# Formulaic Sequences and Writing Development in Portuguese as a Second Language

**Larissa Goulart**
*Northern Arizona University*

**Abstract:** Formulaic sequences are an important aspect of writing development. Studies of language processing have shown that multiword units are processed as a single unit by native speakers. Little research, however, has been conducted on formulaic sequences in Portuguese (Sardinha, Teixeira, and Ferreira 2014; Ferreira 2014), and none of those have described the development of the use of formulaic sequences across learners of Portuguese. This study examines the use of lexical bundles, sequences of three or more words that frequently occur in a corpus, in written texts across five proficiency levels (N=572). Three-word bundles were identified based on their dispersion in each subcorpora. The extracted bundles occurred in at least 4% of the texts in the corpus. These bundles were then classified based on their structure in noun, verb, preposition, adverb, clause, and pronoun-based bundles. The results of this study indicate that lower-level learners tend to use fewer bundle types but more bundle tokens. In addition, pronoun-based bundles are more frequent in lower-level texts, while clausal and adverb-based bundles are more frequent in advanced levels. The results can be used to inform the teaching and material development of Portuguese as a second language.

**Keywords:** Formulaic sequences, learner development, lexical bundles, Portuguese as a second language, second language writing

## Introduction

A significant part of linguistic fluency depends on correctly understanding and producing multiword expressions, such as clusters, collocations, and lexical bundles. Research on multiword processing has shown that native speakers of a language process these expressions as a single element. Ellis (1996), for instance, argues that multiword sequences are processed by working memory the same way as single words are processed. Sinclair (1991) also said that formulaic units of language are a "single choice, even though they might appear to be analyzable into segments" (110). In recent years, several researchers have explored how these formulaic sequences develop in learners' language, identifying the patterns of language use associated with each proficiency level (see Goulart 2019 for a review of studies of collocational patterns). This paper addresses the use of a specific type of formulaic sequences, lexical bundles, across different levels of Portuguese learners.

Lexical bundles are sequences of three or more words that occur more frequently than expected by chance in a corpus (Biber, Johansson, Leech, Conrad, and Finegan 1999). Several studies have explored the use of lexical bundles across learners' developmental levels (e.g., Chen and Baker 2016; Staples et al. 2013) or compared learners to native speakers (Ädel and Erman 2012). The results of these studies reveal different patterns of language development. In Chen and Baker (2016), lower-level learners used more bundles associated with spoken discourse, while higher-level learners used more bundles associated with written discourse. In Staples et al. (2013), lower-level learners used bundles more frequently than learners at more advanced levels; nevertheless, most of these bundles appeared in the writing prompt. These studies on formulaic language across levels of development can be used to inform teaching and assessment of learners' language. Most of this research, however, has examined learner development in texts written in English. The study presented here investigates the use of lexical bundles across five levels of development in a language other than English, specifically Portuguese.

Previous research studies investigating lexical bundles in Portuguese have examined the use of lexical bundles in textbooks (Ferreira 2014) and across different registers (Sardinha, Teixeira, and Ferreira 2014). The results of the first study have shown that textbook language is more formulaic than naturally occurring language, and the results of the second study have shown that academic registers have fewer bundles than political speeches and other routinized registers. Even though these studies contribute to our understanding of the language students will encounter when learning and reading in Portuguese, they do not help us understand the language learners produce. Therefore, this research aims to examine the frequency and structure of lexical bundles across proficiency levels.

## Lexical Bundles

Biber et al. (1999: 990) defined lexical bundles as "recurrent expressions that do not have idiomatic meaning or a specific grammatical function." Lexical bundles are identified based on their frequency in a corpus. In other words, these sequences of words are markedly more frequent in a specific corpus. Biber and Conrad (1999: 188) mention that, even though lexical bundles do not have grammatical functions, they have strong structural correlates. This correlation allows researchers to categorize bundles according to structural patterns, such as prepositional, phrasal, and clausal bundles. These classifications can, in turn, inform a systematic comparison between bundles in two or more corpora (Pan, Reppen and Biber 2016).

Biber et al. (1999: 996) classified their bundles based on the first element of each sequence. This initial classification revealed linguistic patterns that characterized the registers investigated. For instance, bundles containing pronouns only occurred in conversations, and bundles starting with noun-phrases were more common in academic discourse. This approach proved to be useful for research comparing groups of text varieties; therefore, several studies have continued to use structural classification as part of their bundle analysis (i.e., Cortes 2008; Sardinha et al. 2014).

To date, most of the research on lexical bundles has focused on English. A few exceptions are Cortes (2008), who compared bundles in English and Spanish writing in history articles; Tracy-Ventura, Cortes, and Biber's study (2007), which contrasted the use of bundles in Spanish conversation and academic prose; Kim's research (2009) on the use of bundles in Korean academic prose and conversation; Granger (2014), who investigated the differences in French and English stem bundles in parliamentary debates and newspaper editorials; and the previously mentioned studies with Portuguese bundles (Sardinha et al. 2014; Ferreira 2014). These studies revealed some challenges for formulaic research in languages other than English, such as the length of bundle size and the limitations of cross-language comparisons. Thus, further studies in Romance languages are needed in order to further our understanding of how these structures work in languages other than English.

This study focuses on examining language variation through lexical bundle types and tokens across proficiency levels in Portuguese. Token counts represent the overall number of lexical bundles found in the corpus, while type counts represent the number of bundles found in the corpus. Therefore, in the sentence, *eu gosto de todos os meus amigos, porque os meus amigos gostam de dançar* (I like my friends because my friends like to dance), there are three bundle tokens and two bundle types (*os meus amigos* and *eu gosto de*). It is hoped that the results of this study will give us an insight into language development. The following research questions guided this study:

1. What differences, if any, are there in the number of types and tokens of lexical bundles across proficiency levels?
2. What differences, if any, are there in the structural types of lexical bundles across proficiency levels?

## Method

### PEAPL Corpus

In order to answer these research questions, the University of Coimbra subcorpus of the Written Productions of Portuguese as a Second Language corpus (PEAPL) was used. This subcorpus contains 624 texts written by 458 international students enrolled in the Portuguese for Foreigners Program at the University of Coimbra. These students came from 50 different countries and had 39 different first languages (see Martins, Ferreira, Sitoe, Abrantes, Janssen, Fernandes, Silva, Lopes, Pereira, and Santos 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). Table 1 presents the number of texts and words in each subcorpora.

*Table 1. PEAPL Subcorpora*

| **Level** | **N of Texts** | **N of Words** | **Min** | **Max** | **Mean Length** |
|---|---|---|---|---|---|
| A1 | 81 | 14,752 | 107 | 324 | 182.12 |
| A2 | 100 | 20,252 | 100 | 475 | 202.52 |
| B1 | 247 | 70,360 | 104 | 669 | 264.80 |
| B2 | 89 | 27,905 | 124 | 519 | 313.50 |
| C1 | 55 | 14,595 | 131 | 456 | 265.30 |
| *Total* | *572* | *147,864* | *100* | *669* | *258.50* |

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. We can also see that the total number of texts in the table does not match the number of texts in the whole corpus. The reason for this is that texts with less than 100 words were excluded from this analysis.

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 all five levels have responded to the three topics. Table 2 illustrates how these topics are distributed in the corpus.

*Table 2. Written Topics*

| **Levels** | **Self (N of texts)** | **Society (N of texts)** | **Environment (N of texts)** |
|---|---|---|---|
| A1 | 67 | 4 | 10 |
| A2 | 61 | 4 | 35 |
| B1 | 125 | 42 | 80 |
| B2 | 33 | 18 | 38 |
| C1 | 20 | 15 | 20 |
| *Total* | *306* | *83* | *183* |

Table 2 shows that the most common topic in the three initial levels was related to the self. More advanced levels write more texts dealing with the topic of the environment. Overall, prompts related to the individual's opinions are the most common across the five different levels. 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 Identification and Classification

In lexical bundle research in Romance languages, different bundle sizes have been explored. Even though these previous studies have discussed the results of extracting 3-, 4-, or 5-word bundles (Cortes, 2008; Granger 2014), none of them have proposed a definite bundle size for Portuguese or any other Romance language. In previous research centered on Portuguese (Ferreira, 2004; Sardinha et al. 2004), both 3- and 4-word bundles have been adopted. Therefore, both bundle sizes were piloted at the initial stages of this research. After this preliminary analysis, the researcher decided that a 3-word bundle would be more informative for this study. First, 3-word bundles were more frequent than 4-word bundles in this corpus. Second, when extracting 4-word bundles, several 3-word bundles that contained relevant grammatical information were excluded from the bundle list. Third, 4-word bundles contained repetitions of the same 3-word bundles but with variables slots (i.e., *eu gosto de* *); hence, analyzing 3-word bundles resulted in obtaining the same grammatical information as 4-word bundles. Finally, the researcher found that analyzing 4-word bundles in short texts as these would limit the number of bundles extracted from each text.

After considering bundle size, frequency and range were examined. Tracy-Ventura et al. (2007: 219) emphasize that lexical bundles are identified empirically based on both their frequency and their dispersion in the corpus. Considering frequency, previous research on lexical bundles has adopted cut-off points varying from 10 to 40 occurrences per million. Nevertheless, in a small corpus such as PEAPL, establishing a high threshold would result in a list of fewer than 10 bundles. This would not provide much information about developmental patterns found in learners of Portuguese. Therefore, instead of setting a frequency threshold, the main criteria for bundle extraction was range, as each bundle had to occur in at least 4% of the texts in the corpus. This guarantees that the bundles extracted are not representative of one author's idiolect. In addition, care was taken to guarantee that bundles did not cross sentence boundaries.

In order to answer research question two, the bundles were also classified structurally. For this classification, the categorization scheme used in previous studies (e.g., Cortes 2008; Pan et al. 2016) was adapted to match learners' language. One of the main modifications made in this scheme is that aside from the first element, the second element is also examined in the case of clausal bundles. Table 3 describes the coding scheme and provides examples from the corpus.

*Table 3. Structural Classification*

| Classification | Definition | Examples |
|---|---|---|
| Pro | Bundle starts with a personal pronoun | eu gosto de |

| | | |
|---|---|---|
| Adv | Bundle starts with an adverb | aqui em Portugal |
| VP | Bundle starts with a verb, or a negator followed by a verb | moro em Coimbra,<br>não gosto de |
| NP | Bundle starts with a noun, article followed by a noun phrase, a coordinator at the phrase level, or infinite verbs | casa da minha,<br>as ruas de,<br>e meus amigos,<br>viver no campo |
| PP | Bundle starts with a preposition | para as minhas |
| Clausal | Bundle contains a subordinator or coordinator at the clausal level in the first or second position | porque gosto de,<br>e quero estar,<br>caminhar porque quero |

In addition to modifying the way bundles were classified, categories were added to this scheme (Adv and Pro). Most studies of lexical bundles have investigated academic language where pronouns were not common, therefore, they did not include this category. The same occurs for adverbs which frequently appear in learners' texts. These categories were included upon an initial piloting of the classification. Finally, bundles were classified upon examination of their concordance lines. Special attention was paid to bundles with "e" as these could be used in phrasal or clausal bundles. After bundle extraction and classification, bundle frequency was normed per 1,000 words. Antconc (Anthony 2019) was used to conduct bundle extraction and classification. Antconc is a freely available software used for corpus analysis. It has many functionalities, including word-lists, concordancers, keywords, and n-grams that can be used for linguistics research, as well as classroom activities.

## Results and Discussion

A total of 356 bundles were extracted from the corpus. The upper-intermediate subcorpus contained the highest number of bundles (N=107), followed by intermediate students (N=74), beginners (N=60), and elementary (N=60). The complete list of bundles is presented in Appendix B. Overlapping bundles, such as "eu gosto de" and "de viver no" were not combined because none of the bundles extracted had a 100% overlap. In other words, "viver no campo" could also be combined with "gosto muito de." Therefore, merging overlapping bundles would exclude important structural information from the analysis. Figure 1 depicts the number of bundles of different bundle forms used in each subcorpora (types) and Figure 2 depicts the overall frequency of bundles in each subcorpora (tokens) normalized per 1,000.

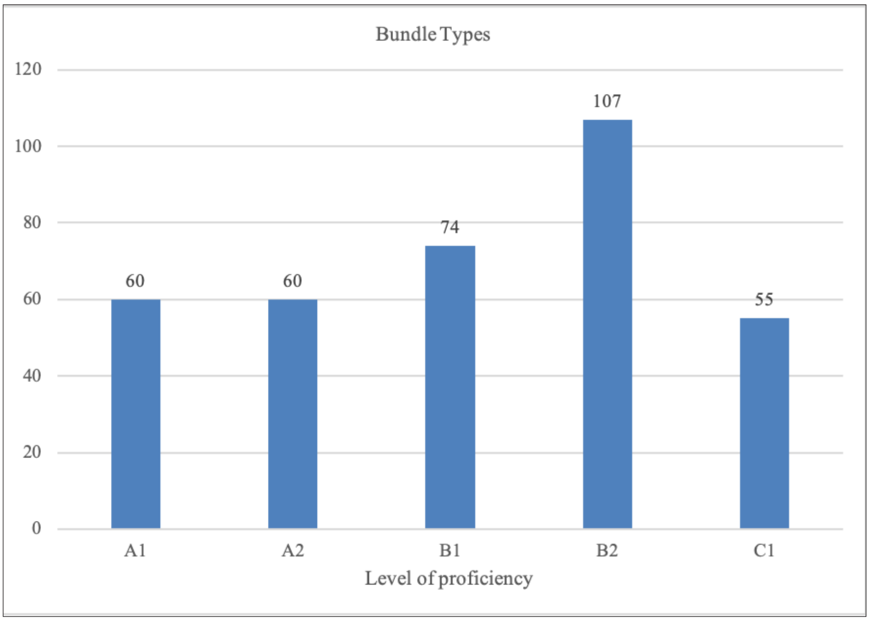


**Figure 1. Lexical Bundle Types**

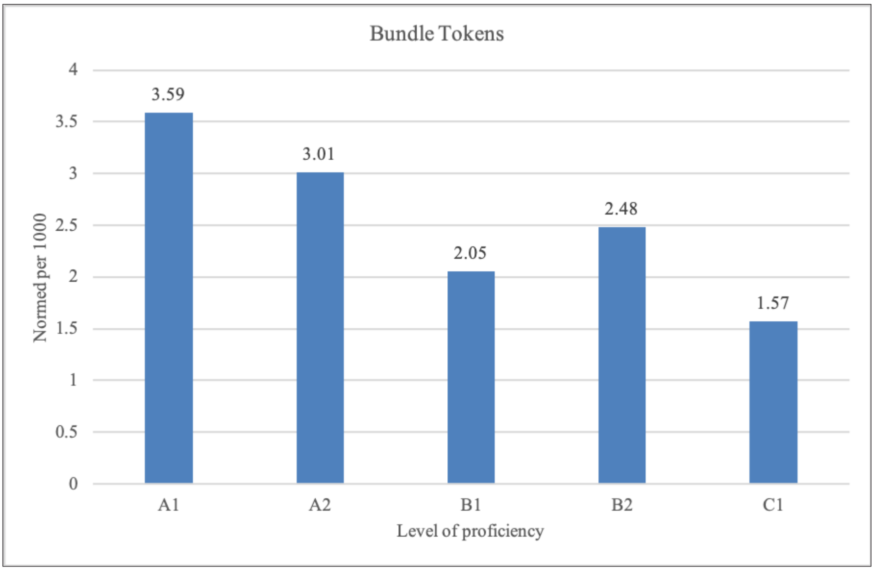


**Figure 2. Lexical Bundle Tokens**

Figures 1 and 2 show that upper-intermediate students use more bundle types, but beginner and elementary students use more bundle tokens. This indicates that beginner and elementary learners tend to repeat the same bundle form several times in their writing, as Excerpt 1 exemplifies. In this short excerpt, we can see that the student uses “os meus amigos” twice, avoiding the use of referential devices, such as personal pronouns.

> Excerpt 1: Compro sapatos e visito *os meus amigos*. Eu gosto de dançar com *os meus amigos*. (turco.a1.50.33.1j)[1]

It is worth noting that we find the opposite pattern in upper and intermediate learners. In these cases, the number of bundle types is high, but the normalized token count is low. This suggests that students at these levels have a greater repertoire of bundles; thus, avoiding unnecessary repetitions of the same bundle. Excerpt 2 exemplifies this pattern. In this excerpt, the student uses two synonyms, *morar* and *viver,* in order to avoid repetition.

> Excerpt 2: Desde 1988, ano que os meus olhos viram a luz por vez primeira, sempre *morei na cidade*...Poderia afirmar que sim, gosto muito de *viver na cidade*. (espanholgalego.b2.72.69.3q)

While this analysis of token/types occurrences across levels already reveals patterns of variation, a detailed analysis of bundle structures at each level will provide a more comprehensive account of learners' development.

### Structural Patterns of Lexical Bundles in Beginner Writing

Beginner learners of Portuguese used 60 different bundle types, and these bundles are mainly noun phrase-based (N=19) and prepositional phrase-based (N=16). These are also the most frequent bundle tokens found in the beginners' corpus. See Figures 3 and 4 for a graphic representation of the structural patterns found in this subcorpus.

Not surprisingly, most of the noun and preposition-based bundles refer to concrete objects, people, and places that are related to the topic of the writing (i.e., *meus amigos*, *minha família*, or *em Portugal*). The following excerpts exemplify these patterns:

> Excerpt 3: mas no último *fim-de-semana* nós fomos para a praia a *Figueira da Foz* (alemao.a1.37.1.1a)

> Excerpt 4: Agora eu estudo em Portugal mas moro na Turquia com *a minha família*. Tenho duas irmãs. A irmã mais velha é casada. (turco.a1.24.1.1a)

[1] Filenames are given for all excerpts. These represent speakers first language, followed by CERF level, ID number, and prompt number.

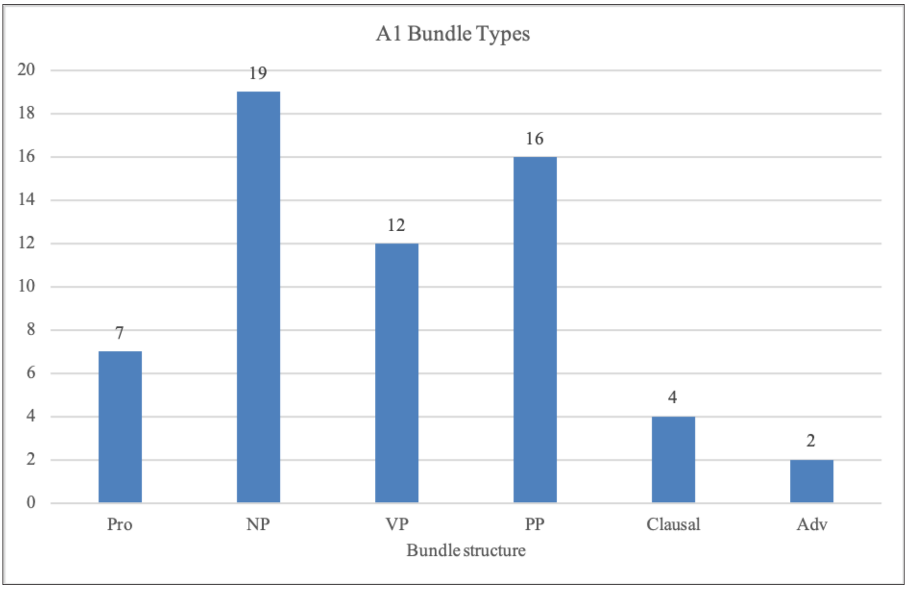


**Figure 3. Lexical Bundle Types for Beginner Levels**

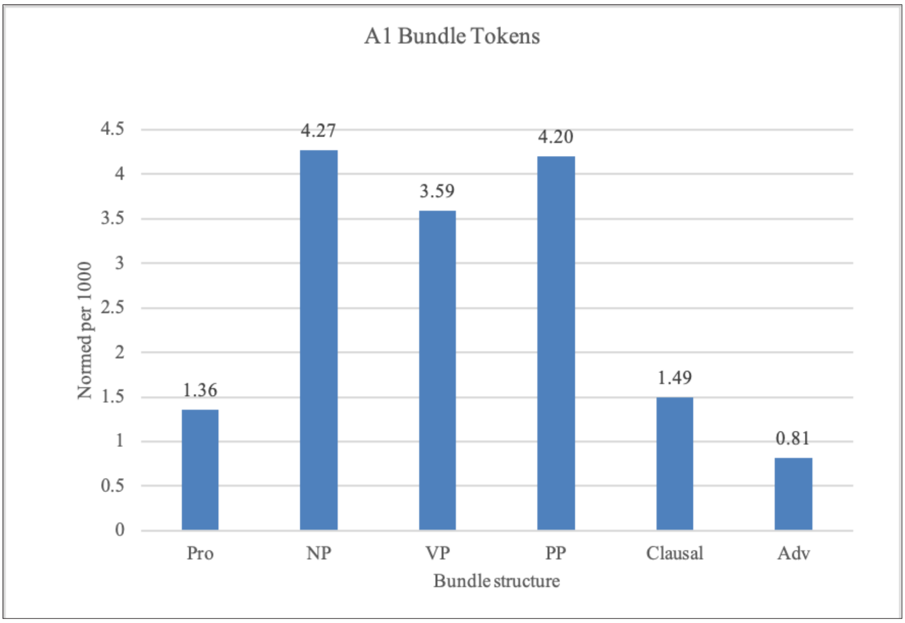


**Figure 4. Lexical Bundle Tokens for Beginner Levels**

It is also worth noting that verb-based bundles are somewhat frequent in beginner writing. As we can see, this type of bundle represents 3 out of the 10 most frequent bundle types in beginner writing. The verb based-bundles in Table 4 also show that *gostar* is a prolific bundle at this beginner stage, as it is part of 5 of the 10 most frequent bundles.

*Table 4. Top 10 Bundles in Beginner Writing*

| Frequency | Range | Bundle | Structure |
|---|---|---|---|
| 31 | 24 | eu gosto muito | Pro |
| 30 | 23 | gosto muito de | VP |
| 28 | 20 | eu gosto de | Pro |
| 19 | 16 | os meus amigos | NP |
| 16 | 14 | a minha família | NP |
| 15 | 11 | com os meus | PP |
| 14 | 12 | a minha mãe | NP |
| 14 | 11 | fim de semana | NP |
| 13 | 13 | moro em Coimbra | VP |
| 13 | 8 | gosto de fazer | VP |

Finally, even though pronoun-based bundles are not the most frequent structure at beginner levels of writing, they are more frequent than in other levels. In this subcorpus, 7 out of the 60 bundles identified contained pronouns, and all of them were first-person pronouns (see excerpts 5 and 6 for examples). This indicates that at beginner levels students are not confident with pronoun omission. The fact that students do not omit the first-person pronoun might also relate to how their first languages set the null subject parameter. Nevertheless, such an investigation is out of the scope of the present paper.

Excerpt 5: No fim-de-semana *eu gosto muito* de apanhar o autocarro para ir nas pequenas cidades do Portugal. (italiano.a1.59.33.j1)

Excerpt 6: *Eu gosto muito* de nadar. (polaco.a1.46.33.1j)

The excerpts presented in this section not only reflect the major language patterns found in this subcorpus, but also a trend in the type of sentences used. We can see from these excerpts that students use simple sentences, without subordinate or coordinate devices, opting for several simple sentences (see Excerpt 4) instead of elaborated sentences.

### Structural Patterns of Lexical Bundles in Elementary Writing

In total, 60 bundles were extracted from the elementary corpus. Most of these bundles were noun-based (N=20), followed by prepositional bundles (N=11). Interestingly, even though the number of prepositional bundle types is almost half the number of noun-based bundles, their token count is almost the same. Figure 5 presents the number of bundle types and Figure 6 presents the number of bundle tokens at this level of development.

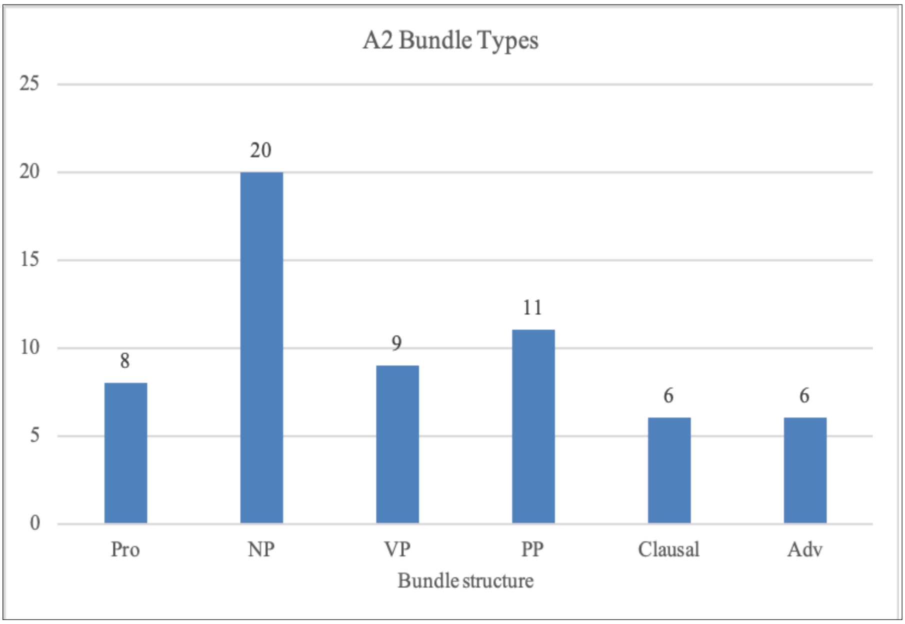


**Figure 5. Lexical Bundle Types for Elementary Levels**

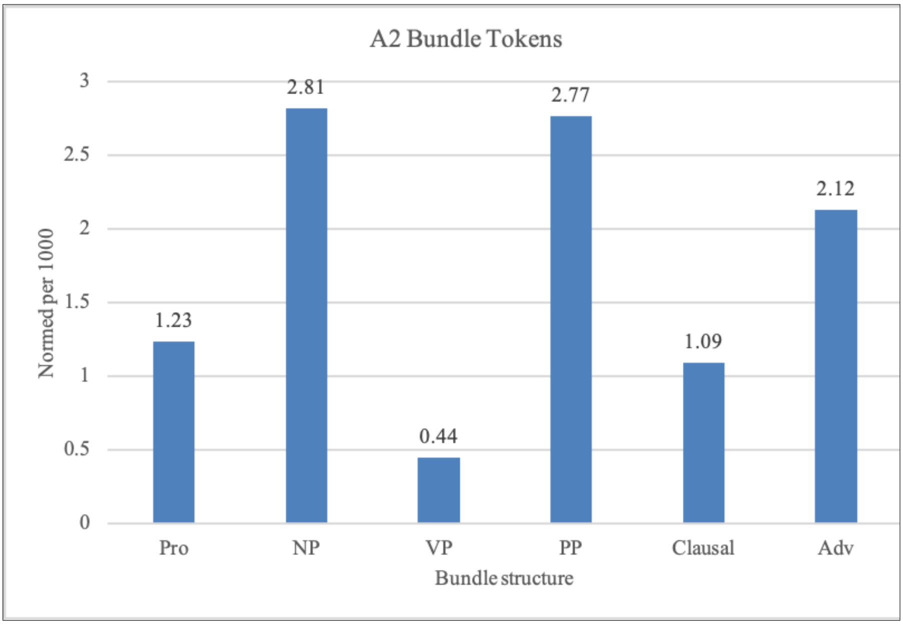


**Figure 6. Lexical Bundle Tokens for Elementary Levels**

Figure 5 shows that both prepositional and verb-based bundles have few types, i.e., forms, that are repeated frequently in these texts. The reason for this is that "gosto muito de" is overwhelmingly more frequent (39 out of 95) than other verb-based bundles, and the same happens with the preposition-based bundle "com os meus" (35 out of 72). Excerpt 7 presents this repetition of "gosto muito de" in the same text, while Excerpt 8 illustrates how these two bundles co-occur in several texts at this level.

> Excerpt 7: Gosto de vestir saias e vestidos pois não *gosto muito de* calças, que acho incómodos. Eu *gosto muito de* desportos, aproveito quase todos os invernos para fazer snowboarding. (alemão.a2.34.1.1a)

> Excerpt 8: *Gosto muito de* passar tempo com *os meus amigos*. (alemao. A2.37.1.1a)

As we can see from Table 5, bundles related to likes and dislikes are still the most frequent ones at an elementary level. Nevertheless, noun-based bundles are the majority of bundle types among the top 10 bundles, indicating a move from verb-based bundles.

*Table 5. Top 10 Bundles in Elementary Writing*

| **Frequency** | **Range** | **Bundle** | **Structure** |
|---|---|---|---|
| 39 | 26 | gosto muito de | VP |
| 36 | 27 | eu gosto de | Pro |
| 35 | 29 | com os meus | PP |
| 28 | 23 | eu gosto muito | Pro |
| 25 | 18 | os meus amigos | NP |
| 24 | 19 | a minha família | NP |
| 23 | 19 | os meus pais | NP |
| 16 | 12 | com a minha | PP |
| 15 | 9 | meios de transporte | NP |
| 14 | 12 | a minha mãe | NP |

### Structural Patterns of Lexical Bundles in Intermediate Writing

There were 74 bundle types in the intermediate corpus: most of them were noun-based bundles (N=21), preposition-based bundles (N=15), and verb-based bundles (N=14). At the intermediate level, these noun and prepositional bundles show more variation in their topic, with more bundles referring to time (i.e., *fim de semana* and *no meu tempo*) in conjunction with the bundles associated with family and likes seen in previous levels. This could be a reflection of learners' increase in vocabulary size at this proficiency level.

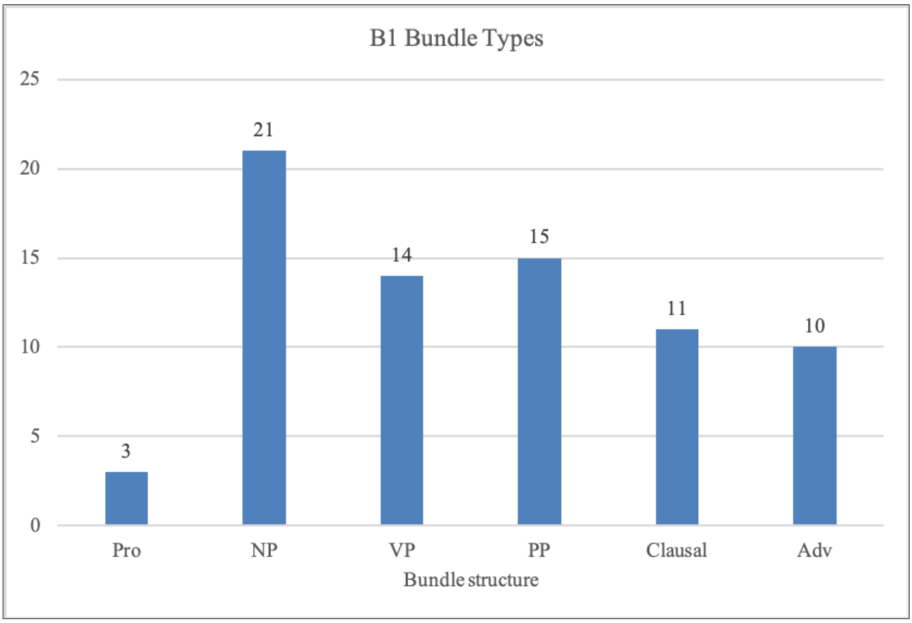


**Figure 7. Lexical Bundle Types for Intermediate Levels**

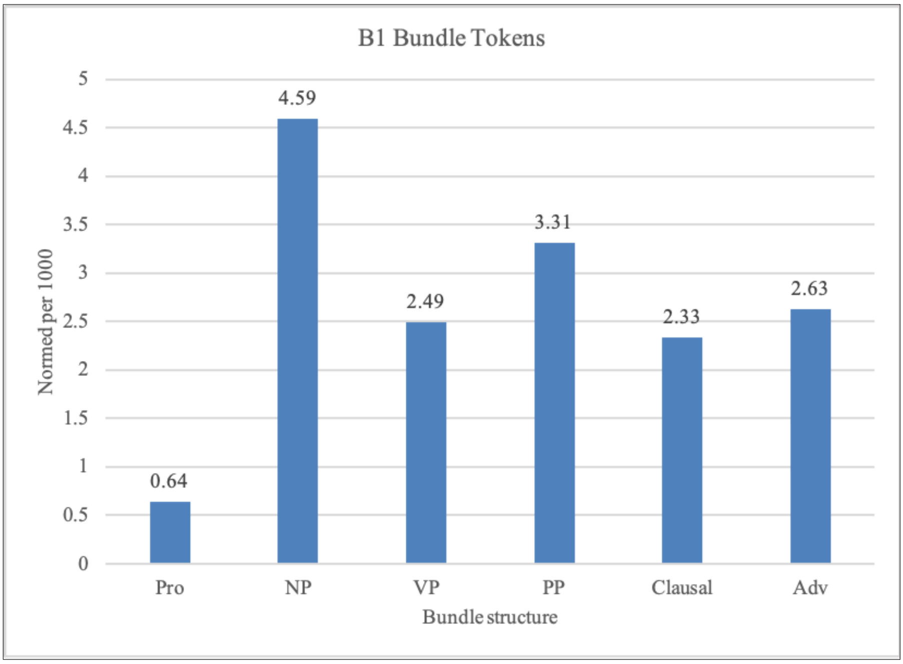


**Figure 8. Lexical Bundle Tokens for Intermediate Levels**

Figures 7 and 8 depict the number of bundle types and tokens, respectively, for each structure in the intermediate corpus. Based on these figures, we can see a decline in the number of pronoun-based bundles and an increase in the number of clausal- and adverbial-based bundles when compared to the beginner and elementary levels.

The decline in the number of pronoun-based bundles and the increase and verb-based bundles can be explained by the omission of first-person pronouns in these texts (Excerpt 9) and by the occurrence of impersonal structures (Excerpt 10). In Excerpt 9, it is evident that this student not only omits the pronoun but also develops the sentence using a *that*-clause, very differently from the excerpts seen at the beginner level.

> Excerpt 9: Todos os países têm suas particularidades. No caso de Chile, *acho que a* sua geografia é muito curiosa. (espanhol.b1.51.50.2l)

> Excerpt 10: Então, uma coisa que se podia fazer para que fosse mais agradável viver lá, era aumentar o número dos policiais. *É verdade que* não é um dos bairros piores de Coimbra... (alemao.b1.16.33.1j)

Among the top 10 most frequent bundles, we see a greater variety of structures, with the first clausal bundle appearing among the most frequent bundles. Bundles with the verb *gostar* are still abundant, yet this is likely a result of the number of texts with topics related to self and preferences.

*Table 6. Top 10 Bundles in Intermediate Writing*

| **Frequency** | **Range** | **Bundle** | **Structure** |
|---|---|---|---|
| 85 | 58 | gosto muito de | VP |
| 54 | 40 | gosto de fazer | VP |
| 51 | 35 | os meus amigos | NP |
| 46 | 32 | meu tempo livre | NP |
| 42 | 37 | eu gosto muito | Pro |
| 41 | 31 | com os meus | PP |
| 32 | 28 | aqui em Coimbra | Adv |
| 31 | 17 | eu gosto de | Pro |
| 30 | 27 | e por isso | Clausal |
| 30 | 21 | viver no campo | NP |

### Structural Patterns of Lexical Bundles in Upper-Intermediate Writing

In the upper-intermediate level, 107 bundles were extracted. This is the highest number of bundles at any level in this corpus. Most of these are prepositions and noun-based bundles. Differently from previous levels, however, prepositional bundles are more frequent in both the number of types and tokens. Figures 9 and 10 also show an extension of the pattern found in intermediate bundles; that is, there is a decline in pronoun bundles and an increase in verb- and adverb-based bundles.

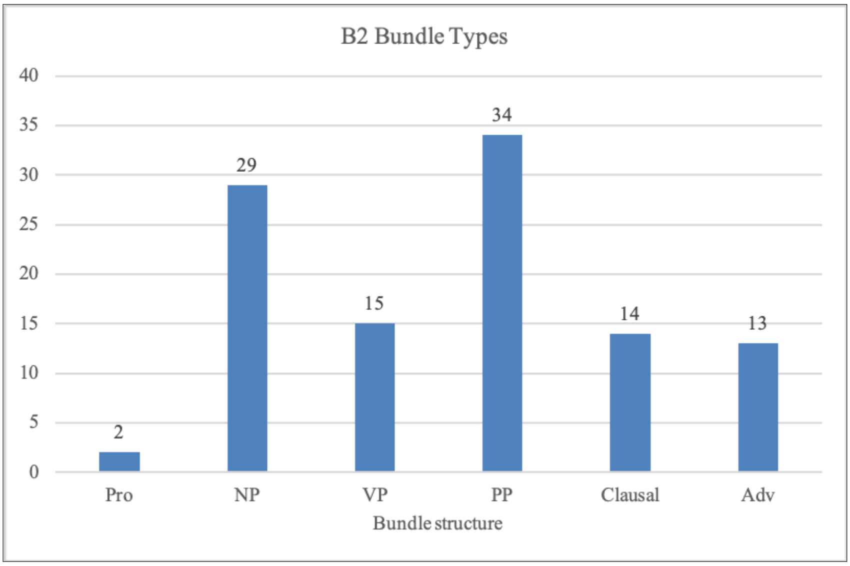


**Figure 9. Lexical Bundle Types for Upper-Intermediate Levels**

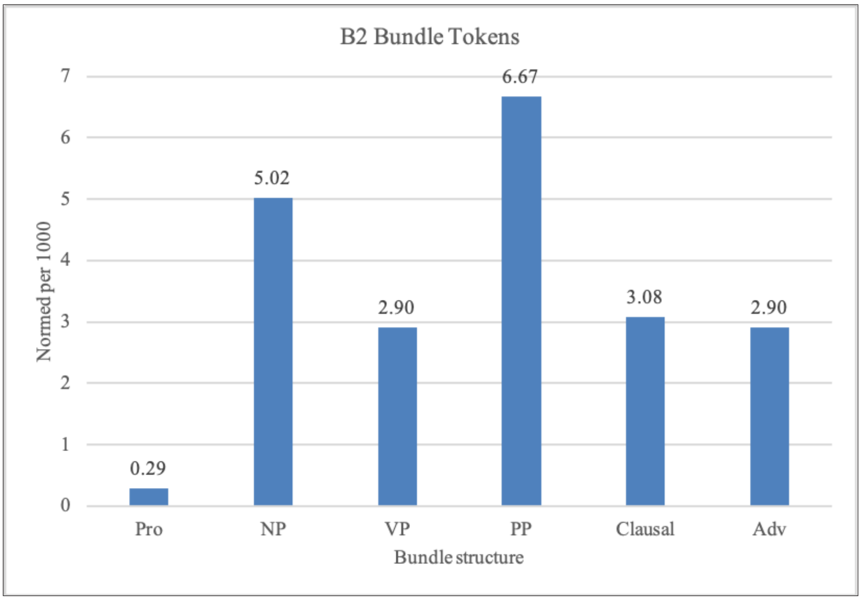


**Figure 10. Lexical Bundle Tokens for Upper-Intermediate Levels**

It is worth highlighting that we see the highest number of types and tokens of clausal bundles at this level. Nevertheless, in at least one instance, this is due to a fixed expression, which can be seen in Excerpt 11 where all bundles with "que esteja" are part of the construction "espero que esteja bem."

> Excerpt 11: *Espero que esteja* tudo bem contigo e que os exames corram bem. (eslovaco.b2.12.6.1b)

At this level, we also see an increase in the number of infinitive verb forms, as Table 7 exemplifies in the first and third row. Excerpt 12 illustrates the use of infinitive construction in these upper-intermediate texts. It is worth contrasting this excerpt to the ones at lower levels. Where we would find sentences composed of a subject, one verb, and a simple complement at the beginner level, here we find dependent clauses and elaborated verb-phrase constructions.

> Excerpt 12: Eu prefiro *viver na cidade* e ir descansar no campo quando tenho férias o no [*sic*] fim-de-semana. Quando tiver mais idade, acho que vou *viver no campo*. (lituano.b2.8.69.3q)

*Table 7. Top 10 Bundles in Upper-Intermediate Writing*

| **Frequency** | **Range** | **Bundle** | **Structure** |
|---|---|---|---|
| 31 | 20 | viver na cidade | NP |
| 24 | 18 | vida no campo | NP |
| 24 | 17 | viver no campo | NP |
| 18 | 17 | a vida no | NP |
| 15 | 11 | é um país | VP |
| 14 | 13 | há muito tempo | VP |
| 13 | 12 | muito tempo que | Adv |
| 11 | 11 | tudo bem contigo | Adv |
| 11 | 10 | tempo que não | Clausal |
| 10 | 9 | e por isso | Clausal |

### Structural Patterns of Lexical Bundles in Advanced Writing

The advanced level in the corpus contained 55 bundles. These bundles were mainly noun and preposition-based bundles. As figures 11 and 12 illustrate, we see a decline in the number of types and tokens of verb-, clausal-, and adverbial-based bundles. Nevertheless, results associated with the advanced level should be taken with caution, since this subcorpus was considerably smaller than the previous ones.

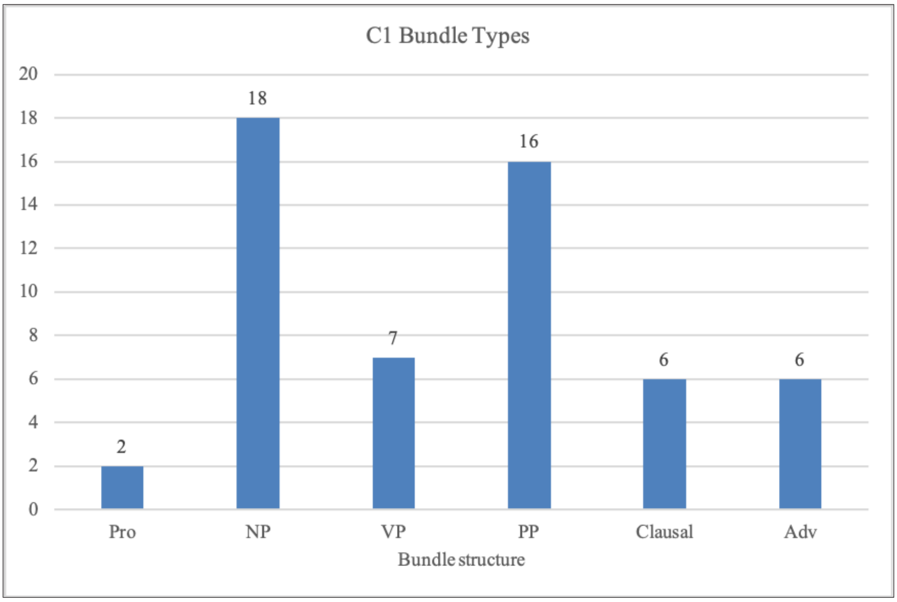


**Figure 11. Lexical Bundle Types for Advanced Levels**

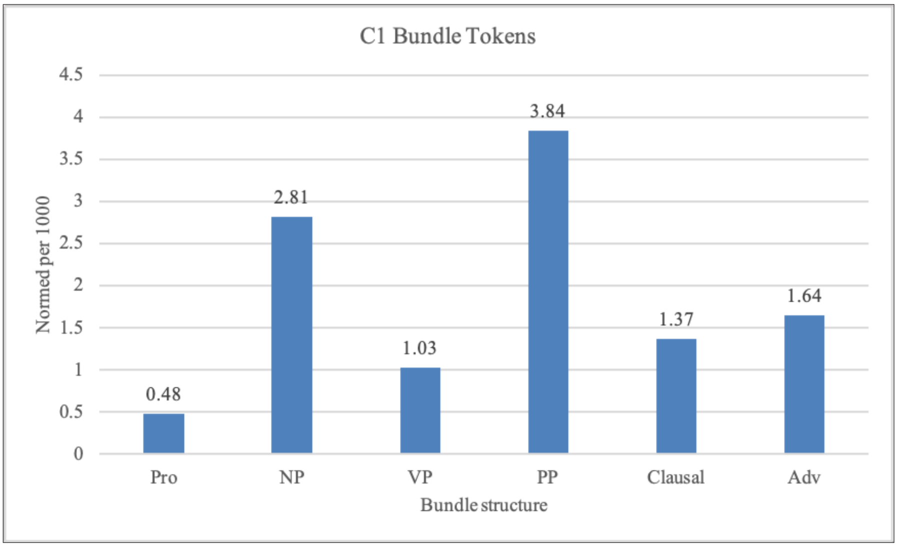


**Figure 12. Lexical Bundle Tokens for Advanced Levels**

Among the top 10 bundles used in advanced writing, we find fewer instances of *gostar*-related bundles and more diverse structures that had already appeared at the intermediate and upper-intermediate levels (*viver no campo*, *viver na cidade*). At this level, we also see learners using references to places (*aqui em Portugal*).

*Table 8. Top 10 Bundles in Advanced Writing*

| Frequency | Range | Bundle | Structure |
|---|---|---|---|
| 12 | 9 | gosto muito de | VP |
| 10 | 6 | viver no campo | NP |
| 7 | 5 | a minha vida | NP |
| 7 | 4 | de viver no | PP |
| 7 | 5 | gosto de viver | VP |
| 6 | 4 | vida no campo | NP |
| 6 | 5 | viver na cidade | NP |
| 6 | 6 | é um país | VP |
| 5 | 4 | a minha casa | NP |
| 5 | 4 | aqui em Portugal | Adv |

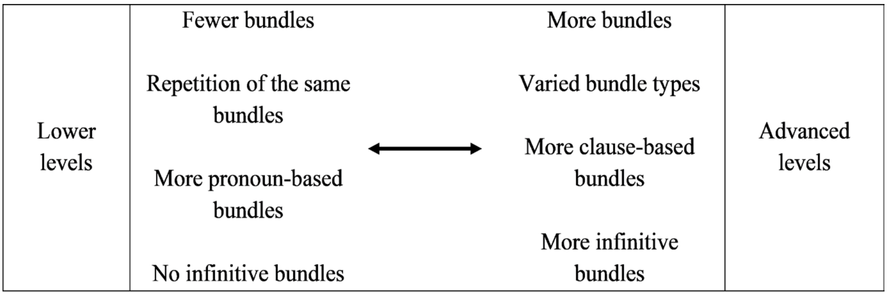
### Structural Patterns of Lexical Bundles Across Levels

This detailed analysis of frequency and structural patterns of lexical bundles across proficiency levels revealed some patterns of language acquisition. First, the number of bundles increases with proficiency levels, with the exception of the advanced level. Nevertheless, the pattern found in the advanced level might simply be a result of a small sample size. Second, patterns of structural development include a decline in the use of pronoun-based bundles in advanced writing and an increase in the use of clausal-, adverb-, and verb-based bundles. This suggests that at lower levels students use fewer instances of dependent clauses and noun-modification resources. Along the same lines, it is worth noting the increase in the use of preposition-based bundles, as these also indicate noun-modification structures.

Finally, by examining the example sentences, we can see that lower-level learners of Portuguese tend to have simpler sentences in beginner levels (A1 and A2) and more elaborate sentences at intermediate levels (B1 and B2). However, further research on this topic is still needed before deducing that these sentences are more complex. Figure 13 summarizes the development patterns found in these lexical bundles.

## Conclusion

This study aimed to examine learner development in Portuguese writing across five levels of proficiency: beginner, elementary, intermediate, upper-intermediate, and advanced. The results of this study have shown that there are substantial differences in the use of lexical bundles across levels. These patterns of development were summarized in Figure 13. The results of this study, though limited by the corpus size, could inform the teaching of Portuguese as a second language. Teachers could use these as a resource to develop materials based on the grammatical patterns found in this study, rather than focusing on the specific lexical choices, since these are closely related to the topic of the writings. As an example, teachers can create activities where the students have to fill in the gap by selecting synonym verbs to avoid the repetition of *gostar*. Advanced students could also examine the texts and identify bundles associated with the genre structure.



| | | | | |
|---|---|---|---|---|
| Lower levels | Fewer bundles<br>Repetition of the same bundles<br>More pronoun-based bundles<br>No infinitive bundles | ⟷ | More bundles<br>Varied bundle types<br>More clause-based bundles<br>More infinitive bundles | Advanced levels |

**Figure 13. Summary of Findings**

This research is not without its limitations—the main one being the effect of the topic in the bundles extracted. Future studies should, if possible, have students write the same prompt across levels, or collect texts on varied topics, so that topic would not play such a large role in the bundles extracted. Future studies could also replicate this study in larger corpora. This is necessary to determine if the patterns found in the PEAPL corpus can be extrapolated for other samples of Portuguese learners. In addition, this paper aimed to take into account language background as a variable, so future studies should control for learner's L1 whenever possible. For instance, the use of personal pronouns might relate to students' first language, and that could be further explored in a study controlling for L1. This study also looked solely into the structural patterns found in lexical bundles. However, future research could examine the functional patterns of learner writing and correlate them to the structural patterns found in this study.

Finally, the researcher hopes that this exploratory study can motivate future endeavors describing the language produced by learners of Portuguese as a second language, as studies such as these are extremely helpful when producing teaching materials and planning Portuguese classes.

## Works Cited

Ädel, Annelie, and Britt Erman. "Recurrent Word Combinations in Academic Writing by Native and Non-Native Speakers of English: A Lexical Bundles Approach." *English for Specific Purposes*, vol. 31, no. 2, 2012, pp. 81-92.

Anthony, Laurence. AntConc (Version 3.5.8) [Computer Software]. Waseda University, 2019, https://www.laurenceanthony.net/software.

Biber, Douglas et al. *Longman Grammar of Spoken and Written English*. Longman, 1999.

Cortes, Viviana. "A Comparative Analysis of Lexical Bundles in Academic History Writing in English and Spanish." *Corpora*, vol. 3, no. 1, 2008, pp. 43-57.

Chen, Yu-Hua, and Paul Baker. "Lexical Bundles in L1 and L2 Academic Writing." *Language Learning & Technology*, vol. 14, no. 2, 2010, pp. 30-49.

Ellis, Nick. "Sequencing in SLA: Phonological Memory, Chunking, and Points of Order." *Studies in Second Language Acquisition,* vol. 18, no. 1, 1996, pp. 91-126.

Ferreira, Telma. "Lexical Bundles in Brazilian Portuguese." *Working with Portuguese Corpora*, 2014, pp. 131-147.

Goulart, Larissa. "The Use of Collocations Across Proficiency Levels: A Literature Review." *BELT-Brazilian English Language Teaching Journal*, vol. 10, no. 2, 2019, pp. 1-15.

Granger, Sylviane. "A Lexical Bundle Approach to Comparing Languages: Stems in English and French." *Languages in Contrast*, vol. 14, no. 1, 2014, pp. 58-72.

Kim, Youjin. "Korean Lexical Bundles in Conversation and Academic Texts." *Corpora*, vol. 4, no. 2, 2009, pp. 135-165.

Martins, Cristina, et al. *Corpus de produções escritas de aprendentes de PL2 (PEAPL2): Subcorpus Português língua estrangeira*. CELGA-ILTEC, 2019.

Pan, Fan et al. "Comparing Patterns of L1 Versus L2 English Academic Professionals: Lexical Bundles in Telecommunications Research Journals." *Journal of English For Academic Purposes*, vol. 21, 2016, pp. 60-71.

Sardinha, Tony Berber, Rosana Teixeira, and Telma Ferreira. "Lexical Bundles in Brazilian Portuguese." *Working with Portuguese Corpora*, 2014, pp. 33-68.

Sinclair, John. *Corpus, Concordance, Collocation*. Oxford UP, 1991.

Staples, Shelley, et al. "Formulaic Sequences and EAP Writing Development: Lexical Bundles in the TOEFL iBT Writing Section." *Journal of English For Academic Purposes*, vol. 12, no. 3, 2013, pp. 214-225.

Tracy-Ventura, Nicole, et al. "Lexical Bundles in Spanish Speech and Writing." *Working with Spanish Corpora*, 2007, pp. 217-231.

## 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)
*Write a text where you introduce yourself, speak about your physical characteristics, your family life, your house, your likes and wishes. If you do not want to talk about yourself, you can make it up!*

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)
*Write a letter to a friend who you have not talked to in a long time. Remind him of past moments that you experienced together and tell him about your personal and professional life right now.*

Fale daquilo que gosta de fazer nos tempos livres. (33.1J)
*Talk about things you like to do in your free time.*

*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)
*Each country has a different culture and geography. Describe your country, noting the peculiarities of each region, the main monuments and highlight some of the most frequent habits of your culture.*

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)
*Certainly, you already had an opportunity to connect with people from cultures different than yours. Talk about an episode that you remember, the difficulties you felt, the differences and similarities between both cultures and the experience you shared.*

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)
*There is, certainly, meals that you like very much and others that you hate. Talk about this and what your family and friends think about this.*

*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)
*Do you like to live in the city? Do you think that, if you could, you would like to live in the countryside? Think about the advantages and disadvantages of living in the city or in the countryside. Write about this.*

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)
*Talk about means of transportation. Talk about the ones you have already used, and the ones you would like to travel on. If you want, you can talk about a trip you have done.*

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)
*Talk about the neighborhood where you live. Mention if you like it and if you think that there are things that could be changed to make it more pleasant to live there.*

**Appendix B**

| **Bundle** | **Level** | | | | |
|---|---|---|---|---|---|
| a minha família | A1 | A2 | B1 | B2 | C1 |
| aqui em portugal | A1 | A2 | B1 | B2 | C1 |
| com a minha | A1 | A2 | B1 | B2 | C1 |
| fim de semana | A1 | A2 | B1 | B2 | C1 |
| a minha mãe | A1 | A2 | B1 | B2 | |
| com os meus | A1 | A2 | B1 | B2 | |
| os meus amigos | A1 | A2 | B1 | B2 | |
| todos os dias | A1 | A2 | B1 | B2 | |
| gosto muito de | A1 | A2 | B1 | | C1 |
| eu gosto de | A1 | A2 | B1 | | C1 |
| nos tempos livres | A1 | A2 | B1 | | C1 |
| com o meu | A1 | A2 | B1 | | |
| eu acho que | A1 | A2 | B1 | | |
| eu gosto muito | A1 | A2 | B1 | | |
| gosto de fazer | A1 | A2 | B1 | | |

| Bundle | Level | | | | |
|---|---|---|---|---|---|
| e a minha | A1 | A2 | | | |
| e o meu | A1 | A2 | | B2 | |
| com meus amigos | A1 | A2 | | | |
| em Coimbra e | A1 | A2 | | | |
| eu chamo me | A1 | A2 | | | |
| figueira da foz | A1 | A2 | | | |
| meus amigos e | A1 | A2 | | | |
| o meu pai | A1 | A2 | | | |
| a minha vida | A1 | | B1 | B2 | C1 |
| é uma cidade | A1 | | B1 | B2 | C1 |
| há muito tempo | A1 | | B1 | B2 | C1 |
| estou a estudar | A1 | | B1 | B2 | |
| com os amigos | A1 | | B1 | | |
| ir ao cinema | A1 | | B1 | | |
| a minha casa | A1 | | | | C1 |
| e gosto muito | A1 | | | | C1 |
| minha casa é | A1 | | | | C1 |
| na semana passada | A1 | | | | C1 |
| a língua portuguesa | A1 | | | | |
| agora estou em | A1 | | | | |
| ao fim de | A1 | | | | |
| beijinhos e até | A1 | | | | |
| casa de banho | A1 | | | | |
| de Coimbra e | A1 | | | | |
| e até breve | A1 | | | | |
| é muito bom | A1 | | | | |
| e os meus | A1 | | | | |
| em casa de | A1 | | | | |
| em Portugal porque | A1 | | | | |
| estou em Coimbra | A1 | | | | |
| eu e o | A1 | | | | |
| eu moro em | A1 | | | | |

| Bundle | Level | | | | |
|---|---|---|---|---|---|
| eu tenho de | A1 | | | | |
| gosto de ir | A1 | | | | |
| gosto de ver | A1 | | | | |
| ir à praia | A1 | | | | |
| moro em Coimbra | A1 | | | | |
| na faculdade de | A1 | | | | |
| na universidade de | A1 | | | | |
| o meu marido | A1 | | | | |
| para a praia | A1 | | | | |
| pessoas da rua | A1 | | | | |
| por aqui estou | A1 | | | | |
| porque é muito | A1 | | | | |
| Universidade de Coimbra | A1 | | | | |
| aqui em Coimbra | | A2 | B1 | B2 | C1 |
| para mim é | | A2 | B1 | B2 | C1 |
| mais ou menos | | A2 | B1 | B2 | |
| os meus pais | | A2 | B1 | B2 | |
| acho que é | | A2 | B1 | | C1 |
| meu tempo livre | | A2 | B1 | | |
| não gosto de | | A2 | B1 | | |
| que eu gosto | | A2 | B1 | | |
| também gosto de | | A2 | B1 | | |
| eu não gosto | | A2 | | | C1 |
| meio de transporte | | A2 | | | C1 |
| a minha irmã | | A2 | | | |
| ano e sou | | A2 | | | |
| da minha família | | A2 | | | |
| e acho que | | A2 | | | |
| é um pouco | | A2 | | | |
| em Portugal eu | | A2 | | | |
| eu sou um | | A2 | | | |
| eu sou uma | | A2 | | | |

| Bundle | Level | | | | |
|---|---|---|---|---|---|
| eu tenho uma | | A2 | | | |
| mas eu não | | A2 | | | |
| meios de transporte | | A2 | | | |
| meus pais e | | A2 | | | |
| não é muito | | A2 | | | |
| no fim de | | A2 | | | |
| o meu curso | | A2 | | | |
| o meu irmão | | A2 | | | |
| os meios de | | A2 | | | |
| os meus colegas | | A2 | | | |
| por isso eu | | A2 | | | |
| quase todos os | | A2 | | | |
| que é muito | | A2 | | | |
| que não é | | A2 | | | |
| sou uma pessoa | | A2 | | | |
| sou uma rapariga | | A2 | | | |
| tenho o cabelo | | A2 | | | |
| um pouco mais | | A2 | | | |
| de viver na | | | B1 | B2 | C1 |
| de viver no | | | B1 | B2 | C1 |
| e por isso | | | B1 | B2 | C1 |
| é um país | | | B1 | B2 | C1 |
| na minha opinião | | | B1 | B2 | C1 |
| o que é | | | B1 | B2 | C1 |
| viver na cidade | | | B1 | B2 | C1 |
| viver no campo | | | B1 | B2 | C1 |
| a possibilidade de | | | B1 | B2 | |
| acho que a | | | B1 | B2 | |
| ao mesmo tempo | | | B1 | B2 | |
| muito tempo que | | | B1 | B2 | |
| no campo é | | | B1 | B2 | |
| tempo que não | | | B1 | B2 | |

| **Bundle** | **Level** | | | | |
|---|---|---|---|---|---|
| da minha vida | | | B1 | | C1 |
| o que eu | | | B1 | | C1 |
| as coisas que | | | B1 | | |
| coisa que eu | | | B1 | | |
| coisas que gosto | | | B1 | | |
| do meu bairro | | | B1 | | |
| é por isso | | | B1 | | |
| é verdade que | | | B1 | | |
| estou a fazer | | | B1 | | |
| fins de semana | | | B1 | | |
| gosto muito da | | | B1 | | |
| há muitas coisas | | | B1 | | |
| meus tempos livres | | | B1 | | |
| muito de fazer | | | B1 | | |
| muito de viver | | | B1 | | |
| no meu tempo | | | B1 | | |
| nos meus tempos | | | B1 | | |
| o bairro onde | | | B1 | | |
| o meu bairro | | | B1 | | |
| o meu tempo | | | B1 | | |
| outra coisa que | | | B1 | | |
| que as pessoas | | | B1 | | |
| que é uma | | | B1 | | |
| que gosto de | | | B1 | | |
| que gosto muito | | | B1 | | |
| que se chama | | | B1 | | |
| também gosto muito | | | B1 | | |
| tempo livre é | | | B1 | | |
| tudo o que | | | B1 | | |
| uma coisa que | | | B1 | | |
| dia a dia | | | | B2 | C1 |
| hoje em dia | | | | B2 | C1 |

| Bundle | Level | | | | |
|---|---|---|---|---|---|
| na cidade é | | | | B2 | C1 |
| na minha vida | | | | B2 | C1 |
| vida no campo | | | | B2 | C1 |
| a cidade e | | | | B2 | |
| a cidade é | | | | B2 | |
| a falta de | | | | B2 | |
| a maior parte | | | | B2 | |
| a última vez | | | | B2 | |
| a vida na | | | | B2 | |
| a vida no | | | | B2 | |
| as pessoas são | | | | B2 | |
| bem contigo e | | | | B2 | |
| centro da cidade | | | | B2 | |
| cidade é muito | | | | B2 | |
| cidade e o | | | | B2 | |
| cidade ou no | | | | B2 | |
| como por exemplo | | | | B2 | |
| da cidade e | | | | B2 | |
| da vida no | | | | B2 | |
| de todos os | | | | B2 | |
| do meu país | | | | B2 | |
| do que no | | | | B2 | |
| e colegas de | | | | B2 | |
| é muito mais | | | | B2 | |
| e o campo | | | | B2 | |
| em Portugal a | | | | B2 | |
| espero que esteja | | | | B2 | |
| está a correr | | | | B2 | |
| esteja tudo bem | | | | B2 | |
| estilo de vida | | | | B2 | |
| estou em Portugal | | | | B2 | |
| eu estou a | | | | B2 | |

| **Bundle** | **Level** | | | | |
|---|---|---|---|---|---|
| eu sei que | | | | B2 | |
| milhões de habitantes | | | | B2 | |
| na cidade e | | | | B2 | |
| na cidade ou | | | | B2 | |
| na república checa | | | | B2 | |
| não nos vemos | | | | B2 | |
| nasci numa cidade | | | | B2 | |
| no campo a | | | | B2 | |
| no campo com | | | | B2 | |
| no campo e | | | | B2 | |
| no campo mas | | | | B2 | |
| no campo também | | | | B2 | |
| no centro da | | | | B2 | |
| no outro lado | | | | B2 | |
| numa cidade e | | | | B2 | |
| o meu país | | | | B2 | |
| os teus pais | | | | B2 | |
| ou no campo | | | | B2 | |
| para a minha | | | | B2 | |
| para o campo | | | | B2 | |
| perto da cidade | | | | B2 | |
| por causa de | | | | B2 | |
| por causa do | | | | B2 | |
| por outro lado | | | | B2 | |
| posso dizer que | | | | B2 | |
| quase todas as | | | | B2 | |
| que a vida | | | | B2 | |
| que é mais | | | | B2 | |
| que esteja tudo | | | | B2 | |
| que não nos | | | | B2 | |
| que no campo | | | | B2 | |
| que para mim | | | | B2 | |

| Bundle | Level | | | | |
|---|---|---|---|---|---|
| queima das fitas | | | | B2 | |
| tenho a certeza | | | | B2 | |
| todo o mundo | | | | B2 | |
| tudo bem contigo | | | | B2 | |
| última vez que | | | | B2 | |
| vale a pena | | | | B2 | |
| vida na cidade | | | | B2 | |
| viver numa cidade | | | | B2 | |
| a maioria dos | | | | | C1 |
| andar a pé | | | | | C1 |
| às vezes é | | | | | C1 |
| bem o que | | | | | C1 |
| da minha casa | | | | | C1 |
| de ti e | | | | | C1 |
| do sul anos | | | | | C1 |
| e a sua | | | | | C1 |
| e gosto de | | | | | C1 |
| equipa de basquetebol | | | | | C1 |
| família e amigos | | | | | C1 |
| gosto de viver | | | | | C1 |
| muito grande e | | | | | C1 |
| norte do país | | | | | C1 |
| os colegas de | | | | | C1 |
| os companheiros de | | | | | C1 |
| ponto de vista | | | | | C1 |
| por causa das | | | | | C1 |
| português na minha | | | | | C1 |
| que na cidade | | | | | C1 |
| todas as noites | | | | | C1 |