Compression-based Language Complexity under Register Variation in Portuguese

Felipe Ribas Serras and Marcelo Finger
{frserras,mfinger}@ime.usp.br
Institute of Mathematics, Statistics and Computer Science
University of São Paulo, São Paulo, Brazil

Abstract

Compression-based language complexity metrics show promise as holistic parameters for measuring linguistic complexity across intra- and cross-linguistic scenarios. Yet, their sensitivity to specific forms of linguistic variation requires further experimental validation. We examine the sensitivity of this metric family to register variation in Portuguese, a phenomenon already established for English. We refine the validation process found in previous literature by introducing a more granular statistical analysis to evaluate both the individual and joint sensitivity of these metrics to register variation at the sentence level. Our results confirm they are highly sensitive to functional variation in Portuguese, exhibiting the same structural morphosyntactic trade-off consistent with that observed in English and in cross-linguistic studies.

1 Introduction

Linguistic complexity is a fundamental concept for both linguistic analysis and the development of language technologies. In linguistics, complexity appears in sociolinguistics as a factor distinguishing manifestations of the same language across different situations (Biber and Conrad, 2009). It also features in linguistic typology as a fundamental variable of hypotheses such as the equicomplexity hypothesis — that all languages are equally complex — and the morphosyntactic trade-off hypothesis — that the greater the morphological complexity of a language, the lower its syntactic complexity, and vice versa (Hockett, 1958; Bentz et al., 2023).

In Natural Language Processing (NLP), the concept of complexity intersects with readability and its cognitive counterpart, comprehension, in the development of tools for democratizing access to written language through algorithms for textual simplification and elaboration (Leal and Aluísio, 2024).

The applicability of complexity across intra-linguistic and inter-linguistic contexts has increasingly revealed its potential as a cross-cutting linguistic parameter. This provides a basis for improving our understanding of language adaptation mechanisms across different scopes, potentially shedding light on the universals governing language (Szmrecsanyi, 2009; Biber, 1995).

Existing tools available for measuring complexity — such as Coh-Metrix (McNamara et al., 2014) for English and NILC-Metrix (Leal et al., 2024) for Portuguese — rely on a multivariate approach, allowing the computation of several distinct yet complementary complexity features. These tools have been instrumental to complexity analysis and the development of complexity-oriented technologies within their target languages; however, despite their wide scope and interpretability, many of its metrics depend on pre-existing structures, such as syntax trees and morphological segmentations, thus requiring specific computational resources. This complicates the use of complexity as a general parameter, particularly in cross-linguistic scenarios: many languages lack these resources, and where they exist, implementation differences hinder comparability.

In this context, compression-based metrics (also called Kolmogorov complexity metrics), developed in Juola (2008); Ehret and Szmrecsanyi (2016), offer a universal alternative. They are text-based and function consistently across languages without relying on tools like parsers. Grounded in information theory, these metrics define complexity as the information content of a text, estimated via bit-level compression using off-the-shelf compression algorithms like `gzip`. To isolate the complexity of specific levels, such as syntax or morphology, they apply random distortions primarily affecting the target level, and measure the impact on the resulting compressed file size.

Experimental results have demonstrated the sen-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 808–818

sitivity of compression-based metrics to different forms of linguistic variation, including variation across different languages and families (Juola, 2008; Serras et al., 2024a) as well as within the same language (Ehret, 2021). These findings often align with previous theoretical assumptions, allowing for new perspectives on classic hypotheses. A challenge, however, arises due to the linguistic generality of these assertions: Ehret (2021) shows that these metrics are sensitive to variation between registers (established situational varieties within the same language) for various registers of the English language, even maintaining in the intra-linguistic scenario a trade-off across registers between morphological and syntactic complexities, as proposed by Hockett (1958) for the cross-linguistic scenario. However, the question remains: does this variation hold true for other languages, such as Portuguese? If so, to what extent does it occur and how does its behavior differ from English?

Extending these results and answering these questions are essential for understanding the universality of these phenomena and how the complexity parameters involved can be used to inform holistic theories about the functioning of human language. Accordingly, the aim of this work is to extend these results to the Portuguese language, specifically answering the following questions: (i) Do compression-based metrics vary by register in Portuguese? (ii) Does this variation maintain a trade-off between syntactic and morphological complexity, as observed for English? (iii) How do these variation patterns resemble or differ from those in English?

To answer these questions, we focus on the sentence level, since experimental results indicate that (i) the variation in linguistic behavior at the sentence level is significant between registers of Portuguese (Serras et al., 2024b, 2026), (ii) this variation occurs mainly in terms of informational density (Biber, 1988; Serras et al., 2026), and (iii) at least at the cross-linguistic level, the dynamics between morphological and syntactic complexity occur within scopes similar to the sentence, with inter-sentential complexity remaining constant among different linguistic manifestations (Juola, 2008).

We seek not only to build upon previous results in other languages, helping to establish a more robust foundation for the use of compression-based complexity metrics as a general linguistic parameter, but also to offer a fresh perspective by extending these results and evaluating these phenomena at the sentence level. Furthermore, we aim to provide greater confidence to the Portuguese NLP community regarding the application of compression-based metrics, promoting its use in tasks addressing linguistic variation and complexity, such as simplification and elaboration, readability assessment, and the detection and classification of Portuguese situational varieties.

Although we initially expected the studied complexity metrics to function in Portuguese similarly to English, this cannot be guaranteed without experimental verification. Results suggest that some languages, such as Sino-Tibetan languages (Shcherbakova et al., 2023), may not exhibit the same morphosyntactic trade-off behavior observed for many languages. Furthermore, register variation interacts intricately with the unique cultural environment of each language and may accompany the age and hegemony of literate institutions (Biber, 1995). Thus, we put the hypothesis of similar behavior in Portuguese to the test, establishing a methodology distinct from prior works, focused on the sentence level, clearly differentiating the joint and individual capabilities of each metric, while trying to maintain maximum statistical rigor.

The remainder of this paper is organized as follows: Section 2 details the methodology, including the complexity metrics, dataset, experimental setup, and statistical procedures; Section 3 presents our results and analysis, while an in-depth discussion is provided in Section 4; Section 5 addresses limitations and future steps; and Section 6 offers our concluding remarks.

2 Methodology

In this section we describe the compression-based complexity metrics under study (Section 2.1), the Portuguese language datasets used (Section 2.2), the experimental setup (Section 2.3), and the statistical procedures applied to analyze the results (Section 2.4).

2.1 Compression-Based Metrics

Compression-based linguistic complexity metrics measure structural linguistic complexity. These metrics quantify informational redundancy and irregularity within linguistic representations across various levels of linguistic analysis (Ehret, 2021).

The assumption behind these metrics is that the complexity of a sequence of symbols can be equated with the amount of information it contains,

or with the size of the smallest representation required to reconstruct it. Highly redundant and regular sequences permit greater lossless compression, reflecting inherently lower information density and complexity (Juola, 2008).

The theoretical limit of this compression is known as Kolmogorov complexity. Compression-based metrics approximate this limit using data compression algorithms. These algorithms reduce a sequence of symbols S to a compact version $K(S)$. In practice, compression is not optimal, but it efficiently approximates the minimum representation, with its size $|K(S)|$ serving as an estimate of the true complexity (Grünwald and Vitányi, 2008).

Within this framework, the overall linguistic complexity of a text T is estimated via the size of its compressed representation, $|K(T)|$. To account for the inherent dependence of the compressed size on the original text length $|T|$, an adjusted complexity score proposed by (Ehret and Szmrecsanyi, 2016) is usually adopted. This involves fitting a linear regression between the length of the samples $|T|$ and their corresponding compressed sizes $|K(T)|$. The residuals of this model are then utilized as the measure of overall complexity, as in Equation 1.

$$C_o(T) = Res(|T|, |K(T)|) \tag{1}$$

To assess complexity at a specific linguistic level l, such as syntax s or morphology m, it is necessary to estimate the portion of information transmitted through that level. To do so, a random artificial distortion is applied to the text to specifically affect the redundancy and regularity patterns of the target level ($T \rightarrow T^l$); the size of the compressed distorted text $|K(T^l)|$ is then compared to the size of the original compressed text $|K(T)|$. Syntactic and morphological complexities[1] are subsequently calculated as in Equations 2 and 3, respectively. The way these metrics are computed reflect the way each of these levels is expected to react to the distortion process. A more in-depth discussion can be found in Ehret and Szmrecsanyi (2016).

$$C_s(T) = \frac{|K(T^s)|}{|K(T)|} \tag{2}$$

[1]We acknowledge the debate regarding the suitability of "Syntactic" and "Morphological Complexity" as metric names. Serras et al. (2024a) proposes "Extra-Word" and "Intra-Word Complexity" as more appropriate descriptors. While we recognize the validity of these reservations, we retain the standard nomenclature to ensure consistency with previous literature.

$$C_m(T) = -\frac{|K(T^m)|}{|K(T)|} \tag{3}$$

Deletion is the most commonly used distortion process, in which 10% of the sample is randomly deleted. When syntax is the target level, 10% of the words in the sample are deleted, while for morphology, 10% of the characters within the words are deleted.

The syntactic complexity metric C_s is a a measure of the number of word order rules and the resulting rigidity of word arrangement within the sample. For morphology, C_m estimates the diversity of word forms present in a sample, encompassing both lexical diversity and the variety of inflectional and derivational mechanisms.

We adopted Serras et al. (2024a)'s implementation of compression-based complexity metrics. Beyond the metrics explored so far, this package includes an alternative approach to morphological complexity C^*_m proposed by Juola (2008). Instead of applying deletion-based distortion, this variant replaces words with unique numerical indices, maintaining relative word order while eliminating internal structure($T \rightarrow T^{m*}$). It also employs a distinct combination method of resulting file sizes (Equation 4). As the implementation supports both `gzip` and `bzip2` compressors, we chose to experiment with both.

$$C^*_m(T) = \frac{|K(T)|}{|K(T^{m*})|} \tag{4}$$

2.2 Data and Text Types

Various forms of linguistic variation result in differences in complexity. In this study, following Ehret (2021), we focus on structural complexity differences stemming from intra-linguistic variation at the register level.

Registers are varieties of a language characterized by pervasive differences in linguistic features, which reflect a functional response to the specific situational context of production. Registers constitute a wide-ranging category of variation encompassing diverse levels of generality, where more general registers comprise more specific ones (Biber and Conrad, 2009).

To focus on sentence-level distinctions, this study utilizes the `carol-domain-sents` dataset (Serras et al., 2026)[2], comprising sentences ex-

[2]https://huggingface.co/datasets/carolina-c4ai/carol-domain-sents

tracted from a deduplicated version of the Carolina Corpus (Finger et al., 2025), a general corpus of contemporary Brazilian Portuguese. Each sentence is annotated with several categories, including `source typology` — broadly corresponding to the concept of register defined above. Given its original design for sentence classification, the dataset encompasses three splits, which were merged for this experiment.

We constructed a balanced subset by sampling 3,000 sentences from each of the 20 registers with the highest sentence counts in `carol-domain-sents`, including "ACTIVITIES / EXPERIENCE SHARING", "ARTICLE", "EDUCATIONAL RESOURCES", "JURIDICAL REPORT", "NEWS", "OPEN COURT HEARING", "PRECEDENTS BULLETIN", "SCIENTIFIC NEWS", "SUBTITLE", "TRAVEL GUIDE", "TWEET", "USER PAGE", "VIRTUAL DISCUSSION", and "VOCABULARY ENTRY". Following an exploratory analysis, we excluded the registers "PROPOSAL OF BINDING PRECEDENT", "APPELLATE DECISION RECORDS", "JURIDICAL TOPIC PUBLICATION", "STUDY OF PRECEDENTS", "JURIDICAL SPEECH" and "REQUEST FOR PROPOSALS" from the "Juridical" domain, due to frequent sentence segmentation and PDF extraction errors, which could compromise the reliability of the computed complexity values.

2.3 Experimental Setup

Our experiments consisted of computing compression-based complexity metrics for various samples of each register, using the values obtained to investigate whether, as in English, these metrics are also sensitive to linguistic variation at the register level in Portuguese.

Since we focus on the sentence level, it would be natural to use each individual sentence as a sample for calculating complexity; however, compression algorithms require a minimum volume of data to ensure pattern recognition and achieve compression that approximates Kolmogorov complexity. For this reason, we adopted the following alternative setup: for each register, we constructed pseudo-documents formed by uniformly sampling and concatenating sentences from that register. Following Ehret (2016)'s recommendation, we fixed the size of the pseudo-documents at a constant value n — since the resulting texts do not share the same propositional content — setting $n = 250$ sentences and generating $m = 250$ pseudo-documents per register.

The rationale for this setup is that, by randomly sampling sentences from different documents and combining them in random order, we disrupt any inter-sentential patterns. Thus, the complexity measures reflect a combination of the complexities of the individual sentences and the relationships between sentences that share the same register.

A similar sampling procedure is performed in Ehret (2016); however, to our knowledge, it preserves the relative order and single-document origin of sampled sentences, in contrast to our approach.

For each register, the pseudo-documents and their respective complexity values, computed as described, are used to examine the distinctiveness of the registers based on the metrics. In the next Subsection, we describe the statistical procedures applied to these data to answer our research questions.

2.4 Statistical Procedures

As mentioned early, in this work we introduce a more granular statistical analysis to evaluate both the individual and joint sensitivity of compression-based linguistic complexity metrics to register variation in Portuguese. We assessed several metrics: morphological and syntactic complexities via deletion (C_m and C_s), and the variant of Juola (2008) for morphological complexity via substitution C_m^* (See Section 2.1). These include two versions for each metric: one using `gzip` as the underlying compression algorithm and another using `bzip2`.

We divided our statistical analysis into two stages: (i) **univariate analysis**, in which we evaluated the ability of each complexity metric to distinguish between registers, and (ii) **multivariate analysis**, in which we evaluated the combined power of these metrics to explain linguistic variation across registers.

In the univariate stage, we conducted an Analysis of Variance (ANOVA) (Huberty and Olejnik, 2006), a statistical method widely used to investigate whether a numerical variable, the complexity metric, differs significantly across groups defined by a categorical variable, the registers.

The test is based on the computation of the F-statistic and its associated p-value. The F-statistic measures the ratio of between-group variability (registers) MS_B to within-group variability MS_W (Equation 5). High F-values indicate that the means differ across registers and that register in-

fluences the mean of the complexity metric under evaluation.

$$F = \frac{MS_B}{MS_W} \tag{5}$$

In addition to F, we computed the statistic η^2 (Huberty and Olejnik, 2006), which quantifies the proportion of total variance in the complexity metric SS_T explained by register differences SS_B (Equation 6). Values of η^2 range from 0, indicating that register explains none of the variation, to 1, indicating that register explains all variation, and as a rule of thumb $\eta^2 > 0.14$ is considered a large effect size (Cohen, 2013). Since F is sensitive to sample size while η^2 is not, we employed the latter to assess practical significance independent of sample size.

$$\eta^2 = \frac{SS_B}{SS_T} \tag{6}$$

In the second stage, we conducted a Multivariate Analysis of Variance (MANOVA) (Huberty and Olejnik, 2006), to analyze the combined discriminatory power of the evaluated complexity metrics. Given the multivariate nature of the analysis, we utilized summary statistics; unlike the univariate case, several summary statistics are available. These statistics are derived from the eigenvalues λ_i of the matrix product $\mathbf{E}^{-1}\mathbf{H}$, where $\mathbf{H}$ denotes the hypothesis matrix and $\mathbf{E}$ the error matrix. We computed the following summary statistics:

- **Wilks' Lambda (Λ):** Represents the ratio of error variance to total variance, defined as $\Lambda = \prod_{v=1}^{r}(1+\lambda_v)^{-1}$. Values closer to zero indicate stronger group differences.
- **Pillai's Trace (U):** Represents the sum of explained variance proportions: $U = \sum_{v=1}^{r} \frac{\lambda_v}{1+\lambda_v}$. It's considered the most robust statistic to assumption violations.
- **Hotelling-Lawley Trace (T):** The sum of the ratio of explained variance to error variance for each dimension, calculated as $T = \sum_{v=1}^{r} \lambda_v$.
- **Roy's Greatest Root (θ):** It is defined by the largest eigenvalue, $\theta = \lambda_{\max}$, representing the maximum possible difference between groups.

To conclude the multivariate analysis, we performed a Linear Discriminant Analysis (LDA) as a follow-up to interpret the MANOVA results, identifying the dimensions that best separate the registers and determining the contribution of each complexity metric to these distinctions (Huberty and Olejnik, 2006). Prior to these analyses, we removed redundant complexity metrics (correlation $\rho > 0.8$) based on a correlation analysis, which also allowed us to examine morphosyntactic trade-offs.

3 Results

In line with our methodology, we divided the results into two stages: **univariate analysis** (Section 3.1), and **multivariate analysis** (Section 3.2).

3.1 Univariate Analysis

Before applying ANOVA to the experimental results described in Section 2, we verified the assumptions for each complexity metric evaluated. ANOVA has three basic assumptions: **independence of observations**, **normality of residuals**, and **homogeneity of variance** (Huberty and Olejnik, 2006).

Independence of observations Each complexity value was obtained from a different pseudo-document, composed of distinct sentences in distinct orders. Although, by construction, there is a small overlap between the sentences comprising the pseudo-documents of the same register — approximately 8% between any two pseudo-documents — we consider this overlap insufficient to violate independence, since complexity is computed over the entire pseudo-document.

Normality of Residuals We visually inspected QQ Plots (Quantile-Quantile Plots) of the residuals for all 6 complexity metrics. These plots compare the empirical distribution of residuals to a theoretical normal distribution. For a normal distribution, the points align closely with the reference diagonal. We observed, for all cases, points distributed around the diagonal without significant systematic deviations, indicating that the residual distributions are approximately normal.

Homogeneity of Variance We used box plots to examine the homogeneity of variance. For all metrics, we observed distributions centered on 0 on the Y-axis, with similar spreads between registers. We also computed the maximum ratio between variances s^2_{max}/s^2_{min}. Based on the rule of thumb of $s^2_{max}/s^2_{min} \leq 3$, deletion metrics satisfy the assumption. Replacement-based metrics do not;

however, the test remains robust in these cases due to the balanced design. We consider this sufficient to proceed, noting that results for deletion metrics are likely more reliable (Dean et al., 2017).

All QQ plots and box plots examined are available in our repository[3]. Based on the evidence, we conclude that the assumptions are satisfied and that ANOVA is applicable.

Table 1 presents the test results, including the F-statistic and the η^2 statistic for each complexity metric. In all cases, we obtained high F-values with statistically significant p-values ($p < 0.01$), confirming significant variation in complexity across registers for all the complexity metrics evaluated.

Metric	F-Statistic	P-Value	η^2
C_m (bzip2)	5988.982	< 0.01	0.957
C_m (gzip)	5957.920	< 0.01	0.957
C_m^* (bzip2)	3756.886	< 0.01	0.933
C_m^* (gzip)	1578.462	< 0.01	0.855
C_s (bzip2)	78.399	< 0.01	0.226
C_s (gzip)	67.924	< 0.01	0.202

Table 1: ANOVA Results

The η^2 values are also substantial, indicating that discriminatory power is not an artifact of sample size and showing that the grouping into registers substantially explains the observed variance in complexity metrics.

Notably, for both statistics, the discriminatory power of morphological complexity metrics is systematically higher than that of syntactic complexity metrics. This will be discussed in Section 4.

3.2 Multivariate Analysis

3.2.1 Assumptions

As in the previous section, we assessed the MANOVA assumptions before performing the analysis. In addition to the assumption of **data independence**, which follows that of ANOVA, MANOVA assumes the **multivariate normality of metrics** and the **homogeneity of covariance matrices**. Furthermore, it is recommended to **avoid multicollinearity** among the metrics.

Multivariate normality We evaluated this property by comparing the experimental points against the theoretical quantiles of the χ^2 distribution. For an approximately multivariate normal distribution, we expect the experimental points to align with the main diagonal with minimal fluctuation. We observed this expected behavior for each register, with less than 4% of the points in each group identified as outliers. We also repeated the tests with automatic outlier removal, yielding results equivalent to those reported. Based on this, we consider the distribution to be sufficiently normal.

Homogeneity of Covariance Matrices We adopted Huberty and Olejnik (2006)'s suggestion to use the difference between the log-determinants of the covariance matrices as a practical heuristic for estimating homogeneity. Small differences suggest that homogeneity is acceptable. After removing highly correlated metrics, we observed a maximum difference of 2.4. Although Huberty and Olejnik (2006) explicitly states that there is no standard reference value, we consider this result acceptable, as it indicates that the determinants across groups are of similar order of magnitude. Regardless, it is worth noting that because our groups are balanced, MANOVA is robust to violations of this assumption.

However, unlike the univariate case, which relies solely on the F-statistic, the multivariate case involves several summary statistics (Section 2.4) with varying levels of robustness to assumption violations. Of all the statistics used, Pillai's trace (U) is the most robust to violations (Bobbitt, 2021), making it the most reliable statistic in this regard. Similar to the univariate case, the diagnostic plots regarding the assumptions can be found in our repository.

3.2.2 Correlation and Trade-off

The final MANOVA assumption verified was the absence of multicollinearity, i.e., that there are no redundant pairs within the set of quantitative variables used in the analysis. The standard approach to address this is to compute the correlation between the complexity metrics and remove those with high absolute correlation.

This procedure is particularly relevant because one of the main patterns observed in compression-based complexity metrics in English is the morphosyntactic trade-off: an inverse relationship between syntactic and morphological complexity across registers. This phenomenon, posited by Hockett (1958) for the cross-linguistic scenario, was confirmed by Ehret (2021) for the intra-linguistic context within English.

[3] https://github.com/frserras/register-complexity-pt

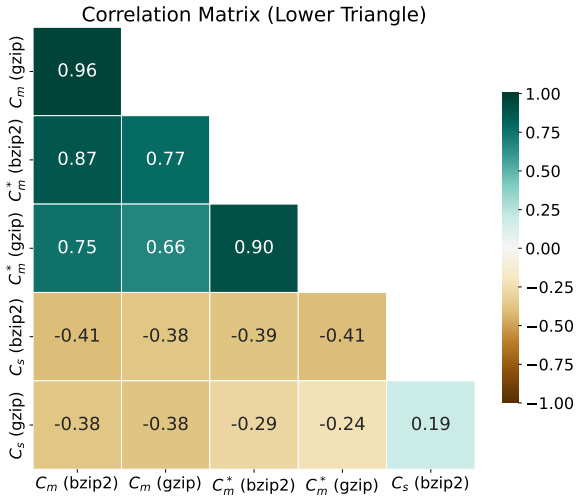

Figure 1: Pearson Correlation between complexity metrics.

Thus, computing correlations enables us to answer three questions: (i) which metrics should be removed to satisfy the assumption of non-multicollinearity in MANOVA; (ii) whether there is agreement among the metrics of morphological complexity and among those of syntactic complexity; and (iii) whether we can observe, in Portuguese as in English, the morphosyntactic trade-off at the register level.

Figure 1 presents the Pearson correlations between complexity metrics. We observed positive correlations among metrics within the same linguistic level — morphology and syntax — and negative correlations between morphological and syntactic metrics. These relationships are illustrated in Figure 2, which shows the morphosyntactic trade-off across Portuguese registers.

To satisfy the assumption of non-multicollinearity, we removed C_m^* and C_m (gzip), eliminating pairs with a correlation $\rho > 0.8$. This concludes the verification of assumptions, allowing us to perform MANOVA on this reduced set of metrics.

3.2.3 MANOVA Results

We applied MANOVA to the experimental results, yielding the summary statistics shown in Table 2. All statistics indicate significant differentiation among registers. Notably, Pillai's trace (U), the most robust statistic, confirms this discriminatory power.

These results demonstrate that, consistent with the univariate case, the variation among registers is also reflected in the multivariate set of complexity metrics. Consequently, there is a significant systematic difference across the evaluated Portuguese registers regarding morphosyntactic complexity.

Metric	Value	P-Value
Wilks' lambda (Λ)	0.035	< 0.01
Pillai's trace (U)	1.128	< 0.01
Hotelling-Lawley trace (T)	22.855	< 0.01
Roy's greatest root (θ)	22.667	< 0.01

Table 2: Manova Results

Following the significant MANOVA results, we performed Linear Discriminant Analysis (LDA) (Section 2.4). LDA identifies a linear transformation of the complexity metrics that projects data into the space that best separates the registers. While MANOVA is explanatory in nature, LDA is predictive, allowing us to assess how accurately we can classify registers based on compression-based morphosyntactic complexity metrics.

Table 3 presents the linear transformation obtained through LDA. We observe a clear separation between the optimal dimensions: they load primarily on either morphological or syntactic metrics. Notably, dimensions favoring syntactic metrics (LD2 and LD3) do so with much lower magnitude than those focused on morphological metrics. This not only corroborates previous results regarding (i) the morphosyntactic trade-off and (ii) the greater discriminatory power of morphological complexity — which is also corroborated by the high proportion of explained variance captured by the primary dimension, LD1 — but also demonstrates that these phenomena are crucial for distinguishing Portuguese registers.

Metric	LD1	LD2	LD3
C_m (bzip2)	4.908	-0.592	-0.154
C_s (bzip2)	0.141	-1.000	0.572
C_s (gzip)	-0.007	-0.544	-0.981

Table 3: Linear Discriminant coefficients. Explained Variance per Component: LD1: 99.2%, LD2: 0.5%, LD3: 0.3%.

4 Discussion

Our results demonstrate that, as in English, compression-based complexity metrics are sensitive to register variation in Portuguese, both individually and collectively, and that situational linguistic variation leads to systematic differences in morphosyntactic complexity.

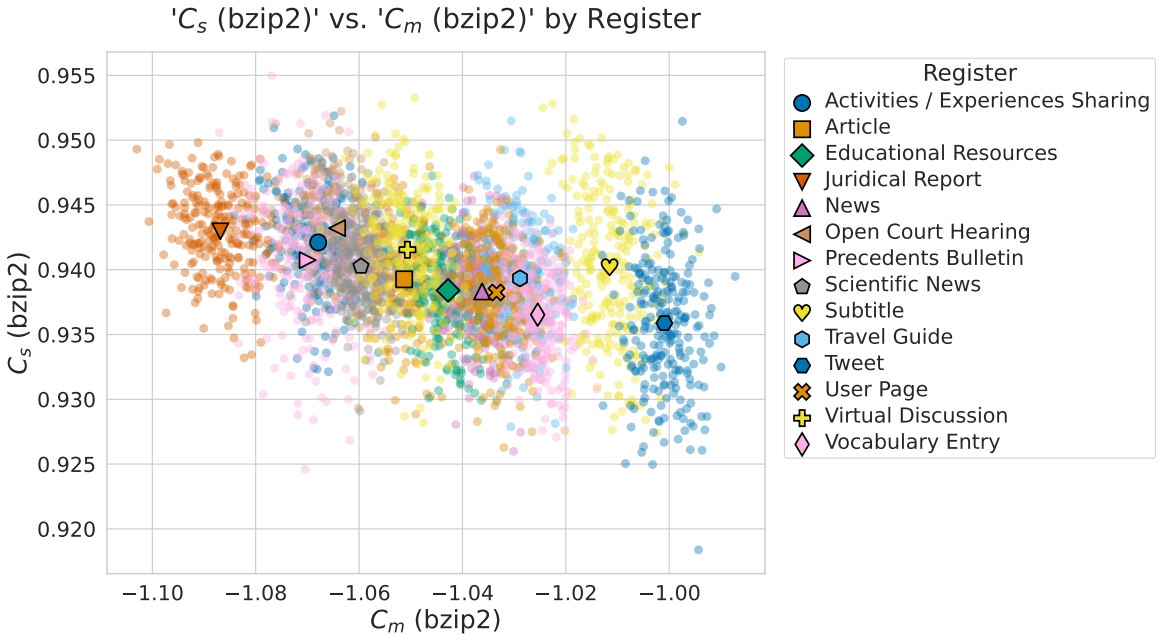

Figure 2: Observed Trade-off between C_m and C_s for `bzip2`

Furthermore, the morphosyntactic trade-off pattern observed in English also holds true for Portuguese registers. This suggests that, as hypothesized in Ehret (2021), intra-linguistic variation functions to maintain morphosyntactic equilibrium, pointing to a widespread phenomenon across languages, as evidenced here in Portuguese.

We observed that morphological complexity is significantly more discriminative than syntactic complexity for Portuguese. Since the results for English lack equivalent variance tests and our dataset comprises fewer registers, it is difficult to determine for now whether this is a specific property of Portuguese or a general phenomenon.

Given that Portuguese has high morphological expressiveness compared to English, it is possible that this is indeed a feature of the language. This would imply that, while the trade-off is universal, it aligns with the morphosyntactic profile of each language, utilizing its specific expressive mechanisms to differentiate situational variants.

Alternatively, if the greater predictive power of morphological complexity is a general property and not language-specific, this suggests that register variation mainly relies on changes in complexity at the lexical and intra-lexical levels to distinguish different functional varieties. This is logical, given that while the grammar of a language is a closed system, its lexicon is an open system, allowing for greater creativity and offering greater flexibility for functional linguistic variation. However, our current results are insufficient to validate one of these explanations and rule out the other.

In general, our results demonstrate that compression-based complexity metrics are valuable tools for the parameterization of linguistic variation in Portuguese, supporting their reliable application in future research. Furthermore, the high discriminatory power observed underscores the applicability of these metrics for technologies addressing intra-linguistic variation and complexity in Portuguese. These include complexity measurement, text simplification or elaboration, and register classification tasks, where the compression metrics can serve as either primary or supplementary features.

This is not to say that the complexity metrics studied here can function independently and without appropriate treatment as decision criteria for classifying registers or as direct measures of readability. The works discussed throughout this paper unambiguously characterize both the functional variation of language and the difficulty humans have in processing it as complex and multidimensional phenomena. Here, we establish the theoretical and experimental basis necessary for these metrics to be reliably integrated into more complex models aimed at both applications.

5 Future Steps

Despite the positive results achieved, our study has limitations. Among them, we highlight: (i) the predominance of written data over originally spoken registers, especially because the distinction between spoken and written language is known to be crucial in terms of complexity (Biber, 1988; Ehret, 2021); (ii) a smaller number of registers compared to experiments conducted in English; (iii) the limitation of the tests used, which are only capable of detecting general differences among registers rather than specific pairwise differences; (iv) the lack of explainability of the studied metrics, which cannot provide direct association with linguistic phenomena, (v) the limitation of the compression method regarding the critical volume of data.

These limitations present opportunities for future work, whether through expanding the experiments by introducing spoken registers from other sources, conducting post-hoc analyses to identify differences between registers in a more granular way, systematically comparing compression-based metrics with more explainable linguistic complexity features and baselines, evaluating the capability of models built on these metrics in complexity-related NLP tasks and benchmarks, or by experimenting with alternative methods for extending the results to the sentence level. We could also replicate these experiments using statistical tests with fewer assumptions to increase the robustness of our results.

6 Conclusions

In this paper, we explore the sensitivity of compression-based linguistic complexity metrics to intra-linguistic variation in Portuguese. In Section 1, we posed the following research questions, which now we revisit:

Do compression-based metrics vary by register in Portuguese? Yes. Our tests show that the studied complexity metrics vary systematically across registers. However, the sensitivity of morphological complexity metrics was much higher than that of syntactic complexity metrics.

Does this variation exhibit a trade-off between syntactic and morphological complexity, as observed for English? Yes. We observe a substantial trade-off between these metrics, consistent with English.

How do these variation patterns resemble or differ from those in English? We observed a similar trade-off with a linear trend, consistent with English. It is unclear whether the greater discriminative power of morphological complexity over syntactic complexity is a phenomenon specific to Portuguese.

We hope that our results will help build a broader understanding of the studied complexity metrics. adding to pre-existing results and providing experimental foundation for the use of compression-based complexity metrics in Portuguese.

Limitations

As mentioned in Section 5, this study has the following limitations: (i) the predominance of written data over originally spoken registers, (ii) a smaller number of registers compared to experiments conducted in English; (iii) the limitation of the tests used, which are only capable of detecting general differences among registers rather than specific pairwise differences; (iv) the lack of explainability of the studied metrics, which cannot provide direct association with linguistic phenomena, (v) the limitation of the compression method regarding the critical volume of data.

Acknowledgments

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. This work was partly supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8.248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in TIC 13, DOU 01245.010222/2022-44. Marcelo Finger was partly supported by the São Paulo Research Foundation (FAPESP) (grants 2023/00488-5 (SPIRA-BM), 2022/11254-2 (EMU)); and the National Council for Scientific and Technological Development (CNPq) (grant PQ 302963/2022-7). Felipe Ribas Serras was supported by IBM through a scholarship managed by FUSP (Support Foundation for the University of São Paulo) (Project 3541) and by a PPI-SOFTEX scholarship managed by FUSP (Project 3970). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001. The authors used genera-

tive AI tools to assist with writing (paraphrasing and language refinement) and code development. All AI-generated suggestions were verified by the authors.

References

Christian Bentz, Ximena Gutierrez-Vasques, Olga Sozinova, and Tanja Samardžić. 2023. Complexity trade-offs and equi-complexity in natural languages: A meta-analysis. *Linguistics Vanguard*, 9(s1):9–25.

Douglas Biber. 1988. *Variation across Speech and Writing*, 1st edition. Cambridge University Press.

Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*, 1st edition. Cambridge University Press.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*, 1st edition. Cambridge University Press.

Zach Bobbitt. 2021. What is Pillai's Trace? (Definition & Example). `https://www.statology.org/pillais-trace/`. Statology. Accessed: 2026-01-06.

Jacob Cohen. 2013. *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Angela Dean, Daniel Voss, and Danel Draguljić. 2017. *Design and Analysis of Experiments*. Springer Texts in Statistics. Springer International Publishing, Cham.

Katharina Ehret. 2016. *An information-theoretic approach to language complexity: variation in naturalistic corpora*. Ph.D. thesis, Albert-Ludwigs-Universität Freiburg.

Katharina Ehret. 2021. An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory*, 17(2):383–410.

Katharina Ehret and Benedikt Szmrecsanyi. 2016. *An information-theoretic approach to assess linguistic complexity*, pages 71–94. De Gruyter, Berlin, Boston.

Marcelo Finger, Maria Clara Paixão De Sousa, Cristiane Namiuti, Vanessa Martins Do Monte, Aline Silva Costa, Felipe Ribas Serras, Mariana Lourenço Sturzeneker, Miguel De Mello Carpi, Mayara Feliciano Palma, and Gabriela Alves Lachi. 2025. Building Carolina: Metadata for Provenance and Typology in a Corpus of Contemporary Brazilian Portuguese. *Cadernos de Linguística*, 6(4):e812.

Peter D. Grünwald and Paul M.B. Vitányi. 2008. Algorithmic information theory. In Pieter Adriaans and Johan van Benthem, editors, *Philosophy of Information*, Handbook of the Philosophy of Science, pages 281–317. North-Holland, Amsterdam.

C.F. Hockett. 1958. *A Course in Modern Linguistics*. A Course in Modern Linguistics. Macmillan.

Carl J. Huberty and Stephen Olejnik. 2006. *Applied MANOVA and Discriminant Analysis*, 1st edition. Wiley Series in Probability and Statistics. Wiley.

Patrick Juola. 2008. *Assessing linguistic complexity*, pages 89–108. John Benjamins Publishing Company.

Sidney Evaldo Leal and Sandra Maria Aluísio. 2024. Complexidade textual e suas tarefas relacionadas. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 3rd edition, book chapter 25. BPLN.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2024. NILC-Metrix: Assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, 58(1):73–110.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*, 1st edition. Cambridge University Press.

Felipe Serras, Miguel Carpi, Matheus Branco, and Marcelo Finger. 2024a. Analysing and validating language complexity metrics across South American indigenous languages. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 152–165, Bangkok, Thailand. Association for Computational Linguistics.

Felipe Ribas Serras, Miguel de Mello Carpi, Mariana Lourenço Sturzeneker, Mayara Feliciano Palma, Aline Silva Costa, Vanessa Martins do Monte, Cristiane Namiuti, Maria Clara Ramos Morales Crespo, Maria Clara Paixão de Sousa, and Marcelo Finger. 2026. Automatic analysis and classification of discourse domains in brazilian portuguese. *Linguamática*, 17(2):131–171.

Felipe Ribas Serras, Mariana Sturzeneker, Miguel de Mello Carpi, Mayara Feliciano Palma, Maria Clara Ramos Morales Crespo, Aline Silva Costa, Vanessa Martins Do Monte, Cristiane Namiuti, Maria Clara Paixão de Souza, and Marcelo Finger. 2024b. Exploring computational discernibility of discourse domains in Brazilian Portuguese within the Carolina corpus. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 255–265, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Olena Shcherbakova, Volker Gast, Damián E. Blasi, Hedvig Skirgård, Russell D. Gray, and Simon J. Greenhill. 2023. A quantitative global test of the complexity trade-off hypothesis: the case of nominal and verbal grammatical marking. *Linguistics Vanguard*, 9(s1):155–167.

Benedikt Szmrecsanyi. 2009. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of english. *Language Variation and Change*, 21(3):319–353.