classification of lexical bundles

	Humanities	Life Sciences
Discourse organizer		
Framing	de acordo com o/a/os/as; a partir de um/uma; por meio de uma	de acordo com o/a/os/as; material e métodos o; a análise dos dados; trata-se de um estudo
Contrast	na medida em que	na medida em que
Explanatory	no que diz respeito; no que se refere; com o objetivo de; que diz respeito à; que se refere à	no que se refere; com o objetivo de; que se refere à/ao; no que diz respeito; na maioria das vezes; com o intuito de; tendo em vista que; que diz respeito à; os dados foram coletados
Resultative	uma vez que a/o	uma vez que a/o; para o desenvolvimento da/de; com a finalidade de; que a maioria dos
Referential		
Entity	do/no mercado de trabalho	do/o/pelo ministério da saúde; da universidade federal de/da; sistema único de saúde; da escola de enfermagem; pelo comitê de ética; curso/s de graduação em; escola de enfermagem da/de; do departamento de enfermagem; de enfermagem da universidade; aprovado pelo comitê de; do curso de enfermagem; do conselho nacional de; do sistema único de; conselho nacional de saúde; foi aprovado pelo comitê; do curso de graduação; do sistema de saúde; no mercado de trabalho; associação brasileira de enfermagem; enfermagem da universidade federal; curso de enfermagem da
Methods	do/no rio de janeiro; rio grande do sul; estado de São Paulo; com os estados unidos	de/a coleta de dados; de ética em pesquisa; consentimento livre e esclarecido; termo de consentimento livre; comitê de ética em; de consentimento livre e; o termo de consentimento; ética em pesquisa da; os critérios de inclusão
Place		estado de são paulo; do/no estado de são; do rio de janeiro; rio grande do sul; universidade de são paulo; do rio grande do; universidade federal do rio; da faculdade de medicina; unidades básicas de saúde; unidade de terapia intensiva
Subjects		dos/os profissionais de saúde; da/a equipe de enfermagem; da/a assistência de enfermagem; da/a equipe de saúde; dos/os profissionais de enfermagem; profissionais de saúde e
Time	ao longo do tempo; ao mesmo tempo em; mesmo tempo em que; a partir dos anos; segunda metade do século	durante o período de; ao mesmo tempo em; mesmo tempo em que
Topic	a existência de uma	dos/nos/os/de/ao serviços de saúde; na/da/de/a qualidade de vida; de/a promoção de/da saúde; do/o cuidado de enfermagem; da/na área de saúde; de graduação em enfermagem; do/o processo de trabalho; de saúde da família; de/a educação em saúde; de atenção à saúde; de sólidos solúveis totais; teor de sólidos solúveis; do processo de enfermagem; serviços de saúde e; qualidade de vida e; estratégia saúde da família; de saúde da população; atenção primária à saúde; fatores de risco para
Stance		
	do ponto de vista; fato de que; a idéia de que; ponto de vista da; fato de que a; ao contrário do que	do ponto de vista; o fato de que

Source: the author.