Information Theory and Linguistic Variation: A Study of Brazilian and European Portuguese

Diego Alves
Saarland University / Saarbrücken, Germany
diego.alves@uni-saarland.de

Abstract

We present a general analysis of the lexical and grammatical differences between Brazilian and European Portuguese by applying entropy measures, including Kullback-Leibler divergence and word order entropy, across various linguistic levels. Using a parallel corpus of BP and EP sentences translated from English, we quantified these differences and identified characteristic phenomena underlying the divergences between the two varieties. The highest divergence was observed at the lexical level due to word pairs unique to each variety but also related to grammatical distinctions. Furthermore, the analysis of parts-of-speech (POS), dependency relations, and POS tri-grams provided information concerning distinctive grammatical constructions. Finally, the word order entropy analysis revealed that while most of the syntactic features analysed showed similar patterns across BP and EP, specific word order preferences were still apparent.

1 Introduction

Portuguese, a Romance language from the Indo-European family, is the eighth most spoken language in the world according to Eberhard et al. (2024), and the most spoken language in the Southern Hemisphere. It is the official language of eight countries: Angola, Brazil, Cape Verde, Equatorial Guinea, East Timor, Guinea-Bissau, Mozambique, and Sao Tome and Principe. However, it is spoken, as the native language, by more than 99% of the population only in Portugal and Brazil. According to Instituto Camões (2021), in 2021, Portuguese was spoken by around 280 million people.

Due to its population size and increasing economic importance, the Brazilian variety of Portuguese has expanded its influence throughout the twentieth and twenty-first centuries. The impact of this variety can be seen, for instance, in the field of Natural Language Processing (NLP), where many tools and language models have been specifically developed for Brazilian Portuguese (e.g., BERTimbau (Souza et al., 2020) and Albertina 100M PTBR (Santos et al., 2024)). Despite Portugal's smaller population, the European variety of Portuguese has maintained its prestige and significant importance within the Lusophone community, especially in the NLP field (Branco et al., 2023). Unfortunately, other varieties still lack representativeness, particularly in the NLP field, as described by Alves (2024).

Since the colonization period, the Portuguese spoken in Brazil has evolved differently from European Portuguese, influenced by various factors, across multiple linguistic levels, including lexical, grammatical, and phonological. The analysis of these differences have been object of a large variety of linguistic works, and the detection of these varieties is a current topic in the NLP community (e.g., VarDial Shared Task 2023 (Aepli et al., 2023)).

Despite efforts to unify the Portuguese varieties (e.g., the Orthographic Agreement (Pinto, 2012)), the emphasis on differences appears to be the main trend on social media, where users from various regions engage in endless discussions about the most correct ways to express themselves. One example is the amount of discussion generated by the BBC article (BBC, 2024), in which the linguist Fernando Venâncio states that in a few decades, Brazil will be speaking Brazilian, a different language from Portuguese.

While linguistic papers usually focus on specific linguistic differences, often without employing corpus-based analysis, many NLP works tend to concentrate solely on improving tools for specific applications, giving little attention to the analysis of these differences.

Therefore, the aim of this paper is to provide a general overview of the lexical and grammatical differences between Brazilian and European Portuguese, using information theory measures such and parallel corpus. Our objective is to quantify

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 9–19
January 19, 2025.

these differences and, through qualitative analysis, identify the main lexical and grammatical aspects responsible for the observed variations. Moreover, we aim to show how efficient these methods are in the identification of typical lexical and grammatical features of different varieties of the same language.

The remainder of the paper is organized as follows. In Section 2 we discuss related work on Brazilian and European varieties of Portuguese. Sections 3 and 4 presents our methods and results. We conclude with a summary and outlook (Section 5).

2 Related Work

As previously mentioned, purely linguistic works comparing European and Brazilian Portuguese tend to focus on specific linguistic phenomena. For example, the work by Kato and Martins (2016) describes what the authors refer to as a major difference in the grammar of the two varieties, namely the placement of clitic pronouns. Moreover, they also propose an analysis concerning information focus, as well as contrastive and emphatic focus.

The difference between Wh-questions in both varieties was examined diachronically by De Paula (2017), revealing a clear temporal evolution with noTable differences in word order patterns (e.g., WhV versus WhSV).

An interesting study regarding the lexical level was conducted by Silva (2010). The authors compared Brazilian and European Portuguese at the lexical level using uniformity measures developed by Geeraerts et al. (1999). They focused on the lexical fields of clothing and soccer, identifying a divergence only in the clothing category. The authors examined 21 pairs of synonyms to calculate the uniformity measures. In contrast, our approach is broader, as our measures allow us to identify divergent terms without relying on a pre-established list and can also be used to identify typical grammatical patterns in each variety.

Many other studies focused on intonational and phonological aspects (cf. Frota et al. (2015); Escudero et al. (2009); Frota and Vigário (2001)), which are not the focus of our analysis.

The focus of NLP studies on Portuguese varieties is typically on detecting the correct variety, as seen in the 2023 and 2024 VarDial Shared Tasks (Aepli et al., 2023; Chifu et al., 2024). Besides specific shared tasks organized for this purpose, variety detection is also the subject of other studies, such as the system proposed by (Castro et al., 2016), which focuses on tweets from both varieties and achieves an accuracy of 0.93.

Another valuable application of NLP tools for different varieties of Portuguese was presented by (Cortes et al., 2024). The authors focused on the localization task (i.e., adapting linguistic and cultural material between different locales). Using large language models, they achieved considerable success in adapting machine translation to Brazilian and European Portuguese.

Regarding the use of information theory to describe language variation, Degaetano-Ortlieb and Teich (2018) presented a data-driven diachronic analysis of scientific English, detecting periods of linguistic change in terms of lexical and grammatical features. Their approach is based on relative entropy (Kullback-Leibler Divergence), comparing temporally adjacent periods and sliding along the timeline from past to present. In this paper, our aim is to adopt a similar approach; however, instead of conducting a diachronic analysis, we propose to compare different varieties of Portuguese synchronically.

Entropy measures are also relevant for comparing different languages in terms of word order patterns. In typology, Levshina (2019) used entropy in quantitative studies of word order variation, measuring it at different levels of granularity. Additionally, Montemurro and Zanette (2011) applied entropy measures to demonstrate that the impact of word order on language structure is a statistical linguistic universal. However, these typological studies do not address potential changes in word order across different varieties of the same language. Thus, our approach aims to use word order entropy measures to detect syntactic variation between European and Brazilian Portuguese, to assess whether these differences should be considered in typological studies involving Portuguese.

3 Methods

3.1 Data

For our comparative analysis, we utilized the FRMT dataset (Riley et al., 2023), which comprises paired sentences in European and Brazilian Portuguese. The sentences for each variant are translations of original English sentences carried out by translators specializing in the respective Portuguese variants. Notably, the curators of the FRMT dataset intentionally selected English sen-

tences that required distinct, non-optional translations for each Portuguese variant.

In this study, we concatenated all the texts from FRMT repository, omitting the original English sentences, thus creating a parallel corpus of aligned sentences in European and Brazilian Portuguese, totaling 5,478 sentences. The token distribution is presented in Table 1.

Variety	Number of Tokens
European Portuguese	138,355
Brazilian Portuguese	135,873

Table 1: Distribution of tokens in the FRMT dataset regarding European and Brazilian varieties.

Although the size of the chosen corpus is limited, it has the advantage of providing parallel sentences for both varieties, thereby minimising potential lexical and grammatical biases that can occur in less homogeneous corpora. Moreover, since part of this corpus was designed to highlight lexical differences between the varieties, it is useful for testing the efficacy of our methods in identifying these differences.

Our analysis focus on lemmas, parts-of-speech, and syntactic relations. Thus, both corpora were parsed using the Portuguese model of the Stanza parser Qi et al. (2020). The model used was trained with the Bosque corpus[1] which contains both Brazilian and European varieties. No manual verification of the annotations was made, however, in the qualitative analysis of the differences between the varieties, it was possible to notice that the parser provided coherent results.

3.2 Relative Entropy

To quantify the lexical and grammatical differences between the varieties of Portuguese, we used relative entropy, specifically Kullback-Leibler Divergence (KLD; Kullback and Leibler (1951)). This method compares probability distributions by calculating the number of extra bits required to encode a data set A using a model based on data set B for a given set of elements X, as described by equation 1.

$$D_{KL}(A\|B) = \sum_{x \in X} A(x) \log \left(\frac{A(x)}{B(x)} \right) \quad (1)$$

In our case, A and B correspond to the varieties of Portuguese. Regarding the elements X, we conducted the following analysis:

1. Lemmas
2. Parts-of-Speech (POS)
3. Dependencies Relations (deprel)
4. Parts-of-Speech tri-grams

Therefore, the idea is to analyse the lexical discrepancies using the lemmas, and to use the other analysis to examine the grammatical differences regarding both varieties.

KLD provides a measure regarding the extent of divergence between corpora and highlights the features most strongly linked to these differences.[2]

Thus, for each feature X, we can measure the divergence between the two corpora. Additionally, by using pointwise KLD, i.e., the individual KLD for each feature (lemmas, POS, deprels, and POS tri-grams), we can identify the specific features that are more typical for one variety or the other, with a p-chi value < 0.001.

Due to the asymmetric characteristic of the KLD, we are interested in both directions, i.e., the number of extra bits required to encode the Brazilian Portuguese dataset based on data from the European Portuguese (D_{KL}(BP||EP)) and vice-versa (D_{KL}(EP||BP)).

3.3 Word Order Entropy

To analyse possible word order differences regarding Brazilian and European Portuguese, we use the word order entropy measure as established by Levshina (2019). The entropy is calculated for 18 different word order patterns, using POS and deprels to define them. The list of different patterns can be seen in Table 2 as defined by Levshina (2019).

The entropy measure correspond to the one defined by Shannon (1948). It reflects the variation in word order across the twenty-four dependencies and co-dependencies outlined in Table 2. For each word order pattern in the corpus, the entropy was calculated using the formula presented in (2):

$$H(X) = \sum_{i=1}^{2} P(X_i) \log(P(X_i)) \quad (2)$$

[1] https://github.com/UniversalDependencies/UD_Portuguese-Bosque

[2] Discrepancies in vocabulary size are addressed using Jelinek-Mercer smoothing with a lambda value of 0.05 (see Zhai and Lafferty (2004) and Fankhauser et al. (2014)).

Type	Label	Dependent	Head
Nominals heads	nsubj_Pred	Subject (noun or pronoun)	root
	nobj_pred	Direct object (noun or pronoun)	root
	obl_pred	Oblique phrase	root
	nmod_noun	Nominal dependent (noun or pronoun)	Noun
Co-dependent nominals	nsubj_obj	Nominal subject and object	-
	obj_obl	Nominal object and oblique phrase	root
Modifiers and heads	nummod_Noun	Numeric modifier	Noun
	amod_Noun	Adjectival modifier	Noun
	advmod_V-Adj	Adverbial modifier or Adjective	Verb
Function words and heads	det_Noun	Determiner	Noun
	adp_Noun	Adposition	Noun
	aux_Verb	Auxiliary	Verb
	cop_pred	Copula	Any nominal
	mark_ccomp/advcl	Subordinators	Predicate of complement clause
Clauses	csubj_pred	Clausal subject	Predicate of the main clause
	ccomp_pred	Clausal complement	Predicate of the main clause
	acl_Noun	Adjectival clause	Noun
	advcl_pred	Adverbial clause	Predicate of the main clause

Table 2: Description of the 18 syntactic features chosen for the word order entropy analysis.

Here, X is a binary variable representing two possible word orders, $P(X_i)$ refers to the probability of one of these orders, i.e., its relative proportion in a given corpus. When one word order has a proportion of 1 and the reverse order has a proportion of 0, or vice versa, the entropy H is 0, indicating no variation. Conversely, if both word orders have a proportion of 0.5, the entropy reaches its maximum value of 1.

We calculated the entropy measures for all 18 patterns in both European and Brazilian Portuguese to determine whether there is significant word order variation across the different syntactic relations listed in Table 2.

4 Results

4.1 Relative Entropy

As explained earlier, we calculated the Kullback-Leibler divergence for four sets of features, covering both lexical and grammatical levels: lemmas, parts of speech, dependency relations, and parts-of-speech trigrams. The overall results are presented in Figure 1.

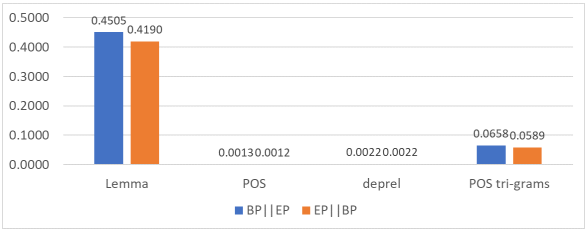

Figure 1: Overall KLD results regarding lemmas, parts-of-speech, dependency relations, and parts-of-speech tri-grams.

The results show that the greatest divergence occurs at the lexical level, followed by POS tri-grams. In contrast, the usage of dependency relations and POS do not differ significantly, with values very close to zero.

At the lexical level (i.e., lemmas), more bits are required to encode the BP corpus using the EP model than vice versa, suggesting that BP has a more complex vocabulary, at least regarding our limited dataset.

In terms of POS tri-grams, we observe the same phenomenon; however, both divergence measures

are close to zero, indicating a less pronounced discrepancy compared to the lexical level.

Besides the overall divergence analysis, we also examined the pointwise KLD for each feature to identify the most typical elements of each variety.

At the lexical level, out of the 18,439 lemmas extracted from both corpora, 271 showed a significant KLD measure (positive for either EP or BP) with a p-chi value below 0.001. Tables 3 and 4 shows the 30 most typical tokens for each corpus.

Lemma	KLD - BP
ele	0.0066
esse	0.0060
usar	0.0059
ônibus	0.0057
equipe	0.0051
trem	0.0044
ela	0.0040
tela	0.0036
abacaxi	0.0030
suco	0.0030
pedestre	0.0028
garota	0.0028
mouse	0.0027
banheiro	0.0026
eles	0.0025
terno	0.0024
pois	0.0024
Prêmio	0.0022
sorvete	0.0020
motorista	0.0020
US$	0.0020
gol	0.0020
videogame	0.0019
conectar	0.0019
grampeador	0.0019
isso	0.0018
usuário	0.0016
paletó	0.0016
prêmio	0.0015
controle	0.0015

Table 3: Lemmas with statistically valid differences regarding pointwise KLD for Brazilian portuguese.

The pointwise KLD measure effectively captures the lexical specificities of each variety. Since we are using a parallel corpus, it is easy to identify the pairs of words that express the same meaning in the different varieties. The lexical differences can be classified into different classes.

Lemma	KLD - EP
este	0.0143
o	0.0122
a	0.0084
utilizar	0.0082
autocarro	0.0053
equipa	0.0052
condução	0.0048
sumo	0.0046
ter	0.0034
ecrã	0.0034
telemóvel	0.0032
golo	0.0032
comboio	0.0031
ananá	0.0028
pequeno-almoço	0.0028
0	0.0027
rapariga	0.0023
peão	0.0021
agrafador	0.0019
rato	0.0019
E.U.A.	0.0019
carta	0.0018
regressar	0.0017
registar	0.0017
utilização	0.0017
se	0.0017
normalmente	0.0017
Prémio	0.0017
isto	0.0016
videojogo	0.0016

Table 4: Lemmas with statistically valid differences regarding pointwise KLD for European portuguese.

First, ortographic variations: even though the Ortographic Agreement (Pinto, 2012) proposed to unify the orthographies of the different varieties of Portuguese, some words can still be written in more than one form. In our analysis, we can identify: *económico* (EP) / *econômico* (BP) (economic); *facto* (EP) / *fato* (BP) (fact); *prémio* (EP) / *prêmio* (BP) (prize).

Additionally, regarding synonyms: different words are used by the various varieties to express the same meaning. In some cases, the word may also exist in the other variety but is not necessarily used in the same contexts. For example: *autocarro* (EP) / *ônibus* (BP) (bus); *fato* (EP) / *paletó* (BP) (suit); *gelado* (EP) / *sorvete* (BP) (ice cream).

Finally, concerning grammatical choices: the

lexical analysis also indicates specific grammatical preferences for each variety, and these cases present the highest KLD values. For instance, the demonstrative adjectives *este* and *esse* (this) are used to distinguish proximity. *Este* is used when the object is closer to the speaker, while *esse* is used when the object is closer to the other interlocutor. However, this differentiation is becoming less common in BP, where the form *esse* is increasingly preferred, as described by Meira and Guirardello-Damian (2018). Moreover, we can observe the presence of the third-person singular pronoun in the BP variety, which relates to the loss of the pro-drop property due to verbal simplification in BP (cf. Duarte (2000)). Two other interesting phenomena can be noted: the typicality of the definite article *o* in EP, due to its obligatory usage with possessive determiners in this variety (cf. Castro (2006)), and the preposition *a*, also in EP. In Brazil, it has mostly been replaced by the preposition *em* when combined with movement verbs (Gil and da Silva, 2023). Furthermore, the typicality of *a* in EP is due to its use as a subordinating conjunction, combined with an infinitive to express an ongoing action, whereas in BP, the gerund is typically used (cf. Hricsina (2014)).

Besides the cases mentioned above, there are other particularly interesting lexical differences: the typical use of the explicative or conclusive conjunction *pois* in BP, which is also part of the EP vocabulary. In our corpus, when *pois* is used in BP, the most common equivalent in EP is *porque*. Also, there is a clear preference in the usage of the verb *usar* (to use) in BP, while, in EP, the typical choice is *utilizar* (to utilise). This result should be confirmed with a larger corpus as it could just imply a preference of the translators who composed the data used in this study.

It is important to note that, due to the limited size of the corpus, while many lexical differences can be identified, it does not encompass the full extent of the lexical specificities of both varieties. The texts are restricted to a particular register (written Portuguese), so a transcribed spoken corpus could be used to complement this lexical analysis.

Regarding the pointwise KLD values for the parts-of-speech, we identify the following significative differences:

- BP: Pronouns, symbols, and adverbs
- EP: Determiners

The typicality of pronouns in BP can be attributed to the loss of the pro-drop phenomenon, as previously mentioned. Symbols appear more prominently in the Brazilian corpus, with the use of US$ or R$ instead of *dólares* and *reais*, which are more common in the European Portuguese data. The frequency of adverbs is quite similar in both corpora; however, differences arise in the choice of adverbs used. For instance, *então* is more typical in BP, while *contudo* is more representative of EP. Additionally, the specificity of determiners in EP can be attributed to their more frequent use before possessive determiners in this variety.

Regarding the dependency relations, the following statistically significant differences were found:

- BP: advmod, nummod, nsubj
- EP: acl:relcl, aux, det, mark, expl, iobj

Analyzing the corpora qualitatively, it is evident that, in some cases, the adverbial modifier used in BP is replaced by adjectival constructions in EP (e.g., *abaixo* in BP (below) and *inferiores* in EP (inferior)). The use of numerical modifiers in BP is prominent in temporal constructions. For example, *no dia 6 de setembro* in BP (on the 6th of September) and *a 6 de setembro* in EP (on the 6th of September). In BP, the token *6* is labeled as a numerical modifier (nummod), whereas in EP, it is labeled as oblique (obl). The frequent use of nominal subjects in BP was expected, given the loss of the pro-drop phenomenon in this variety.

Regarding the more representative dependency relations in EP, the typicality of determiners can be attributed to the greater use of articles in this variety, as previously explained. Additionally, the prevalence of the "mark" relation is due to constructions involving *a + infinitive* to express ongoing actions, whereas BP typically uses the gerund.

In EP, adnominal relative clauses (acl) are often replaced by adverbial clauses (advcl) or adnominal clauses (acl) in BP. For example, *que enfrentava* (who was facing) in EP becomes *enfrentando* (facing) in BP, and *reformas que visavam* (reforms that aimed) in EP is replaced by *reformas com o objetivo de melhorar* (reforms with the objective of improving) in BP.

The expletive (expl) relation in Portuguese is used to mark reflexive pronouns with pronominal verbs. EP clearly shows a preference for these types of verbs. For instance, *demitiu-se* (he quit) and *divorciou-se* (he divorced) are common in EP,

while BP favors constructions like *renunciou* (he resigned) and *é divorciado* (he is divorced).

Regarding the indirect object (iobj), there is no clear preference for specific constructions in EP compared to BP. The corpus reveals various instances where the iobj is replaced by a direct object, often due to different verb choices, which require different arguments.

Finally, the auxiliary (aux) relation is more typical in EP within compound verb phrases (e.g., *tendo sido* (he has been) and *depois de ter sido* (after having been)), whereas in BP, the auxiliary is often omitted, with only the participle or infinitive used directly (e.g., *sido* and *depois de ser*).

Regarding the analysis of POS 3-grams, Tables 5 and 6 present the 15 POS patterns most typical for BP and EP.

Lemma	KLD - BP
DET-NOUN-NUM	0.0028
VERB-ADP-DET	0.0021
ADP-SYM-NUM	0.0020
PRON-VERB-ADP	0.0018
AUX-VERB-ADP	0.0017
SYM-NUM-NUM	0.0016
PRON-VERB-DET	0.0014
NUM-ADP-NUM	0.0012
NOUN-ADP-PRON	0.0012
DET-NOUN-ADV	0.0011
ADJ-ADP-NOUN	0.0011
ADV-AUX-VERB	0.0010
PRON-ADV-VERB	0.0010
CCONJ-PRON-VERB	0.0009
NOUN-ADV-AUX	0.0008

Table 5: POS 3-grams with statistically valid differences regarding pointwise KLD for Brazilian portuguese.

The typical tri-grams for the different varieties confirm the grammatical patterns already identified in the examination of POS and dependency relations.

It is possible to identify the typical usage of two determiners in European Portuguese (EP), specifically the article and possessive determiner, in patterns such as DET-DET-NOUN and VERB-DET-DET. Additionally, we can observe the verbal construction formed by VERB-SCONJ-VERB (e.g., *estar a fazer* (to be doing)). This analysis also reveals the syntactic preference of EP for placing oblique and direct object clitic pronouns after the verb (e.g., VERB-PRON-ADP, NOUN-VERB-PRON), while in Brazilian Portuguese (BP), these pronouns are usually placed before the verb (cf. Kato and Martins (2016)).

Lemma	KLD - EP
DET-DET-NOUN	0.0067
ADP-DET-DET	0.0034
VERB-DET-DET	0.0025
VERB-PRON-ADP	0.0023
AUX-VERB-DET	0.0015
DET-DET-ADJ	0.0014
PRON-ADP-DET	0.0013
NOUN-VERB-PRON	0.0011
NUM-NUM-NUM	0.0011
ADP-ADP-DET	0.0010
VERB-SCONJ-VERB	0.0009
DET-NOUN-SCONJ	0.0009
AUX-ADJ-ADP	0.0009
ADP-NUM-NUM	0.0007
VERB-PRON-ADV	0.0007

Table 6: POS 3-grams with statistically valid differences regarding pointwise KLD for European portuguese.

Regarding the BP, the patterns PRON-VERB-ADP and PRON-VERB-ADV indicate two different phenomena, the more typical usage of pronouns as nominal subjects and the usage of clitic pronouns positioned before the verbs (also identified in patterns such as PROPN-PRON-VERB). Moreover, the typicality of the gerund is also observed (e.g., AUX-VERB-ADP). We can also identify a preference in BP for the usage of constructions such as VERB-ADP (e.g., *a ele* (to him)), being replaced by a clitic pronoun in EP (e.g., *lhe*).

Overall, the KLD analysis at different linguistic levels allows for the identification of a myriad of typical features (both lexical and grammatical) for each variety. The overall KLD indicates that most differences occur at the lexical level. However, by using pointwise KLD, we can examine specific grammatical preferences more closely.

4.2 Word Order Entropy

As described in Section 3, in addition to the KLD analysis, we also calculated word order entropy values for a set of 18 syntactic features for both varieties of Portuguese, as listed in Table 2. Figure 2 presents the ensemble of results.

Most of the 18 syntactic features display similar entropy values for both EP and BP. Several features, such as adp_NOUN, aux_Verb, mark_ccomp/advcl,

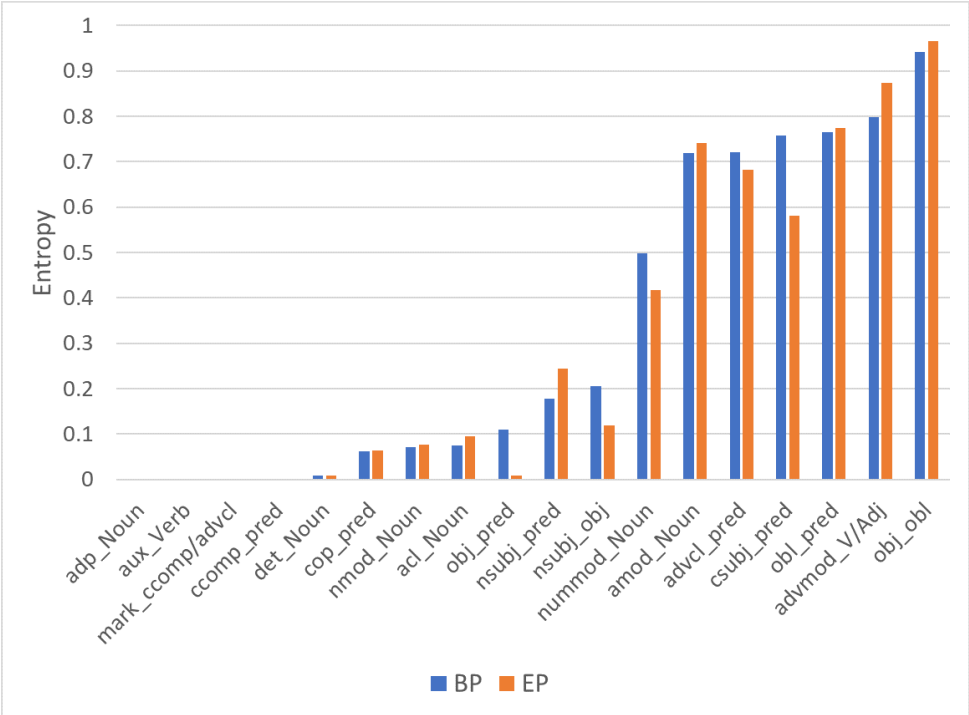

Figure 2: Word order entropy values for the 18 syntactic features described in Table 2 for Brazilian (BP) and European Portuguese (EP).

ccomp_pred, and det_Noun, show entropy values close to 0, indicating a relatively fixed word order (e.g., auxiliary verbs consistently precede the main verb). Other features exhibit values ranging from 0.3 to 0.8, reflecting some flexibility in word order. The feature with the entropy value closest to 1 is obj_obl, which indicates a strong preference in both varieties for placing the direct object before the oblique argument.

Focusing on the features where discrepancies between the varieties can be observed, we notice that obj_pred, nsubj_pred, nsubj_obj, nummod_Noun, and csubj_pred show the most divergence between the varieties.

The difference in word order between the direct object and the predicate (root) can be attributed to the previously discussed variation in the position of clitic objects. While nominal objects are consistently placed after the verb in both varieties, pronominal objects are typically positioned before the verb in BP and after the verb in EP. Thus, the entropy in this case is closer to 0 for EP and higher for BP, indicating more variability in the word order.

A qualitative analysis of the corpus showed that, regarding the nsubj_pred feature, EP present more sentences with the root preceding the nominal subject. For example, *Como afirmou Galeno* (EP) and *Como Galeno disse* (As Galeno said), thus having a higher entropy value.

The difference in entropy values for the nsubj_obj feature can be attributed to subordinate clauses where the relative pronoun *que* precedes the nominal subject of the clause. This structure occurs more frequently in BP, though it is also possible in EP. The variation may be explained by the translator's verb choice, i.e., in EP, the construction sometimes required an oblique complement, whereas in BP, a direct object was necessary.

BP exhibits a word order entropy for nummod_Noun close to 0.5, while for EP, this measure is lower, indicating a slightly more fixed word order. This can be explained by the higher frequency of expressions such as *meados dos anos 2000* (in the middle of the 2000s) and *no dia 6 de setembro* (on the 6th of September) in BP, where the nouns (i.e., *anos* and *dia*) are often included. In EP, however, these nouns tend to be omitted.

Finally, the last dependency relation showing a significant difference between the varieties of Portuguese is csubj_pred. The results indicate a more fixed ordering in BP (i.e., the clausal subject typically follows the predicate). In the corpus, examples from EP, such as *Proteger e melhorar o património bibliográfico do país são dois...* (To protect and improve the bibliographic patrimony of the country are two...), show that the token labeled

as csubj (e.g., the verb *proteger*) appears before the predicate *dois*. In BP, however, this sentence is restructured with nouns instead of verbs: *A proteção e aprimoramento do legado bibliográfico do país são outros dois...* (The protection and improvement of the bibliographic legacy of the country are two others...), thus replacing the clausal subject with a nominal one.

The word order entropy analysis revealed specific syntactic phenomena that differ between BP and EP. While some of these word order tendencies can be attributed to inherent linguistic characteristics of the varieties (e.g., the position of clitic objects), others may come from stylistic choices made by the translators who created the corpora. A more extensive analysis using larger corpora could further complement and refine our findings.

5 Conclusion and Future Work

In this paper, we provided a general overview of the lexical and grammatical differences between Brazilian and European Portuguese. By applying entropy measures (i.e., Kullback-Leibler divergence and word order entropy) across various linguistic levels to a parallel corpus of BP and EP sentences translated from English, we quantified these differences and identified the most characteristic phenomena underlying these divergences.

Regarding KLD, the highest divergence was observed at the lexical level. The lexical analysis not only allowed us to identify word pairs that differ between the two varieties but also revealed specific grammatical preferences, such as the loss of the pro-drop phenomenon in BP. Additionally, the analysis of POS, dependency relations, and POS tri-grams enabled a more detailed examination of the grammatical constructions typical to each variety (e.g., the use of the gerund and the position of clitic objects).

Finally, the word order entropy study showed that, while the majority of the 18 features analyzed exhibited similar results, specific word order preferences were still observed between the varieties.

For future work, we aim to expand this analysis using larger corpora to verify whether the tendencies identified in this study (e.g., the order of clausal subject and predicate) can be confirmed. Additionally, as the methods used here can be applied to studies of linguistic variation in general, we plan to extend this analysis to other varieties of Portuguese. We also intend to complement our study with other information-theoretic measures, such as surprisal to help us identify what would be the most unexpected words and grammatical constructions in each variety when processed with a model trained with a different one.

6 Limitations

While this study provides an overview of the lexical and grammatical differences between Brazilian and European Portuguese, it does not encompass the regional linguistic varieties found within Brazil and Portugal. Additionally, due to the limited size of the dataset and its specific register, this analysis may not capture all existing differences. As mentioned in the paper, some linguistic phenomena observed may be attributed to the stylistic preferences of the translators, rather than representing typical characteristics of the varieties themselves.

7 Ethical Considerations

The dataset used for this study is publicly available and curated by Riley et al. (2023). We are committed to maintaining transparency in our methodology and findings throughout this research. Each result is accompanied by examples derived from a qualitative analysis of the corpora, allowing readers to understand the context and significance of our findings. Additionally, we have explicitly addressed potential biases and inconsistencies within the dataset and our analysis in the text, acknowledging their implications for the interpretations drawn from our study.

8 Acknowledgments

This research is funded by *Deutsche Forschungsgemeinschaft* (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the vardial evaluation campaign 2023. *arXiv preprint arXiv:2305.20080*.

Diego Fernando Válio Antunes Alves. 2024. An evaluation of portuguese language models' adaptation to african portuguese varieties. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 544–550.

BBC. 2024. Língua portuguesa: um dos mais ricos patrimônios da humanidade. Accessed: 2024-10-09.

António Branco, Sara Grilo, and Joao Silva. 2023. *Language Report Portuguese*, pages 195–198.

Ana Castro. 2006. *On possessives in Portuguese*. Universidade NOVA de Lisboa (Portugal).

Dayvid Castro, Ellen Souza, and Adriano L.I. De Oliveira. 2016. Discriminating between brazilian and european portuguese national varieties on twitter texts. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 265–270.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletic Haddad, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. Vardial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–15. The Association for Computational Linguistics.

Eduardo G Cortes, Ana Luiza Vianna, Mikaela Martins, Sandro Rigo, and Rafael Kunst. 2024. Llms and translation: different approaches to localization between brazilian portuguese and european portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 45–55.

Mayara Nicolau De Paula. 2017. A comparative diachronic analysis of wh-questions in brazilian and european portuguese. *Diadorim*.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Maria Eugênia Lamoglia Duarte. 2000. The loss of the'avoid pronoun'principle in brazilian portuguese. *Brazilian portuguese and the null subject parameter.-(Editionen der Iberoamericana. Serie B, Sprachwissenschaft; 4)*, pages 17–36.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the World*, 27 edition. SIL International, Dallas, Texas. Online version: `http://www.ethnologue.com`.

Paola Escudero, Paul Boersma, Andréia Schurt Rauber, and Ricardo AH Bion. 2009. A cross-dialect acoustic description of vowels: Brazilian and european portuguese. *The Journal of the Acoustical Society of America*, 126(3):1379–1393.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and Visualizing Variation in Language Resources. In *LREC*, pages 4125–4128.

Sónia Frota, Marisa Cruz, Flaviane Svartman, Gisela Collischonn, Aline Fonseca, Carolina Serra, Pedro Oliveira, and Marina Vigário. 2015. Intonational variation in portuguese: European and brazilian varieties. *Intonation in romance*, (1):235–283.

Sónia Frota and Marina Vigário. 2001. On the correlates of rhythmic distinctions: The european/brazilian portuguese case.

Dirk Geeraerts, Stefan Grondelaers, and Dirk Speelman. 1999. Convergentie en divergentie in de nederlandse woordenschat: een onderzoek naar kleding-en voetbaltermen.

Maitê Moraes Gil and Augusto Soares da Silva. 2023. A study on the conceptual structure of the use of prepositions in the complement of goal-oriented motion verbs in brazilian portuguese. *Cognitive Semantics*, 9(1):73–102.

Jan Hricsina. 2014. Substituição do gerúndio pela construção a+ infinitivo no português europeu (estudo diacrónico). *Studia Iberystyczne*, (13):383–401.

Instituto Camões. 2021. Português no mundo. Accessed: 2024-10-09.

Mary Aizawa Kato and Ana Maria Martins. 2016. European portuguese and brazilian portuguese: an overview on word order. *The handbook of Portuguese linguistics*, pages 15–40.

Solomon Kullback and Richard A Leibler. 1951. On Information and Sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology*, 23(3):533–572.

Sérgio Meira and Raquel Guirardello-Damian. 2018. Brazilian-portuguese: Noncontrastive exophoric use of demonstratives in the spoken language. *Demonstratives in cross-linguistic perspective*, 14:116.

Marcelo A Montemurro and Damián H Zanette. 2011. Universal entropy of word ordering across linguistic families. *PLoS One*, 6(5):e19875.

Paulo Feytor Pinto. 2012. *Novo acordo ortográfico da língua portuguesa*. Leya.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Parker Riley, Timothy Dozat, Jan A Botha, Xavier Garcia, Dan Garrette, Jason Riesa, Orhan Firat, and Noah Constant. 2023. Frmt: A benchmark for few-shot region-aware machine translation. *Transactions of the Association for Computational Linguistics*, 11:671–685.

Rodrigo Santos, João Rodrigues, Luís Gomes, João Silva, António Branco, Henrique Lopes Cardoso, Tomás Freitas Osório, and Bernardo Leite. 2024. Fostering the ecosystem of open neural encoders for portuguese with albertina pt-* family. *Preprint*, arXiv:2403.01897.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Augusto Soares da Silva. 2010. Measuring and parameterizing lexical convergence and divergence between european and brazilian portuguese. *Advances in cognitive sociolinguistics*, 45:41.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.