OPEN ACCESS

ESCOLA DE HUMANIDADES
PUCRS

LETRAS DE HOJE
Studies and debates in linguistics, literature and Portuguese language
Letras de hoje Porto Alegre, v. 55, n. 4, p. 477-495, out.-dez. 2020
e-ISSN: 1984-7726 | ISSN-L: 0101-3335

http://dx.doi.org/10.15448/1984-7726.2020.4.38377

PORTUGUÊS COMO LÍNGUA ADICIONAL: LINGUÍSTICA E TRADUÇÃO

Lexical Bundles across levels of Proficiency in Portuguese as a Second Language: an examination of bundle function

Pacotes Lexicais e níveis de proficiência em Português como Segunda Língua: uma investigação da função de pacotes lexicais

Paquetes léxicos y niveles de dominio del portugués como segundo idioma: una investigación de la función de los paquetes léxicos

Marine Laísa Matte[1]
orcid.org/0000-0002-4702-6967
marine.matte@ufrgs.br

Larissa Goulart[2]
orcid.org/0000-0002-4190-4589
larissa.goulart@nau.edu

Recebido: 05/6/2020
Aprovado: 07/12/2020
Publicado: 09/02/2021

Abstract: Formulaic sequences are known for being measures of foreign language fluency for learners. Research in language processing suggests that native speakers as well as learners process these sequences as a single word (ELLIS, 1996). Nevertheless, little is known about the use of formulaic sequences in Portuguese and, even fewer studies have examined the use of formulaic sequences in learners of Portuguese. Therefore, in this study, we sought to investigate the textual function of lexical bundles extracted from a corpus of learners of Portuguese as a Second Language (PSL). Lexical bundles are sequences of three or more words that occur with larger than expected frequency in a specific corpus. In this study, we used corpus linguistics tools to extract lexical bundles that occur frequently at two levels of proficiency – beginner and intermediate – in Portuguese. These bundles were, then, classified according to their textual function. Results indicate that beginner level students use more bundles associated with concrete references, while intermediate learners use more bundles associated with textual organization and stance. This study contributes to the description of Portuguese acquisition at these two levels of proficiency. In addition, the results can foster classroom activities with which the PSL teachers introduce new functions of lexical bundles to students. Finally, we hope that this study motivates more research describing the language used at different stages of Portuguese acquisition.

Keywords: Portuguese as a Second Language. Corpus Linguistics. Formulaic Language. Lexical Bundles. Second Language Writing.

Resumo: Sequências formulaicas são conhecidas por serem uma medida de fluência em língua estrangeira. Pesquisas em processamento da linguagem indicam que tanto falantes nativos quanto aprendizes processam essas sequências de palavras como uma única unidade (ELLIS, 1996). No entanto, pouco se sabe sobre o uso de sequências formulaicas em Português e, menos ainda, sobre como aprendizes de Português desenvolvem o uso de sequências formulaicas em diferentes níveis. Portanto, neste estudo, investigamos a função textual de pacotes lexicais extraídos de um corpus de aprendizes de Português como Segunda Língua (PSL). Pacotes lexicais, ou lexical bundles, são sequências de três ou mais palavras que ocorrem mais do que o esperado em um determinado corpus. Para isso, utilizamos ferramentas da linguística de corpus para extrair pacotes lexicais que ocorrem frequentemente em dois níveis de proficiência – iniciante e intermediário – em Português. Esses pacotes foram, então, classificados de acordo com a sua função textual. Os resultados indicam que aprendizes em níveis iniciais utilizam mais pacotes lexicais que se referem a objetos e sujeitos concretos enquanto aprendizes no nível intermediário utilizam mais pacotes lexicais de referência ao texto e posicionamento pessoal. Esse estudo contribui para a descrição da língua utilizada por aprendizes de Português nestes

[1] Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil
[2] Northern Arizona University (NAU), Flagstaff, AZ, USA

dois níveis. Além do mais, os resultados encontrados podem fomentar atividades em sala de aula em que professores de PSL apresentem pacotes lexicais com funções diversas aos alunos. Por fim, esperamos que esse estudo motive mais pesquisas que busquem descrever os diferentes estágios de aquisição de PSL.

Palavras-chave: Português como Segunda Língua. Linguística de Corpus. Linguagem Formulaica. Pacotes Lexicais. Escrita em segunda língua.

Resumen: Se sabe que las cadenas de fórmulas son una medida de la fluidez en un idioma extranjero. Las investigaciones sobre el procesamiento del lenguaje indican que tanto los hablantes nativos como los estudiantes procesan estas cadenas de palabras como una sola unidad (ELLIS, 1996). Sin embargo, se sabe poco sobre el uso de secuencias de fórmulas en portugués y menos aún sobre cómo los estudiantes portugueses desarrollan el uso de secuencias de fórmulas en diferentes niveles. Por lo tanto, en este estudio, investigamos la función textual de paquetes léxicos extraídos de un corpus de estudiantes de portugués como segunda lengua (PSL). Los paquetes léxicos, o paquetes léxicos, son secuencias de tres o más palabras que ocurren más de lo esperado en un corpus dado. Para ello, utilizamos herramientas de lingüística de corpus para extraer paquetes léxicos que ocurren con frecuencia en dos niveles de competencia, principiante e intermedio, en portugués. A continuación, estos paquetes se clasificaron según su función textual. Los resultados indican que los estudiantes en los niveles iniciales usan más paquetes léxicos que se refieren a objetos y temas concretos, mientras que los estudiantes en el nivel intermedio usan más paquetes léxicos que se refieren al texto y al posicionamiento personal. Este estudio contribuye a la descripción del idioma utilizado por los estudiantes de portugués en estos dos niveles. Además, los resultados encontrados pueden incentivar las actividades de aula en las que los profesores de PSL presentan paquetes léxicos con diferentes funciones a los estudiantes. Finalmente, esperamos que este estudio motive más investigaciones que busquen describir las diferentes etapas de la adquisición de PSL.

Palabras clave: portugués como segunda lengua. Lenguaje del cuerpo. Lenguaje de fórmulas. Paquetes léxicos. Escrito en segundo idioma.

Introduction

There has been extensive research on formulaic sequences (WRAY, 2013), especially on how important and difficult they are to learners of any foreign language, regardless of their proficiency level (PAQUOT; GRANGER, 2012). Under the overarching term *formulaic language,* we find several different instances of words sequences, such as, collocations, idioms, lexical phrases, and lexical bundles, the latter being the object of study of the present investigation.

Considering that mastering formulaic sequences - including lexical bundles - is intimately related to language proficiency, it is imperative to understand how language learners use these linguistic features across levels of development. However, we know very little about what linguistic patterns, namely lexical bundles (LBs), learners of Portuguese use since most studies examining the use of LB have described the use of these structures across levels in English as a second language (L2).

Ferreira (2014) has investigated how LBs in Portuguese appear in textbooks, Sardinha, Teixeira and Ferreira (2014) have focused on LBs in different registers, and Goulart (in press) has analyzed their structure. Nevertheless, these studies are scarce, thus, the urgent need for further exploring LBs in Portuguese. Differently from Goulart (in press), who has analyzed the structural patterns across levels of development, this study focuses on the functional patterns of the LBs previously found in that study and relates both the structure and function of these sequences of words. Having said that, it is our hope to contribute to a further understanding of both structure and function of LBs in Portuguese.

This study is divided into five sections, being this introduction the first one, followed by a description of what lexical bundles are and some findings of previous research on the topic. Then, on section three, the corpus is described, as well as the methods. The results accompanied by the discussion are presented in section four, and the fifth and last section is dedicated to the conclusion.

Lexical bundles

Biber et al. (1999) define LBs as a frequent and recurring sequence of words in a given text, and also as building blocks in a discourse. The authors' understanding of LBs have inspired subsequent corpus-driven studies related to several types of multiword sequences. However, not any sequence can be qualified as a LB, as some sequences might account for individual authors' writing style. In order to be considered a LB these sequences have to occur with a representative frequency in a given corpus. Previous studies have used varied frequency thresholds in order to

identify LB sequences. Most studies have varied between occurrences of 20 to 40 times per million words (BIBER; BARBIERI, 2007; CHEN; BAKER, 2010; CORTES, 2004). In addition, researchers tend to adopt a specific range in order to identify LBs. Hyland (2008), for instance, considers LBs appearing in 10% of the texts in a corpus.

Precedent studies have classified LBs based on their structure and function. Structural classification is related to the correlation between bundles in terms of structure. Cortes (2004), for instance, identified two possible structures in history bundles: noun phrases and prepositional phrases. Further structural taxonomies can be found in Cortes (2008) and Pan *et al.* (2016). It is worth pointing out that classifications of bundles, both in terms of structure and function, help us group them and observe patterns that were not possible taking only frequency into consideration.

Previous functional classifications of LBs' functional patterns are found in Biber, Conrad and Cortes (2004) whose categories follow the bundle's discourse function: stance, discourse organizer, referential and conversational bundles. This framework was later adapted by Hyland (2008) in order to fit his corpus of research articles. For the author, the categories are research-oriented, text-oriented, and participant-oriented. Nevertheless, even though Biber, Conrad and Cortes (2004) and Hyland's (2008) frameworks have been thoroughly used to the functional classification of LBs, Biber, Conrad and Cortes (2004) recommend that the functional classification should emerge from the bundles extracted in a specific study.

Most studies adopting a lexical bundle approach have focused on academic discourse, by describing the bundles used in research articles or abstracts in different disciplinary groups. Research on LBs describing language development across learner levels have resulted in conflicting patterns. On one hand, Chen and Baker (2016) found that learners at lower levels of proficiency tend to use more bundles associated with conversation. A similar pattern was found in Staples et al. (2013), for whom lower-level learners use bundles more frequently than their more advanced counterparts, but these bundles are used in the prompts

Few studies have investigated lexical bundles in languages other than English. Tracy-Ventura, Cortes and Biber (2007) compared conversation and academic prose produced by Spanish users. Cortes (2008), still focusing on Spanish, contrasted bundles in History written articles with the writing of English users. Korean and French was investigated by Kim (2009) and Granger (2014) respectively, the former in terms of patterns of bundles in academic prose and conversation, while the later examined French and English stem bundles in two genres: parliamentary debates and newspaper editorials. More recently, Navarro-Gil and Caro (2019) investigated L1 Spanish writing dissertations in L2 English in contrast with published research articles in L1 English in linguistics and medicine. Sardinha, Teixeira and Ferreira (2014) and Ferreira (2014), as already mentioned, are some of the few studies that have analyzed bundles in Portuguese.

Based on this outline of previous research on the use of LBs, this study aims at answering two questions:

> 1) What differences, if any, are there in the types and tokens of lexical bundles in beginner and intermediate levels?
>
> 2) To what extent do the functions of the bundles extracted vary at each level of proficiency?

Method

The Corpus of Written Productions of Portuguese as a Second Language

In order to answer the research questions posed above, the University of Coimbra subcorpora of the Written Productions of Portuguese as a Second Language corpus (PEAPL) was used. In its entirety, this subcorpus contains 624 texts written by 458 international students who were enrolled in the program of Portuguese for foreigners at University of Coimbra (UC). These

students came from 50 different countries and had 39 first languages (see MARTINS et al. 2019 for a comprehensive description of the corpus). These students were enrolled in classes that represented levels of the Common European Framework of Reference for Language (CEFR): beginner (A1), elementary (A2), intermediate (B1), upper-intermediate (B2) and advanced (C1). For the purposes of this study, the two beginner levels (A1 and A2) and the two intermediate levels (B1 and B2) were combined to form a beginner and intermediate corpus. The advanced subcorpus was excluded from the analysis due to its small size. In addition, texts with less than 100 words were excluded from the analysis.

TABLE 1 – PEAPL Subcorpora

Level	N of texts	N of words	min	max	mean length
Beginner	181	35,004	100	475	193.39
Intermediate	336	98,265	104	669	292.45
Total	*517*	*133,269*	*100*	*669*	*257.77*

Source: Elaborated by the authors

As we can see from **Table 1**, the corpus reflects the population of students enrolled in the Portuguese for foreigners' program at UC; thus, it is not balanced by level of proficiency. The texts included in the corpus were a response to nine stimuli presented in Appendix A. These stimuli emerged from three broad topics: the self (i.e. talk about your likes and dislikes), society (i.e. talk about your culture), and the environment (i.e. talk about your neighborhood). Students in both levels responded to the three topics. **Table 2** illustrates how these topics are distributed in the corpus.

TABLE 2 – Written topics

Levels	Self	Society	Environment
Beginner	128	8	45
Intermediate	158	60	118
Total	*286*	*68*	*163*

Source: Elaborated by the authors

Table 2 shows that most of the texts in the corpus were written as a response to texts related to the individual. Nevertheless, environment related topics become more frequent at the intermediate level. In this section, the corpus and subcorpora used for the analysis were described. In the following section, the method for bundle identification and classification will be presented in detail.

Bundle extraction

This study draws on previous findings of a research examining learner language development in lexical bundles (see GOULART, in press). Therefore, bundle size and bundle extraction followed this previous investigation. Three-word bundles were selected as the most appropriate bundle size due to the fact that these are short texts, varying from 100 to 600 words. In addition, upon initial analysis, it was determined that four-word bundles resulted in variable slots at the final bundle position (*eu gosto de **); thus, three-word bundles resulted in the same grammatical and functional information as four-word bundles.

For extraction criteria, the researchers piloted

different solutions, in order to guarantee that these bundles were representative of the two levels being investigated. Tracy-Ventura, Cortes and Biber (2007) highlight that lexical bundles are based on their frequency and dispersion in the corpus. That is, not all sequences of 3 or more words can be called a lexical bundle. Previous studies have adopted frequency thresholds varying from 10 to 40 times per million. Upon initial analysis, it became clear that adopting a high frequency threshold would result in a limited number of bundles (10 or less) for each subcorpora, resulting in a partial and limited analysis. In addition, for the purposes of this study, dispersion was more critical than frequency. When examining the patterns of language development, the researchers wanted to guarantee that the bundles found were representative of that level, rather than on the learner's idiolect. Therefore, bundles had to occur in at least 5% of the texts in each subcorpora in order to be extracted. This guaranteed that the bundles had a frequency of at least 12 occurrences in each subcorpora, without compromising the number of bundles extracted. Bundles were extracted using the n-gram function on Antconc. After bundle extraction, their raw frequency was normalized by a thousand.

Bundle classification

This study seeks to explore specifically how bundle functions vary across two levels of proficiency in Portuguese. Previous studies had already examined structural development but lacked an analysis of functional development along with a correlation between function and form. Even though it is not the focus of this study, bundle structure was classified according to the categories presented in **Table 3.**

TABLE 3 – Structural Classification

Structure	Definition	Examples
VPs	verb-phrases, negator + verb phrases, adverbs followed by verb phrases	gosto muito de, também gosto de
NPs	noun-phrases, adverbs followed by noun phrases	os meus pais, aqui em Coimbra
PPs	prepositional phrases	na universidade de
Clause	clausal coordinators, conjunctions, and subordinators.	e quero o
Pro	pronoun	Eu gosto de

Source: Elaborated by the authors

These categories will not be detailed here as they have already been extensively discussed in Goulart (in press). In brief, these categories classify the structure of the bundles in noun, verb, preposition, pronoun and clausal-based bundles. The criteria for classification are usually the first element of the bundle, with the exception of special cases, such as adverb-based bundles, and clausal bundles.

It is worth noting that, although Hyland's (2008) and Biber, Cortes and Conrad's (2004) categories have been thoroughly used in previous studies, a functional taxonomy should emerge from the bundles found in the corpus, rather than imposed on the data. After an initial survey of the data, the following functional taxonomy was created for the bundles extracted in this corpus.

TABLE 4 – Functional Classification

Function	Definition	Examples
Referential	Bundles that refer to one of the following categories: Time People Place Other	todos os dias a minha família em Coimbra eu os meios de
Stance	Bundles in which the student expresses a position in relation to a topic.	eu acho que muito de viver
Description	Bundles in which the student characterizes herself or an object.	estou a estudar chamo-me xxxx tenho o cabelo
Textual	Bundles used to connect two ideas in the text. In addition, bundles that are markers of the genre being written were included in this category (i.e. *tudo bem contigo*).	por outro lado ao mesmo tempo por isso que

Source: Elaborated by the authors

Table 4 illustrates the functional categories used in this study. Two broad categories were modelled after Biber, Conrad and Cortes (2004): referential and stance. Nevertheless, the subcategories of referential bundles were included a priori in the data analysis. The reason for doing so is that in the initial survey of the bundles it was clear that at both levels, most bundles were referential; thus, including subcategories could give us a more detailed analysis of development. Hence, referential bundles were classified in references to place, people, time and others. Stance bundles refer to a student's expression of opinion towards themselves or others. Bundles of description were added to this taxonomy upon the realization that a considerable number of bundles in both levels focused on describing a person or an object; therefore, this category could provide us with relevant patterns of development. In addition, bundles labeled as textual refer to either structures that are markers of a specific genre (i.e. *tudo bem contigo*), or discourse-oriented devices (i.e. *e por isso*). This category is somewhat similar to discourse organizers, but broader in their criteria of inclusion.

Bundles were classified after carefully reading the concordance lines associated with them, and discussion of disagreements related to bundle classification. Following the classification, the raw frequency of each category was normalized per thousand.

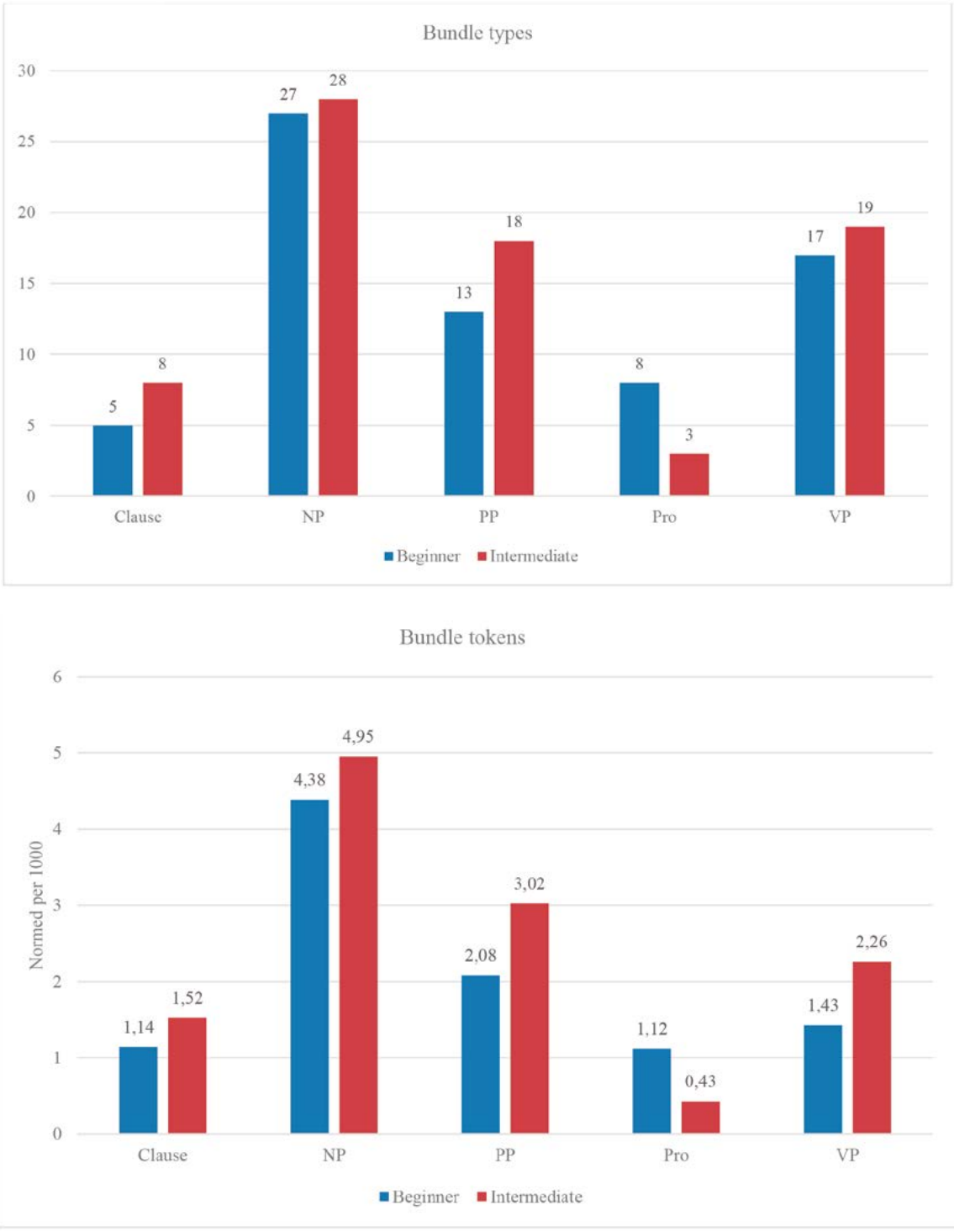
Results and discussion

Using the threshold for dispersion explained in the previous section, 146 bundles were extracted, 70 in the beginner subcorpus and 76 in the intermediate subcorpus. Out of these 146 bundles, 24 occur in both corpora. Appendix B presents a complete list of the bundles identified and the subcorpora that they occur. Overall, the beginner corpus contains 10.15 lexical bundles per thousand words and the intermediate subcorpus contains 12.18 bundles per thousand words. In this section, we will briefly introduce the results of the structural patterns found across levels. Then, the functional patterns for each level will be discussed and compared. Finally, the relationship between functional and structure will be examined.

The structural types of lexical bundles across levels

As explained in the section above, the structural classification used in a previous investigation of the same corpus was adapted to combine the A1 and A2 corpus into our beginner corpus, and the B1 and B2 corpus in our intermediate corpus. **Figure 1** depicts the structural patterns for types and tokens of bundles across the two levels investigated in this paper.

Figure 1 – Structural bundles types and tokens.



Source: Elaborated by the authors

A detailed analysis of the structural patterns of development found across levels can be found in Goulart (in press). Here a summary of the structural patterns is presented in order to inform the later comparison between form and function across learner level. **Figure 1** indicates that clause and verb-phrase bundles are more frequent in texts produced by intermediate learners, while pronoun-based bundles are more frequent in beginner learners. This figure also shows that there are no sizable differences between noun-phrase bundles across levels. Excerpts 1 and 2 illustrate these patterns of structural development across level.

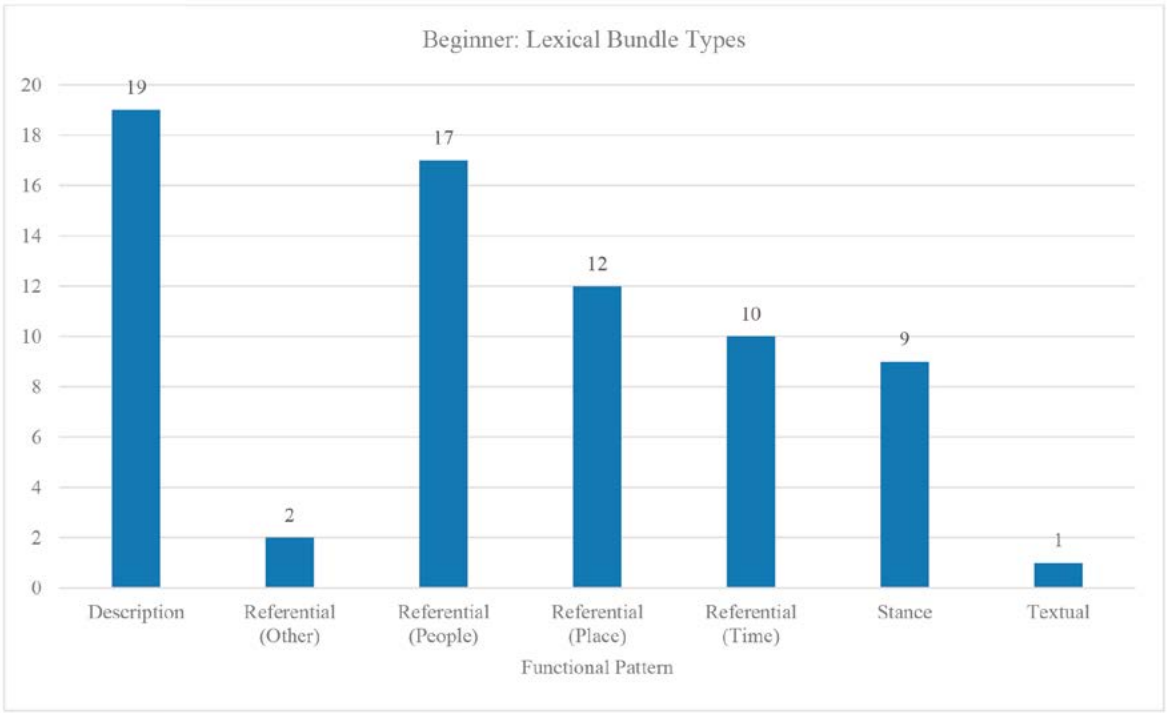
> **Excerpt 1**: Compro sapatos e visito *os meus amigos*. Eu gosto de dançar com *os meus amigos*. (turco.a1.50.33.1j)
>
> **Excerpt 2**: No imaginário colectivo a cidade representa um lugar caótico e *ao mesmo tempo* fascinante, cheio de coisas para ver, para experimentar, para comprar. É o sítio em que cada dia há alguma coisa nova para fazer. (italiano. b2.52.69.3q)

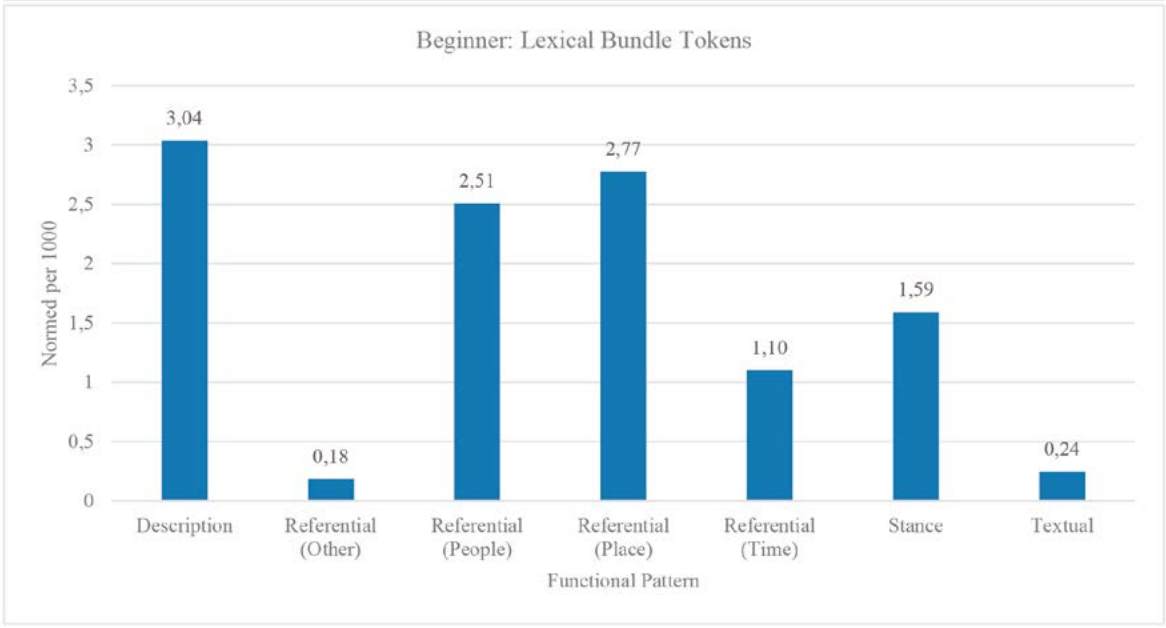
As we can see in **Excerpt 1**, students at lower levels do not use many connectors, therefore, we find few instances of dependent clauses in this subcorpus. We can also see the repetition of the same bundle in which referential devices would be appropriate. It is worth noting that there is no elaboration on this sentence. That is, in **Excerpt 1**, the "sapatos" are not characterized nor the "amigos". In contrast, in **Excerpt 2**, we have a coordinated sentence, indicating that students at this level are familiar with connectors. In addition, we see a list and characterization of several elements (lugar, coisa, sítio). These two excerpts exemplify the patterns found in the quantitative analysis. Lower-level learners do not have the vocabulary or the structural devices to elaborate further in their writings, meanwhile intermediate learners have a larger vocabulary and knowledge of grammatical structures to give their text further details.

The functional types of lexical bundles across levels

Out of the 70 bundles extracted in the beginner level, most of them (N=19) were bundles of description, followed by bundles of reference to people (N=17), place (N=12), time (10), and stance (N=9). Only one bundle was classified as textual and two were classified as reference to other elements. Considering the number of tokens used for each functional pattern, stance and referential bundles are slightly more frequent than suggested by the number of types. This indicates that learners use the same bundle repeatedly to convey this function. **Figure 2** illustrates both the number of types and tokens across functional patterns.

Figure 2 – Beginner level bundles by types and tokens.





Source: Elaborated by the authors

Bundles of description are considerably more frequent in texts written at lower levels. As we can see in **Excerpt 3a**, in which an elementary student responds to the prompt 1.1a (write a text where you introduce yourself...), we have two instances of "Eu sou" and one instance of "Chamo-me". While these are appropriate forms to respond to the prompt, we can see in **Excerpt 3b** how an advanced student responds to the same prompt.

> **Excerpt 3a**: *Eu sou uma* rapariga Erasmus da Roménia. *Chamo-me XXXXX* e estudo direito. *Eu sou uma* rapariga magra, com cabelo preto, comprido e de olhos azuis. (romeno.a2.74.1.1a)
>
> **Excerpt 3b**: Sou uma menina francesa com origens portuguesas. Decidi viajar e viver em Portugal para descobrir as minhas raízes. De facto, meu pai nunca me ensinou a língua portuguesa por causa de um drama que aconteceu com seus pais, meus avós. Ao final decidi conhecer mais sobre essa cultura por meus próprios meios. (frances.c1.28.1.1a).

In **Excerpt 3b**, we see a more detailed presentation, focusing on the student's motivation to learn Portuguese, rather than the student's physical appearance. Another example of descriptions are the bundles associated with "gostar" as we can see in Excerpt 4. In such cases, the student describes the things that she likes and dislikes. Overall, the bundle *gosto muito de* is the most frequent in the corpus.

> **Excerpt 4**: Quando tenho tempo livre, *gosto de fazer* algumas coisas. Eu *gosto de sair* com *o meu namorado* e também com *os meus amigos*. (italiano.a1.58.33.1j)

Excerpt 4 also illustrates the role of referential bundles at beginner levels. Usually, these *gostar* bundles are accompanied by concrete referential bundles of people and places. While in Excerpt 4, we see examples of people, in Excerpt 5, we can see how bundles that refer to places are used by writers in lower levels.

> **Excerpt 5**: Mas também gosto de *ir ao cinema*, ao restaurante e jantar fora com os meus pais ou com os meus amigos. (italiano.a1.55.33.1j)

It is worth noting that referential bundles of concrete subjects (people and places) are more frequent than bundles of abstract matters, such as time. When we look at the number of bundle tokens, we see that time references are not that common at lower levels and most time-reference

bundles are fragments of two constructions *nos fins de semana* and *no meu tempo livre*.

Turning to stance bundles, we have several instances of bundles containing the emphasizer *muito,* which were classified as a show of stance. The other type of stance bundle found in the corpus are *acho*-based bundles. Excerpt 6 illustrates the use of *eu acho* in this subcorpus.

> **Excerpt 6**: Quando vejo a pianista pela primeira vez, *acho que é* muito bonita, elegante e muito charmosa. As mãos dela em cima da máquina, movem-se rapidamente. (chines.a2.09.33.1j)

As we can see, *acho* is used to convey a student's personal opinion without supporting arguments. At this level, there is only one instance of textual bundle (*por isso eu*), which suggests that these are not frequent at this level of writing.

Table 5 presents the top 10 most frequent bundles in the beginner level corpus, along with their functional classification. As we can see, most of the top 10 bundles refer to people or are descriptions of the author and the author's preference.

TABLE 5 – Top 10 most frequent bundles in the beginner corpus

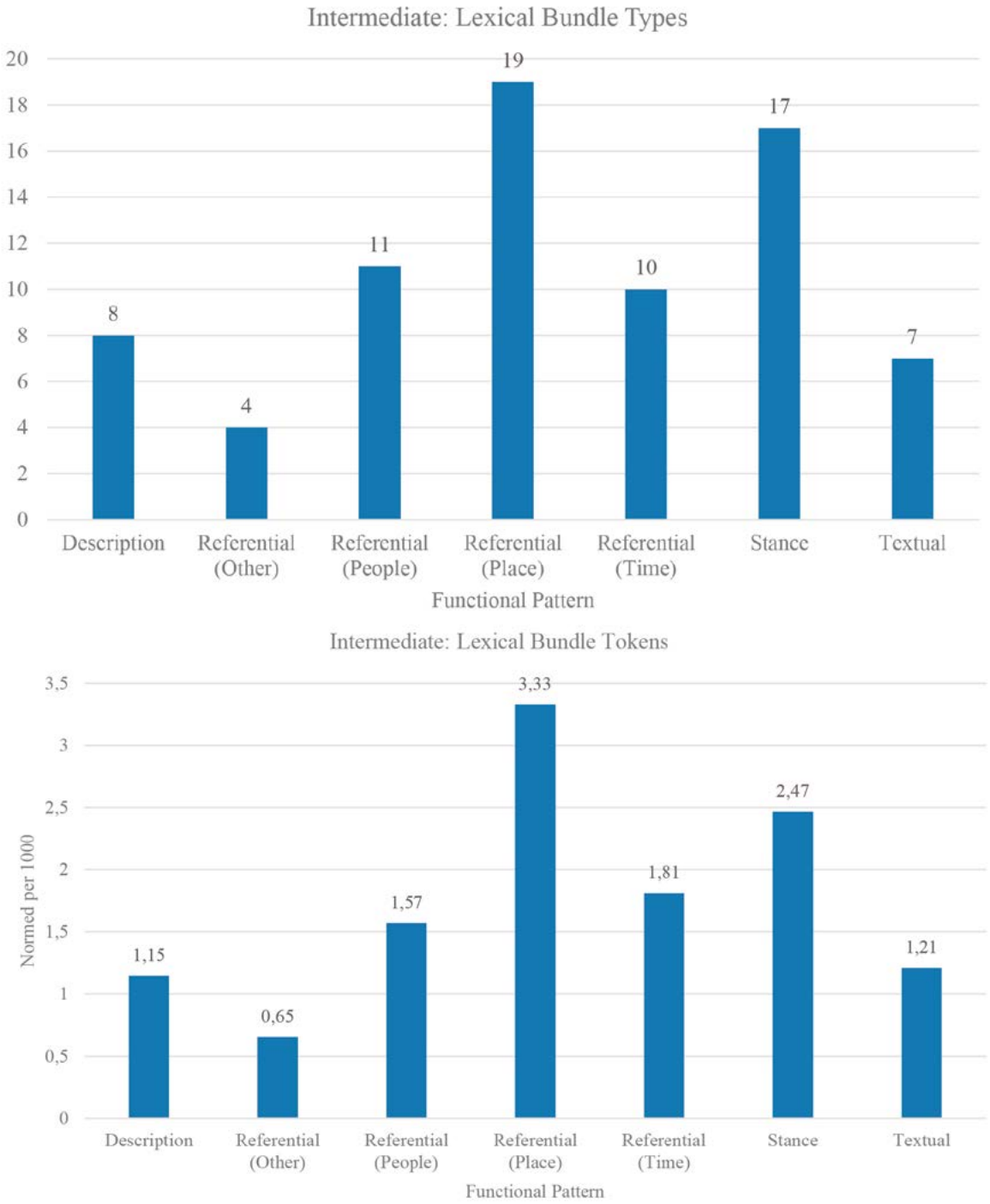
Frequency	Range	Bundle	Function
69	49	gosto muito de	Stance
64	47	eu gosto de	Description
59	47	eu gosto muito	Stance
42	41	chamo-me xxxxx	Description
40	33	a minha família	Referential
28	24	a minha mãe	Referential
27	23	os meus pais	Referential
26	22	com a minha	Referential
23	19	fim de semana	Referential
23	15	gosto de fazer	Description

Source: Elaborated by the authors

At the intermediate level, out of the 76 bundles extracted 19 were bundles referring to a place, followed by 17 stance bundles, 11 referring to people, and 10 referring to time. The number of bundle tokens seems to follow the same pattern as the number of types, suggesting that the amount of bundle types for each bundle token is more balanced at the intermediate level than at the beginner level. This means that while at the beginner level we could see repetitions of the same bundle form skewing the count of bundle tokens, at the intermediate level we do not see this repetition take place.

Figure 3 illustrates the patterns of bundle types and tokens found at the intermediate level. It is clear from this visual representation that there are considerable contrasts between beginner and intermediate learners. First, we can notice a steep decline in the number of description bundles, along with a somewhat noticeable decline in the number of bundles referring to people. In contrast, stance and textual bundles present a considerable increase in both types and tokens used.

Figure 3 – Intermediate level bundles by types and tokens.



Source: Elaborated by the authors

Differently from the beginner corpus, place referential at the intermediate level are not preceded by *gostar*-bundles. They either refer to the place where the student is talking from (see Excerpt 7) or to comparisons between the city and the countryside (see **Excerpt 8**).

Excerpt 7: De qualquer modo, *aqui em Coimbra*, certamente, tenho mais tempo livre do que na Alemanha.

Excerpt 8: Na minha opinião *a vida no campo é* muito mais saudável do que *a vida na cidade*. A natureza, a tranquilidade, o contacto com a terra e com as pessoas é muito mais intenso. (espanholgalego.b2.72.69.3q)

In **Excerpt 8**, we can see that instead of using the verb *gostar* to express preferences, students use *na minha opinião*. That is, we see an increase in the repertoire of devices students use to indicate preferences. We can also notice that the use of place referential bundles might be an outcome of the writing prompt "do you like to live in the city?".

Na minha opinião also represents one of the 17 stance bundle types found in the intermediate corpus. In a similar manner to the beginner level, bundles with *muito* are a large part of the stance bundles identified at this level. Nevertheless, there is an increase in different bundle forms, as Excerpt 9 exemplifies.

> **Excerpt 9**: *Eu acho que é* realmente muito interessante e muito bom que tenha tido *a possibilidade de* conhecer pessoas me (sic) através dos jantares na república. (alemão.b1.121.6.1b)

In Excerpt 9, we see both the use of *a possibilidade de* as well as *eu acho que* combined. This exemplifies how students at intermediate levels are aware of functions and use different bundle forms to express these functions.

Textual bundles at an intermediate level tend to have elements of discourse organizers (i.e. *ao mesmo tempo, por outro lado, e por isso).* Excerpt 10 and 11 illustrates the use of *ao mesmo tempo* and *e por isso* at intermediate levels. We can see that these bundles are used to give an order to the facts being discussed (Excerpt 10) and to describe the relationship between two sentences (Excerpt 11).

> **Excerpt 10**: Adoro as aulas da cultura e literatura brasileira. O meu curso da língua portuguesa é muito divertido e, *ao mesmo tempo*, prático. (polaco.b1.58.6.1b)

> **Excerpt 11**: Durante a sua história, a região foi governada pela Igreja até a unificação da Itália *e por isso* a gente do povo ganhou acentuados sentimentos anti-clericais. (italiano.b1.116.50.2l)

Table 6 presents the 10 most frequent bundles at an intermediate level. We can see that similar to the beginner level most of them are referential bundles. Nevertheless, the forms are strikingly different. In addition, we can see that there is only one bundle of description among the most frequent bundles, which is in contrast to the patterns found at the beginner level.

TABLE 6 – Top 10 most frequent bundles in the intermediate corpus

Frequency	Range	Bundle	function
88	61	gosto muito de	stance
61	38	viver na cidade	referential
57	41	os meus amigos	referential
55	41	gosto de fazer	description
54	38	viver no campo	referential
47	37	com os meus	referential
47	33	meu tempo livre	referential
45	40	eu gosto muito	stance
40	39	há muito tempo	referential
40	36	e por isso	textual
37	33	aqui em Coimbra	referential

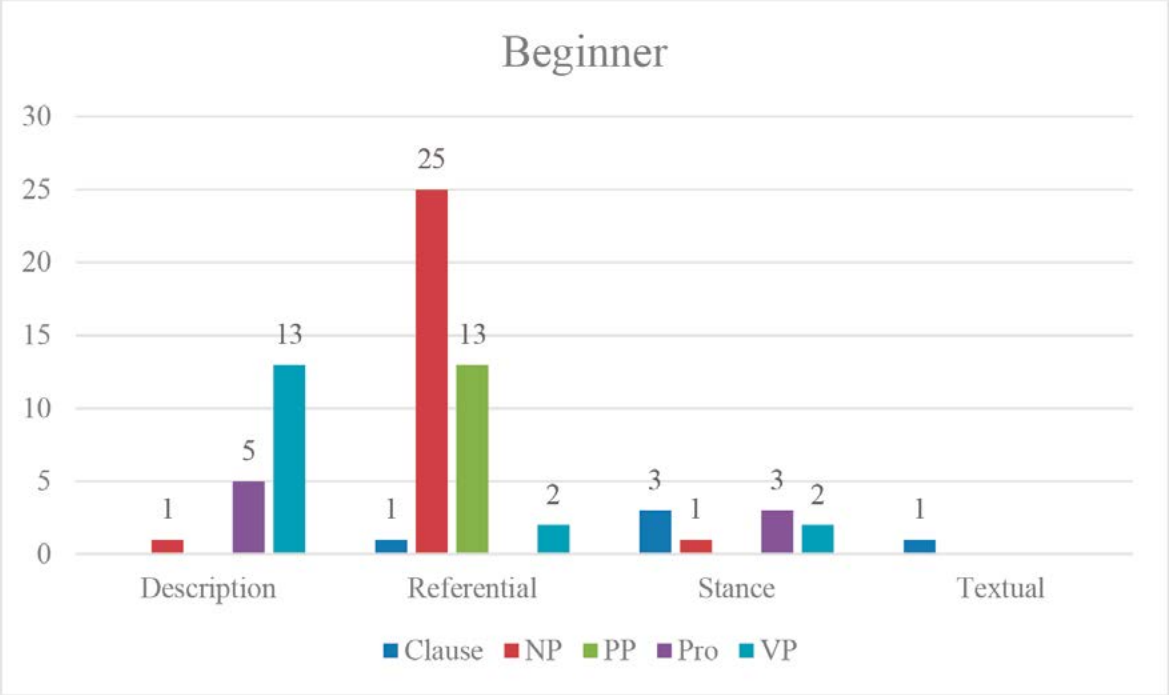
Source: Elaborated by the authors

Contrasting the functional patterns found of both beginner and intermediate levels, we can see that learners of Portuguese at beginner stages tend to use more bundles of self-description (*me chamo x, eu sou uma)* and likes and dislikes (*eu gosto de)*. In addition, beginner writers use more bundles that refer to people and places and these usually occur after *gostar*-bundles (*eu gosto da Universidade de Coimbra,* or *Eu gosto da comida da minha mãe).* We can see that, even though the writing prompts are the same, beginner level learners focus on concrete subjects close to the individual. At the intermediate level we see that learners use more bundles of reference to places (*aqui em Coimbra)* as a response to a specific writing prompt. In addition, we can notice that intermediate learners use more textual bundles (*por isso que, ao mesmo tempo)* and bundles of stance (*eu acho que).* Upon bundle examination, it is possible to notice that these learners use more bundles about abstract ideas and focus on text organization, using bundles that help make clear for the reader the relationship between clauses in a paragraph.

The relationship between forms and function across levels

In this section, we examine the possible relationship between form and function at these two levels of proficiency. For this comparison, we have combined all referential bundles into a single variable. In addition, we only considered bundle type. **Figure 4** illustrates the patterns found for beginner levels.

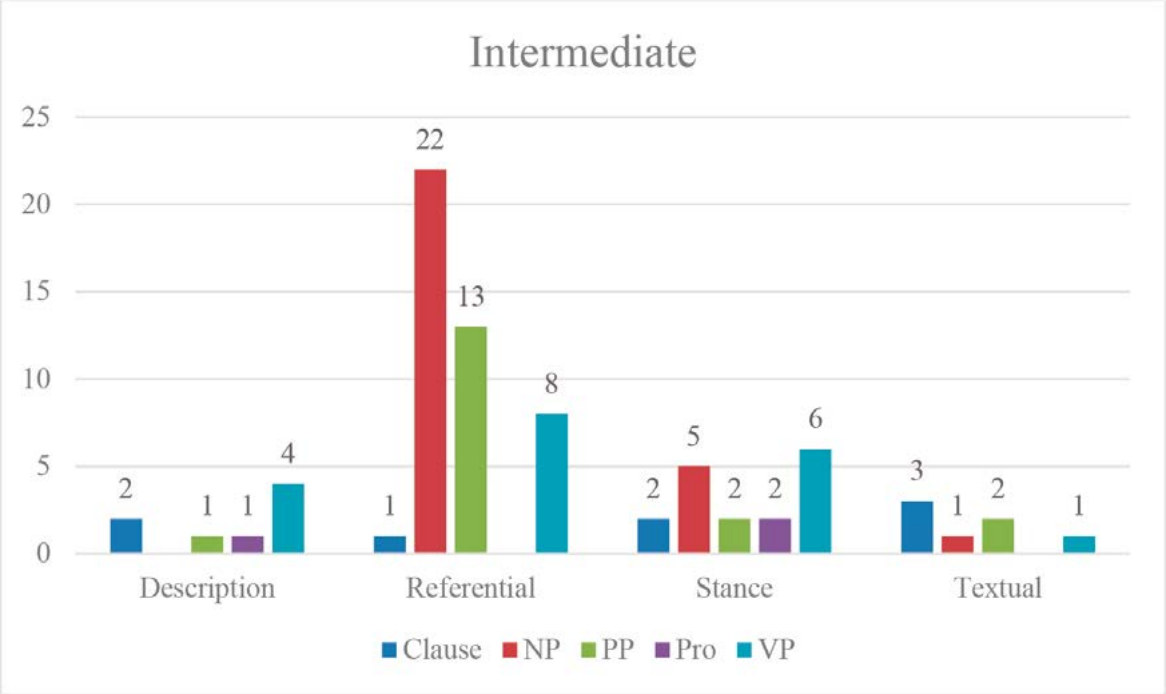
Figure 4 – Form and function relation in the beginner corpus.



Source: Elaborated by the authors

Not surprisingly we can see that most noun-based and preposition-based bundles have referential functions, which can be explained by the high frequency of bundles containing family members (*o meu pai)* or prepositions followed by places (*no campo ou)*. It is interesting to notice that most bundles of description at the beginner level take the form of verb-phrase bundles. This can be explained by the high frequency of bundles starting with *gostar.* Out of the 19 description bundles found at the beginner level, five of them start with *gosto*.

Figure 5 – Form and function relation at intermediate level.



Source: Elaborated by the authors

Turning to the intermediate level, we see the same occurrence of noun and preposition-based bundle being used as referential bundles. It is worth noting, however, the dispersed pattern we see with structure type in stance bundles. Other studies have found stance bundles to be mainly clausal and verb-based, which is not the case here. Of course, this could be explained by the limited number of words in the corpus, but it could also indicate that these structures need to be analyzed in more structural detail to determine whether these bundles contain clausal elements.

The high frequency of verb-based bundles with a referential function is due to infinitive bundles being counted as verb-based bundles. That is, bundles such as *viver no campo* and *viver na cidade* have been added to the verb-phrase count, thus resulting in the patterns we see here.

It is worth pointing out that aside from PPs and NPs as referential bundles, there is not a clear relationship between form and function across these two levels. This is a surprising finding since most studies that have looked at this relationship have found patterns of association. This finding might relate to the fact that these are learners of Portuguese, who still have to increase their repertoire of formulaic sequences in order to show these patterns in their texts.

Conclusion

The goal of this study was to examine the functional patterns of lexical bundles across two levels of Portuguese learners, beginner and intermediate. In order to achieve this goal, we used the PEAPL corpus from the University of Coimbra. Texts written by two proficiency levels were examined. The results indicate that lower-level learners tend to use more bundles of description and reference, while higher level learners tend to use more bundles associated with stance and textual organization.

The results of this study, though limited by the corpus size, could inform teachers of Portuguese as a second language in their material development. For instance, these teachers could give the list of bundles identified in this study and ask their students to highlight these structures in their own texts, or to create exercises in which students have to complete blanks with the correct functional bundle. In addition, we hope that the

results of this brief analysis can motivate teachers to implement small corpus projects in their classrooms, so as to learn about the language patterns of their specific group of students.

Finally, we hope that this study motivates future research describing the language development of Portuguese as a second language. These studies are necessary to inform the practice of teachers and materials developers based on empirical findings, rather than the teacher's intuitions. Future studies could replicate the methodology adopted in this study with a larger corpus, taking in account more levels of proficiency, or looking at the influence of learner's L1. In addition, future research should not be limited to formulaic language. Researchers could examine the use of specific grammatical features across levels of development.

References

BIBER, D.; BARBIERI, F. Lexical bundles in university spoken and written registers. *English for specific purposes,* [*s. l.*], v. 26, n. 3, p. 263-286, 2007. https://doi.org/10.1016/j.esp.2006.08.003.

BIBER, D.; CONRAD, S.; CORTES, V. If you look at... lexical bundles in academic lectures and textbooks. *Applied Linguistics*, [*s. l.*], v. *25*, p. 371-405, 2004. https://doi.org/10.1093/applin/25.3.371.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *Longman grammar of spoken and written English* (Vol. 2). London: Longman, 1999.

CHEN, Y. H.; BAKER, P. Lexical bundles in L1 and L2 academic writing. *Language learning & technology,* [*s. l.*], v. 14, n. 2, p. 30-49, 2010.

CORTES, V. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, [*s. l.*], v. *23*, n. 4, p.397-423, 2004. https://doi.org/10.1016/j.esp.2003.12.001.

CORTES, V. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, [*s. l.*], v. 3, n. 1, p.43-57, 2008. https://doi.org/10.3366/E1749503208000063.

ELLIS, N. C. Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in second language acquisition*, [*s. l.*], v. 18, n. 1, p. 91-126, 1996. https://doi.org/10.1017/S0272263100014698.

FERREIRA, T. D. L. S. B. Idiomaticity in a Coursebook for Teaching Brazilian Portuguese as a Foreign Language. *Working with Portuguese Corpora*, [*s. l.*], p.131-155, 2014.

GOULART, L. Formulaic sequences and writing development in Portuguese as a Second Language. *Spanish and Portuguese Review*, [*s. l.*], v. 6, in press.

GRANGER, S. A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast*, [*s. l.*], v. 14, n. 1, p. 58-72, 2014. https://doi.org/10.1075/lic.14.1.04gra.

HYLAND, K. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes,* [*s. l.*], v. 27, n. 1, p. 4-21, 2008. https://doi.org/10.1016/j.esp.2007.06.001.

KIM, Y. Korean lexical bundles in conversation and academic texts. *Corpora,* [*s. l.*], v. 4, n. 2, p. 135-165, 2009. https://doi.org/10.3366/E1749503209000288

MARTINS, C.; FERREIRA, T.; SITOE, M.; ABRANTES, C.; JANSSEN, M.; FERNANDES, A.; SILVA, A.; LOPES, I.; PEREIRA, I.; SANTOS, J. *Corpus de Produções Escritas de Aprendentes de PL2* (PEAPL2): Subcorpus Português Língua Estrangeira. Coimbra: CELGA-ILTEC, 2019.

NAVARRO GIL, N.; MARTÍNEZ CARO, E. Lexical bundles in learner and expert academic writing. *Bellaterra journal of teaching and learning language and literature*, [*s. l.*], v. 12, n. 1, p. 65-90, 2019. https://doi.org/10.5565/rev/jtl3.794.

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* [*s. l.*], v. 21, p. 60-71, 2016. https://doi.org/10.1016/j.jeap.2015.11.003.

PAQUOT, M.; GRANGER, S. Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, [*s. l.*], v. 32, p. 130-149, 2012. https://doi.org/10.1017/S0267190512000098.

SARDINHA, T. B.; TEIXEIRA, R.; FERREIRA, T. Lexical bundles in Brazilian Portuguese. *Working with Portuguese corpora,* [*s. l.*], p. 33-68, 2014.

STAPLES, S; EGBERT, J.; BIBER, D.; MCCLAIR, A. Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for academic purposes,* [*s. l.*], v. 12, n. 3, p. 214-225, 2013. https://doi.org/10.1016/j.jeap.2013.05.002.

TRACY-VENTURA, N.; CORTES, V.; BIBER, D. (2007) Lexical bundles in Spanish speech and writing. *In:* PARODI, Giovanni (ed.) *Working with Spanish Corpora.* London: Continuum, 2007. p. 217-231.

WRAY, A. Formulaic language. *Language Teaching*, [*s. l.*], v. 46, n. 3, p. 316-334, 2013. https://doi.org/10.1017/S0261444813000013.

Marine Laísa Matte

Aluna de doutorado em Linguística Aplicada (LA) na Universidade Federal do Rio Grande do Sul, instituição em que também realizou o seu mestrado em LA e graduação em Letras – Português/Inglês. Atualmente, é professora do curso de Letras da Universidade do Vale do Taquari.

Larissa Goulart

Aluna de doutorado em Linguística Aplicada na Northern Arizona University. Possui mestrado em Ensino de Língua Inglesa com foco em Inglês para fins Específicos pela Universidade de Warwick e graduação em Letras – Português/Inglês pela Universidade Federal do Rio Grande do Sul.

Mailing address:

Marine Laísa Matte

Universidade Federal do Rio Grande do Sul

Av. Paulo Gama, 110

Farroupilha, 90040-060

Porto Alegre, RS, Brasil

Larissa Goulart

Northern Arizona University

S San Francisco St, 1899, 86011

Flagstaff, AZ, Estados Unidos

Appendices

Appendix A

Estímulo

O indivíduo

Escreva um texto em que se apresente, em que fale das suas características físicas, da sua vida familiar, da sua casa, dos seus gostos e dos seus desejos. Se não quiser falar de si, pode inventar! (1.1A)

Escreva uma carta a um amigo que não vê há muito tempo. Recorde momentos passados em conjunto e fale-lhe da sua vida pessoal e profissional actuais. (6.1B)

Fale daquilo que gosta de fazer nos tempos livres. (33.1J)

A sociedade

Todos os países são diferentes a nível cultural e geográfico. Descreva o seu país, observando as particularidades das suas regiões, os principais monumentos e saliente alguns dos hábitos mais frequentes da sua cultura. (50.2L)

Certamente já teve oportunidade de contactar com pessoas de cultura diferente da sua. Fale de um episódio que lhe recorde esse momento, das dificuldades sentidas, das diferenças e semelhanças encontradas entre as duas culturas e das experiências que partilharam. (52.2L)

Há, certamente, comidas de que gosta muito e há outras que detesta. Fale disto e daquilo que pensam os seus familiares e amigos sobre o assunto. (55.2M)

O meio ambiente

Gosta de viver na cidade? Acha que, se pudesse, gostaria mais de vir no campo? Pense em vantagens e desvantagens de viver na cidade ou no campo. Escreva sobre isso. (69.3Q)

Fale de meios de transporte. Fale daqueles em que já viajou e daqueles em que gostaria de viajar. Se quiser, pode contar uma viagem que tenha feito. (75.3S)

Fale do bairro onde mora. Diga se gosta dele e se acha que há coisas que podiam mudar para que fosse mais agradável lá viver. (77.3T)

Appendix B

Bundle	Beginner	Intermediate
a minha família	*	*
a minha mãe	*	*
aqui em coimbra	*	*
aqui em portugal	*	*
com a minha	*	*
com o meu	*	*
com os amigos	*	*
e a minha	*	*
e o meu	*	*
estou a estudar	*	*
eu acho que	*	*
eu gosto de	*	*
eu gosto muito	*	*
fim de semana	*	*
gosto de fazer	*	*
gosto muito de	*	*
há muito tempo	*	*
ir ao cinema	*	*
mais ou menos	*	*
meu tempo livre	*	*
meus tempos livres	*	*
nos tempos livres	*	*
os meus pais	*	*
todos os dias	*	*
a minha casa	*	
a minha irmã	*	
ao fim de	*	
casa de banho	*	
chama se xxxxx	*	
chamo me xxxxx	*	
com meus amigos	*	
de coimbra e	*	
e acho que	*	
e às vezes	*	
e os meus	*	
em coimbra e	*	
em coimbra eu	*	
em portugal eu	*	
estou em coimbra	*	
eu chamo me	*	

Bundle	Beginner	Intermediate
eu moro em	*	
eu não gosto	*	
eu sou uma	*	
eu tenho uma	*	
gosto de ir	*	
gosto de jogar	*	
gosto de ver	*	
livres eu gosto	*	
meios de transporte	*	
meus amigos e	*	
meus pais e	*	
moro em coimbra	*	
na faculdade de	*	
na universidade de	*	
não é muito	*	
não gosto de	*	
no fim de	*	
o meu irmão	*	
o meu namorado	*	
o meu pai	*	
os meios de	*	
os meus colegas	*	
por isso eu	*	
que é muito	*	
que eu gosto	*	
sou uma pessoa	*	
também gosto de	*	
tempos livres eu	*	
tenho o cabelo	*	
universidade de coimbra	*	
a minha vida		*
a possibilidade de		*
a vida no		*
acho que a		*
acho que é		*
ao mesmo tempo		*
as coisas que		*
as pessoas são		*
centro da cidade		*
com os meus		*
da cidade e		*

Bundle	Beginner	Intermediate
da minha casa		*
da minha vida		*
de viver no		*
e por isso		*
é um país		*
é uma cidade		*
estou em portugal		*
fins de semana		*
gosto muito da		*
há muitas coisas		*
muito de fazer		*
muito de viver		*
muito tempo que		*
na cidade e		*
na cidade é		*
na minha opinião		*
no campo é		*
no meu tempo		*
nos meus tempos		*
o meu país		*
o meu tempo		*
o que é		*
o que eu		*
os meus amigos		*
outra coisa que		*
para mim é		*
por isso que		*
por outro lado		*
posso dizer que		*
que é uma		*
que gosto de		*
também gosto muito		*
tempo que não		*
tudo bem contigo		*
tudo o que		*
uma coisa que		*
vida na cidade		*
vida no campo		*
viver na cidade		*
viver no campo		*
viver numa cidade		*