JoSS
Journal of Speech Sciences

DOI: 10.20396/joss.v14i00.20586

TEXTUAL PATTERNS AND MACHINE LEARNING CLASSIFICATION IN ACADEMIC WRITING: A LINGUISTIC ANALYSIS OF THESES AND DISSERTATIONS FROM A BRAZILIAN GRADUATE PROGRAM

PAIXÃO, CRYSTTIAN ARANTES[1*]

[2]Professor of the Federal University of Bahia – ORCID: https://orcid.org/0000-0002-3809-4490

Abstract: *This study investigates linguistic patterns in academic texts produced within the Graduate Program in Linguistic Studies (PosLin) at the Federal University of Minas Gerais. A corpus comprising 1,270 documents, 730 master's dissertations and 540 doctoral theses was compiled and analyzed using computational linguistic techniques. Exploratory analyses included the extraction of unigrams, bigrams, trigrams, and the classification of the most frequent tokens into morphological categories (nouns, verbs, adjectives and adverbs). Despite the shared institutional context and research tracks, subtle differences in lexical and structural features were observed between the two academic levels. To evaluate whether these differences could support automated classification, machine learning models were trained on bag-of-words representations of the texts. Gradient Boosting emerged as the most effective algorithm, achieving an AUC of 0.989 with only the 1,000 most frequent tokens, demonstrating that high classification accuracy can be reached without extensive computational overhead. The results show that textual analysis combined with supervised learning can effectively distinguish academic genres within a single graduate program. Furthermore, the approach holds potential for broader applications in genre classification, fake news detection, and discourse analysis. This study also reinforces the importance of continued research in computational linguistics for underrepresented languages such as Brazilian Portuguese, especially in the context of formal and academic writing.*

Keywords: Computational Linguistics; Brazilian Portuguese; Corpus Linguistics.

[*] Corresponding author: crysttian@gmail.com

1 Introduction

Brazil currently has more than 5,000 graduate programs, encompassing approximately 7,000 master and doctoral courses, according to the Coordination for the Improvement of Higher Education Personnel (CAPES in Portuguese) (1). Among the country's leading programs is the Graduate Program in Linguistic Studies (PosLin) at the Faculty of Letters of the Federal University of Minas Gerais. Established in 1973, the program rates 6 by Capes, reflecting its academic excellence in the Brazilian graduate education system. PosLin has played a key role in training professionals who work both in Brazil and abroad. The program is structured around three research tracks: Theoretical and Descriptive Linguistics, Text and Discourse Linguistics, and Applied Linguistics (2).

A textual linguistic analysis enables the systematic investigation of language patterns, aiming to understand how language operates across different levels and contexts. Such analyses encompass phonetics, morphology, syntax, semantics, pragmatics, and textual and discourse analysis.

Collections of documents, particularly those with defined textual structures, are essential for advancing studies that apply computational models for a variety of purposes. When referring to a structured and organized set of texts in Linguistics, it is designated as a *corpus*. If the dataset consists of a collection, it is defined as *corpora*. The texts in these collections may originate from diverse sources, such as audio transcriptions, literary works, journalistic materials, digital data, and others. Soares, Yamashita and Anzanello (3) developed a parallel corpus comprising 240,000 documents to train translation models. Kauffmann (4) applied multidimensional and canonical analyses to the fiction of Machado de Assis to explore the author's textual style. Another study, conducted by Owa (5), employed lexical multidimensional analysis and topic modeling to examine a corpus of academic articles in English and Portuguese. In research on metaphor identification, Nunes (6) used a corpus composed of bilingual journalistic texts in Spanish and Portuguese. Regardless of the dataset´s language, text collections provide critical data for advancing research, whether as a corpus or multiple corpora. Notably, Kutuzov, Kopotev, Sviridenko and Ivanova (7), as well as Yilmaz and Römer (8) analyzed texts in Russian and Ukrainian, and in English.

This study aimed to apply selected methodologies to investigate textual features using a corpus composed of dissertations and theses produced within PosLin. The linguistic analyses focused on determining the frequency of n-grams, as well as identifying the most frequently used morphological classes: verbs, adverbs, nouns, and adjectives, in the scientific documents.

In addition to linguistic analyses, other methodologies may be employed to study textual patterns within a set of texts. Among these, machine learning–based approaches are particularly interesting. Barbosa (9) used narrative and expository texts in Portuguese to evaluate the performance of three machine learning-based classification algorithms. Another methodology employed was a deep learning neural network for text classification, developed by Cândido (10). In this study, different models were compared, and the size of the analyzed textual dataset was assessed. One approach to analyze textual patterns is through topic modeling, as previously noted by Owa (5) and further demonstrated by Li (1), analyzing news group texts and New York Times articles using classification models. Another study that used academic texts to evaluate the application of classification algorithms was conducted by Humaidi, Sutristo and Laksono (12).

Following the examination of the textual characteristics of the PosLin corpus, the second objective was to apply a machine learning methodology using supervised models that were trained and validated based on the corpus. To conclude the study, the final objective focused on investigating the influence of the number of terms used in training and validating the models.

2 Linguistic Analysis

Linguistic analysis comprises methods that allow evaluation of the forms and uses human language, providing a foundation for observing, describing, and interpreting language use (13). Among the branches of linguistic analysis, Computational Linguistics (CL) stands out.

The CL provides the tools that enable machines to perform analyses and recognize textual patterns (14). The training of machines for textual pattern analysis is one of the foundational processes behind the development of Large Language Models (LLMs). LLMs, such as ChatGPT and Gemini, are capable of handling several linguistic features of Brazilian Portuguese, as noted by de Moraes (15). Therefore, although LLMs are becoming increasingly established, continued research in computational linguistics remains essential, like all natural languages for Brazilian Portuguese, due to its unique linguistic features and dynamic nature of a living language (16). This is one of the main reasons why studies involving language must be continuous, since language is subject to ongoing changes and is deeply shaped by regional, social, and cultural factors.

N-grams are one of the methodologies used in CL. The “n” in n-grams refers to the number of terms that compose each sequence. When n equals 1, the result is unigrams, which correspond to the individual terms of a text. When n is set to 2 or 3, the resulting sequences are bigrams and trigrams, consisting of two and three terms, respectively. N-grams are considered one of the earliest statistical methods for extracting information from texts (14). They are used in the initial processing of texts, enabling the structure of subsequent analyses (17, 18). In the case of unigrams, calculating their frequency provides insight into the most common terms in a text, allowing associations with its content. Unigrams are typically visualized as word clouds; however, they are limited in their ability to fully capture the complexity of the textual content they represent. To complement the information provided by unigrams, bigrams and trigrams capture combinations of terms, offering additional contextual insights. These combinations offer greater semantic context, enabling more effective associations with the textual content they stand for. Bigrams occur more frequently tham trigrams and both are less frequent than unigrams.

Morphological classes, commonly referred to as grammatical categories, represent groups of words that contribute to sentence construction. The four important classes are verbs, adverbs, nouns and adjectives. Nouns serve the function of naming objects. In turn, adjectives are responsible for characterizing nouns. On the other hand, verbs indicate the actions, states, or processes associated with textual elements. Finally, adverbs function to modify verbs, adjectives, and other adverbs.

3 Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI). ML is based on computational models and algorithms that enable machines to acquire knowledge from a dataset. ML models are applied across a wide range of purposes, with particular emphasis on natural language processing

(NLP) (19), including tasks such as translation, text generation, and other forms of textual analysis.

Generally, ML models are classified based on how they are trained, a process commonly referred to as learning. During learning, the model is exposed to an input dataset and produces an output. Learning paradigms are typically categorized as supervised, unsupervised, semi-supervised, reinforcement, and transfer learning (20, 21). In supervised learning, the model is trained on a dataset with labeled inputs and corresponding outputs. Unlike supervised learning, unsupervised learning involves training on datasets without labels, allowing the model to identify patterns on its own. In semi-supervised learning, the model is trained on a dataset in which part of the data is labeled and the rest is unlabeled. In reinforcement learning, the model learns through trial and error. Finally, in transfer learning, a pre-trained model is used as the basis for further fine-tuning.

Another common way to categorize models is based on the kind of task they perform, either classification or regression. Regression models are applied to datasets for estimating continuous values. Unlike regression models, classification models aim to categorize the dataset by following or discovering underlying patterns (22, 23).

In the case of labeled datasets, as may occur in a corpus, supervised classification models are particularly prominent, as demonstrated by Hsu (24). Among the prominent models are Gradient Boosting, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Neural Networks (20, 21, 22, 23). Gradient Boosting combines multiple decision trees to minimize prediction errors. K-Nearest Neighbors classifies data based on the categories of neighboring elements, using a metric distance calculated from the input features. Support Vector Machine employs a hyperplane-based approach to classify data. Logistic Regression is a statistical model that estimates the probability of a given instance belonging to a particular class. Finally, Neural Networks process data using a mathematical architecture that simulates the functioning of biological neurons. These models appear in various studies involving corpus analysis. Hsu (24) reported the application of supervised models for text classification. Britto (25) compared supervised learning techniques in the analysis of textual genres. Almatarneh and Gamallo (26), as well as Allam et al. (27), used supervised models to classify opinions and news articles.

4 Material and Methods

The corpus analyzed in this study was constructed from documents, theses and dissertations, produced as part of the research conducted within PosLin. The PosLin website provides access links to the thesis and dissertation files (28). To perform the analysis, a script was developed in the R programming language (29) to collect the files[2]. The collection was carried out on October 21, 2024, based on 1,276 links; however, after processing, 1,270 files were obtained, comprising 540 theses and 730 dissertations. Six files, consisting only of the initial pages, were removed. Following the file collection, the documents were converted from PDF format to plain text. During the conversion process, tables and charts were transformed into text, while figures were discarded. The plain text files ranged in size from 0.54 to 33.74 MiB, with an average size of 3.07 MiB. Theses ranged from 0.97 to 33.74 MiB, with a mean size of 4.10 MiB,

[2] The scripts are made available at the following link: https://github.com/crysttian/paperjspeech

while dissertations ranged from 0.54 to 9.12 MiB, with an average size of 2.30 MiB. Although the number of theses was lower than that of dissertations, their average size was greater; therefore, all files were included to compose the corpus for analysis.

For initial evaluation of the texts, word clouds (unigrams), bigrams, and trigrams were generated from the processed content. After processing, the text was submitted to a parser for morphological classification. Due to the complexity of the texts, particularly related to encoding, and to simplify processing, the 5,000 most frequent tokens were selected and submitted to the parser for classification. Following this classification, the ten most frequent verbs, nouns, adjectives, and adverbs were identified across all files, and then separately for theses and dissertations.

After characterizing the texts, machine learning models (Gradient Boosting, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Neural Network) were employed to classify the texts in the corpus, using the 50, 100, 300, 500, 1,000, 1,500, 2,000, 5,000, 10,000, 20,000, and 30,000 most frequent tokens. Each term was assigned a weight based on its frequency using a bag-of-words[3] approach (30). To this end, the texts were categorized as dissertations or theses. The dataset was split into 90% for training and the remaining 10% for model validation. To estimate the adjustment coefficients, ten repetitions were performed, with each iteration involving resampling to create new training and testing datasets. As model performance metrics, the area under the ROC curve (AUC), classification accuracy, F1 score, precision, and recall were considered (21). After selecting the best-performing model, a confusion matrix was used to analyze the resulting classification. The software used included Orange Data Mining (version 3.38) (31), R (version 4.3) (29), and RStudio (version 2024.12.1) (32), along with the packages tm (33), slam (34), wordcloud (35), RColorBrewer (36), readr (37), and quanteda (38), as well as the palavras parser (39), licensed to the LEEL[4]. Processing was carried out on a computer equipped with a 9th-generation Intel Core i5 processor (4 cores, 8 threads) and 16 GiB of RAM.

5 Results and Discussion

One of the simplest analyses that can be performed is the calculation of word frequency within the corpus. One way to present this count is through a word cloud, in which the size of each term corresponds to its frequency of occurrence. Three word clouds were generated: one for the entire corpus, including both theses and dissertations, and two separate clouds, one for theses and one for dissertations.

Figure 1 presents the word cloud for the corpus comprising both theses and dissertations. The six most frequent terms are “ser”, “sobre”, “forma”, “língua”, “relação”, and “texto”, followed by others with lower frequencies. In Figure 2a, the word cloud for the theses is shown, with the six most prominent terms being “ser”, “sobre”, “língua”, “forma”, “texto” and “análise.” For the dissertations, the results are shown in Figure 2b. The six most frequent terms are “ser”, “sobre”, “língua”, “forma”, “texto” and “análise.” It is worth noting that the word clouds share common terms, though the frequency with which these words occur varies between the texts. While word clouds indicate term occurrences, they do not allow for conclusive analysis. Terms

[3] It is a textual representation model in which a document is described as a set of words and their respective frequencies of occurrence in the text, disregarding word order and syntactic structure.

[4] Laboratory of Empirical and Experimental Studies of Language (LEEL - Laboratório de Estudos Empíricos e Experimentais da Linguagem) of Federal University of Minas Gerais.

such as “análise”, “relação”, “texto” and “língua” are nouns that reflect the subject of research conducted at PosLin, but their presence alone is not sufficient for deeper interpretation.

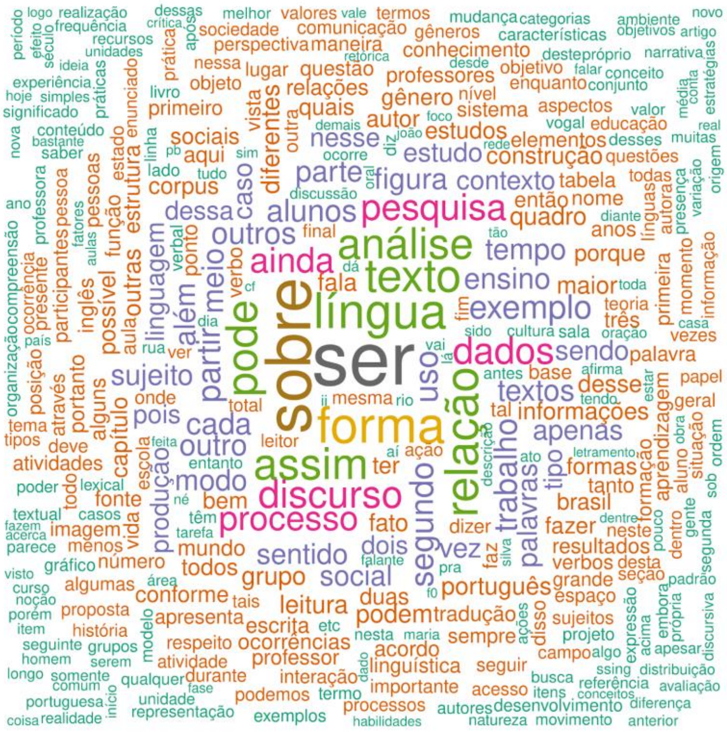
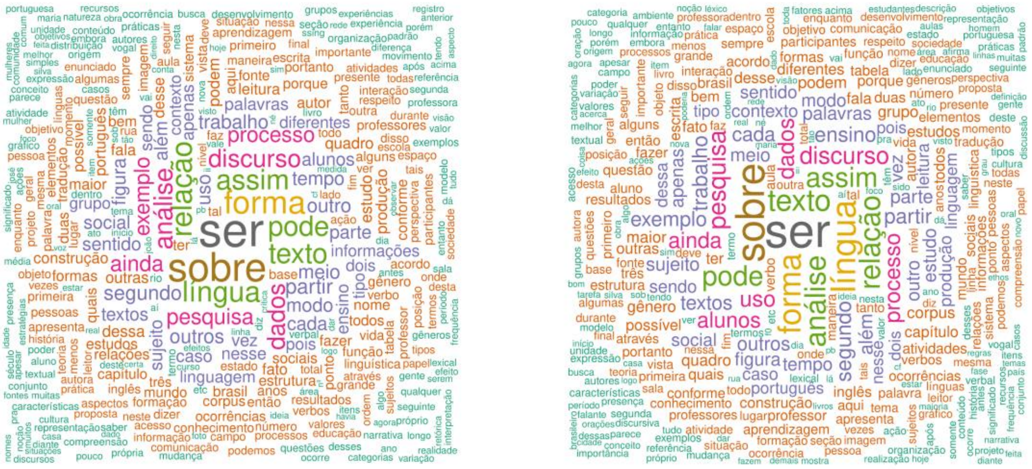


Figure 1: Word cloud of master and doctoral texts.



a) Theses

b) Dissertations

Figure 2: Word cloud for theses and dissertations.

N-grams composed of two (bigrams) or three (trigrams) terms offer an alternative approach for exploring a corpus. To this end, the texts were processed to generate bigrams and trigrams for the entire corpus (dissertations and theses), as well as separately for dissertations and theses. To generate the bigrams and trigrams, a list of the 1,000 most frequent sequences was created, from which the top twenty were selected and analyzed in relation to PosLin's thematic focus areas.

In Figure 3, the bigrams and trigrams for the entire corpus are shown. Several bigrams are directly related to PosLin's research tracks, such as “língua portuguesa”, “língua inglesa”, “itens lexicais” and “estrutura morfológica”, among others. Bigrams already allow a deeper understanding and can be linked to PosLin’s thematic areas. The trigrams also reveal terms related to PosLin’s research subject, with notable examples including “informações etimológicas distribuição”, “registro dicionários bluetau” and “banco dados projetos”. Some trigrams also include terms such as “cemitério” and “sepultamento”, which are associated with specific studies conducted within the program. Other trigrams are related to dictionaries used in the studies. Although most references were removed from the files, some citations remained, notably those mentioning Italian and Portuguese dictionaries. In addition to dictionaries, two authors, Kress and van Leeuwen known for their work in multimodality were identified, as well as references to the Atlas Toponímico do Estado de Minas Gerais (ATEMIG).

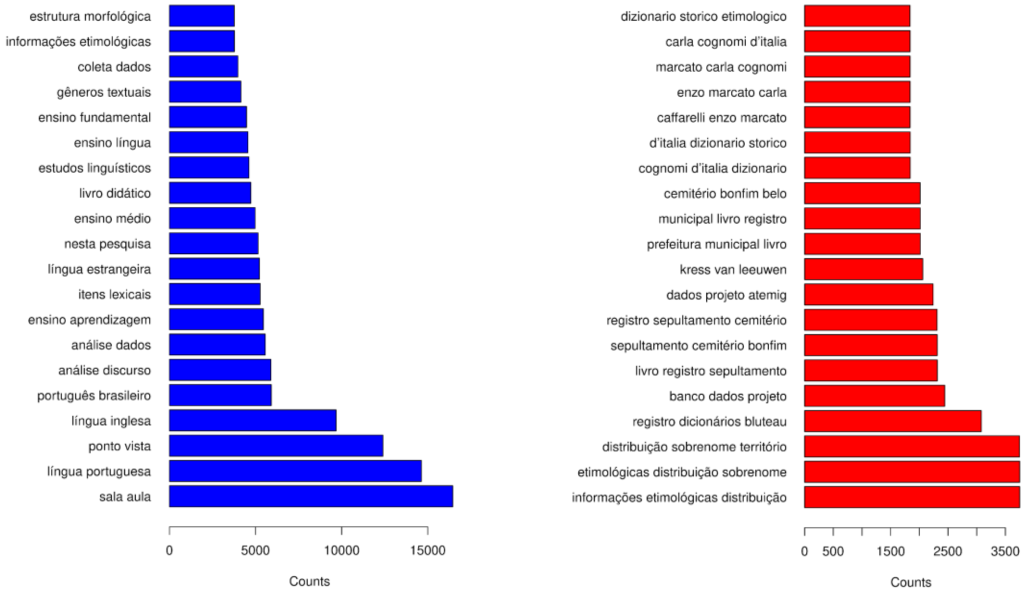


Figure 3: Bigrams and trigrams were extracted from theses and dissertations from PosLin.

In Figure 4, the occurrence of bigrams and trigrams in the theses is shown. As observed in the bigrams from the complete corpus, several terms related to the program’s subject stand out once again, such as “língua portuguesa”, “língua inglesa”, “estrutura morfológica” and “informações etimológicas”, among others. The trigrams are nearly identical to those in the complete corpus, differing only in the order of the terms presented. Figure 5 shows the bigrams and trigrams for the dissertations. The bigrams include terms that differ from those in the theses, with notable examples such as “ensino aprendizagem”, “ensino fundamental” and “livro(s) didático(s)”, among others. The trigrams do not reveal particularly significant terms but include several elements that commonly appear in the texts, notably references to the dictionaries used. As previously mentioned, regular expression rules were applied to remove common structural elements from the documents, such as front matter prior to the table of contents and bibliographic references. However, the results indicate that in some cases these elements were not completely removed, or they may have been explicitly mentioned within the texts. The documents do not follow a standardized structure: some are organized into formal sections, while others are divided into chapters, and they also vary in the encoding used during their creation. As an initial approach, the results suggest that the texts reflect the research conducted within PosLin. Another analysis of the bigrams and trigrams indicates that they are related to language teaching and learning, suggesting that the graduate program explores applied linguistics research within its lines of inquiry.

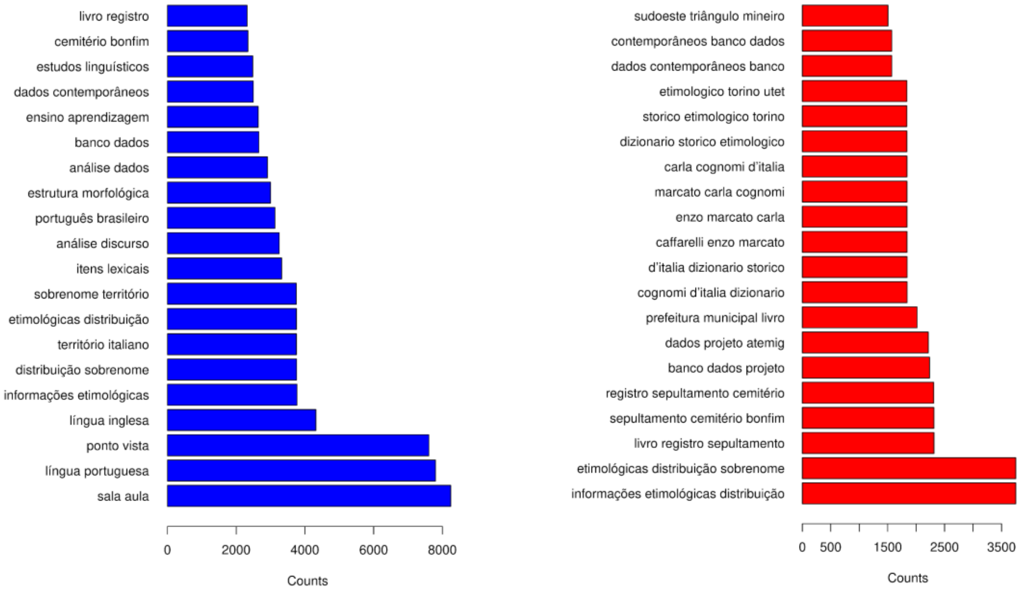


Figure 4: Bigrams and trigrams from PosLin theses.

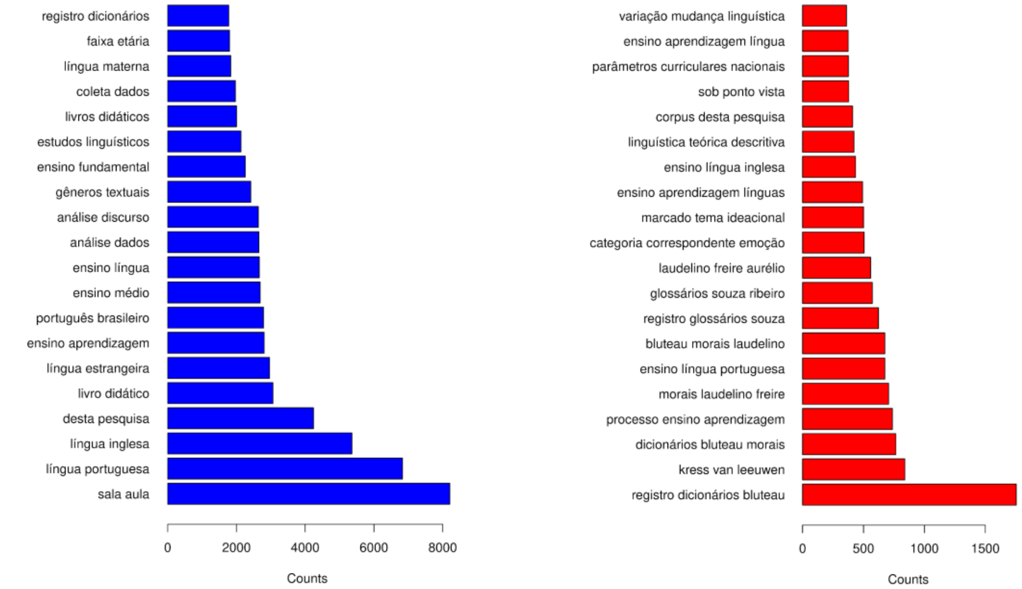


Figure 5: Bigrams and Trigrams from PosLin Dissertations.

The results so far indicate that, although the texts were produced within the same graduate program and share common research tracks, some differences emerged in the n-grams. To deepen the analysis, the 5,000 most common terms in the corpus—comprising both theses and dissertations—were selected and submitted to morphological classification using the palavras parser. Due to the volume of information to be processed and the variability in encoding text, the most frequent terms were selected for classification as an initial approach. Some terms may have been misclassified; however, this was considered acceptable for a preliminary analysis. Four morphological classes were considered (nouns, adjectives, verbs and adverbs) analyzed both for the complete corpus (dissertations and theses) and separately. The results are presented in the following tables: 1 (nouns), 2 (adjectives), 3 (verbs), and 4 (adverbs).

The most frequently used nouns are highlighted in Table 1. An analysis of the top ten nouns shows that only “aluno” and “exemplo” are not shared between theses and dissertations. This does not mean that the terms are absent in one modality, but they occur with lower frequency, as the analysis was limited to the top ten. It is noteworthy that dissertations and theses largely share the same nouns in describing their studies. An analysis of the most frequent nouns reveals that all are closely related to the research themes explored at PosLin. Academic texts tend to feature a higher prevalence of abstract nouns (40). This is supported by the results, as terms such as “língua”, “forma”, “relação”, “análise”, “pesquisa”, “discurso” and “processo” are classified as abstract nouns.

Table 1: Nouns used in PosLin dissertations and theses.

Dissertations	Frequency	Theses	Frequency
texto	77,374	texto	91,112
língua	70,735	forma	90,06
forma	66,229	relação	83,367
relação	57,521	língua	78,926
análise	50,736	discurso	60,295
aluno	47,373	análise	60,126
pesquisa	45,277	processo	55,888
estudo	44,857	estudo	53,599
discurso	43,915	exemplo	52,404
processo	42,165	pesquisa	52163

Table 2 presents the ten most frequently used adjectives. A comparison between dissertations and theses shows that only “belo” and “importante” are not shared among the top ten terms. Once again, an analysis of the occurring terms reveals that adjectives such as “linguístico” and “discursivo” are closely aligned with the research themes pursued at PosLin.

Table 2: Adjectives used in PosLin dissertations and theses.

Dissertations	Frequency	Theses	Frequency
grande	43,601	social	55,902
social	40,168	grande	54,277
linguístico	38,915	linguístico	45,809
possível	22,984	próprio	29,974
próprio	20,735	possível	28,252
novo	20,129	novo	28,072
discursivo	19,588	primeiro	24,549
primeiro	18,483	discursivo	24,390
importante	16,939	belo	22,399
textual	15,860	importante	20,316

An analysis of the ten most frequently used verbs shows that “partir” and “escrever” are the only ones not shared between the corpora (Table 3). The remaining verbs appear in both types of texts. These verbs are commonly used across various discourse genres, particularly in academic writing (41, 42).

Table 3: Verbs used in PosLin dissertations and theses.

Dissertations	Frequency	Theses	Frequency
ser	123,233	ser	156,052
poder	107,161	poder	134,473
fazer	76,229	fazer	99,119
apresentar	56,618	ver	69,949
ver	49,439	apresentar	65,721
ter	43,782	ter	53,155
falar	42,867	falar	51,818
escrever	33,652	dizer	42,511
dizer	32,655	partir	39,214
dar	32,064	dar	38,392

Finally, we examined the ten most frequently used adverbs, listed in Table 4. The analysis shows that the same adverbs appear in both theses and dissertations, differing only in frequency. The use of adverbs is common in genres such as academic, journalistic, and argumentative writing (43). Therefore, their occurrence is prominent in the analyzed corpus. Most adverbs are shared between theses and dissertations, as shown in Table 4, differing only in frequency.

Table 4: Adverbs used in PosLin dissertations and theses.

Dissertations	Frequency	Theses	Frequency
assim	48,818	assim	61,908
ainda	32,369	ainda	42,651
além	25,605	meio	35,768
meio	25,132	além	32,090
então	24,367	apenas	29,209
apenas	23,709	então	28,351
aqui	21,895	aqui	27,616
bem	20,630	bem	25,709
sempre	15,554	sempre	20,513
primeiro	15,457	tanto	20,030

An analysis of the four morphological classes reveals higher frequencies in doctoral theses, which can be attributed to their greater average length compared to dissertations. Theses are typically more extensive and complex, as they are required to address original research problems (44), whereas dissertations may revisit previously explored topics.

The analysis of n-grams and the four morphological classes provides evidence that the texts explore similar themes aligned with PosLin’s research tracks. Due to the high degree of similarity in the terms used, a compelling question arises: is it possible to automatically classify a document as either a thesis or a dissertation based solely on its textual content? To investigate this, a machine learning approach was employed to determine whether a model could be trained to accurately classify a text as either a thesis or a dissertation.

The development of a machine learning classifier was based on a binary labeling of the files as either “thesis” or “dissertation.” The texts were preprocessed by converting all content to lowercase and removing stopwords, numbers, and punctuation. As a result, only the frequency of each token in the text was considered. Each term was assigned a weight based on its frequency using a bag-of-words approach. Once the terms and their weights were computed for each file, the corpus was split into 90% for model training and 10% for validation, with ten repetitions applied during the training and validation process. The models used included Gradient Boosting, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, and Neural Network.

Table 5 presents the estimated values of model fit quality for the corpus, considering the 50, 1,000, and 30,000 most frequent tokens from the corpus of dissertations and theses. Regardless of the number of frequent tokens, the model that achieved the best fit values (AUC) was Gradient Boosting (gray color in table). Note that for the 50 most frequent tokens, the AUC was 0.829, whereas for 1,000 tokens was 0.989, and for 30,000 tokens was estimated at 0.985. Also in Table 5, when analyzing only the AUC values for Gradient Boosting, there is a significant difference between the AUC for 50 tokens (0.829) compared with 1,000 and 30,000 tokens (0.989 and 0.985). However, between 1,000 and 30,000 tokens the difference is small (0.004), and the remaining metrics (CA, F1, Precision, and Recall) are identical. The processing time for 1,000 and 30,000 tokens is significant, ranging from minutes to hours, respectively. This fact justifies choosing the corpus size based on the 1,000 most frequent tokens. It should also be noted that increasing the number of tokens is a determining factor for improving the fit quality of the other models, though their results remain inferior to Gradient Boosting.

Table 5: Performance metrics for the models Gradient Boosting (GB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Neural Network (NN), using the 50, 1,000, and 30,000 most frequent tokens in the corpus, with weights calculated using the bag-of-words approach. The metrics used were AUC (Area Under the ROC Curve), CA (Classification Accuracy), F1 Score, Prec (Precision), Recall, and MCC (Matthews Correlation Coefficient). The best results for each token count are highlighted.

Tokens	Model	AUC	CA	F1	Prec	Recall
50	KNN	0.781	0.737	0.733	0.736	0.737
	GD	0.829	0.762	0.761	0.761	0.762
	SVM	0.522	0.628	0.540	0.720	0.628
	LR	0.801	0.735	0.732	0.734	0.735
	NN	0.596	0.587	0.588	0.589	0.587
1,000	KNN	0.786	0.708	0.680	0.742	0.708
	GD	0.989	0.949	0.949	0.949	0.949
	SVM	0.615	0.436	0.312	0.525	0.436
	LR	0.779	0.735	0.733	0.733	0.735
	NN	0.625	0.644	0.640	0.640	0.644
30,000	KNN	0.744	0.660	0.600	0.730	0.660
	GD	0.985	0.949	0.949	0.949	0.949
	SVM	0.819	0.692	0.651	0.749	0.692
	LR	0.631	0.650	0.642	0.645	0.650
	NN	0.587	0.629	0.621	0.623	0.629

Figure 6 displays the AUC (Area Under the ROC Curve) values as a function of corpus size for the models Gradient Boosting (GB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Neural Network (NN) in classifying texts as either dissertations or theses. The results indicate that as the number of tokens increases, only the AUC values for GB remain consistently high, while the other models show slight fluctuations.

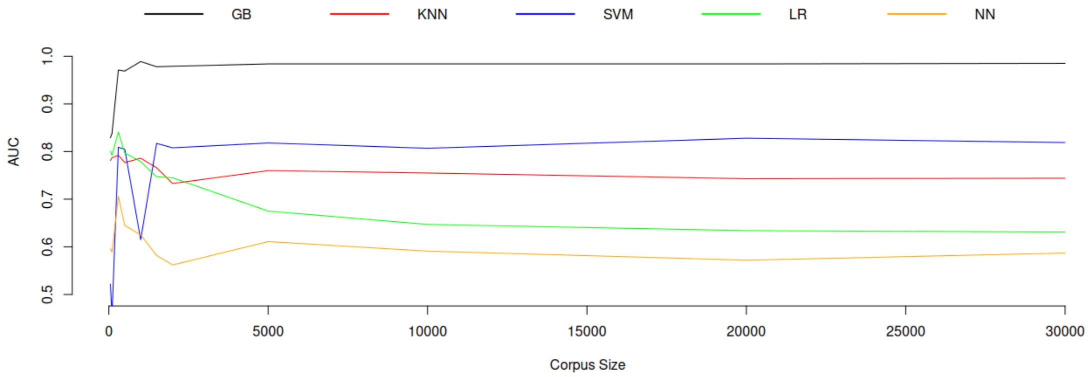


Figure 6: Area under the ROC curve (AUC) as a function of corpus size for the models Gradient Boosting (GB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Neural Network (NN).

Figure 7a shows that the AUC for the GB model tends to stabilize, in contrast to the other models, as the number of tokens increases. Figure 7b isolates the AUC trend, clearly demonstrating that as the number of tokens rises, the AUC approaches stability, reaching a peak value of 0.989. Using the 300 most frequent terms, the GB model achieved an AUC of 0.971. With 1,000 terms, the AUC was estimated at 0.989, while with 1,500 terms it was 0.978. For token counts ranging from 5,000 to 30,000, AUC values remained between 0.984 and 0.985. Depending on the analysis objective, the marginal gain of 0.014 may not justify the computational cost, as processing time ranges from minutes for 300 tokens to several hours for 30,000. The results indicate that, for classifying texts as dissertations or theses, the GB model achieves an AUC of 0.989 using the top 1,000 terms, with an average processing time of just 13 minutes, which is already sufficient for practical purposes. This finding is consistent with similar studies. Noguti, Vellasques, and Oliveira (45) analyzed legal texts using a limited dataset, and Lu (46) successfully classified documents with a reduced number of tokens. For the analyses, we considered the AUC. However, other accuracy measures may also be used, such as the F1 Score. In this study, they were not employed because the values were similar, but we recommend always taking them into account.

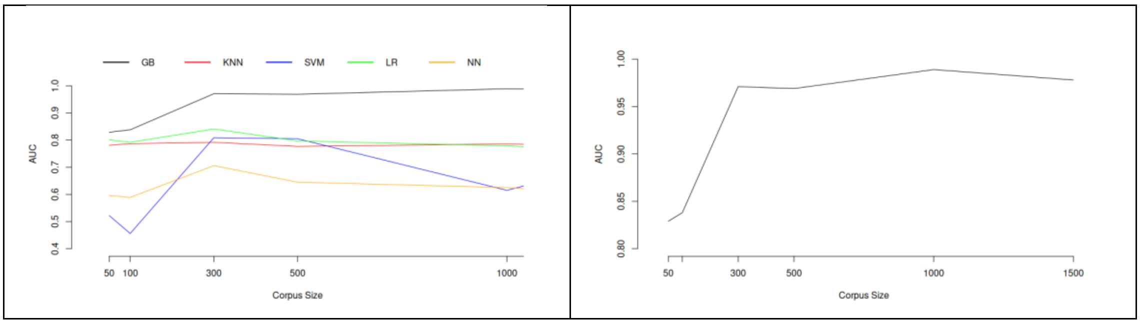


Figure 7: Area under the ROC curve as a function of corpus size for the models Gradient Boosting (GB), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression (LR), and Neural Network (NN) — left (a); and for GB only — right (b).

Table 6 presents the confusion matrix for the test sample, which included 1,270 files. Of the 730 dissertations, 691 were correctly classified, while 39 were misclassified as theses (false negative). Among the 540 theses, 514 were correctly classified and 26 were misclassified. These misclassifications may have resulted from the textual similarity between dissertations and theses, as both tend to follow a formal scientific writing style. An analysis of the misclassified documents reveals that they exhibit textual features similar to those of the class to which they were assigned. For instance, this may occur when a dissertation written with characteristics of a thesis, using the same most frequent terms, which in this case would lead the model to an incorrect classification. The same may occur with a thesis that, when written, presents the characteristic terms of a dissertation and consequently misclassified.

Table 6: Confusion matrix for the application of the Gradient Boosting model to 1,270 files (540 theses and 730 dissertations).

		Expected		
		Theses	Dissertation	Total
Predicted	Theses	514	26	540
	Dissertations	39	691	730
	Total	553	717	1270

The results indicate that, based on documents exhibiting a linguistic pattern, machine learning models can be effectively employed as classifiers. The linguistic pattern determines the quality of model training and application. Consequently, the methodology presented may be extended to other corpora addressing topics such as fake news, disinformation, and hate speech. Nonetheless, comparable performance may not be attainable across different datasets (corpora). It should be noted that the texts analyzed here follow a formal scientific writing style. Despite the similarities observed in token, bigram, trigram, and part-of-speech frequencies (nouns, adjectives, verbs, and adverbs), the models were still able to achieve robust classification performance.

6 Conclusions

The analyses were conducted using a corpus composed of dissertations and theses from PosLin. Through textual analyses, it was possible to identify linguistic patterns by means of n-grams and morphological classes that reflected the research lines of the program. The application of machine learning methodologies enabled the classification of the texts (dissertations and theses) with high performance, with Gradient Boosting emerging as the model that provided the best fit quality. Furthermore, it was found that the number of terms analyzed can be reduced without significant loss of performance, thereby optimizing computational resources during processing. Future studies will focus on analyzing abstracts instead of full texts, in order to assess whether this approach allows for classifications equal to or even superior to those previously obtained.

7 Acknowlegments

I am grateful to Deise Prina Dutra, Tommaso Raso, and Adalgisa Ribeiro Torres for their academic support, guidance, and valuable feedback on this work. This study is dedicated to the researcher and lifelong professor Heliana Ribeiro de Mello, who introduced me to Linguistics and with whom I had the privilege of sharing moments of learning, conversation, advice, and friendship.

REFERENCES

1. Brazil. Ministry of Education (MEC), CAPES. Stricto sensu graduate programs in Brazil surpassed 350,000 enrollments in 2023. [Internet]. [cited 2025 Aug 25]. Available from: https://tinyurl.com/yc2f5e28.

2. Programa de Pós-Graduação em Estudos Linguísticos - POSLIN [Internet]. [cited 2025 Aug 25]. Available from: http://www.poslin.letras.ufmg.br/

3. Soares F, Yamashita GH, Anzanello MJ. A Parallel Corpus of Theses and Dissertations Abstracts. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. 2018 [cited 2025 Aug 25];11122 LNAI:345–52. Available from: https://tinyurl.com/mrasvzhv

4. Kauffmann CH. Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis. 2020 Apr 30 [cited 2025 Aug 25]; Available from: https://tinyurl.com/mrb85unk

5. Owa DLM. Estudo comparativo entre análise multidimensional lexical e modelagem de tópicos. 2021;

6. Nunes W da C. A identificação de metáforas em corpus jornalístico comparável bilíngue: estudo contrastivo espanhol/português. 2023 May 30 [cited 2025 Aug 25]; Available from: https://tinyurl.com/2te87dwu

7. Kutuzov A, Kopotev M, Sviridenko T, Ivanova L. Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints. 2016 Apr 18 [cited 2025 Aug 25]; Available from: https://arxiv.org/pdf/1604.05372

8. Yilmaz S, Römer U. A corpus-based exploration of constructions in written academic English as a lingua franca. 2020 Feb 20;59–88.

9. Barbosa GA, Batista HHN, Miranda P, Santos J, Isotani S, Cordeiro T, et al. Aprendizagem de Máquina para Classificação de Tipos Textuais: Estudo de Caso em Textos escritos em Português Brasileiro. Simpósio Brasileiro de Informática na Educação (SBIE) [Internet]. 2022 Nov 16 [cited 2025 Aug 26];920–31. Available from: https://tinyurl.com/2mtvfpdf

10. Cândido ECR. Um estudo comparativo de redes neurais profundas para classificação automática de texto. 2020 Feb 14 [cited 2025 Aug 25]; Available from: https://tinyurl.com/4wk3k2ns

11. Li X. Text classification using topic modelling and machine learning [Internet]. Nanyang Technological University; 2024 [cited 2025 Aug 25]. Available from: https://hdl.handle.net/10356/176723

12. Humaidi MH, Sutrisno, Laksono PW. Implementation of Machine Learning for Text Classification Using the Naive Bayes Algorithm in Academic Information Systems at Sebelas Maret University Indonesia. E3S Web of Conferences [Internet]. 2023 Dec 18 [cited 2025 Aug 26];465:02048. Available from: https://tinyurl.com/m7pvthw2

13. Fiorin JL, editor. Introdução à Linguística I: objetos teóricos. 6th ed. São Paulo: Contexto; 2010. 226.

14. Jurafsky D, Martin JH. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models [Internet]. 3rd ed. 2025. Available from: https://web.stanford.edu/~jurafsky/slp3/

15. Moraes LC, Silvério IC, Marques RAS, Anaia BC, de Paula DF, de Faria MCS, et al. Linguistic ambiguity analysis in large language models (LLMs). Texto Livre [Internet]. 2025 [cited 2025 Aug 25];18:e53181. Available from: https://tinyurl.com/yrf845h8

16. Bagno Marcos. Preconceito linguístico. 56th ed. Parábola Editorial; 2015 [cited 2025 Aug 26]. 352.

17. Goldberg Y. Neural Network Methods for Natural Language Processing. 2017 [cited 2025 Aug 25]; Available from: https://link.springer.com/10.1007/978-3-031-02165-7

18. She X, Zhang D. Text Classification Based on Hybrid CNN-LSTM Hybrid Model. Proceedings - 2018 11th International Symposium on Computational Intelligence and Design, ISCID 2018. 2018 Jul 2;2:185–9.

19. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

20. Bishop CM. Pattern Recognition and Machine Learning. 1st ed. New York: Springer; 2006. (Information Science and Statistics).

21. Géron Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2022 [cited 2025 Aug 25]; Available from: https://tinyurl.com/bdexmfra

22. Carvalho ACPLF, Menezes ÁG, Bonidia RP. Ciência de Dados – Fundamentos e Aplicações. São Paulo: LTC; 2024. 376.

23. Morettin PA, Singer J da M. Estatística e ciência de dados. 2025 ;[citado 2025 ago. 26]

24. Hsu BM. Comparison of Supervised Classification Models on Textual Data. Mathematics. 2020 Aug;8:851.

25. Britto FA, Ferreira TC, Nunes LP, Parreiras FS. Comparing Supervised Machine Learning Techniques for Genre Analysis in Software Engineering Research Articles [Internet]. 2021 [cited 2025 Aug 25]. p. 63–72. Available from: https://aclanthology.org/2021.ranlp-1.8/

26. Almatarneh S, Gamallo P. Comparing Supervised Machine Learning Strategies and Linguistic Features to Search for Very Negative Opinions. Information 2019, Vol 10, Page 16 [Internet]. 2019 Jan 4 [cited 2025 Aug 25];10(1):16. Available from: https://www.mdpi.com/2078-2489/10/1/16/htm

27. Allam H, Makubvure L, Gyamfi B, Graham KN, Akinwolere K. Text Classification: How Machine Learning Is Revolutionizing Text Categorization. Information 2025, Vol 16, Page 130 [Internet]. 2025 Feb 10 [cited 2025 Aug 26];16(2):130. Available from: https://www.mdpi.com/2078-2489/16/2/130/htm

28. Programa de Pós-Graduação em Estudos Linguísticos - POSLIN [Internet]. [cited 2025 Aug 25]. Available from: http://www.poslin.letras.ufmg.br/bancodefesas.php

29. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. Vienna, Austria; 2023. Available from: https://www.R-project.org/

29. Harris ZS. Distributional Structure. WORD [Internet]. 1954 Aug [cited 2025 Aug 25];10(2–3):146–62. Available from: https://tinyurl.com/2wnaefsh

31. Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinović M, et al. Orange: Data Mining Toolbox in Python. Journal of Machine Learning Research [Internet]. 2013;14(Aug):2349–53. Available from: http://jmlr.org/papers/v14/demsar13a.html

32. Posit team. RStudio: Integrated Development Environment for R [Internet]. Boston, MA; 2025. Available from: http://www.posit.co/

33. Feinerer I, Hornik K, Meyer D. Text Mining Infrastructure in R. Journal of Statistical Software [Internet]. 2008 Mar 31 [cited 2025 Aug 25];25(5):1–54. Available from: https://tinyurl.com/ye27uru5

34. Hornik K, Meyer D, Buchta C. slam: Sparse Lightweight Arrays and Matrices [Internet]. 2024. Available from: https://CRAN.R-project.org/package=slam

35. Fellows I. wordcloud: Word Clouds [Internet]. 2018. Available from: https://CRAN.R-project.org/package=wordcloud

36. Neuwirth E. RColorBrewer: ColorBrewer Palettes [Internet]. 2022. Available from: https://CRAN.R-project.org/package=RColorBrewer

37. Wickham H, Hester J, Bryan J. readr: Read Rectangular Text Data [Internet]. 2024. Available from: https://CRAN.R-project.org/package=readr

38. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software [Internet]. 2018;3(30):774. Available from: https://quanteda.io

39. Bick E. The Parsing System ``Palavras'': Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. In: Proceedings of the International Conference on Computational Processing of the Portuguese Language (PROPOR 2000). Évora, Portugal: Springer; 2000. p. 35–45. (Lecture Notes in Artificial Intelligence; vol. 2721).

40. Swales JM. Genre Analysis: English in Academic and Research Settings. Cambridge, UK: Cambridge University Press; 1990.

41. Neves MH de M. Gramática de usos do português. 2nd ed. São Paulo: Editora UNESP; 2011.

42. Rawlins JD, Eckstein G, Hanks E, Lester EW, Wilde L, Bartholomew R. Intentional function and frequency of reporting verbs across six disciplines: A cluster analysis. International Journal of English for Academic Purposes: Research and Practice [Internet]. 2024 Mar 6 [cited 2025 Aug 26];4(1):47–71. Available from: https://tinyurl.com/ycyrhvzn

43. Larsson T, Callies M, Hasselgård H, Laso NJ, Vuuren S van, Verdaguer I, et al. Adverb placement in EFL academic writing: Going beyond syntactic transfer. International Journal of Corpus Linguistics [Internet]. 2020 Aug 28 [cited 2025 Aug 26];25(2):156–85. Available from: https://tinyurl.com/37h8kyu8

44. Fernandes ICS. Marcadores discursivos e efeitos de sentido: além das fronteiras dos estudos sobre coesão. Estudos Linguisticos [Internet]. 4º de abril de 2016 [citado 27º de agosto de 2025];42(3):1073-87. Disponível em: https://revistas.gel.org.br/estudos-linguisticos/article/view/915

45. Noguti MY, Vellasques E, Oliveira LES. A Small Claims Court for the NLP: Judging Legal Text Classification Strategies With Small Datasets. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics [Internet]. 2024 Sep 9 [cited 2025 Aug 25];1840–5. Available from: http://arxiv.org/abs/2409.05972

46. Lu J, Henchion M, Bacher I, Namee B mac. A Sentence-level Hierarchical BERT Model for Document Classification with Limited Labelled Data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [Internet]. 2021 Jun 12 [cited 2025 Aug 25];12986 LNAI:231–41. Available from: https://arxiv.org/pdf/2106.06738