Exploring Computational Discernibility of Discourse Domains in Brazilian Portuguese within the Carolina Corpus

Felipe Ribas Serras[1], **Mariana Lourenço Sturzeneker**[1], **Miguel de Mello Carpi**[1], **Mayara Feliciano Palma**[1], **Maria Clara Ramos Morales Crespo**[2], **Aline Silva Costa**[3], **Vanessa Martins do Monte**[1], **Cristiane Namiuti**[4], **Maria Clara Paixão de Sousa**[1], **Marcelo Finger**[1]
[1]University of São Paulo, São Paulo, Brazil
[2]University of Bologna, Bologna, Italy
[3]Federal Institute of Education, Science and Technology of Bahia, Vitória da Conquista, Brazil
[4]State University of Southwestern Bahia, Vitória da Conquista, Brazil
{frserras,mariana.sturzeneker,miguel}@ime.usp.br

Abstract

In this study, we explore the computational discernibility of Portuguese language discourse domains using a balanced sample from the Carolina corpus, including its five largest domains: *Juridical*, *Entertainment*, *Journalistic*, *Virtual* and *Instructional*. We analyze discernibility across three levels: degree of duplication, linguistic features distribution, and separability within semantic embedding spaces. We found clear quantitative differences between domains at all levels, compatible with expected qualitative properties. Our analysis shows that these domains can be distinguished based on various computable text properties, and suggests a consistent complexity scale between them. We identify the distinguishing properties and their potential benefits for NLP tasks. Additionally, we provide domain-balanced and deduplicated versions of Carolina for future research.

1 Introduction

The recent wave of large language models has boosted the amount of resources available for NLP in Portuguese, generating a robust and competitive foundation of computational assets. Models such as BERTimbau (Souza et al., 2020) , Albertina (Rodrigues et al., 2023), Sabiá (Pires et al., 2023) and *corpora*/datasets such as BRWaC (Wagner Filho et al., 2018), Oscar (Suárez et al., 2019, 2020), ClueWeb22 (Overwijk et al., 2022) and the Carolina Corpus (Sturzeneker et al., 2022; Crespo et al., 2023) are just a few examples.

Most of these resources are general, treating the Portuguese language as a homogeneous whole, without focusing on specific dialects, registers, or domains. For a language that historically has comparatively low computational resources, such as Portuguese, it is coherent to seek the development of transversal tools, aiming to meet multiple needs in different fields of application.

However, to comprehend the diversity of varieties of the Portuguese language, one needs to look past the production of general resources by means of a systematic exploration of those resources and tools. This type of experimentation comes in response to an already old perception in linguistics: that the division of a language into sub-languages is a way of making it operationally useful (Catford, 1965).

Such explorations can vary, both in the dimension of variation of the language they focus on (e.g. dialect, idiolect, norm, user-medium relationship, genre, type, domain, etc.) (Gregory, 1967) and in the methodological approach adopted to explore it, generating different possible combinations.

In this work, we focus on the *domain of discourse* as a dimension of language variation. Discourse domains are typological variations within a natural language characterized by properties, structures, and conventions determined by the context and/or communicative situation in which the texts occur (Gregory, 1967; Douglas, 2004).

The success of large language models, that intensified resource development, has also solidified a performance-based development approach. Feasibility studies and explorations of the meaning of an application are abandoned in favor of training and evaluating models for the application in question. Good performance metrics are often read as proof of the feasibility of the task, without carrying out deeper analysis.

In this study, we aim to challenge this approach by assessing the computational feasibility of distinguishing domains of discourse in Portuguese. This serves as a preliminary investigation before delving into the development of models for domain discernibility. Hence, in this study, we address the following questions:

1. Are discourse domains computationally discernible?

2. If so, which properties differentiate them?

3. What approaches for discerning domains in NLP tasks are experimentally supported?

To address these questions, we employ data from the Carolina Corpus, a general corpus of Portuguese made of open texts extracted from the internet. Each text has a header that contains various information, including three typological annotations: broad type, type declared by the source, and discourse domain (Crespo et al., 2023). We focus on the latter, using the others only when necessary to better understand our results. We evaluate the discernibility of different discourse domains under the following aspects: level of duplication, distribution of linguistic features, and separability in semantic embedding spaces.

This paper is organized as follows. In section 2 we present relevant related works; in section 3 we detail the dataset used in our analysis; in section 4 we present and justify the different aspects used to analyze discernibility, as the general methodology adopted across the different levels of analysis; in section 5 we present the analysis of domain discernibility under processes of deduplication; in section 6 we present the analysis of discernibility under linguistic features extracted by computational models, and in section 7 we present the analysis of domain discernibility within semantic embedding spaces. In section 8 we present our conclusions, contributions, and future steps.

2 Related Works

This study offers a unique analytical exploration of computational differentiation of discourse domains in Portuguese, as far as we know. In this section, we list other works that coincide with ours in certain aspects.

Regarding the construction of resources, some authors prefer building diverse corpora, like Williams et al. (2018), while others create datasets for specific typologies, such as Koreeda and Manning (2021).

In contrast to generalist models, certain studies concentrate on domain-specific computational models (Fonseca et al., 2016; Gu et al., 2021; Lee et al., 2020; Beltagy et al., 2019; Zhou et al., 2013; Serras and Finger, 2022; Viegas, 2022; Polo et al., 2021; de Colla Furquim and de Lima, 2012). Often, these domain-specific models are built by adapting generalist models to specific application domains, so-called domain adaptation techniques.

Text classification models considering linguistic information, are proposed in various works (Johnson et al., 2002; Gonçalves and Quaresma, 2005). Kessler et al. (1997), for instance, addresses specifically the issue of genre classification in Portuguese using linguistic features.

Multiple works delve into text complexity, including Juola (2008), Szmrecsanyi (2016), and Ehret and Szmrecsanyi (2019). Leal et al. (2023) provides a set of complexity metrics for Portuguese, with some overlap with the linguistic features used here.

3 Data

Our data source was the Carolina Corpus, an open and curated digital collection of Portuguese documents, developed for training large language models and facilitating linguistics research. In Carolina's version 1.2 Ada, typological information is organized into three distinct metadata entries: broad typology, source typology, and domain. *Broad typology* represents a methodological division based on how data was segmented during analysis and retrieval. *Source typology* refers to the text's typology as declared in the source from which the document was extracted, it tends to be specific and non-standardized. *Domain* represents the discourse domain of the text, annotated by the Carolina team using a pre-defined system applied over the different examined sources.

Regarding discourse domain, our primary tag of interest, corpus documents are categorized into ten distinct groups: Instructional (41.8%), Juridical (23.8%), Entertainment (14.7%), Journalistic (10.6%), Virtual (7.4%), Academic (0.51%), Commercial (0.43%), Legislative (0.38%), Literary (0.19%) and Pedagogical (0.096%).

The five primary Carolina domains, collectively representing around 98.4% of the corpus tokens, are defined below. The source types contained within each of these domains are listed to enhance comprehension of their composition:

- *Instructional*: texts distributed in spaces designed for instructing and educating readers, such as virtual encyclopedias. The source typologies contained within this domain are: *vocabulary entry*, *educational resource*, *help documentation* and *travel guide*;

- *Juridical*: documents distributed within the Brazilian Judiciary branch. It encompasses a

very diverse list of source typologies, i.e. *appellate decision records*, *request for proposals*, *study of precedents by minister*, *topical publication*, *report*, *open court hearing*, *speech*, *proposal of binding precedent*, *minutes*, *constitution annotated*, *precedents bulletin*, *biography*, *glossary*, *resolution*, *court members information* and *treaty*;

- *Entertainment*: texts distributed within platforms designed for entertainment purposes. This domain consists of a single source typology: *subtitles*;
- *Journalistic*: texts distributed within news platforms and related environments. The source typologies within this domain are *news*, *scientific news*, *article*, *opinion* and *journalistic blog*;
- *Virtual*: texts distributed solely within native virtual environments, such as social media platforms. The source typologies contained in this domain are *user page*, *discussion*, *tweet*, *activities organization and experiences sharing*, *personal blog* and *faq*.

The sources of documents within each domain can be found in the corpus provenance tags concerning each document. General provenance information is also available on Carolina's homepage[1]. Carolina developers are dedicated to incorporating new domains into the corpus and achieving a balance between existing domains. This ensures the possibility of repeating our experiments in the future with new domains and a more balanced dataset.

4 Methodology

Our analysis of discernibility was divided into three distinct approaches: degree of duplication, distribution of linguistic features, and separability in embedding spaces. This division was chosen to accommodate the multidimensional nature of discourse domains and the selection of these approaches was based on anticipated differences in language conventions between domains, specifically:

- the use of technical terms, formulaic language, and phatic expressions. These variations directly influence the degree of document duplication within each domain;
- the vocabulary usage and its characteristics, leading to morphological and syntactic differences, evident through morphosyntactic features analysis;
- the subject matter covered in the texts, affecting the average semantics of documents. This potential difference between texts could be detected by employing a separability analysis over semantic embedding spaces.

The focus of this work is on distinguishing discourse domains within a **computational scope**. Consequently, our analysis is consistently mediated through computational tools, namely Onion (Section 5), spaCy (Section 6), and NILC embeddings (Section 7).

To extract the data for discernibility analysis used across the three approaches, we created a smaller balanced version of Carolina in terms of domains, named Carol·$\mathcal{B}$: Balanced Carolina Subcorpus[2]. Carol·$\mathcal{B}$ contains a similar number of tokens from each of Carolina's largest domains: *Instructional*, *Juridical*, *Entertainment*, *Journalistic*, and *Virtual*. In total, the sub-corpus has 304,205,653 tokens, approximately 60,8M tokens per domain.

We randomly sampled documents of different domains until we meet the number of tokens of the smallest domain (*Virtual*). Sampling was performed in order to also keep balanced the source types[3] within each domain, maintaining a maximum representation of the internal diversity of all selected discourse domains.

5 Discernibility through Deduplication

Our approach to evaluating the degree of textual duplication between the documents of a domain was to use a deduplication tool. We understand *deduplication* as the process of removing unoriginal content from a corpus, and, consequently, *deduplicated* is a text or a corpus after the performance of deduplication.

Here, we used Onion (ONe Instance ONly) (Pomikálek, 2011)[4] as our deduplication tool. Onion is a computational tool that determines if

[1] https://sites.usp.br/corpuscarolina/documenta/1-2-ada/repositorios-2023

[2] The links to all data and source code developed for this study are available at this list: https://github.com/stars/frserras/lists/domain-discernibility-carolina

[3] Information on the typology of the text as declared in the source from which it was extracted. See Crespo et al. (2023).

[4] Onion is available at: https://corpus.tools/wiki/Onion.

each text in a *corpus* is completely or partially duplicated and removes duplicates. A duplicate content threshold $\mathcal{T} \in [0, 1]$ can be provided, where $\mathcal{T} = t$ means that only the documents with $10t\%$ or more of repeated n-grams will be considered as duplicated and consequently removed. To classify a n-gram as repeated, Onion compares the texts' n-grams with a list of previously processed n-grams. Thus, it takes into account the order in which the documents are presented to it.

We used Onion to compare the duplicate removal rate of whole documents within each domain of the corpus. We used the default settings for all parameters except for $\mathcal{T}$[5], and repeated the deduplication process 5 times, each with a different randomized order of documents. For each domain, we computed the mean and variance over the random orderings of the *density of removed tokens* $\mathcal{D}$ for different values of the *minimum originality required* for a text to be kept $\mathcal{O}$, defined in equations 1 and 2. The obtained curves can be seen in Figure 1.

$$\mathcal{D} = \frac{\text{\# removed tokens}}{\text{\# tokens in the domain}} \tag{1}$$

$$\mathcal{O} = 1 - \mathcal{T} \tag{2}$$

When analyzing Figure 1's curve behavior, it's clear that some domains are more susceptible to changes in $\mathcal{O}$, e.g. *Juridical* and *Entertainment*. This likely stems from the nature of the domains: *Juridical* texts can be very similar in structure and contain more standardized and repetitive language; while *Entertainment* texts on Carolina are mainly subtitles of kids' movies and TV series and probably make use of repetitive and simplified language, with thematic superposition between the episodes of the same TV series. The *Entertainment* and *Juridical* domains also contain the largest documents, therefore, when Onion lists and compares n-grams, larger average documents likely affect deduplication rates.

Two variables of high interest are the densities of removed tokens when the required originality is minimum $\mathcal{D}|_{\mathcal{O}=0}$ and maximum $\mathcal{D}|_{\mathcal{O}=1}$. They represent the density of documents that are completely duplicated and the density of documents that are not completely original, respectively.

[5]We also experimented with the size n of each n-gram, but as no meaningful variation was observed, we adopted the default $n = 5$. Pomikálek (2011, p. 80) analyzed the impact of the n-gram length on his work and concluded that any n-gram configuration should work well, with few "pathological" exceptions.

Juridical and *Entertainment* domains have the highest $\mathcal{D}|_{\mathcal{O}=1}$, which follows the behavior patterns previously mentioned. The *Instructional* domain has the third higher $\mathcal{D}|_{\mathcal{O}=1}$. This can also be explained by the fact that some encyclopedic texts, which constitute a large part of the texts in this domain, follow a more structured pattern. The values of $\mathcal{D}|_{\mathcal{O}=1}$ for each domain are also shown in Figure 1.

Virtual is the only domain with a meaningful $\mathcal{D}|_{\mathcal{O}=0}$, but other domains have also some degree of absolute duplication according to Onion. In Table 1, we exhibit the absolute number of removed tokens per domain when $\mathcal{O} = 0$ and the equivalent number of removed tokens when we consider exact copies in detriment of Onion *criteria*. The only domains with exact copies are *Virtual* and *Journalistic*. Noticeably, *Virtual* contains the most exact copies. Analyzing the duplicates, we came across several examples of phatic language and functional texts, e.g. greeting tweets in the *Virtual* domain, and posts notifying readers that a column would not be posted on that day, in the *Journalistic* domain.

The randomized order of texts minimally impacted the results, evident in the subtle variance indicated by lighter shading in each graph line. Specifically, *Juridical* and *Virtual* domains exhibited higher variance, yet there are discernible consistent patterns in the curves, underscoring the robustness of Onion as a deduplication tool.

Figure 1 and our analysis demonstrate the discernibility of domains based on internal duplication degrees. To facilitate various future applications, we capitalized on this exploration and developed "Carol·($\mathcal{D}$+$\mathcal{B}$): Deduplicated and Balanced Carolina Sub-corpus". Carol·($\mathcal{D}$+$\mathcal{B}$) was created by reducing duplication of the Carolina corpus using varying $\mathcal{T}$ values for each domain: $\mathcal{T} = 0$ for *Instructional*, $\mathcal{T} = 0.1$ for *Journalistic*, $\mathcal{T} = 0.5$ for *Entertainment*, and $\mathcal{T} = 0.8$ for *Juridical*. This process yielded token counts of 62,766,935, 68,543,795, 60,880,758, and 81,863,020 per domain, respectively. The methodology outlined in Section 4 was then reapplied to construct another balanced sub-corpus incorporating the deduplicated domains.

6 Discernibility through Linguistic Features

Linguistic theories provide a wide variety of features according to which one can describe linguistic

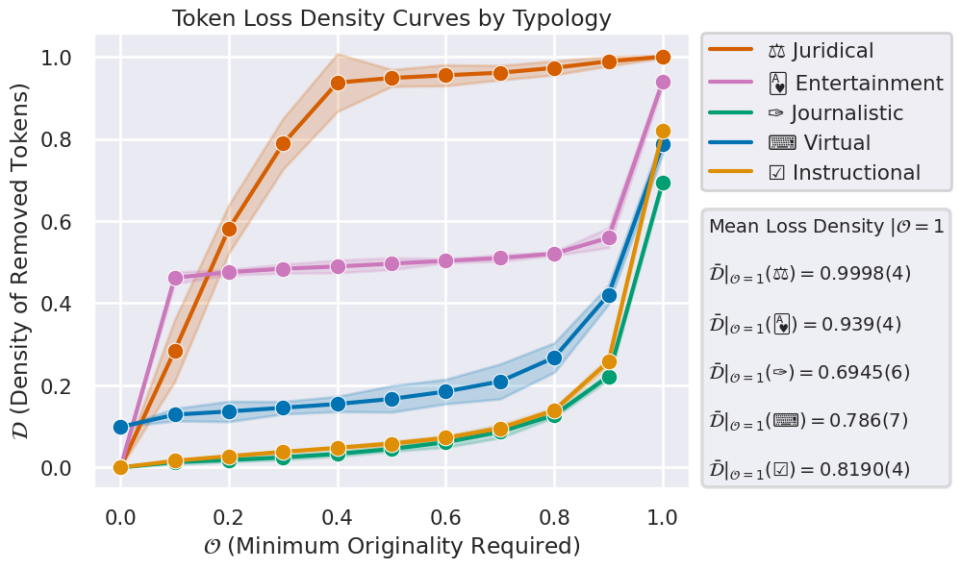

Figure 1: Token loss density curves by domain.

Table 1: Comparison between tokens removed by Onion and exact duplicates.

	Removed tokens (Onion)	**Removed tokens (Exact duplicates)**
Instructional	519	0
Entertainment	1.139	0
Journalistic	4.913	1.019
Juridical	0	0
Virtual	6.002.616	32.322
Total	**6.009.187**	**33.341**

properties and structures. Features based on linguistic theories have a well-established theoretical basis and standard semantic interpretation, which enables in-depth analysis. However, they often require annotation by specialists, which is costly and unfeasible for large *corpora*.

Computational models trained to annotate texts according to these features are a possible alternative. These models possess inherent errors. However, by analyzing a substantial amount of texts and applying statistical techniques to annotated feature values, we can effectively mitigate and estimate the analysis error, thereby formally ensuring their reliability.

In this work, we are interested in the **computational** discernibility of textual domains. So the use of computational models also guarantees that the linguistic features used to discern them are, at least, approximately computable. This allows conclusions to be drawn about the discernibility of discourse domains in computational contexts.

For feature annotation, we use the pre-trained models from the *spaCy* package[6]. These are state-of-the-art models for Portuguese that allow the extraction of a diverse set of linguistic features. Formally, we define a feature $\mathcal{F}_j$ as in 3, where $\mathbb{U}$ is a set of text units over which $\mathcal{F}_j$ is computed (e.g. words, n-grams, sentences), $\mathbb{F} = \{f_i\}$ is the set of values f_i that $\mathcal{F}_j$ can take, and $c_i \in 2^{\mathbb{U}}$ is the annotation context. A model is then a computable approximation $\hat{\mathcal{F}}_j$ of $\mathcal{F}_j$. The features used in our analysis and their respective sets $\mathbb{U}$ and $\mathbb{F}$ are represented in Table 2.

$$\mathcal{F}_j : \mathbb{U} \times 2^{\mathbb{U}} \to \mathbb{F}; (u_i, c_i) \mapsto \mathcal{F}_j(u_i, c_i) = f_i \quad (3)$$

Due to the size of the *corpus* and models, we analyzed a sample $\mathcal{S}$ of 1% of Carol·$\mathcal{B}$. For the features for which $\mathbb{F} = \mathbb{N}$, statistics were obtained from aggregation over the whole $\mathcal{S}$ set. For the other features, we applied a partitioning technique: $\mathcal{S}$ was divided into 10 partitions s_l and the distribution of the values of each feature $\mathcal{F}_j$ was computed independently over each partition s_l for each domain D_k .

For each feature $\mathcal{F}_j$ we compute the average probability over the partitions s_l of $\mathcal{F}_j$ being f_i if the discourse domain is D_k, represented by

[6] https://spacy.io/

Table 2: Linguistic Features evaluated in this work.

Feature	$\mathbb{U}$	$\mathbb{F}$
Tokens per Sentence	Sentence	$\mathbb{N}$
Characters per Token	Token	$\mathbb{N}$
Stop Words per Sentence	Sentence	$\mathbb{N}$
Tokens per Sentence	Sentence	$\mathbb{N}$
Punctuation Symbols per Sentence	Sentence	$\mathbb{N}$
Morphological Number	Token	{SING, PLUR, $\varnothing$}
Morphological Case	Token	{NOM, DAT, ACC, $\varnothing$}
Morphological Gender	Token	{MASC, FEM, $\varnothing$}
Morphological Tense	Token	{PRES, PAST, IMP, FUT, $\varnothing$}
Morphological Mood	Token	{SUB, IND, CND, $\varnothing$}
Named Entity Type	Token Sequence	{ORG, MISC, LOC, PER $\varnothing$}
Part-of-Speech	Token	{SCONJ, VERB, PROPN, PRON, CCONJ, ADV, AUX, ADJ, DET, NOUN, ADP, INTJ, NUM, X, PUNCT, SYM}

$\bar{\mathcal{P}}_j(f_i|D_k)$ and defined in equation 4. We use the standard error $\sigma_j(f_i|D_k)$ as the correspondent error.

$$\bar{\mathcal{P}}_j(f_i|D_k) = \frac{1}{|\mathcal{S}|}\sum_{s_l} \mathcal{P}(\mathcal{F}_j = f_i|\mathcal{D} = D_k) \quad (4)$$

To discern between domains, we compare $(\bar{\mathcal{P}}_j(f_i|D_k), \sigma_j(f_i|D_k))$ for each pair of distinct discourse domains. We perform the Student's T-Test for each pair and only report the differences between pairs of domains where the p-value associated with the test is $p \leq 0.03$, i.e. we only report the cases in which the confidence of the difference between domains is higher than 97% [7].

This procedure allowed us to conclude that several of the linguistic features evaluated are distinctive in relation to discourse domains. Below, we present the main differences observed by feature family.

Numerical Features ($\mathbb{F} = \mathbb{N}$)

This feature set consistently demonstrates discernible differences across domains. Specifically, *Juridical* documents exhibit greater average length, employ larger words, and contain a higher number of punctuation marks and stop-words per sentence. Regarding the average value of these features, the *Juridical* domain is followed by *Journalistic* or *Instructional*, *Virtual*, and *Entertainment* texts, which showcase the lowest averages. The distribution of tokens per sentence, illustrated in Figure 2, demonstrates these patterns.

The recurring pattern observed across various domains, where characteristics consistently exhibit a certain order, suggests a hierarchical structure among these domains. One possible way to explain this behavior is in terms of "language complexity": *Juridical* texts use more intricate language, resulting in longer words and sentences. Conversely, *Entertainment* texts tend to employ simpler constructs, resulting in smaller numerical features values.

Morphological Features

Tense

In documents within the *Virtual* and *Entertainment* domains, the use of the present tense is more prevalent, in texts within the *Instructional* and *Journalistic* domains the past tense is the most used and within the *Juridical* domain the future tense is dominant.

The domains previously associated with less linguistic complexity predominantly use the present tense. This observation suggests a correlation: domains with simpler sentence structures typically employ simpler verb tense formations. Specifically, in Carol$\cdot\mathcal{B}$, where the *Virtual* and *Entertainment* domains consist mainly of tweets and subtitles, respectively, the prevalence of the present tense can be rationalized by the nature of these texts, focusing mainly on current events. Conversely, the preeminence of past tense in *Instructional* and *Journalistic* texts aligns with their characteristic reporting of events from the past. Lastly, the usage of the future tense in *Juridical* texts can be attributed to the prescriptive nature of judicial decisions, often dictating conditions and actions to be followed in the future.

Case

In documents within the *Virtual* and *Entertainment* domains, the use of the nominative case is domi-

[7] For analysis convenience, we have displayed here only a representative subset of the pertinent distributions, with complete data and plots accessible through our repositories.

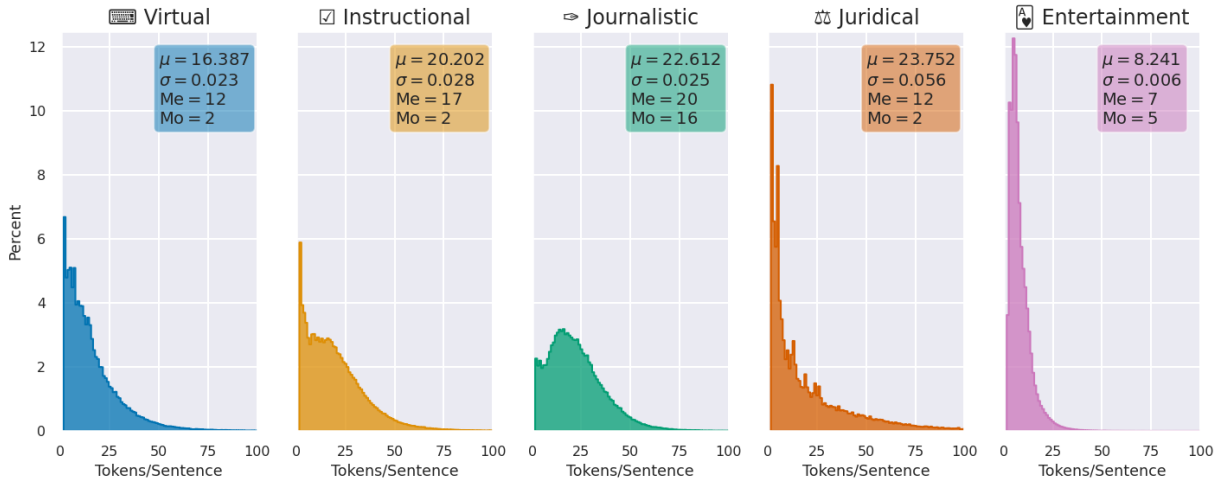

Figure 2: Distribution of sentence lengths across domains.

nant, while *Juridical* and *Instructional* texts use mainly the accusative case. Here, *Journalistic* texts exhibit a relative balance between noun cases. Again, the separation between the domains seems consistent with the ordering observed for the previous features.

Other Morphological Features

Other morphological features, i.e. gender, number, and mood are not visually distinctive between discourse domains. This is likely because word gender is mostly arbitrary with little semantic charge. Similarly, while word number can convey meaning, there is no clear reason to expect that a given domain refers to more plural entities than another.

Part-of-Speech Tags and Named Entity Types

The overall distribution of Part-of-Speech (PoS) tags over different domains is illustrated in Figure 3. For the majority of PoS tags, when we order the discourse domains by the relative importance of the tag, the observed order of the domains is *Juridical, Instructional, Journalistic, Virtual, Entertainment* or the exact opposite. In some cases *Juridical* and *Instructional* are swapped, but only when they are not discernible using the T-test, i.e. even in these cases the described pattern is still statistically compatible with the obtained data.

This ordering is respected by the following PoS tags: *SCONJ*, *VERB*, *PROPN*, *PRON*, *ADV*, *AUX*, *ADJ*, and *ADP*. Ignoring the PoS tags that are very underrepresented in the dataset (*INTJ*, *X*, *PUNCT*, and *SYM*), the only exceptions to this ordering are *DET*, *NOUN* and *NUM*. The order is compatible with the overall scale that was observed in previous features, suggesting that, in fact, the discourse domains within the Carolina Corpus follow some kind of spectrum. However, PoS tags indicate that this may be related not only to language complexity but also to the mode of speech (see Gregory (1967)).

Named Entity Types also exhibit distinct distributions across domains. *Entertainment* texts predominantly mention people and have few references to places and organizations, contrasting with *Juridical* texts. On the other hand, *Journalistic* texts emphasize organizations and show fewer miscellaneous named entities. *Instructional* documents, in comparison, do not notably deviate from other discourse domains with regard to this feature.

This section's analysis shows clear differences between discourse domains, indicating that computational differentiation is possible based on these linguistic features. Additionally, it reveals intriguing patterns within the corpus domains, offering insights into the underlying nature of discourse domains in Portuguese.

7 Discernibility through Embeddings

Word embeddings are vector space representations of lexical meaning, derived from algorithms based on the distributional principle and trained on extensive corpora. They are valuable tools in computational semantics tasks, capturing useful semantic relationships between words, like synonymy, antonymy, and similarity (Jurafsky and Martin, 2009).

Given its representational capacity, it is expected

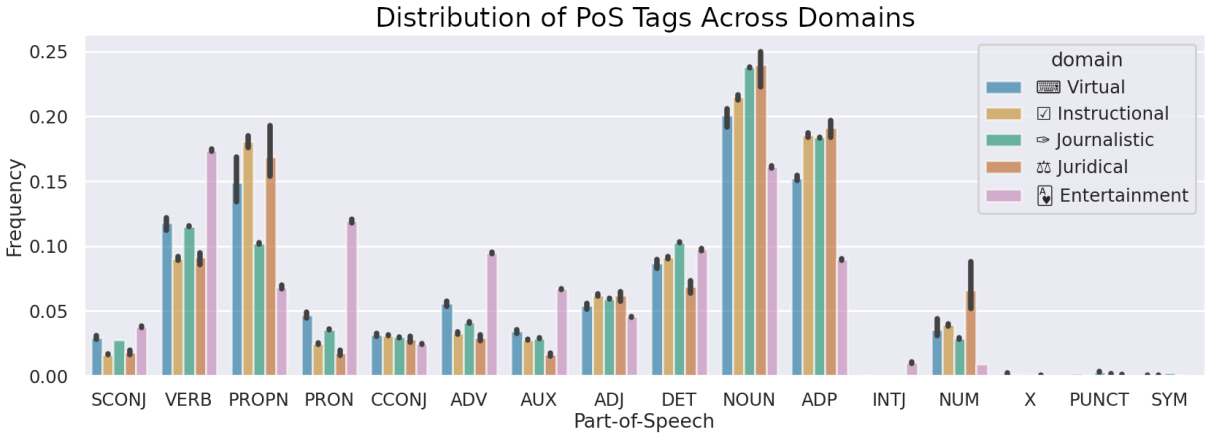

Figure 3: Distribution of PoS tags across domains.

that these spaces can reveal differences between domains, at the semantic level. To explore this, we assess the discernibility of discourse domains using NILC-Embeddings (Hartmann et al., 2017), a celebrated static embedding repository for Portuguese. We explore GLOVE (Pennington et al., 2014), SKIP-GRAM, and CBOW (Mikolov et al., 2013) embeddings with both 50 and 100 dimensions. We compute two metrics: Silhouette scores between discourse domains and the count of out-of-vocabulary (OOV) tokens from each domain.

The Silhouette score is a metric for measuring separability in vector spaces, commonly applied in clustering (Rousseeuw, 1987). We compute the average silhouette[8] between all domains and for each pair of domains, using a random sample of $20,000$ sentences from each domain[9]. Figure 4(a) exhibits the results for CBOW-100.

In all embedding spaces, we observed silhouette scores consistent with that shown in figure 4(a): when calculated between all domains simultaneously, the silhouette takes on a small and sometimes negative value, meaning low separability. Furthermore, the pairs with the lowest and highest silhouette values remain consistent, corresponding to opposite positions on the scale *Juridical*, *Instructional*, *Journalistic*, *Virtual*, *Entertainment*. Meanwhile, adjacent pairs on the same scale occupy the middle of the distribution. This is, surprisingly, the same domain ordering obtained in previous sections.

In summary, while domains collectively lack clear native discernibility in explored embedding spaces, pairwise semantic distinctions exist, aligning with the scale of domains observed in previous analyses.

Additionally, we noted a consistent decrease in average silhouette with higher-dimensional embedding spaces. The CBOW models, at both lengths, were the only ones to exhibit a positive silhouette between the set of all domains, indicating greater domain separability in the CBOW space, compared to others. Hence, this family of embeddings can be more suitable for models of discourse domains classification.

Figure 4(b) illustrates the counts of out-of-vocabulary tokens in the sample sentences for each domain. These counts serve as a metric of how well the semantic field of each domain is represented by these embedding spaces.

We see that domains differ significantly in their count of OOV tokens, suggesting that domain-specific embedding models, leveraging specialized vocabularies, could enhance embedding applicability to domain-specific tasks. Interestingly, domains at opposite ends of the previously observed domain ordering exhibit the most substantial OOV token count differences, e.g. *Entertainment-Juridical*.

Furthermore, the OOV counts for each domain can be roughly explained by the distribution of domains in the corpora used for training the embeddings (See Hartmann et al. (2017)). Entertainment texts have fewer tokens than Journalistic and Instructional texts in the single-genre parts of the training corpora, which can explain its OOV counts. Virtual and Juridical domain OOV counts are less clear, as they do not explicitly appear in the single-genre corpora used for training, but can be contained in the mixed-genre corpora. Further analysis

[8]We use cosine distance as the distance metric.

[9]Sentence embeddings are derived through the mean of the constituent token embeddings.

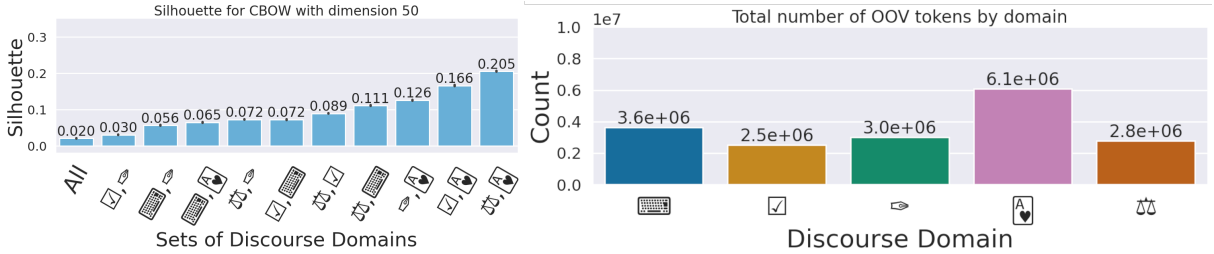

Figure 4: (a) Silhouette with CBOW 50d for each set of discourse domains. (b) Number of OOV tokens by domain.

is required to fully grasp these counts, but they seem to generally align with domain distribution in the training set, as expected.

Generally, Carolina's main discourse domains appear distinguishable in embedding spaces without any transformation, highlighting possible semantic differences between domains. Further examination using clustering and classification algorithms employed over these embedding spaces could provide deeper insights into their underlying capacity to separate discourse domains.

8 Conclusions

In this work, we evaluated the possibility of the computational discernibility of discourse domains, using data from the Carolina corpus. We analyzed discernibility under three distinct approaches: duplication, linguistic features, and embeddings. We now return to the questions presented in Section 1 and try to answer them briefly in light of our results:

1. **Are discourse domains computationally discernible?** Yes. The evaluated domains are highly discernible in our sample. Additionally, most detected differences seem to align with their position in the scale (*Juridical*, *Instructional*, *Journalistic*, *Virtual*, *Entertainment*), what may be linked to language complexity or discourse mode, requiring further investigation.

2. **If so, which properties differentiate them?** Properties such as degree of duplication, sentence and word length, part-of-speech tags, and verbal tense are distinctive. Furthermore, many domains are relatively pairwise distinguishable in semantic embeddings spaces.

3. **What approaches for discerning domains in NLP tasks are experimentally supported?** Given the observed differences, models of deduplication, part-of-speech tagging, tokenization and segmentation, named entity recognition, and embedding generation are some of which could benefit from distinctions between discourse domains.

Several further research directions are possible. We highlight: (i) the development of domain-specialized NLP models, (ii) a more in-depth exploration of inter-domain text deduplication, (iii) an in-depth study of the relation between textual complexity and linguistic differences observed between domains, and (iv) the training of discourse domain classification and clustering models.

In addition to our analysis and source code, our main contributions include producing and providing balanced and deduplicated versions of the Carolina Corpus, as well as the methodology created and adopted in this paper, which provides metrics that computationally discern the discourse domains and can be used to differentiate a diverse set of language varieties in large corpora.

Acknowledgements

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the University of São Paulo, the São Paulo Research Foundation (FAPESP) (grant #2019/07665-4) and by the IBM Corporation. Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants #2015/21880-4, #2014/12236-1); and the National Council for Scientific and Technological Development (CNPq) (grant PQ 303609/2018-4). Felipe Ribas Serras, Mariana Lourenço Sturzeneker and Maria Clara Ramos Morales Crespo were supported by FUSP (Support Foundation for the University of São Paulo) (Project 3541).This work was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

John Cunnison Catford. 1965. *A linguistic theory of translation*, volume 31. Oxford University Press London.

Maria Clara Ramos Morales Crespo, Maria Lina de Souza Jeannine Rocha, Mariana Lourenço Sturzeneker, Felipe Ribas Serras, Guilherme Lamartine de Mello, Aline Silva Costa, Mayara Feliciano Palma, Renata Morais Mesquita, Raquel de Paula Guets, Mariana Marques da Silva, et al. 2023. Carolina: a general corpus of contemporary brazilian portuguese with provenance, typology and versioning information. *arXiv preprint arXiv:2303.16098*.

Luis Otávio de Colla Furquim and Vera Lúcia Strube de Lima. 2012. Clustering and categorization of brazilian portuguese legal documents. In *International Conference on Computational Processing of the Portuguese Language*, pages 272–283. Springer.

Dan Douglas. 2004. Discourse domains: The cognitive context of speaking. *Studying speaking to inform second language learning*, 8:25–47.

Katharina Ehret and Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in sla production data. *Second Language Research*, 35(1):23–45.

E Fonseca, L Santos, Marcelo Criscuolo, and S Aluisio. 2016. Assin: Avaliacao de similaridade semantica e inferencia textual. In *Computational Processing of the Portuguese Language-12th International Conference, Tomar, Portugal*, pages 13–15.

Teresa Gonçalves and Paulo Quaresma. 2005. Is linguistic information relevant for the classification of legal texts? In *Proceedings of the 10th international conference on Artificial intelligence and law*, pages 168–176.

Michael Gregory. 1967. Aspects of varieties differentiation. *Journal of linguistics*, 3(2):177–198.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Rodrigues, and Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In *Anais do XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 122–131, Porto Alegre, RS, Brasil. SBC.

David E. Johnson, Frank J. Oles, Tong Zhang, and Thilo Goetz. 2002. A decision-tree-based symbolic rule induction system for text categorization. *IBM Systems Journal*, 41(3):428–437.

Patrick Juola. 2008. Assessing linguistic complexity. *Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam, Netherlands*.

D. Jurafsky and J.H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall series in artificial intelligence. Pearson Prentice Hall.

Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. 1997. Automatic detection of text genre. *arXiv preprint cmp-lg/9707002*.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. *Language Resources and Evaluation*, pages 1–38.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ramon Pires, Hugo Abonizio, Thales Rogério, and Rodrigo Nogueira. 2023. Sabiá: Portuguese large language models. *arXiv preprint arXiv:2304.07880*.

Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia,

and Renato Vicente. 2021. Legalnlp–natural language processing methods for the brazilian legal language. *arXiv preprint arXiv:2110.15709*.

Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Phd thesis, Masaryk University, Faculty of Informatics.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt. *arXiv preprint arXiv:2305.06721*.

Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Felipe R Serras and Marcelo Finger. 2022. verbert: automating brazilian case law document multilabel categorization using bert. *arXiv preprint arXiv:2203.06224*.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Mariana Sturzeneker, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte, and Cristiane Namiuti. 2022. Carolina's methodology: building a large corpus with provenance and typology information. In *DHandNLP@ PROPOR*, pages 53–58.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache.

Benedikt Szmrecsanyi. 2016. An informationtheoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, 57:71.

Charles Felipe Oliveira Viegas. 2022. Jurisbert: Transformer-based model for embedding legal texts.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Xujuan Zhou, Xiaohui Tao, Jianming Yong, and Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. In *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 557–562.