Proposta recebida em Fevereiro 2025 e aceite para publicação em Novembro 2025.

Análise e Classificação Automática de Domínios Discursivos no Português do Brasil

Automatic Analysis and Classification of Discourse Domains in Brazilian Portuguese

Felipe Ribas Serras
Universidade de São Paulo

Miguel de Mello Carpi
Universidade de São Paulo

Mariana Lourenço Sturzeneker
Universidade de São Paulo

Mayara Feliciano Palma
Universidade de São Paulo

Aline Silva Costa
Instituto Federal de Educação, Ciência e Tecnologia da Bahia

Vanessa Martins do Monte
Universidade de São Paulo

Cristiane Namiuti
Universidade Federal do Sudoeste da Bahia

Maria Clara Ramos Morales Crespo
Universidade de São Paulo

Maria Clara Paixão de Sousa
Universidade de São Paulo

Marcelo Finger
Universidade de São Paulo

Resumo

Este artigo trata da identificação dos domínios discursivos *Jurídico*, *Entretenimento*, *Jornalístico*, *Fórum Virtual* e *Instrucional* do português brasileiro no nível sentencial, amostrados do *corpus* Carolina. Avaliamos propriedades gramaticais, lexicais e semânticas. Demonstramos que os domínios são discerníveis e se organizam em uma escala consistente que associamos à distinção oral-envolvido vs. literato-informacional a partir da comparação com outros trabalhos. Treinamos classificadores *Transformer* em um novo *dataset* de sentenças para identificação de domínios, alcançando alta performance. Os padrões de erro dos modelos correlacionam-se com a escala identificada, indicando a captura desta dimensão de variação. Disponibilizamos publicamente os *datasets* e modelos produzidos.

Palavras chave

domínios discursivos; Português Brasileiro; classificação de sentenças; modelos *transformer*; variação linguística

Abstract

This paper addresses the identification of the *Juridical*, *Entertainment*, *Journalistic*, *Virtual*, and *Instructional* discourse domains of Brazilian Portuguese at the sentence level, sampled from the Carolina corpus. We evaluate grammatical, lexical, and semantic properties. We demonstrate that the domains are discernible and organized into a consistent scale, which we associate with the oral-involved vs. literate-informational distinction based on comparison with other works. We trained *Transformer* classifiers on a new sentence dataset for domain identification, achieving high performance. The models' error patterns correlate with the identified scale, suggesting the models captured this dimension of variation. The datasets and models developed in this study are publicly available.

Keywords

discourse domains; Brazilian Portuguese; sentence classification; transformer models; linguistic variation

1. Introdução

A heterogeneidade é uma característica intrínseca da linguagem: todas as línguas humanas apresentam uma enorme capacidade de variação. Quando pensamos nessa variação linguística, são as diferenças entre dialetos que prontamente nos ocorrem. Entretanto, a linguagem humana é muito mais plástica do que pode parecer à primeira vista: uma mesma pessoa pode, num intervalo de segundos, redigir um documento jurídico e fazer um *post* numa rede social, mudando completamente as características da linguagem produzida em resposta ao espaço daquela produção. Este tipo de variação contextual pode ser descrito como variação em registro, gênero ou domínio discursivo — cujas diferenças discutiremos na Seção 2 — e é uma parte intrínseca e fundamental da competência da linguagem humana.

O entendimento desses fenômenos de variação é, portanto, essencial no desenvolvimento de teorias e modelos da linguagem (Catford, 1965). Entretanto, essas distinções têm se tornado críticas também no desenvolvimento de tecnologias da linguagem: se antes a adaptação de um modelo generalista a um domínio específico para a realização de tarefas daquele domínio — a chamada *domain adaptation* — era suficiente para abordar a variação linguística no PLN, a implantação de modelos gerativos em grande escala como assistentes virtuais universais tem mudado esse cenário, já que agora esses modelos precisam acompanhar a plasticidade linguística dos indivíduos que almejam assistir.

Entretanto, a incorporação da variação linguística como parâmetro, seja na linguística tradicional ou no desenvolvimento da tecnologias da linguagem, exige: (i) o entendimento acerca de que formas, em quais dimensões e escopos esse tipo de variação se dá, através de quais características se manifesta e quais tarefas de processamento de linguagem impacta; (ii) recursos que permitam estudar esses fenômenos de variação, identificar diferentes variedades e representar e selecionar as variedades de interesse para cada aplicação.

Nessa linha, nosso propósito neste artigo é explorar a maneira como esse tipo de variação situacional se dá no português brasileiro, desenvolvendo no processo recursos linguístico-computacionais abertos que facilitem a distinção das variedades estudadas, buscando facilitar a elaboração de recursos e modelos sensíveis à variação linguística no futuro.

Mais especificamente, adotamos o *domínio de discurso* ou *domínio discursivo* como a dimensão de variação linguística sobre a qual nos debruçamos. Exemplos de domínios discursivos são textos jurídicos, de entretenimento, jornalísticos, de fóruns virtuais e textos instrucionais.

Uma literatura significativa se debruça sobre a caracterização e diferenciação computacional das variedades situacionais da língua portuguesa. Em especial, trabalhos como Kauffmann (2005), Sardinha et al. (2014b) e Sardinha (2017) realizam a análise multidimensional, como proposta por Biber (1988), de diferentes registros do português, a partir da anotação semiautomática de características linguísticas e da identificação de padrões subjacentes de coocorrência dessas características. Nossa pesquisa possui diversas semelhanças com os trabalhos dessa linha, entretanto focamos no escopo sentencial, no nível específico dos domínios discursivos, analisamos características linguísticas de forma individualizada, ao invés de seus padrões de coocorrência, propomos novas etapas de análise, e acoplamos o estudo da variação e da diferenciabilidade ao desenvolvimento de classificadores automáticos para tal.

Apesar dessas distinções — ou em razão delas — vemos nossos experimentos como complementares aos experimentos prévios da literatura e acreditamos que esses resultados podem, em conjunto com os anteriores, fornecer um entendimento mais profundo acerca da variação linguística situacional no português, bem como serem usados na construção de uma base sólida para o desenvolvimento de recursos sensíveis à variação.

Organizamos este trabalho em duas partes complementares: na primeira investigamos a diferenciabilidade computacional dos domínios do português, com foco no escopo sentencial, buscando responder às seguintes questões:

Q1 Os domínios discursivos são computacionalmente discerníveis nesse escopo?

Q2 Em caso afirmativo, quais propriedades os diferenciam?

Q3 Quais problemas e tarefas de PLN podem se beneficiar da diferenciação de domínios nesse escopo?

Para responder a essas perguntas, usamos dados do *corpus* Carolina, um *corpus* geral do português, com foco nas variedades brasileiras, composto de textos abertos extraídos da internet. O *corpus* inclui anotação acerca do *domínio do discurso* de cada texto (Sturzeneker et al., 2022; Crespo et al., 2023). Avaliamos a discernibilidade dos cinco domínios mais prevalentes entre os dados — *Jornalístico*, *Instrucional*, *Jurídico*, *Fórum Virtual*[1] e *Entretenimento* — sob os seguintes aspectos: nível de duplicação, distribuição de propriedades linguísticas, diferenças vocabulares e separabilidade em espaços de *embeddings* semânticos.

Na segunda parte, nos debruçamos sobre o desenvolvimento de modelos para a classificação automática de domínios discursivos em português brasileiro, modelada como um problema de classificação de sentenças, como ilustrado na Figura 1. Produzimos um *dataset* para treinar e aperfeiçoar diversos modelos estado-da-arte para esta tarefa, comparando seus desempenhos à luz da nossa investigação anterior sobre diferenciabilidade dos domínios.

[1]Por questões de espaço, abreviamos *Fórum Virtual* para *Virtual*, quando convém.

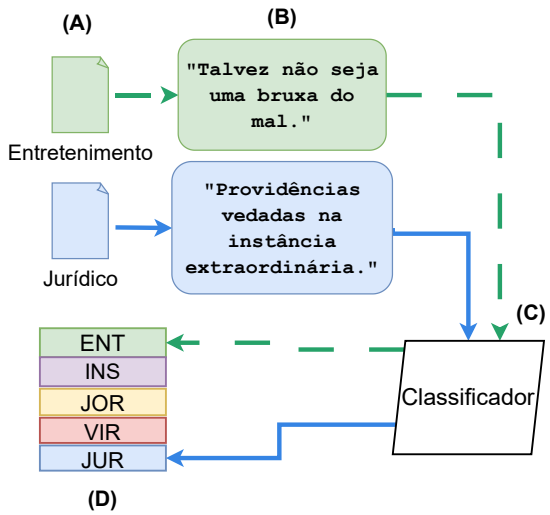


Figura 1: Ilustração da classificação de domínios em nível de sentenças: (A) Documentos de um domínio específico; (B) Extração de sentenças; (C) Análise por classificador da sentença; (D) Recuperação do domínio original.

Além das nossas análises, são resultados desta pesquisa três *subcorpora* do *corpus* Carolina, voltados para a diferenciação de domínios discursivos do português brasileiro, e cinco modelos de classificação automática de domínios.[2]

Este artigo está organizado da seguinte forma: na Seção 2, apresentamos a fundamentação teórica que sustenta nosso trabalho; na Seção 3 os trabalhos correlatos ao nosso; na Seção 4, detalhamos os *datasets* construídos e utilizados; na Seção 5, apresentamos e justificamos a metodologia geral adotada nas duas etapas do trabalho; na Seção 6 são apresentados os procedimentos e resultados obtidos para a investigação de discernibilidade dos domínios; na Seção 7 apresentamos os resultados e análise referentes ao desenvolvimento de classificadores de domínio e na Seção 8, apresentamos nossas conclusões e futuras direções de trabalho.

2. Fundamentação Teórica

Neste trabalho, abordamos a variação intralinguística no português brasileiro, com foco nas variedades chamadas de domínios discursivos ou de domínios do discurso, especificamente aqueles presentes no *Corpus* Carolina. A definição de trabalho durante a construção desse *corpus* é a de que **domínios do discurso** são variedades em uma língua natural, caracterizadas por propriedades e estruturas linguísticas que refletem a **situação comunicativa** dos **espaços** em que os textos são produzidos e veiculados, exemplos sendo textos jurídicos, de entretenimento, jornalísticos, *etc.* Nesta seção confrontamos este com dois conceitos correlatos, registro e esfera discursiva, o que culminará numa definição mais precisa dos domínios estudados e fundamentará os aparatos metodológicos utilizados para diferenciá-los.

No contexto da sociolinguística e dos estudos de registro (Ellis & Ure, 1969; Ghadessy, 1988), as variações de **registro** constituem uma ampla gama de variações de linguagem que surgem em função da **situação comunicativa** em que os falantes se encontram (Biber, 2006). Ferguson (1994) diferencia registros de (i) **dialetos**, que emergem da comunicação recorrente dentro de um mesmo **grupo social**, e de (ii) **gêneros**, que emergem no uso recorrente de um mesmo **tipo de mensagem**.

Registros seriam compostos de três componentes: **parâmetros** que especificam o tipo de situação em que ocorrem, **propriedades ou marcações linguísticas** que diferenciam seus enunciados dos de outros registros, e **funções comunicativas e/ou convenções** que conectariam essas duas componentes (Biber, 1994).

Existem diversos sistemas para descrever os parâmetros situacionais de diferentes registros, e.g. Crystal & Davy (1969), Hymes (1974), Brown & Fraser (1979) e Duranti (1985). A maioria inclui a relação entre os enunciadores, o ambiente e seu nível de compartilhamento, o nível de formalidade, o canal e o modo de comunicação, o propósito da comunicação e o tópico ou tema da enunciação (Biber, 1994). O mesmo acontece com as características linguísticas passíveis de carregar marcações de registro: Ferguson (1994) inclui vocabulário, entonação, formações sintáticas e fonológicas, enquanto Biber (1994) faz uma lista mais detalhada, incluindo características morfológicas, uso de classes morfossintáticas, estratégias de derivação, diferentes tipos de cláusulas e especificação léxica.

A caracterização e a diferenciação dos registros de uma língua, entretanto, não podem ser realizadas a partir da análise isolada de apenas uma dessas características. Diversos trabalhos defendem uma abordagem **multivariável**. Além disso, o cerne da diferença encontra-se, não em estruturas exclusivas, mas na forma como cada registro emprega, com frequências distintas, as estruturas gramaticais globais da língua a qual pertence. Dessa forma, a caracterização e diferenciação de registros pede também uma análise

[2]Esses e outros recursos relacionados a este trabalho podem ser encontrados em: https://github.com/stars/frserras/lists/domain-discernibility-carolina

quantitativa de múltiplas características e dimensões linguísticas, sob uma abordagem **probabilística** (Biber, 1994).

Um caso proeminente na aderência a esses princípios é o da abordagem chamada análise multidimensional, que, como discutiremos na Seção 3, propõe um arcabouço de investigação dos padrões probabilísticos de coocorrência de múltiplas propriedades linguísticas para identificar os arquétipos de variação entre registros (Biber, 2006).

Uma crítica frequente ao conceito de registro é sua generalidade (Crystal & Davy, 1969), já que registros podem ir desde variedades muito específicas como "prosas acadêmicas em artigos de psicologia" até distinções bastante gerais, como a entre "enunciado escrito" e "enunciado falado". Muitos autores da área, entretanto, consideram essa uma vantagem conceitual e se aproveitam dela metodologicamente (Biber, 1994).

Já as **esferas discursivas** são variedades linguísticas associadas a **campos específicos da atividade humana** e que são habitadas por tipos de enunciados característicos e relativamente estáveis, cujas propriedades responderiam às condições e finalidades de comunicação do campo em questão. Exemplos seriam os campos jurídico, religioso e acadêmico (Bakhtin, 2016).

Elas seriam constituídas de **conjuntos de coerções** que regulam a produção da linguagem, de forma que a língua refrate a realidade segundo a lógica do campo de atividade no qual a linguagem é produzida (Grillo, 2006). A esfera, então, condicionaria as relações entre os enunciados, entre os enunciados e seus objetos e entre os enunciados e os enunciadores, acarretando em variações nas escolhas lexicais, fraseológicas, gramaticais e composicionais durante a produção da linguagem (Bakhtin, 2016; Grillo, 2006).

A esfera é entendida como um nível superior numa gradação, estando num patamar mais geral do que os tipos de enunciados que compõem gêneros, e muito mais geral que variações mais pessoais e estilísticas (Grillo, 2006).

É possível observar as semelhanças entre os conceitos de registro e esfera com o de domínio anteriormente apresentado. Para aperfeiçoar nosso entendimento, propomos a seguinte definição de trabalho para os domínios discursivos abordados, posicionando-os na interseção entre registro e esfera: **Domínios do discurso** são tipos específicos de registros de uma língua, ou seja, são variedades que emergem a partir de **situações comunicativas** recorrentes e cujas estruturas e características linguísticas carregam marcações associadas aos parâmetros que definem tais situações. Domínios ocupam um segmento específico do espectro de generalidade dos registros, estando associados a situações comunicativas relacionadas ao **campo de atividade humana** em que a linguagem é produzida e veiculada, sendo produto das coerções ideológicas deste campo de atividade, de forma similar às esferas discursivas de Bakhtin (2016).

Essa definição nos permite aderir aos princípios metodológicos do estudo de registros mencionados e contextualizar nosso trabalho dentro dessa literatura — o que faremos na próxima seção — ao mesmo tempo em que nos atemos ao espaço de variação mais restrito e bem definido da esfera discursiva, nos aproveitando do conhecimento acerca de suas dinâmicas de manutenção.

3. Trabalhos Relacionados

Nesta seção, apresentamos um panorama de trabalhos correlatos ao nosso, passando de forma breve pelos diferentes tópicos que abordamos ao longo do artigo. Retomaremos posteriormente muitos desses trabalhos para comparação direta e análise cruzada com os nossos resultados, em especial nas Seções 6.5 e 7.2.

Para facilitar a leitura, dividimos a apresentação dos trabalhos correlatos em duas partes: uma focada na diferenciabilidade de variedades (Subseção 3.1) e uma focada na classificação de variedades (Subseção 3.2), entretanto existe uma forte conexão entre os grupos de trabalhos mencionados nas duas partes.

3.1. Trabalhos acerca da diferenciabilidade de variedades

Como mencionado na Seção 2, alinhamos nosso tratamento teórico e princípios metodológicos à chamada análise de registros, que estuda a variação intralinguística em função do contexto situacional. Dentro dessa área, os trabalhos mais similares ao nosso são os da abordagem da análise multidimensional.

A análise multidimensional (MDA, do inglês *Multidimensional Analysis*), proposta em Biber (1986, 1988) é uma abordagem ao estudo de registros que parte do princípio de que diferentes situações exigem a adaptação funcional da linguagem e de que essas adaptações funcionais subjacentes se manifestariam superficialmente na forma de modificações sistemáticas no uso de propriedades linguísticas, de maneira estocástica porém pervasiva (Biber & Conrad, 2009c).

Numa análise multidimensional típica, utiliza-se um *corpus* com documentos anotados por registro, que é então etiquetado de forma automática segundo uma grande quantidade de propriedades léxico-gramaticais, resultando em representações multidimensionais de cada documento. Essas representações são reduzidas, através de análise fatorial, para um espaço de menor dimensionalidade. Nesse processo, as relações de coocorrência das propriedades que mais explicam a variabilidade dos dados são conjugadas nessas dimensões de variação. O método também retorna a importância — positiva ou negativa — de cada uma das propriedades para cada uma das dimensões. Esses dados são usados por especialistas para construir uma interpretação funcional para cada uma dessas dimensões de variação, num processo iterativo de análise que leva em consideração a maneira como diferentes documentos e registros se posicionam a cada uma dessas dimensões (Biber & Conrad, 2009c).

Os conceitos básicos da análise multidimensional foram propostos em Biber (1986) e foram posteriormente estendidos em Biber (1988), que se tornou a referência fundacional para a técnica. Biber (1989) apresentou formas de construção de tipologias textuais derivadas da MDA, identificando registros com comportamento coeso, e Biber & Finegan (1989) aplicou a técnica a dados diacrônicos do inglês, identificado as tendências funcionais de evolução de diferentes registros ao longo da história, investigação que teve continuidade nos trabalhos de Biber & Finegan (1992) e Biber & Finegan (2001). Biber (1995), por sua vez, deu passos para além do inglês, conjugando resultados de aplicações da MDA a diferentes línguas e mapeando tendências interlinguísticas sistemáticas da variação de registro.

Além da versão original da MDA, que foca em características léxico-gramaticais, variantes foram propostas, aplicando as mesmas técnicas a diferentes tipos de características. Delfino (2021) identifica, entre essas variantes, (i) a análise multidimensional léxica, que foca na distribuição de itens léxicos para a identificação de campos temáticos e construtos ideológicos dos textos (Sardinha & Fitzsimmons-Doolan, 2025); (ii) a análise multidimensional colocacional, que foca nas distribuições de tuplas de palavras, identificando padrões nos efeitos de *priming* ou pré-ativação dos itens léxicos (Zuppardi, 2020); e (iii) a chamada análise multidimensional semântica, que foca em anotações acerca dos campos semânticos das palavras usadas.

Além da língua inglesa, houve esforços para aplicar a MDA a outras línguas, incluindo Coreano, Somali e Tuvaluano de Nukulaelae (Biber, 1995), Espanhol (Biber et al., 2006) e Russo (Katinskaya & Sharoff, 2015). O mais interessante é que a metodologia se mostrou, ao mesmo tempo, sensível às especificidades de cada língua e capaz de captar fenômenos transversais comuns da variação dos registros, havendo uma forte equivalência nas dimensões primárias de modelos MDA de diferentes línguas.

Para a língua portuguesa, em específico, diversos trabalhos baseados na metodologia MDA foram desenvolvidos. Kauffmann (2005) foi pioneiro em trazer essa abordagem para o português, mas com um número reduzido de características analisadas e com foco em um registro específico: textos jornalísticos. Sardinha et al. (2014a,b) expandem o escopo desse modelo original, construindo um modelo MDA amplo para o português brasileiro, sobre um conjunto bem maior de registros, assemelhando-se ao trabalho original de Biber (1988) para a língua inglesa.

Posteriormente, Sardinha (2017) construiu uma tipologia dos registros da língua portuguesa sobre esse modelo MDA do português, agrupando os registros de acordo com suas similaridades funcionais, aos moldes do que foi feito para o inglês em Biber (1989), mencionado anteriormente.

Outras aplicações da MDA em português focaram no desenvolvimento de modelos para registros específicos, visando estudar as características e dinâmicas de variação internas desses registros, como é o caso de Kauffmann (2022) — que deu continuidade ao trabalho de Kauffmann (2005) sobre textos jornalísticos — de Faria Marques (2023) para textos jurídicos e Kauffmann (2020) para textos literários.

Assim como esses últimos, outros trabalhos, sejam da linha da análise multidimensional, ou estudos linguísticos mais tradicionais, se propuseram a descrever as características de variedades linguísticas similares aos cinco domínios discursivos que estudamos neste trabalho. Entre elas destacamos (segmentados por domínio):

- *Jornalístico*: Kauffmann (2005), Kauffmann (2022), Biber & Conrad (2019);
- *Jurídico*: de Faria Marques (2023), Gozdz-Roszkowski (2011), Carapinha (2018), Nauege (2022);
- *Fórum Virtual*: Biber & Egbert (2018), Sardinha (2014), Biber & Conrad (2009b) Biber & Conrad (2009a), Komesu (2005);
- *Instrucional*: Conrad (2001);

- *Entretenimento*: Pinto (2014), Sardinha & Pinto (2017), Pinto (2013), Biber & Conrad (2009b), Biber & Conrad (2019).

Nesse mesmo sentido, Costa (2018) é uma boa referência geral, que busca distinguir elementarmente variedades linguísticas.

Um aspecto interessante da MDA é que diversos modelos MDA apresentam dimensões de variação associadas à distinção entre discurso mais e menos oralizado, ou mais coloquial e mais técnico, usualmente caracterizada por propriedades linguísticas relacionada à compactação informacional do texto (Biber, 1988, 2001; Lefer & Vogeleer, 2016). Na nossa análise observamos fenômenos similares e buscamos relacionálos com o conceito de complexidade de linguagem, em especial com abordagens à complexidade alinhadas à teoria da informação.

Nesse sentido, convém mencionar que vários trabalhos debruçam-se sobre a complexidade textual com esse recorte, incluindo Juola (2008), Ehret & Szmrecsanyi (2016) e Ehret & Szmrecsanyi (2019). Em especial, Ehret (2021) avalia especificamente a variação de métricas de complexidade linguística baseadas em teoria da informação em função dos diferentes registros linguísticos da língua inglesa.

No que tange à complexidade da língua portuguesa, a literatura é vasta e muito bem sintetizada em Leal & Aluísio (2024). Em especial, Leal et al. (2024) fornece um conjunto bastante amplo de métricas de complexidade para o português, com alguma sobreposição com as características linguísticas avaliadas neste trabalho.

3.2. Trabalhos acerca da classificação automática de variedades

Após o estudo da diferenciabilidade dos domínios, a segunda etapa do nosso trabalho trata sobre o desenvolvimento de modelos para a classificação automática de domínios discursivos do português no nível sentencial. Para tal treinamos modelos baseados em *Transformers*, mais especificamente *bert-base*, *bert-large*, *albertina-100m* e *albertina900m* (Souza et al., 2020; Rodrigues et al., 2023).

Vários trabalhos focam no desenvolvimento de modelos especializados em domínios particulares, visando capturar padrões de linguagem e realizar tarefas específicas de domínio, como Fonseca et al. (2016); Gu et al. (2021); Lee et al. (2019); Beltagy et al. (2019); Zhou et al. (2013); Serras & Finger (2021); Polo et al. (2021); de Colla Furquim & de Lima (2012); Costa et al. (2023). Entretanto, como aqui estamos interessados no desenvolvimento de uma ferramenta geral capaz de diferenciar os diversos domínios, adotamos como base os modelos mencionados, que são gerais para a língua portuguesa, abarcando diversos domínios no seu pré-treinamento.

Muitos trabalhos anteriores se debruçaram sobre a tarefa de classificar automaticamente diferentes variedades intralinguísticas, em diferentes línguas. Em alguns casos são variedades diatópicas, como em Dunn (2019); Goswami et al. (2020); Saeed et al. (2024); van der Lee & van den Bosch (2017), em alguns casos em nível de gênero, como em Kuzman et al. (2022); Sharoff (2007); Kuzman et al. (2023a,b); Myntti et al. (2024), e em alguns casos em nível de registro, como em Kutuzov et al. (2016); Laippala et al. (2017); Repo et al. (2023); Rönnqvist et al. (2022); Kyröläinen & Laippala (2023); Myntti et al. (2024). Usualmente esses trabalhos focam na classificação no escopo de documento e os métodos de classificação variam desde modelos clássicos, como modelos Bayesianos e Máquinas de Vetor de Suporte, até modelos da família *Transformers*.

Em português, o foco usualmente é colocado no nível dialetal, com destaque para a diferenciação do português brasileiro do português europeu. Esse é o caso de trabalhos como Zampieri & Gebre (2012); Castro et al. (2016); Zampieri et al. (2016b,a); Sousa et al. (2025). Outros trabalhos focam em outras variedades, como Zampieri et al. (2016a), que se propõem a desenvolver classificadores automáticos para diferenciar períodos históricos da língua portuguesa, e Kessler et al. (1997) e Monte-Serrat et al. (2021), que se propõem a diferenciar gêneros textuais em português brasileiro.

No que tange à classificação, não localizamos em nossa revisão bibliográfica estudos que apresentem a mesma confluência de características deste trabalho, nomeadamente a língua-alvo, a variedade intralinguística sendo classificada e o tipo de modelo usado para tal. Nesse sentido, o resultado que encontramos mais comparável ao nosso advém de uma linha de investigação sobre classificação automática de registros em um escopo multilíngue, que inclui os trabalhos de Laippala et al. (2017, 2019); Repo et al. (2021); Rönnqvist et al. (2021); Laippala et al. (2021); Repo et al. (2023); Myntti et al. (2024); Eskelinen et al. (2024); Henriksson et al. (2024). Essa linha parte do CORE (Biber & Egbert, 2018), um *dataset* de documentos em inglês extraídos da *web* e anotados segundo um sistema hierárquico de registros.

Os pesquisadores se debruçaram sobre a criação *datasets* de diferenciação de registro em outras línguas — Finlandês (Laippala et al., 2019; Skantsi & Laippala, 2023), Francês e Sueco (Laippala et al., 2020; Repo et al., 2021) — usando o mesmo sistema de anotação de registros do CORE. Esses *datasets* foram então combinados ao CORE original para compor um *dataset* multilíngue de classificação de registros, o *Multilingual CORE (MCORE)*.

Posteriormente, textos em outras línguas, incluindo o português, foram extraídos do *dataset* OSCAR (Ortiz Suárez et al., 2019), anotados manualmente segundo o mesmo sistema de anotação e incorporados ao MCORE (Laippala et al., 2022). Esses *datasets*, entretanto, são significativamente menores e são utilizados apenas para fins da avaliação. O conjunto em português, por exemplo, inclui 334 textos (Laippala et al., 2023).

Ao longo dos anos — paralelamente à construção do MCORE — diversos modelos foram testados na tarefa de classificação de registros, comparando cenários monolíngues e multilíngues e utilizando diferentes estratégias. Atualmente os melhores resultados são os obtidos pelo modelo XLM-R, um modelo multilíngue baseado em *Transformers* (Henriksson et al., 2024).

Embora semelhante, esta linha de pesquisa adota um conceito de registro mais amplo e foca em um cenário multilíngue, no qual o português é uma das línguas minoritárias. Diferente da nossa abordagem, a análise é feita no nível do documento, e não da sentença, e busca manter a distribuição natural dos registros encontrados na *web*. Nossa metodologia, por outro lado, consiste em balancear os diferentes registros a partir de um conjunto de dados diverso e com licenças permissivas.

Na Seção 7.2, fazemos uma comparação qualitativa dos nossos resultados aos outros resultados de performance de classificadores de variedades mencionados, tanto aqueles voltados para registros, como aqueles voltados para outras dimensões de variação.

4. Dados

Para responder às perguntas de pesquisa e treinar os modelos mencionados na Seção 1, utilizamos os dados do *Corpus* Carolina, uma coleção digital aberta e curada de documentos em português, com foco no português brasileiro, desenvolvida tanto para o treinamento de modelos de linguagem, quanto para a pesquisa linguística. Utilizamos a *versão 1.2 - Ada* do *corpus*, na qual cada documento é anotado com informações tipológicas, organizadas num cabeçalho TEI (*Text Encoding Initiative*)(TEI Consortium, 2021) em três ambientes de metadados distintos:

- *tipologia ampla*, que representa uma divisão metodológica dos documentos, baseada em como os dados foram segmentados durante a análise, obtenção e extração da *web*;
- *tipologia da fonte*, que contém a informações tipológicas do texto, conforme declaradas na fonte da qual o documento foi extraído, sendo, por este motivo, um ambiente com valores bastante específicos e não padronizados;
- *domínio*, que representa o domínio do discurso do texto, alinhando-se à definição apresentada na Seção 2 e anotado usando um sistema de categorias desenhado a partir da análise das diferentes fontes de documentos que compõem o *corpus*.

Com relação ao domínio do discurso, nosso principal foco de interesse, os documentos do *corpus* estão categorizados em dez grupos distintos, representados na Tabela 1.

Domínio	**Razão de Documentos**
Instrucional	41,8%
Jurídico	23,8%
Entretenimento	14,7%
Jornalístico	10,6%
Fórum Virtual	7,4%
Acadêmico	0,5%
Comercial	0,43%
Legislativo	0,38%
Literário	0,19%
Pedagógico	0,096%

Tabela 1: Distribuição dos documentos por domínio no *corpus* Carolina.

Os cinco domínios principais do Carolina, que representam, juntos, cerca de 98,4% dos *tokens* do *corpus*, são definidos a seguir. As tipologias declaradas na fonte contidas em cada um desses domínios são listadas:

- *Instrucional*: textos distribuídos em espaços feitos para instruir e educar os leitores, como enciclopédias virtuais. As tipologias como declaradas na fonte contidas nesse domínio são: *verbete*, *recursos educacionais*, *documentação de ajuda* e *guia de viagem*;
- *Jurídico*: documentos distribuídos no Poder Judiciário brasileiro. Engloba uma lista muito diversificada de tipologias declaradas na fonte, i.e. *inteiro teor de acórdão*, *edital*,

memória jurisprudencial, *publicação temática*, *relatório*, *audiência pública*, *discurso*, *proposta de súmula vinculante*, *ata*, *constituição anotada*, *boletim de jurisprudência*, *biografia*, *glossário*, *resolução*, *composições plenárias do Supremo Tribunal Federal* e *tratado*;

- *Entretenimento*: textos distribuídos em plataformas concebidas para fins de entretenimento. Esse domínio consiste de uma única tipologia declarada na fonte: *legenda*;
- *Jornalístico*: textos distribuídos em plataformas de notícias e ambientes relacionados. As tipologias declaradas na fonte nesse domínio são *notícia*, *notícia científica*, *artigo*, *blogue jornalístico* e *opinião*;
- *Fórum Virtual*: textos distribuídos exclusivamente em ambientes virtuais nativos, como plataformas de mídia social. As tipologias declaradas na fonte contidas nesse domínio são *página de usuário*, *discussão*, *tweet*, *organização de atividades e compartilhamento de experiências*, *blogue pessoal* e *perguntas faq*.

A Tabela 2 contém exemplos de sentenças extraídas do *corpus* para cada um desses domínios. As fontes dos documentos em cada domínio podem ser encontradas nas *tags* de proveniência do *corpus* referentes a cada documento. Informações gerais sobre a procedência também estão disponíveis no *website* do Carolina.[3]

Para garantir representatividade em ambas as etapas do trabalho, nos restringimos, nesta pesquisa, aos cinco domínios listados, que são os mais bem representados dentro do *corpus*.

5. Metodologia

Nesta seção, apresentamos os detalhes metodológicos das duas partes deste trabalho: na Subseção 5.1 cobrimos a etapa de variação e discernibilidade, enquanto na Subseção 5.2 relatamos a metodologia referente à etapa de classificação.

5.1. Metodologia de Investigação de Variação e Discernibilidade

Nossa análise de variação e discernibilidade englobou quatro abordagens distintas: grau de duplicação, distribuição de propriedades linguísticas, diferenças vocabulares e separabilidade em espaços de *embeddings* semânticos dos domínios estudados. Essa divisão foi feita para acomodar a natureza multifacetada dos domínios de discurso e prover uma análise multivariável das suas diferenças, ao molde das nossas referências de análise de registros (Seção 2). A seleção dessas abordagens baseia-se nas nossas expectativas acerca das diferenças linguísticas entre os domínios:

- diferenças léxicas e convencionais, como o uso de termos técnicos, linguagem formulaica e expressões fáticas influenciam diretamente o grau de duplicação e a distribuição vocabular de documentos em cada domínio;
- distinções na seleção de estruturas gramaticais levariam a diferenças morfológicas e sintáticas, evidentes por meio da análise das características morfossintáticas dos textos;
- diferenças nos temas e tópicos abordados nos documentos, que influenciariam sua distribuição vocabular, mas também seu conteúdo semântico geral. Essa diferença, por sua vez, poderia ser detectada com o emprego de uma análise de separabilidade em espaços de *embeddings* semânticos.

Como declarado anteriormente, o foco deste trabalho é a distinção de domínios discursivos por meio de uma **abordagem computacional**. Nossa análise é, então, consistentemente mediada por ferramentas computacionais, a saber, *Onion* (Seção 6.1), *spaCy* (Seção 6.2) e *NILC embeddings* (Seção 6.4).

Visando ter dados apropriados para a análise de discernibilidade, criamos uma versão menor e balanceada em domínios do *corpus* Carolina, denominada Carol·$\mathcal{B}$.[4] O Carol·$\mathcal{B}$ contém um número semelhante de *tokens* para cada um dos maiores domínios do Carolina: *Instrucional*, *Jurídico*, *Entretenimento*, *Jornalístico* e *Fórum Virtual*. No total, o *subcorpus* contém 304.205.653 *tokens*, aproximadamente $60,8$ milhões de *tokens* por domínio.

Para construir o *subcorpus*, fizemos uma amostragem aleatória de documentos de diferentes domínios até atingirmos o número de *tokens* do menor domínio selecionado (*Fórum Virtual*). A amostragem foi realizada de modo a manter o equilíbrio de tipologias declaradas na fonte dentro de cada domínio.

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

[4] Os links com todos os dados e códigos-fonte desenvolvidos para esse estudo estão disponíveis nesta lista: https://github.com/stars/frserras/lists/domain-discernibility-carolina

Domínio	Sentenças de Exemplo
Instrucional	"Iniciou-se como escritora ao publicar ''13 Contos de Sobressalto'' (1981), e daí em diante escreveu contos, romances e teatro." "Durante a era soviética a cidade abrigou pessoas de toda a União Soviética."
Jurídico	"Por isso a importância de um julgamento como esse, para trazer luz cada vez mais." "Destarte, nego provimento ao recurso."
Entretenimento	"Beth finalmente concordou com o divórcio." "O Griffin também armou um esquema para eliminar o querido primo Herman."
Jornalístico	"As maiores baixas foram registradas pelas ações dos bancos." "Ele já cuidava de um ninho no muro de sua casa e entrou em contato com o grupo para saber como salvar três enxames do muro da casa de um amigo que ia fazer uma reforma."
Fórum Virtual	"Só curiosidade , nem sei se existe página assim. obrigado!" "não dancei funk :("

Tabela 2: Sentenças extraídas dos cinco principais domínios do *corpus* Carolina.

Uma outra versão derivada do *corpus*, balanceada e parcialmente deduplicada, chamada de Carol·($\mathcal{D}$+$\mathcal{B}$) também foi produzida como produto da análise de nível de duplicação dos diferentes domínios.[5]

Esses dados foram utilizados de forma transversal às abordagens do nosso estudo de discernibilidade. Os procedimentos utilizados na análise de cada uma das abordagens e os resultados obtidos por eles estão descritos nas Seções 6.1, 6.2, 6.3 e 6.4.

5.2. Metodologia de Classificação Automática

Na etapa de classificação automática, realizamos o refinamento de quatro modelos pré-treinados, baseados na arquitetura *Transformer* (Vaswani et al., 2017), além do treinamento de modelos *Naive Bayes* e *Support Vector Machine (SVM)* (Jurafsky & Martin, 2009) para fins de comparação.

Apesar de na primeira parte utilizarmos propriedades linguísticas explicáveis para caracterizar a variação entre domínios, treinamos modelos baseados em *Transformers* para desenvolver os classificadores, que recebem os textos brutos e não as *features* estudadas. Fazemos isso, porque, apesar da riqueza e explicabilidade dessas *features* para fins de análise, o conjunto ainda é muito pequeno para permitir uma boa classificação dos textos. Entretanto, na Seção 7.1, conectamos as duas análises e levantamos hipóteses acerca de como a *performance* dos modelos se correlaciona aos resultados de diferenciabilidade que obtivemos na primeira etapa.

Mantivemos a abordagem orientada à sentença da primeira parte do trabalho e modelamos o problema como uma classificação de sentenças, como ilustrado na Figura 1. A escolha torna a tarefa mais desafiadora, pois as marcações linguísticas de registro são mais sutis nesse nível, testando diretamente a diferenciabilidade dos domínios.

A partir do Carol·($\mathcal{D}$+$\mathcal{B}$) , produzimos um *dataset* de classificação de domínios em nível de sentença, nomeado *carol-domain-sents* . Cada exemplo do *dataset* consiste de uma sentença com comprimento entre 4 e 256 caracteres UTF-8, anotada com o respectivo domínio discursivo. O conjunto possui, ao todo, 4.531.553 sentenças, distribuídas entre os domínios de acordo com a Tabela 3.

A escolha de usar o *subcorpus* balanceado na criação do *dataset* de classificação se deve à intenção de equilibrar a representação entre os domínios no treinamento dos classificadores. Além de tornar a tarefa computacionalmente mais desafiadora, essa decisão se justifica pelo fato de que a distribuição do *corpus* Carolina não reflete a distribuição real dos domínios discursivos na *web*. Isso ocorre porque sua construção prioriza a coleta de dados abertos com licenças

[5] A metodologia de criação do Carol·($\mathcal{D}$+$\mathcal{B}$) será apresentada na Seção 6.1.

permissivas, em vez de buscar representar fielmente a distribuição original dos dados, como faz, por exemplo, o *dataset* CORE (Biber & Egbert, 2018; Laippala et al., 2021). Um exemplo claro dessa discrepância é a predominância de documentos do domínio *Jurídico* no Carolina: embora menos comuns na *web* do que textos de fóruns ou redes sociais, eles são mais acessíveis do ponto de vista legal e, por isso, mais presentes no *corpus*. Assim, o viés introduzido pela distribuição do Carolina não favoreceria um classificador aplicado a dados reais. Optamos, então, por uniformizar a distribuição, tentando garantir, ao menos, que os classificadores treinados aprendam a distinguir os diferentes domínios com a mesma capacidade.

Para fins de treinamento, o conjunto de dados *carol-domain-sents* foi particionado em três subconjuntos: busca de hiperparâmetros (*hps*), treinamento (*train*) e avaliação (*test*) de forma aleatória conforme apresentado na Tabela 3. A aleatorização foi aplicada sem nenhuma imposição para balancear a distribuição de sentenças dos diferentes domínios entre os subconjuntos da partição, entretanto como a amostragem foi uniforme e o *carol-domain-sents* foi construído a partir do Carol·$(\mathcal{D}+\mathcal{B})$, que já havia sido balanceado, o resultado final apresentou bom balanceamento. Uma limitação da estratégia empregada é não garantir que sentenças do mesmo documento de origem sejam designadas para o mesmo subconjunto. Contudo, acreditamos que a robustez dos resultados é assegurada pelo grande volume de textos e de sentenças utilizadas no experimento, que mitiga possíveis problemas.

Os dados do *carol-domain-sents* foram utilizados para o treinamento de cinco modelos, detalhados a seguir.

O modelo *Naive Bayes*, escolhido como referência-base para comparação, aproxima uma distribuição multinomial com 30.000 (trinta mil) variáveis. Cada variável correspondendo a um *token* do *dataset*, ignorando-se capitalização. Os *tokens* selecionados como variáveis são os mais frequentes, desde que apareçam em pelo menos cinco exemplos e em no máximo 95% dos exemplos. Esse modelo foi treinado utilizando a partição *train* completa. Um cenário análogo foi adotado para o modelo *SVM*.

Também refinamos quatro modelos baseados na arquitetura *Transformer*: *bert-base*, *bert-large*, *albertina-100m* e *albertina-900m*. Os dois primeiros correspondem a variações do modelo *BERTimbau* (Souza et al., 2020), enquanto que os dois últimos correspondem a modelos pré-treinados em português brasileiro da família *Albertina* (Rodrigues et al., 2023), baseada no modelo *DeBERTa* (He et al., 2021).

Antes do treinamento/refinamento propriamente dito, realizamos uma busca de hiperparâmetros com a biblioteca Optuna (Akiba et al., 2019) sobre a partição *hps*, dividindo-a em 90% para aprendizado e 10% para validação. Para os modelos *Transformer*, realizamos uma busca para a taxa de aprendizado, taxa de decaimento e taxa de aquecimento, onde o modelo é treinado por três épocas sobre cada combinação. Já para os modelos *baseline* buscamos otimizar os parâmetros para a construção do TF-IDF (Sparck Jones, 1972) (frequência mínima para inclusão, frequência máxima para exclusão, n-gramas considerados) e o hiperparâmetro respectivo do modelo, α no caso do modelo *Naive Bayes* e a regularização α do modelo *Support Vector Machine* (Cortes & Vapnik, 1995). Para cada modelo foram verificadas cem combinações de hiperparâmetros escolhidas com o *Tree-Structured Parzen Estimator* (Watanabe, 2025).

Selecionamos os valores de hiperparâmetros para os quais o modelo sendo avaliado alcançou sua melhor acurácia de validação.[6] A Tabela 4 apresenta os hiperparâmetros buscados e seus respectivos intervalos de busca,[7] enquanto as Tabelas 5 e 6 apresentam os valores de hiperparâmetros ótimos obtidos para os modelos *Transformer* e *baselines* respectivamente.

Com esses valores de hiperparâmetros, realizamos o refinamento dos modelos baseados em *Transformer* na partição *train*, utilizando 99.9% dos exemplos para o treinamento em si e 0.01% para validação. Todos os modelos base foram obtidos a partir da biblioteca *Hugging Face* (Wolf et al., 2020) e foram treinados com a biblioteca PyTorch (Ansel et al., 2024), utilizando precisão mista (*float32*, *bfloat16*) com bateladas (*batches*) de 32 exemplos, por cinco épocas. Os modelos *bert-base*, *bert-large* e *albertina-100m* foram treinados com duas placas gráficas NVIDIA GeForce RTX 3090, enquanto que o modelo *albertina-900m* utilizou duas placas gráficas NVIDIA Tesla A100. Para cada modelo, das suas versões ao final de cada época, selecionou-se aquela com acurácia de validação máxima como representante final e avaliou-se sua performance na partição *test*.

[6] Independente da época em que o melhor resultado foi alcançado para os modelos *Transformer*.

[7] O parâmetro "taxa de decaimento" se refere à taxa de decaimento dos parâmetros do otimizador AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9, \beta_2 = 0.999$).

	Número de Sentenças			
Domínio/Conjunto	***carol-domain-sents***	***train***	***hps***	***test***
Jurídico	906.409	713.635	14.371	178.403
Jornalístico	906.271	713.604	14.233	178.434
Instrucional	906.336	713.136	14.298	178.902
Fórum Virtual	906.346	714.235	14.308	177.803
Entretenimento	906.191	713.542	14.153	178.496
Total	4.531.553	3.568.152	71.363	892.038

Tabela 3: Distribuição de sentenças por domínio para cada conjunto de dados.

Modelos	**Parâmetro**	**Intervalo**	**Passo**
Transformers	Tx. aprendizado	10^{-6} a 10^{-4}	log
	Tx. decaimento	0 a 0.1	0.05
	Tx. aquecimento	0 a 0.2	0.01
Baselines	n-grama	1 a 3	1
	Freq. mín.	1 a 10	1
	Freq. máx.	0.5 a 0.95	0.05
	α (*svm*)	10^{-6} a 10^{4}	log
	α (*naive-bayes*)	10^{-2} a 1	log

Tabela 4: Espaço de busca dos hiperparâmetros

Modelo	Aprendizado	Decaim.	Aquecim.
bert-base	2.1×10^{-5}	0.00	0.05
bert-large	5.4×10^{-6}	0.08	0.05
albertina-100m	6.9×10^{-6}	0.07	0.00
albertina-900m	3.6×10^{-6}	0.03	0.20

Tabela 5: Taxas de aprendizado, decaimento e aquecimento ótimas segundo a busca de hiperparâmetros para cada modelo *Transformer*.

6. Discernibilidade Computacional de Domínios Discursivos

Nesta seção, apresentamos os procedimentos e resultados obtidos ao avaliar o nível de discernibilidade computacional e a natureza da variação linguística entre os cinco domínios discursivos estudados. Nossa análise se divide em quatro abordagens: por meio do grau de duplicação, por meio da extração automática de propriedades linguísticas, por meio de diferenças vocabulares e por meio da separabilidade em espaços de *embeddings* semânticos, apresentadas nas Seções 6.1, 6.2, 6.3 e 6.4, respectivamente.

6.1. Discernibilidade por Meio do Grau de Duplicação

Nossa abordagem para avaliar o grau de duplicação textual entre os documentos de um domínio consistiu em aplicar uma ferramenta de deduplicação automática e realizar uma análise sobre os resultados obtidos. Entendemos a *deduplicação* como o processo de remoção de conteúdo não original de um *corpus* e, consequentemente, *deduplicado* é um texto ou um *corpus* após a realização da deduplicação.

Modelo	n-grama	Freq. mín.	F. máx.	α
svm	(1, 2)	2	0.5	$2.97e-5$
naive-bayes	(1, 2)	2	0.6	$5.77e-2$

Tabela 6: Valores ótimos obtidos bela busca de hiperparâmetros para os modelos *Baseline*.

A ferramenta de deduplicação escolhida foi o *Onion* (*ONe Instance ONly*) (Pomikálek, 2011).[8] *Onion* é uma ferramenta que determina se cada texto em um *corpus* está totalmente ou parcialmente duplicado e remove as duplicatas. Um limite de conteúdo duplicado $\mathcal{T} \in [0, 1]$ pode ser fornecido como parâmetro, $\mathcal{T} = t$ significando que somente os documentos com $10t\%$ ou mais de n-gramas repetidos serão considerados duplicados e, consequentemente, removidos. Para classificar um n-grama como repetido, o *Onion* compara os n-gramas dos textos com uma lista de n-gramas encontrados anteriormente, durante o processamento do *corpus*. Dessa forma, a deduplicação é dependente da ordem em que os documentos são processados.

Utilizamos o *Onion* para comparar a taxa de remoção de documentos duplicados em cada domínio do *corpus*, e as configurações padrão para todos os parâmetros, exceto para $\mathcal{T}$. Repetimos o processo de deduplicação cinco vezes, sendo que em cada uma os documentos estavam aleatoriamente ordenados. Para cada domínio, calculamos, ao longo dessas ordenações, a média e a variância da *densidade de tokens removidos* ($\mathcal{D}$) para diferentes valores da *originalidade mínima requerida* ($\mathcal{O}$) para que um texto seja mantido, definidas nas Equações 1 e 2, respectivamente (ver Figura 2).

[8] *Onion* está disponível em: `https://corpus.tools/wiki/Onion`.

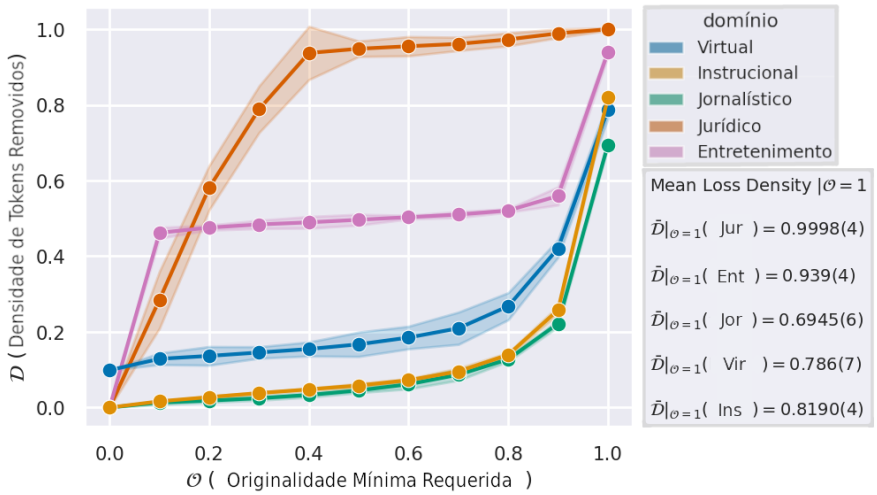


Figura 2: Curvas de densidade de perda de *token* por domínio.

$$\mathcal{D} = \frac{\#\ \textit{tokens}\ \text{removidos}}{\#\ \textit{tokens}\ \text{no domínio}} \tag{1}$$

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

Ao analisar o comportamento das curvas da Figura 2, fica claro que alguns domínios são mais suscetíveis a mudanças em $\mathcal{O}$, por exemplo os domínios *Jurídico* e *Entretenimento*. Isso provavelmente decorre da natureza destes domínios.

Os textos jurídicos podem ser muito semelhantes em sua estrutura e contêm uma linguagem mais padronizada e repetitiva, como vemos nos seguintes exemplos: "`Diante do exposto, nego provimento ao recurso ordinário em habeas corpus.`" e "`Diante do exposto, nego provimento ao recurso por manifesta improcedência...`".

Já os textos de entretenimento do Carolina são legendas de filmes infantis e séries de TV que, muitas vezes, fazem uso de linguagem repetitiva e simplificada, com sobreposição temática entre episódios da mesma série, como observa-se em excertos como "`Precisamos da sua ajuda, Cloud.`" e "`Não, não precisamos de sua ajuda!`".

Os domínios *Entretenimento* e *Jurídico* também contêm os maiores documentos, portanto, quando o *Onion* lista e compara os n-gramas, as médias maiores de *tokens* por documento provavelmente afetam as taxas de deduplicação.

Duas variáveis de grande interesse são as densidades de *tokens* removidos quando a originalidade exigida é mínima $\mathcal{D}|_{\mathcal{O}=0}$ e máxima $\mathcal{D}|_{\mathcal{O}=1}$. Elas representam a densidade de documentos que são completamente duplicados e a densidade de documentos que são totalmente originais, respectivamente. Os valores de $\mathcal{D}|_{\mathcal{O}=1}$ para cada domínio também são mostrados na Figura 2.

Os domínios *Jurídico* e *Entretenimento* têm $\mathcal{D}|_{\mathcal{O}=1}$ mais alto, o que segue os padrões de comportamento mencionados anteriormente. O domínio *Instrucional* tem o terceiro maior $\mathcal{D}|_{\mathcal{O}=1}$. Isso também pode ser explicado pelo fato de que alguns textos enciclopédicos, que constituem grande parte dos textos nesse domínio, seguem um padrão mais estruturado.

Fórum Virtual é o único domínio com $\mathcal{D}|_{\mathcal{O}=0}$ significativo, mas outros domínios também têm algum nível de duplicação absoluta de acordo com o *Onion*. Na Tabela 7, exibimos o número absoluto de *tokens* removidos por domínio quando $\mathcal{O} = 0$ e o número equivalente de *tokens* removidos quando consideramos cópias exatas. Para fazer essa verificação, checamos quais dos textos marcados pelo *Onion* como $\mathcal{O} = 0$ eram de fato cópias exatas de outros textos do domínio. Os únicos domínios com cópias exatas, de acordo com essa verificação, são *Fórum Virtual* e *Jornalístico*. Notavelmente, o domínio *Fórum Virtual* contém o maior número de cópias exatas. Analisando as duplicatas, encontramos vários exemplos de linguagem fática e textos funcionais, por exemplo, *tweets* de saudação no domínio *Fórum Virtual* — como em "`Boa noite (ou madrugada, né?)`" — e, no domínio *Jornalístico*, postagens notificando os leitores de que uma coluna não seria publicada naquele dia, como em "`Excepcionalmente hoje a coluna não será publicada.`".

A ordem aleatória dos textos teve um impacto mínimo nos resultados, evidente na sutil variação indicada pelo sombreamento mais claro em cada linha do gráfico. Os domínios *Jurídico*

e *Fórum Virtual* apresentaram maior variação, mas há padrões consistentes discerníveis nas curvas, ressaltando a robustez do *Onion*.

A Figura 2 e nossa análise apontam, como esperado, para a discernibilidade dos domínios com base nos graus de duplicação interna. Em parte, nossas conclusões são corroboradas por análises prévias da linguagem nesses domínios ou similares. Esse é o caso, por exemplo, do uso de expressões fixas e genéricas dentro da linguagem jurídica (Carapinha, 2018; de Faria Marques, 2023), entretanto nossa análise provê uma perspectiva nova e quantitativa sobre esses mesmos fenômenos.

Para facilitar aplicações futuras, inclusive o treinamento de modelos de classificação automática (Seção 7), desenvolvemos com base nesta exploração o Carol·($\mathcal{D}$+$\mathcal{B}$): *Subcorpus* Deduplicado e Balanceado do Carolina. O Carol·($\mathcal{D}$+$\mathcal{B}$) foi criado utilizando valores $\mathcal{T}$ variados para cada domínio: $\mathcal{T} = 0$ para *Instrucional*, $\mathcal{T} = 0.1$ para *Jornalístico*, $\mathcal{T} = 0.5$ para *Entretenimento* e $\mathcal{T} = 0.8$ para *Jurídico*. Esse processo produziu contagens de *tokens* de 62.766.935, 68.543.795, 60.880.758 e 81.863.020 por domínio, respectivamente. A metodologia descrita na Seção 5 foi, então, reaplicada para construir outro *subcorpus* equilibrado que incorporasse, agora, os domínios deduplicados.

6.2. Discernibilidade por meio de propriedades linguísticas

O uso de propriedades linguísticas nos dá um arcabouço bem estabelecido e ancorado teoricamente para examinar a variação entre domínios. Ao mesmo tempo, a anotação automática de tais propriedades nos permite fazê-lo em larga escala, aplicando técnicas estatísticas de análise a grandes volumes de dados. Propriedades desse tipo também são vastamente utilizadas na literatura de análise de registros, que nos serve de inspiração, como apontado nas Seções 2 e 3.

Além disso, estamos interessados aqui na discernibilidade e na variação entre domínios discursivos do ponto de vista **computacional**. Embora nem todas as propriedades linguísticas relevantes possam ser garantidamente computáveis em sentido estrito, ao partirmos de características linguísticas para as quais já existem modelos computacionais capazes de anotá-las com níveis satisfatórios de precisão, podemos garantir que as conclusões que extraímos sobre a discernibilidade de domínios do discurso refletem o potencial dessas características de serem capturadas por sistemas computacionais em contextos práticos de aplicação.

Para a anotação de propriedades linguísticas, usamos os modelos pré-treinados do pacote *spaCy*.[9] São modelos estado-da-arte para o português, que permitem a extração de um conjunto diversificado de características linguísticas a partir dos textos. Formalmente, definimos uma propriedade linguística $\mathcal{F}_j$ como na Equação 3, onde $\mathbb{U}$ é um conjunto de unidades de texto sobre as quais $\mathcal{F}_j$ é calculada (por exemplo, palavras, n-gramas, sentenças), $\mathbb{F} = \{f_i\}$ é o conjunto de valores f_i que $\mathcal{F}_j$ pode assumir e $c_i \in 2^{\mathbb{U}}$ é o contexto de anotação. Um modelo é, então, uma aproximação computável $\hat{\mathcal{F}}_j$ de $\mathcal{F}_j$. As propriedades usadas em nossa análise e seus respectivos conjuntos $\mathbb{U}$ e $\mathbb{F}$ estão representados na Tabela 8, definições específicas e exemplos para cada propriedade são fornecidos ao longo do texto.

As propriedades avaliadas foram previamente selecionadas visando oferecer uma panorama abrangente da variação linguística entre domínios, em acordo tanto com nossas expectativas de variação, apresentadas na Seção 5.1, como com o que os trabalhos de Ferguson (1994) e Biber (1994) postulam como sendo propriedades passíveis de variação segundo registro.

Embora algumas propriedades tenham sabidamente baixo potencial de diferenciação, elas podem pertencer a grupos de propriedades passíveis de variação e, nesses casos, consideramos importante submetê-las a verificação empírica e reportar os resultados obtidos, sejam positivos ou negativos, a fim de sustentar ou refutar tais suposições com base em evidências. Trabalhos como Biber (1988) e Sardinha et al. (2014b) fazem o mesmo, ao selecionar conjuntos amplos de *features*, reportando posteriormente quais apresentaram ou não papeis significativos para os fenômenos de variação sendo estudados.

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

Devido ao tamanho do *corpus* e dos modelos, analisamos uma amostra $\mathcal{S}$ de 1% do Carol·$\mathcal{B}$. Para as propriedades numéricas, isto é, aquelas para as quais $\mathbb{F} = \mathbb{N}$, as estatísticas foram obtidas a partir da agregação de todo o conjunto $\mathcal{S}$. Para as demais características, aplicamos uma técnica de particionamento: $\mathcal{S}$ foi dividido em 10 partições s_l e a distribuição dos valores de cada propriedade $\mathcal{F}_j$ foi obtida a partir do cômputo independente da propriedade-alvo em cada partição s_l, para cada domínio D_k.

[9] https://spacy.io/

	***Tokens* removidos (*Onion*)**	***Tokens* removidos (Duplicatas exatas)**
Instrucional	519	0
Entretenimento	1.139	0
Journalístico	4.913	1.019
Jurídico	0	0
Fórum Virtual	6.002.616	32.322
Total	**6.009.187**	**33.341**

Tabela 7: Comparação entre *tokens* removidos pelo *Onion* e duplicatas exatas.

Feature	$\mathbb{U}$	$\mathbb{F}$
Tokens por Sentença	Sentença	$\mathbb{N}$
Caracteres por *Token*	*Token*	$\mathbb{N}$
Stop Words por Sentença	Sentença	$\mathbb{N}$
Tokens por Sentença	Sentença	$\mathbb{N}$
Sinais de pontuação por Sentença	Sentença	$\mathbb{N}$
Número Morfológico	*Token*	{SING (Singular), PLUR (Plural), ∅ }
Caso Morfológico	*Token*	{NOM (Nominativo), DAT (Dativo), ACC (Acusativo), ∅ }
Gênero Morfológico	*Token*	{MASC (Masculino), FEM (Feminino), ∅ }
Tempo Morfológico	*Token*	{PRES (Presente), PAST (Passado), IMP (Imperfeito), FUT (Futuro), ∅ }
Modo Morfológico	*Token*	{SUB (Subjuntivo), IND (Indicativo), CND (Condicional), ∅ }
Classe de Entidade Nomeada	Sequência de *Token*	{ORG (Organização), MISC (Miscelânea), LOC (Localização), PER (Pessoa), ∅ }
Parte do discurso	*Token*	{SCONJ (Conj. Subordinativa), VERB (Verbo), PROPN (Nome Próprio), PRON (Pronome), CCONJ (Conj. Coordenativa), ADV (Advérbio), AUX (Verbo Auxiliar), ADJ (Adjetivo), DET (Determinante), NOUN (Substantivo), ADP (Preposição), INTJ (Interjeição), NUM (Numeral), X (Outro/Indefinido), PUNCT (Pontuação), SYM (Símbolo)}

Tabela 8: Propriedades linguísticas avaliadas neste trabalho.

Para cada propriedade $\mathcal{F}_j$, calculamos a probabilidade média sobre as partições s_l de $\mathcal{F}_j$ ser f_i se o domínio do discurso for D_k, representada por $\bar{\mathcal{P}}_j(f_i|D_k)$ e definida na Equação 4. Usamos o erro padrão $\sigma_j(f_i|D_k)$ como o erro correspondente.

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

Para diferenciação entre os domínios, comparamos $(\bar{\mathcal{P}}_j(f_i|D_k), \sigma_j(f_i|D_k))$ para cada par de domínios de discurso distintos. Executamos o teste T de *Student* para cada par e relatamos apenas as diferenças entre pares de domínios em que o valor p associado ao teste é $p \leq 0,03$, ou seja, relatamos apenas os casos em que a confiança da diferença entre os domínios é maior que 97%.[10]

[10]Para permitir a visualização, exibimos um subconjunto representativo das distribuições, estando gráficos complementares disponíveis nos nossos repositórios.

Este processo nos permitiu concluir que várias das propriedades avaliadas são distintivas entre domínios, bem como observar a natureza de sua variação. A seguir apresentamos os resultados obtidos, segmentando-os entre propriedades numéricas e categóricas.

Propriedades numéricas ($\mathbb{F} = \mathbb{N}$)

Esse conjunto de propriedades demonstra, de forma consistente, diferenças perceptíveis entre os domínios. Especificamente, os documentos *jurídicos* apresentam maior comprimento médio (Figura 3), empregam palavras maiores (Figura 4) e contêm um número maior de sinais de pontuação e *stop words* (Figura 5) por sentença. Com relação ao valor médio das propriedades numéricas, o domínio *Jurídico* é seguido por *Jornalístico* e *Instrucional* — que alternam entre si nas posições seguintes — e, após estes, pelos domínios *Fórum Virtual* e *Entretenimento*, que apresentam as médias mais baixas.

O padrão recorrente observado, em que as propriedades exibem sistematicamente uma determinada ordem dos domínios, sugere uma estrutura hierárquica entre esses domínios. Uma maneira possível de explicar esse comportamento é em termos de "complexidade da linguagem": textos jurídicos usam uma linguagem mais complexa, resultando em palavras e frases mais longas. Por outro lado, os textos de entretenimento tendem a empregar construções mais simples, resultando em valores menores de propriedades numéricas.

De fato, resultados advindos de métricas globais de complexidade apontam para a complexidade informacional, ou estrutural, como sendo um fator de distinção significativo entre textos de registros diferentes (Ehret, 2021). Entretanto, o conceito de complexidade é um conceito amplo que pode abarcar uma gama diversa de fenômenos, como extensão textual (Sarti et al., 2021), densidade informacional (Ehret, 2021), quantidades de operações cognitivas durante emissão ou consumo da linguagem (Leal et al., 2022), flexibilização na ordenação dos elementos que compõe a sentença (Bakker, 1998; Sadeniemi et al., 2008). Na subseção 6.5, realizaremos uma análise comparativa com outros trabalhos visando especificar quais fenômenos específicos estão associados à ordenação observada.

Propriedades Categóricas

Além das propriedades numéricas, analisamos também uma série de propriedades categóricas *token* a *token*, divididas em propriedades morfológicas — **Tempo**, **Caso**, **Modo**, **Número** e **Gênero** —, **etiquetas de parte do discurso** e **classes de entidades nomeadas**.

No que tange ao **tempo verbal** (Figura 6), nos documentos dos domínios *Fórum Virtual* e *Entretenimento* o uso do tempo *presente* é mais predominante, nos textos dos domínios *Instrucional* e *Jornalístico* o tempo verbal *passado* é mais usado se comparado aos demais domínios, já no domínio *Jurídico*, o tempo futuro apresenta proeminência relativa.

Os domínios que na análise anterior das propriedades numéricas foram associados a uma menor complexidade linguística, se destacam no uso do tempo *presente*. Essa observação sugere uma correlação: os domínios com estruturas de frases mais simples normalmente empregam formações verbais mais simples, ou mais próximas do prototípico tipológico para aquela categoria gramatical (Croft, 2002). Especificamente, no Carol·$\mathcal{B}$, onde os domínios *Fórum Virtual* e *Entretenimento* consistem principalmente em *tweets* e legendas respectivamente, a frequência do tempo presente pode ser racionalizada pela natureza desses textos, que se concentram principalmente em eventos atuais, estando muitas vezes associados à simultaneidade entre produção e consumo da linguagem. Um exemplo é a sentença a seguir, do domínio *Fórum Virtual*: "`Obrigada por ajudar a controlar as pessoas que editam a discografia da Britney!`".

Por outro lado, a predominância relativa do pretérito nos textos *instrucionais* e *jornalísticos* se alinha com sua característica de relatar eventos do passado, como vemos em "`Após meia hora de show, Pitty tirou seu quimono, tomou uma taça de vinho e entoou composições de batida mais lenta, como ‘‘Me Adora’’ e ‘‘Na Sua Estante’’.`" (*Jornalístico*). Por fim, o uso do tempo verbal futuro em textos *jurídicos* pode ser atribuído à natureza prescritiva das decisões judiciais, que geralmente ditam condições e ações a serem seguidas no futuro, como em "`A CONTRATADA deverá providenciar todos os materiais, equipamentos e acessórios necessários à condução da pré-operação.`".

A Figura 7 representa a distribuição da categoria **caso** através dos domínios. O modelo diferencia nominativo, acusativo e dativo. A análise deve ser feita com ressalvas, já que o caso é uma categoria de uso reduzido no português, onde substantivos não variam em caso, deixando a variação para pronomes, que assumem o caso reto, correspondente ao nominativo, como em "Eu

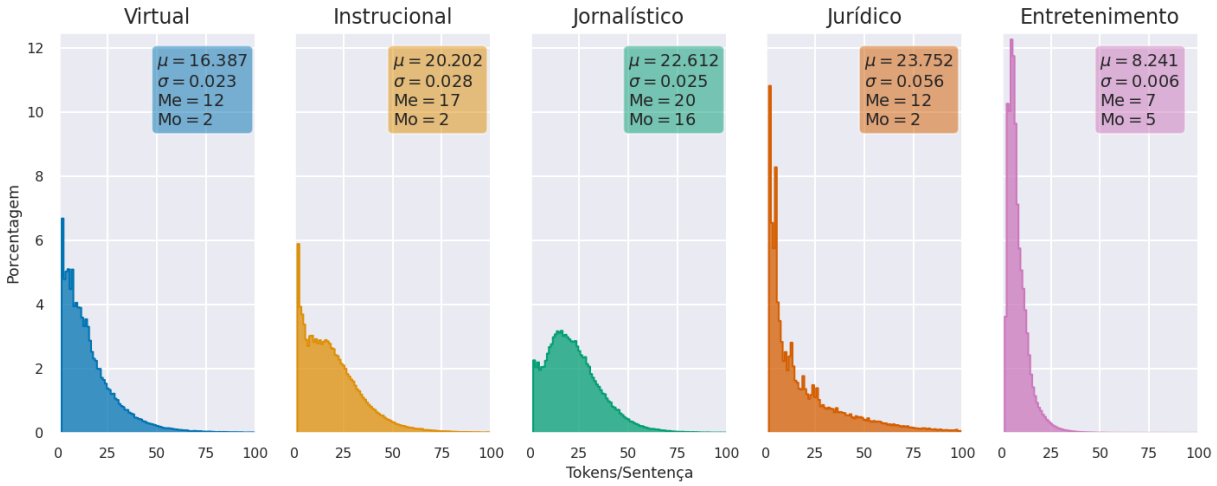


Figura 3: Distribuição do comprimento de sentenças por domínio e respectivas média (μ), erro (σ), moda (Mo) e mediana (Me).

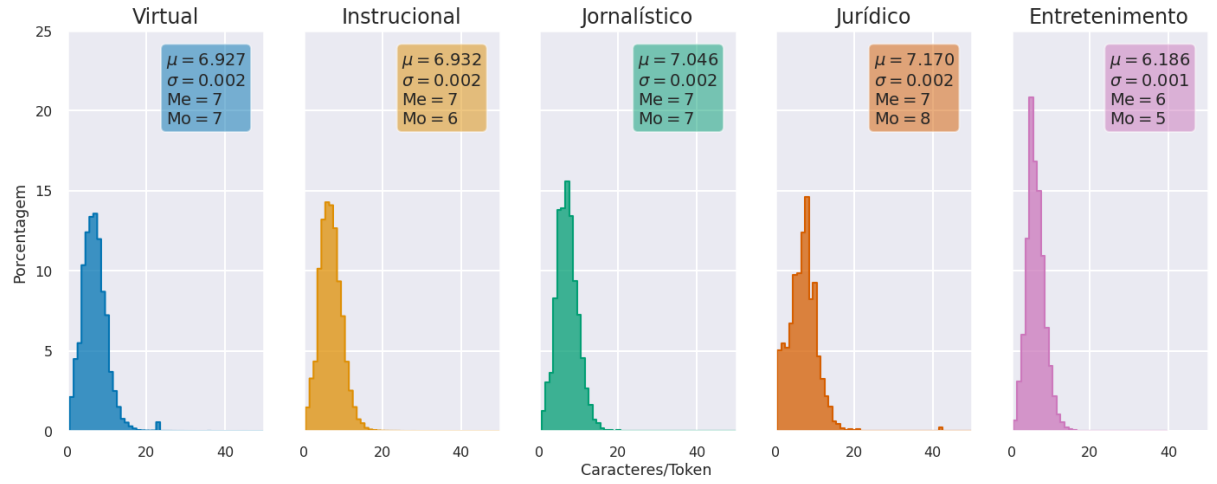


Figura 4: Distribuição do número do comprimento dos *tokens* por domínio e respectivas média (μ), erro (σ), moda (Mo) e mediana (Me).

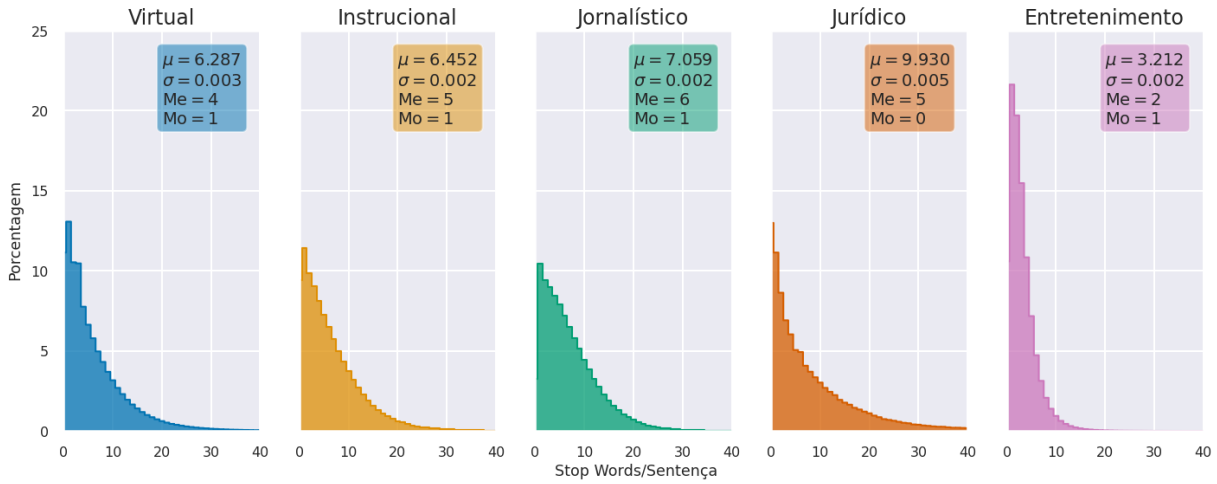


Figura 5: Distribuição da contagem de *stop words* por sentença em cada domínio e respectivas média (μ), erro (σ), moda (Mo) e mediana (Me).

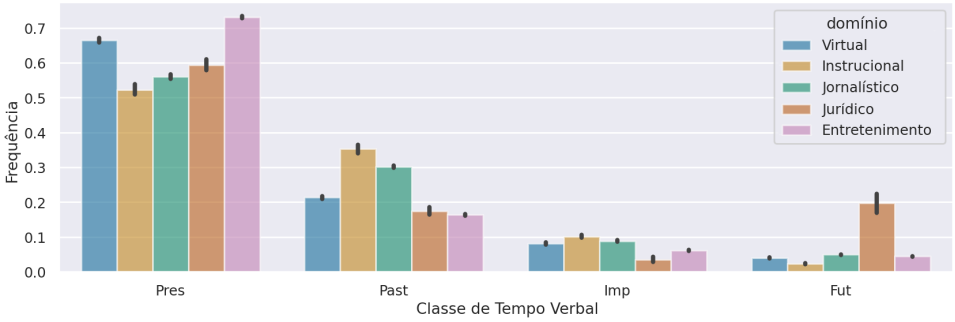


Figura 6: Distribuição das classes de tempo verbal através dos domínios.

`disse pra correr`", ou o caso oblíquo, correspondente ao acusativo, como em "`Você veio por mim`". O dativo, é ainda menos estabelecido na nossa língua, podendo ser associado a objetos indiretos e complementos nominais via preposições, como "a" e "para" e contrações, como em "`Gil interroga Miguel a respeito de que trabalho deu para Anya`."

Nos domínios *Fórum Virtual* e *Entretenimento*, o uso do caso nominativo é dominante, enquanto os domínios *Jurídico* e *Instrucional* usam principalmente o acusativo. Aqui, os textos *Jornalísticos* apresentam um equilíbrio relativo de uso entre os casos. O dativo é sistematicamente menos utilizado em todos os domínios e apresenta pouco valor distintivo. Como para o tempo verbal, a separação observada entre os domínios no que tange ao caso parece consistente com a ordenação observada anteriormente.

As outras propriedades morfológicas avaliadas — **gênero**, **número** e **modo** — não são visivelmente distintas entre os domínios de discurso. Isso ocorre, provavelmente, porque os gêneros das palavras são, em sua maioria, arbitrários, com pouca carga semântica. Da mesma forma, embora o número das palavras transmita significado, não há razão clara para esperar que um determinado domínio se refira a mais entidades plurais do que outro.

A distribuição geral das **etiquetas de parte do discurso**, ou *Part-of-Speech (PoS) tags*, está ilustrada na Figura 8. A Tabela 8 contém o significado de cada *POS tag* utilizada. Exemplos complementares podem ser encontrados na Tabela 16, nos apêndices. Para a maioria das categorias, quando ordenamos os domínios discursivos pela importância relativa da categoria, a ordem observada dos domínios é *Jurídico, Instrucional, Jornalístico, Fórum Virtual, Entretenimento* ou o exato oposto. Em alguns casos, *Jurídico* e *Instrucional* são trocados, mas apenas quando não são discerníveis usando o teste T, ou seja, mesmo nesses casos, o padrão descrito ainda é estatisticamente compatível com essa sequência.

Essa ordenação é respeitada pelas seguintes categorias: *SCONJ*, *VERB*, *PRON*, *CCONJ*, *ADV*, *AUX*, *ADJ*, *ADP* e *NUM*. Ignorando as *PoS tags* que são muito pouco representadas no conjunto de dados (*INTJ*, *X*, *PUNCT*, and *SYM*), as únicas exceções sendo *DET*, *PROPN* e *NOUN*, No caso da classe *NOUN*, entretanto, a ordem relativa dos domínios é equivalente à observada para muitas das propriedades numéricas globais das sentenças.

Essas ordens são compatíveis com a escala observada nas propriedades anteriores, sugerindo que, de fato, os domínios do discurso no *corpus* Carolina seguem um tipo de espectro. No entanto, as *PoS tags* nos permitem ir além da interpretação em termos de complexidade de linguagem, relacionando essa ordenação ao modo discursivo (Gregory, 1967), ou seja, à distinção entre o enunciado escrito e falado. De fato, distinções de modo são fundamentais em praticamente todos os sistemas de classificação de registros linguísticos elencados na Seção 2, e resultados como os de Ehret (2021) têm mostrado uma relação quantitativa clara entre complexidade linguística e modo discursivo. Vários dos modelos de variação em registro, mencionados na Seção 3, como (Biber, 1988) e (Sardinha et al., 2014b), identificam o modo como um elemento essencial de distinção e variação em registro.

Além das etiquetas de parte do discurso e das categorias morfológicas, extraímos também informações acerca das **classes de entidades nomeadas** presentes nos diferentes domínios. O termo "entidade nomeada" denota referentes extralinguísticos mencionados nos textos, podendo ser empresas, instituições, pessoas, lugares, datas, entre outros. Sistemas de extração de informação a partir de textos detectam e classificam essas entidades em classes úteis. No nosso caso, o modelo utilizado diferencia referências às classes *pessoa*, *lugar*, *organização* e *outro*.

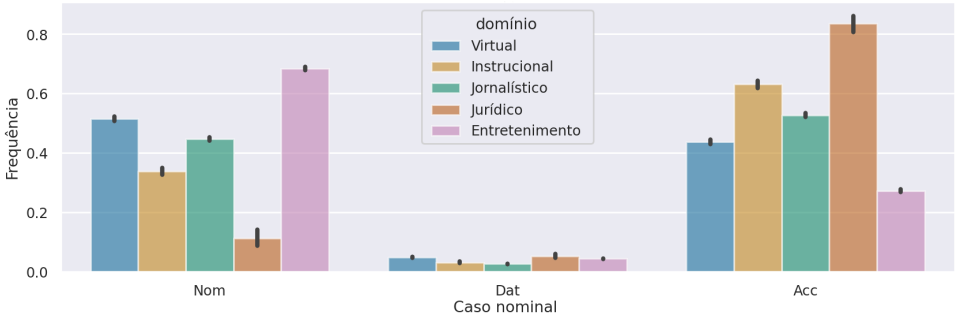


Figura 7: Distribuição das classes de caso nominal através dos domínios.

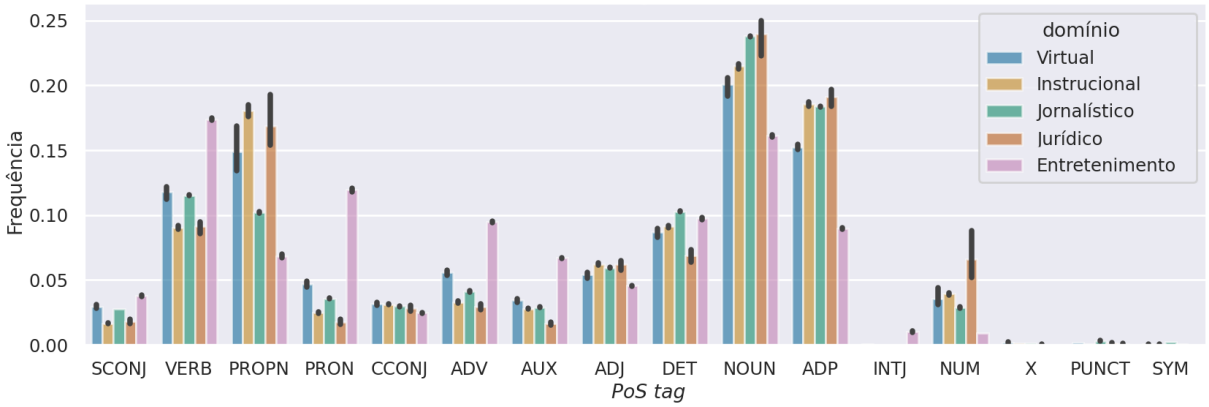


Figura 8: Distribuição de *PoS tags* por domínio.

Como domínios discursivos são, como definimos na Seção 2, variedades relacionadas a campos de atividade humana, avaliar a natureza das menções a entidades externas que compõe esses campos pode ser especialmente interessante para sua caracterização.

De fato, nos resultados obtidos, representados na Figura 9, as classes de entidades nomeadas também apresentam distribuições distintas entre domínios. Os textos em *Entretenimento* mencionam predominantemente pessoas e têm poucas referências a lugares e organizações, em contraste com os textos jurídicos. Por outro lado, os textos jornalísticos enfatizam organizações e apresentam relativamente menos entidades nomeadas diversas. Os documentos instrucionais, em comparação, não se desviam notavelmente de outros domínios de discurso com relação a essa propriedade.

Na nossa amostra os textos de *Entretenimento* são oriundos de legendas de séries e filmes, o que pode explicar a predominância de nomes de pessoas, que parecem ser, em geral, de personagens, como em "`Peterson disse-me sobre você`".

Como mencionado, os textos jurídicos são marcados pelo uso de estruturas fixas e genéricas (Carapinha, 2018; de Faria Marques, 2023). Isso pode explicar a predominância de lugares, já que menções ao local em que o documento é produzido ou no qual a ação se desenrola são marcações comuns nessas estruturas. Instituições oficias do âmbito judiciário, por sua vez, também são largamente mencionadas nesse tipo textual. O seguinte exemplo ilustra ambos os fenômenos: "`Ementa e Acórdão Supremo Tribunal Federal Inteiro Teor do Acórdão - Página 2 de 10 RCL 2517 AGR / RJ Brasília, 05 de agosto de 2014.`".

No domínio *Jornalístico*, o enfoque para organizações pode refletir o papel que organizações diversas tem no espaço público e a função do jornalismo de cobrir suas ações, como em "`Segundo a PM, o rapaz afirmou que foi abordado por um homem, que exigiu o celular.`"

Isso encerra a análise das propriedades linguísticas extraídas automaticamente. Essa análise mostra diferenças claras entre os domínios do discurso dentro da nossa amostra. Além disso, revela padrões intrigantes, oferecendo *insights* sobre a natureza subjacente dos domínios discursivos do português e seus arquétipos de variação.

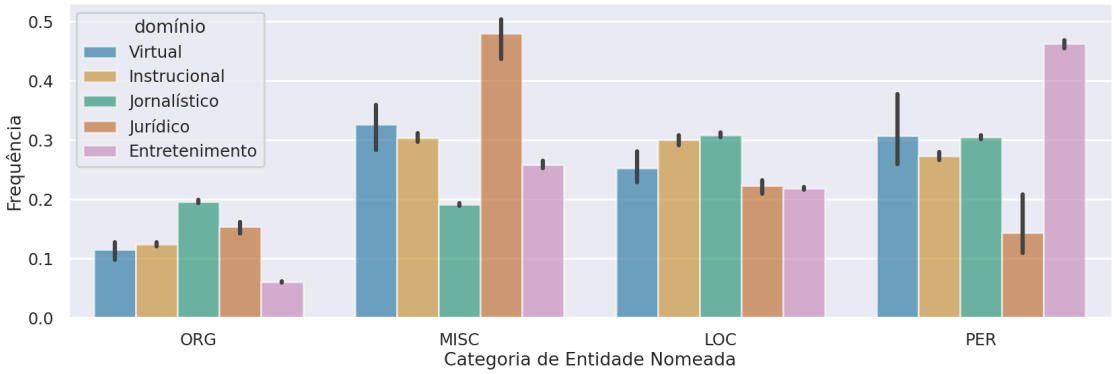


Figura 9: Distribuição das categorias de entidades nomeadas através dos domínios.

6.3. Discernibilidade por meio de Dados Vocabulares

Como apresentado na Seção 3, os primeiros trabalhos acerca da diferenciação de registros, como Biber (1988) e seus derivados imediatos, utilizam propriedades léxico-gramaticais mais similares às que analisamos na última seção. Entretanto, abordagens derivadas como a Análise Multidimensional Léxica (Sardinha & Fitzsimmons-Doolan, 2025) são testemunhos da capacidade do uso de distribuições lexicais na caracterização da variação intralinguística.

Nesta seção, seguimos este princípio, realizando uma análise exploratória das distribuições vocabulares dos diferentes domínios discursivos estudados, investigando os padrões de variação nessa dimensão e o quanto eles corroboram os resultados obtidos nas análises anteriores. Para tal, realizamos um levantamento das palavras mais frequentes em cada um dos domínios representados no Carol·$\mathcal{B}$. As dez mais frequentes por domínio estão representadas na Tabela 9. Realizamos esses levantamentos, tokenizando os textos, padronizando todas as palavras para letras minúsculas e removendo sinais de pontuação e *stop words* em português, de modo a obter uma lista representativa do vocabulário por domínio.

É interessante destacar algumas palavras, cujas frequências podem ser explicadas pela natureza dos sub-registros específicos que compõe os domínios do Carol·$\mathcal{B}$. Esse é o caso, por exemplo da palavra "`the`", que aparece no domínio *Instrucional*. Como esse é composto em larga medida por textos enciclopédicos extraídos da Wikipédia, nos quais geralmente se inclui versões originais do que está sendo traduzido entre parênteses e, já que *stop words* em outros idiomas não foram removidas, a alta frequência é justificada. Da mesma forma, o vocábulo "`1`", no mesmo domínio pode ser explicado pela numeração das referências em páginas da *wiki*.

Também chama a atenção a presença de "`r`" no domínio *Jornalístico*, componente da abreviatura para o Real, "`R$`", sendo valores um tema frequente em textos jornalísticos. Em *Fórum Virtual* também observamos "`utc`" e "`https`", o primeiro sendo explicado pela presença de textos de Fóruns, que contêm o horário da postagem e o segundo pela alta densidade de links nos textos desse domínio. Já no *Jurídico*, é interessante notar a grande quantidade de palavras associadas a legislação: "`lei`", "`art`", abreviação de "`artigo`", e "`nº`", sendo essa última relacionada a números de processos e leis.

É interessante notar que o total de aparições das palavras mais frequentes dos domínios *Entretenimento* e *Jurídico* é significativamente maior do que o total de aparições nos demais domínios, cerca de 2,2 milhões de aparições, sendo o terceiro maior, *Fórum Virtual*, de aproximadamente 1,5 milhões. Esses valores corroboram o que apontamos na Seção 6.1: que esses seriam os dois domínios com mais estruturas repetitivas e com maior conteúdo duplicado.

Além disso, os verbos "`vamos`", "`vai`", "`pode`" e "`vou`" e os advérbios "`aqui`" e "`agora`" figurando entre as palavras mais frequentes no domínio *Entretenimento* reforçam a orientação ao tempo presente nos textos desse domínio. Em contrapartida, nota-se a presença de "`disse`" nas mais frequentes do *Jornalístico*, verbo no pretérito perfeito. A presença dessas palavras corrobora o que observamos a respeito de tempos verbais na Seção 6.2.

Por fim, observa-se na Tabela 10 o grau de sobreposição entre as 200 palavras mais frequentes em cada um dos domínios do Carol·$\mathcal{B}$. Destaca-se a baixa correspondência entre palavras dos domínios *Entretenimento* e *Jurídico* (8,5%) — os domínios mais distantes na escala que observamos na Seção anterior — e a alta correspondência entre os domínios *Fórum Virtual*-*Instrucional* e

Instrucional		Entretenimento		Jornalístico		Jurídico		Fórum Virtual	
palavra	**contagem**	**palavra**	**contagem**	**palavra**	**contagem**	**palavra**	**contagem**	**palavra**	**contagem**
referências	100.229	bem	333.289	anos	117.657	tribunal	360.638	utc	304.144
the	85.242	aqui	262.834	sobre	114.712	federal	286.512	https	195.750
sobre	85.083	vamos	247.195	disse	106.087	nº	279.062	discussão	186.849
anos	82.794	sim	233.832	r	104.588	supremo	270.661	artigo	162.223
cidade	70.292	vai	218.308	brasil	101.509	eletrônico	223.444	sobre	137.649
onde	65.317	agora	196.752	segundo	100.710	conforme	173.053	pra	105.549
durante	60.009	tudo	181.997	ainda	98.594	sob	160.081	página	98.016
1	59.531	pode	170.907	governo	96.684	lei	158.855	artigos	93.333
ligações	58.996	vou	168.111	ano	93.863	art	155.805	aqui	87.324
grande	58.895	então	157.516	presidente	90.364	ministro	153.084	pode	86.921
Total	**726.388**	**total**	**2.170.741**	**total**	**1.024.768**	**total**	**2.221.195**	**total**	**1.457.758**

Tabela 9: As dez palavras mais frequentes por domínio do Carol·$\mathcal{B}$.

Jornalístico-Instrucional — os domínios intermediários na escala observada. Dessa forma, as porcentagens de sobreposição baixa entre *Entretenimento-Jurídico* e a alta entre *Instrucional-Jornalístico* parecem justificadas, corroborando as nossas observações anteriores.

No entanto, chama a atenção a correlação alta entre *Instrucional* e *Fórum Virtual*, o que pode se dever ao fato de ambos possuírem sub-registros distintos, mas extraídos da mesma fonte, a Wikipédia, estando talvez relacionada a essa origem comum a alta similaridade entre as palavras de ambos. Isso sugere que ambientes funcionalmente mistos — como é o caso da Wikipédia, em que diferentes registros são produzidos com funções diferentes, mas em espaço em algum nível compartilhado — podem gerar uma coerção própria do espaço.

Como veremos nas Seções 6.4 e 7, comportamentos similares a esse são observados tanto nas distinções semânticas, como nas matrizes de confusão dos classificadores, indicando as diferenças vocabulares como fator de variação e discernibilidade importante entre os domínios.

6.4. Discernibilidade por meio de Espaços de *Embeddings*

Embeddings são representações em espaços vetoriais de significados lexicais, derivadas de algoritmos baseados na hipótese distribucional e treinados em *corpora* extensos. São ferramentas valiosas em tarefas de semântica computacional, capturando relações semânticas úteis entre palavras, como sinonímia, antonímia e similaridade (Jurafsky & Martin, 2009).

Devido à sua capacidade de representação, é esperado que espaços de *embeddings* possam revelar diferenças entre domínios no nível semântico. Delfino (2021) discute iniciativas focadas no estudo do aspecto semântico da variação intralinguística, no contexto da análise multidimensional, entretanto essas iniciativas utilizam etiquetas de campo semântico — como *Termos gerais e Abstratos*, *Corpo e Indivíduo*, *Artes e ofícios*, etc. — anotadas sobre os itens lexicais e não a análise orientada a espaços de *embeddings*, como a que realizamos aqui.

Para tal, avaliamos a discernibilidade e padrões de variação dos domínios de discurso usando o *NILC-Embeddings* (Hartmann et al., 2017), um repositório de *embeddings* estáticos para o português. Utizamos os *embeddings* dos tipos GLOVE (Pennington et al., 2014), SKIP-GRAM e CBOW(Mikolov et al., 2013) com 50 e 100 dimensões. Calculamos duas métricas: a silhueta entre domínios no espaço de *embeddings* e a contagem de *tokens* fora do vocabulário, ou *out-of-vocabulary (OOV)*, de cada domínio.

A silhueta é uma métrica capaz de mensurar a separabilidade de conjuntos em espaços vetoriais, comumente aplicada em problemas de agrupamento automático (Rousseeuw, 1987). Calculamos a silhueta média[11] entre todos os domínios conjuntamente e para cada par de domínios, usando uma amostra aleatória de 20.000 sentenças de cada domínio.[12] A Figura 10a exibe os resultados para os *embeddings* CBOW-50.

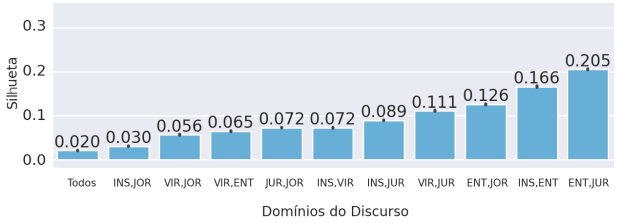
Em todos os espaços de *embeddings*, observamos valores de Silhueta consistentes com os exibidos na Figura 10a: quando calculada entre todos os domínios simultaneamente, a silhueta assume um valor pequeno e, às vezes, negativo, o que representa baixa separabilidade. Além disso, os pares com os valores de silhueta mais baixos e mais altos permanecem consistentes, correspondendo a posições opostas na escala *Jurídico*, *Instrucional*, *Jornalístico*, *Fórum Virtual*, *Entretenimento*. Enquanto isso, os pares adjacentes na mesma escala ocupam o meio da distribuição.

[11] Usamos a distância cosseno como métrica de distância.

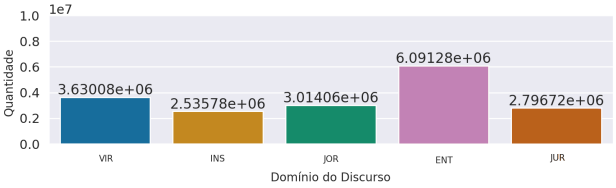
[12] Os *embeddings* de sentenças foram obtidos através da média dos *embeddings* dos *tokens* constituintes.

	Instrucional	Entretenimento	Jornalístico	Jurídico	Fórum Virtual
Instrucional	100,0%	24,5%	50,5%	20,5%	52,0%
Entretenimento	24,5%	100,0%	31,0%	8,5%	45,0%
Jornalístico	50,5%	31,0%	100,0%	24,0%	44,0%
Jurídico	20,5%	8,5%	24,0%	100,0%	17,5%
Fórum Virtual	52,0%	45,0%	44,0%	17,5%	100,0%

Tabela 10: Grau de sobreposição entre as 200 palavras mais frequentes por domínio do Carol·$\mathcal{B}$.



(a) Distribuição de valores de silhueta para diferentes combinações do domínios.



(b) Contagem de *tokens* OOV para cada domínio.

Figura 10: Exemplo de distribuição de Silhuetas e de *Tokens* OOV para os *embeddings* CBOW 50d sobre os domínios discursivos avaliados: *Fórum Virtual* (VIR), *Instrucional* (INS), *Jornalístico* (JOR), *Entretenimento* (ENT), *Jurídico* (JUR).

Essa é a mesma ordenação de domínios obtida nas seções anteriores.

Em resumo, embora os domínios coletivamente não tenham discernibilidade intrínseca clara nos espaços de *embeddings* explorados, existem distinções semânticas par-a-par, alinhadas com a escala de domínios observada nas análises anteriores. Essas distinções se assemelham aos resultados do grau de sobreposição vocabular entre os domínios, como era de se esperar, dada a base vocabular dos *embeddings*.

A Figura 10b ilustra as contagens de *tokens OOV* nas sentenças amostradas de cada domínio. Essas contagens servem como uma métrica de o quão bem o campo semântico de cada domínio é representado por esses espaços de *embeddings*.

Vemos que os domínios diferem significativamente em sua contagem de *tokens OOV*. Isso se junta às nossas observações da seção anterior acerca das diferenças vocabulares entre os domínios, sugerindo que modelos de *embeddings* específicos para cada domínio, modelando os vocabulários especializados, poderiam melhorar a aplicabilidade de tecnologias baseadas em *embeddings* para tarefas específicas sensíveis a esse tipo de variação. É interessante notar, também, que os domínios em extremos opostos da ordem de domínios observada anteriormente apresentam as diferenças mais substanciais na contagem de *tokens OOV*, por exemplo o par *Entretenimento-Jurídico*, o que reflete em algum nível os comportamentos observados para a sobreposição vocabular e pra as silhuetas par-a-par. Isso reforça que a escala que observamos anteriormente dos domínios possui um aspecto semântico significativo.

Além disso, as contagens de *tokens OOV* para cada domínio podem ser explicadas, aproximadamente, pela distribuição dos domínios nos *corpora* usados para treinar os *embeddings* (veja Hartmann et al. (2017)). Os textos de entretenimento têm menos *tokens* do que os textos jornalísticos e instrucionais nas partes de gênero único dos *corpora* de treinamento, o que pode explicar suas contagens de *tokens OOV*. As contagens de *tokens OOV* dos domínios *Fórum Virtual* e *Jurídico* são menos claras, pois não aparecem explicitamente nos *corpora* de gênero único usados para treinamento, mas podem estar contidas nos *corpora* de gênero misto. Seria necessária uma análise mais aprofundada para compreender totalmente essas contagens, mas elas parecem se alinhar, em geral, com a distribuição de domínios no conjunto de treinamento, como esperado. Essa observação serve de testemunho em favor da importância da diversidade intralinguística no treinamento de modelos desse tipo.

Em geral, os principais domínios de discurso do Carolina apresentam diferenças significativas em espaços de *embeddings* sem nenhuma transformação, destacando não só a existência de possíveis diferenças semânticas entre os domínios, como a sensibilidade dos modelos de *embeddings* a essas diferenças. No geral esses comportamentos concordam com aqueles observados para as distribuições vocabulares.

Ao mesmo tempo que essa distinção intrínseca dos próprios espaços é um resultado interessante, o uso de *embeddings* puros, sem aplicação de algoritmos de agrupamento por exemplo, é uma limitação deste trabalho, já que não permite demonstrar o potencial máximo de diferenciabilidade dessas tecnologias. Na Seção 7, entretanto, apresentamos algoritmos de classificação de domínios, que permitirão discussões interessantes nesse sentido.

6.5. Discernibilidade: Discussão e Comparação com a Literatura

A análise apresentada nas subseções anteriores nos permitiu não apenas concluir pela discernibilidade dos domínios discursivos estudados sob diversas variáveis, como nos levou a observar uma escala consistente segundo a qual os domínios parecem se ordenar em diversas *features* linguísticas quantificáveis e categóricas, que parece também influenciar a distribuição dos domínios em outras análises, como a da sobreposição vocabular e nos algoritmos de *embeddings*.

Nas seções anteriores apresentamos uma explicação inicial para tal escala associando-a ao fenômeno de complexidade de linguagem e à distinção entre linguagens mais e menos oralizadas, o chamado modo de discurso, sendo difícil propor uma interpretação mais aprofundada sem dados adicionais.

Entretanto, como mencionado na nossa revisão de literatura, na Seção 3, diversos trabalhos que se debruçam sobre o fenômeno de variação intralinguística em nível de registro adotam a metodologia chamada Análise Multidimensional. Um dos aspectos fundamentais dessa metodologia é a elaboração de interpretações linguísticas funcionais para as dimensões de variação obtidas através dos dados.

Dessa forma, uma maneira de obter uma interpretação mais aprofundada para o padrão observado, ainda dentro do escopo deste trabalho, seria a realização de uma análise comparativa entre a nossa escala de domínios e diferentes modelos de análise multidimensional, verificando a coincidência, ou não, da nossa escala com alguma das dimensões de variação que constituem os modelos. Poderíamos então utilizar a interpretação funcional dessas dimensões para enriquecer a nossa interpretação dos resultados obtidos aqui.

Tal análise comparativa é bastante desafiadora, porque nem as *features* nem as variedades linguísticas avaliadas em diferentes trabalhos coincidem totalmente. Em especial, é comum que vários desses trabalhos separem as propriedades linguísticas que adotamos aqui num número maior de sub-casos, o que impede uma comparação direta, a não ser que todos os sub-casos possuam um comportamento coeso, que possa ser interpretado em conjunto. Dadas essas dificuldades, é importante que tal análise seja entendida como qualitativa e exploratória.

Para realizar essa análise, comparamos a lista de *features* distintivas que seguem nossa escala de domínios com as diferentes *features* que constituem o modelo alvo. Após encontrar as *features* equivalentes, listamos as dimensões do modelo para as quais elas apresentam contribuições significativas e o sinal dessa contribuição, positivo ou negativo. Então, examinamos a lista completa de equivalências em busca de dimensões que apareçam em todas ou quase todas as *features* avaliadas e para as quais o sinal de contribuição da *feature* para essa dimensão coincida com a direção de aumento da *feature* na nossa escala.

Dessa forma, dimensões de variação que coincidem com nossa escala, não só em termos de quais as *features* associadas, mas em qual sua direção de variação, constituiriam bons candidatos ao fenômeno de variação subjacente à escala de domínios observada.

Aplicamos esse processo sobre três modelos multidimensionais distintos: o modelo multidimensional de Sardinha et al. (2014b) para a língua portuguesa e dois modelos pra a língua inglesa — o original de (Biber, 1988) e a nova versão de Biber & Egbert (2018), construída sobre o *corpus* CORE, de registros da web.

Vale ressaltar que, até onde sabemos, o modelo de Sardinha et al. (2014b) é o único modelo multidimensional geral de variação de registros da língua portuguesa atualmente disponível, sendo que os demais mencionados na Seção 3 são modelos voltados para o comportamento interno de domínios individuais, possuindo um escopo mais específico que o nosso. Isso justifica a adição dos modelos em inglês à nossa análise, para ampliar nosso espaço de comparação. Além disso, como mostrado em Biber (1995), diversos fenômenos de variação de registro são transversais ao idioma.

As Tabelas 11, 12 e 13 contêm os resultados da nossa análise comparativa com os modelos multidimensionais de Sardinha et al. (2014b), Biber (1988) e Biber & Egbert (2018), respectivamente.

Em todos os casos, podemos notar que há uma coincidência significativa da nossa escala para com as primeiras dimensões de cada um dos modelos, tanto no que tange às *features* comparáveis, como em seu sinal de variação.

Feature	**Correspondente em Biber (1988)**	**Escala de Domínios**	**Dimensões em Biber (1988) (Polo Respectivo)**
ADV	*Adverbs*	Jur. → Ent.	D1(+) , D3(-)
NOUN	*Nouns, Nominalization*	Ent. → Jur.	D1(-) , D3(+)
ADP	*Prepositions*	Ent. → Jur.	D1(-)
ADJ	*Attributive Adjectives*	Ent. → Jur.	D1(-) , D2(-)
Tamanho Médio das Palavras	*Mean Word Length*	Ent. → Jur.	D1(-)

Tabela 11: Comparação entre as nossas *features* linguísticas e as análogas em Biber (1988), extraídas a partir da reapresentação dos resultados em Biber & Conrad (2009c), excluindo-se D7, a qual os autores não conseguem fornecer uma interpretação funcional sólida.

Feature	**Correspondente em Biber & Egbert (2018)**	**Escala de Domínios**	**Dimensões em Biber & Egbert (2018) (Polo Respectivo)**
Past	*Past Tense Verbs*	Ent. → Jur.	D3(+), D4(+), D5(-)
ADP	*Prepositional Phrases*	Ent. → Jur.	D1(-) , D5(-)
Tamanho Médio das Palavras	*Long Words*	Ent. → Jur.	D3(-), D5(-) , D6(+)
Pres.	*Non-Past-Tense Verbs*	Jur. → Ent.	D1(+) , D5(+)

Tabela 12: Comparação entre as nossas *features* linguísticas e as análogas em Biber & Egbert (2018).

Nos modelos multidimensionais, a ordem das dimensões não é arbitrária, sendo elas organizadas pelo grau segundo o qual explicam a variação observada na amostra. O alinhamento entre a nossa escala e as primeiras dimensões dos modelos significa que essa escala está provavelmente relacionada aos fenômenos de variação mais significativos entre os registros, o que faz sentido dada a pervasividade com que observamos essas escalas de domínios aparecendo em diferentes aspectos das análises.

Esse resultado é particularmente interessante, porque nos mostra que essa dimensão de variação primária — detectada nesses modelos a partir de estudos de coocorrência entre diversas *features*, utilizando algoritmos estatísticos — pode ser percebida numa análise mais simples, a partir da comparação manual das distribuições individuais de cada *feature*, o que corrobora a ubiquidade desses fenômenos de variação primária, mas também indica uma limitação da nossa metodologia, já que não fomos capazes de detectar ou capturar dimensões secundárias de variação, como fazem esses modelos mais intrincados.

No que tange à interpretação funcional dessas dimensões, apesar de os modelos serem distintos, as primeiras dimensões desses três modelos possuem um alto nível de sobreposição, possuindo interpretações funcionais similares.

No caso, tanto de Biber (1988) como de Biber & Egbert (2018), a dimensão primária é entendida como a **dimensão oral-envolvida versus literato-informacional** e está associada à distinção entre registros mais oralizados, que adotam linguagem mais coloquial, com menos densidade informacional e mais marcas da produção simultânea, e registros mais técnicos, que alcançam maior densidade informacional, usualmente através de um processo de revisão mais meticuloso, associado a processos de escrita e elaboração mais longos, visando situações funcionais mais especializadas e restritas.

Quando olhamos para a nossa escala de domínios — *Entretenimento*, *Fórum Virtual*, *Jornalístico*, *Instrucional* e *Jurídico* —, essa pa-

Feature	**Correspondente em Sardinha et al. (2014b)**	**Escala de Domínios**	**Dimensões em Sardinha et al. (2014b) (Polo Respectivo)**
PRON	*Pronouns*	Jur. → Ent.	D1(+), D2(+), D6(+)
VERB	*Verbs*	Jur. → Ent.	D1(+), D4(+), D6(+)
ADV	*Adverbs*	Jur. → Ent.	D1(+), D2(+)
SCONJ	*Subordinating Clause*	Jur. → Ent.	D1(+), D6(+)
NOUN	*Nouns*, *Nominalization*	Ent. → Jur.	D1(-)
Tamanho Médio das Palavras	*Average Word length*	Ent. → Jur.	D1(-)
ADP	*Prepositions*	Ent. → Jur.	D1(-)

Tabela 13: Comparação entre as nossas *features* linguísticas e as análogas em Sardinha et al. (2014b).

rece ser uma interpretação bastante sólida, já que os domínios do início da escala tendem a ser mais coloquiais e oralizados, enquanto textos do final tendem a ser mais densos, técnicos e associados a um tempo maior de preparo e edição.

No trabalho de Sardinha et al. (2014b), a dimensão primária é interpretada exclusivamente como uma dimensão de oposição entre discurso oral e escrito, deixando a distinção entre envolvimento e informacionalidade para uma dimensão posterior. Essa interpretação é um pouco mais restrita que a anterior, mas também bastante compatível com a escala de domínios aqui observada, por argumentos análogos aos que demos para os trabalhos anteriores.

Se observarmos a maneira como os registros mais próximos dos nossos se ordenam nessas dimensões em cada um desses estudos, o que está representado nas Figuras 12a, 11 e 12b, veremos que elas coincidem bastante com a nossa escala.

No caso do modelo de (Biber & Egbert, 2018), uma outra dimensão, a Dimensão 5, constitui também um candidata razoável para a nossa escala, apesar de a Dimensão 1 ser uma candidata mais robusta, dada sua aparição consistente em diferentes modelos. Essa Dimensão 5 é chamada pelos autores de **Dimensão *Irrealis* versus Narração informacional** e é uma dimensão que contrasta discursos mais hipotéticos, voltados a comparações ou condições sobre eventos hipotéticos, contra a narração mais factual de eventos tidos como certos.

Considerando nossa escala de domínios, esta também é uma interpretação plausível. Pode ser o caso, inclusive, dessa dimensão constituir um aspecto secundário da escala e variação dos domínios, combinado-se à dimensão oral-envolvida vs literato-informacional mencionada anteriormente.

O mesmo acontece na comparação com o modelo de Biber (1988), onde, além da dimensão primária, a Dimensão 3 também aparece relacionada de forma consistente a algumas das *features* distintivas em comum. Essa dimensão trata da forma como o discurso estrutura e organiza suas referências, contrastando referências dependentes da situação, no polo negativo e estruturas referenciais mais elaboradas e mais descoladas da contextualização imediata, no polo positivo. Nesse caso, essa interpretação também está de acordo com as expectativas em relação a cada um desses domínios e, novamente, pode ser entendida como uma dimensão secundária associada à dimensão primária: discursos mais oralizados tendem a apelar mais para referências imediatas e compartilhadas entre interlocutores, enquanto discursos escritos e mais informacionais precisam, com mais frequência, de referências elaboradas e independentes do contexto pragmático imediato.

No geral, nossa análise nos mostrou que muito da avaliação original, de associar a escala observada dos domínios ao modo de discurso e à complexidade de linguagem ia numa direção promissora, mas permitiu, ao mesmo tempo, enriquecer essa interpretação. Assim, vimos que a

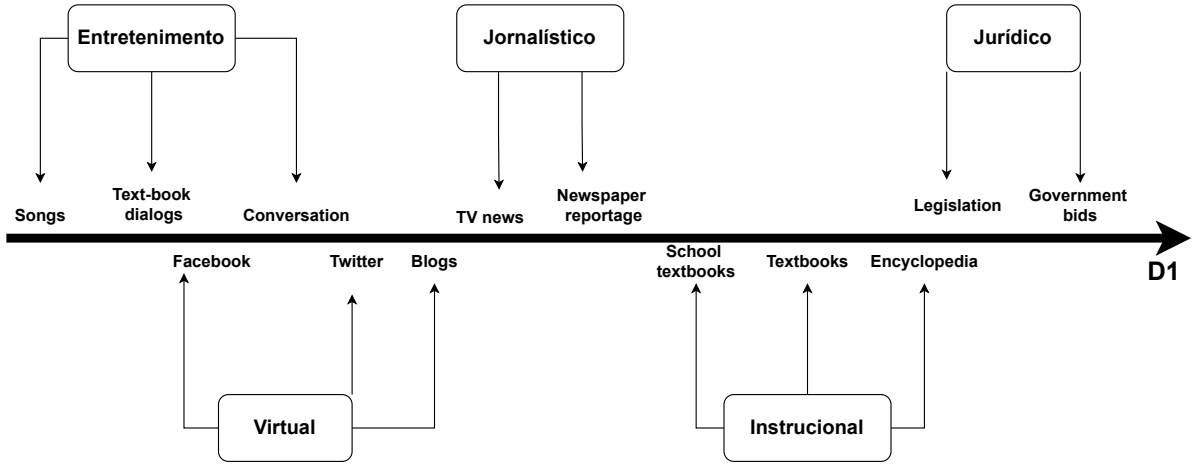
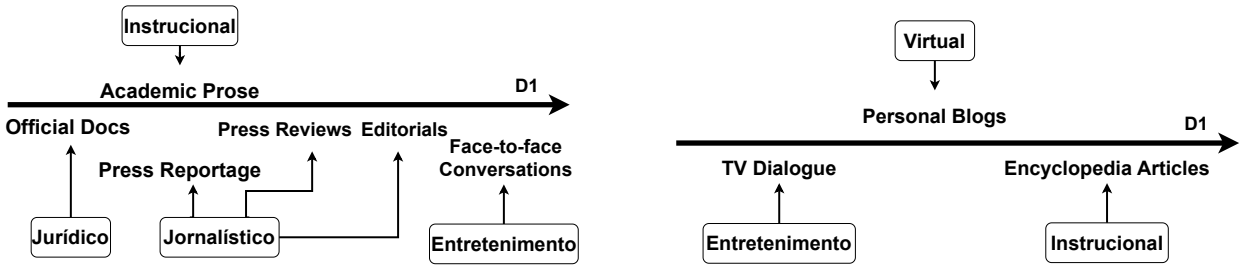


Figura 11: Ordenação Relativa, ao longo da Dimensão de variação 1 dos sub-registros mais similares às componentes dos nossos domínios em Sardinha et al. (2014b): *Songs* (Letras de Música), *Textbook dialogs* (Diálogos de livro-texto), *Facebook*, *Conversation* (Conversas), *Twitter*, *Blogs*, *TV news* (Notícias televisivas), *Newspaper reportage* (Reportagem Jornalística), *School textbooks* (Livros-texto escolares), *Textbooks* (Livros-texto), *Encyclopedia* (Enciclopédia), *Legislation* (Legislação), *Government Bids* (Licitações).



(a) Comparação com (Biber, 1988). **(b)** Comparação com (Biber & Egbert, 2018).

Figura 12: Comparação da ordenação relativa dos domínios com os sub-registros de (a) Biber (1988) e (b) Biber & Egbert (2018) ao longo da Dimensão 1, incluindo *TV Dialogue* (Diálogos de Programas Televisivos), *Personal Blogs* (Blogs Pessoais), *Encyclopedia Articles* (Artigos de Enciclopédia), *Official Docs* (Documentos Oficiais), *Academic Prose* (Prosa Acadêmica), *Press Reportage* (Reportagem Jornalística), *Press Reviews* (Resumo de Notícias), *Face-to-face Conversations* (Conversas cara a cara).

escala observada deve estar relacionada principalmente a fenômenos de oralidade, simultaneidade de produção, tecnicidade, compactação informacional e complexificação estrutural do discurso, bem como, à distinção entre o discurso hipotético e informacional e ao nível de elaboração referencial.

7. Classificação Automática de Domínios Discursivos

As análises apresentadas até aqui revelam que os domínios discursivos são diferenciáveis em nível sentencial, bem como esclareceram muitos dos padrões de variação presentes na nossa amostra. Nosso próximo passo consistiu no treinamento de modelos de classificação automática dos domínios do português nesse nível linguístico. Esse processo teve dois objetivos principais: (i) pôr à prova, de forma prática, a hipótese de diferenciabilidade nesse escopo e (ii) desenvolver recursos computacionais que possam beneficiar aplicações dependentes dessa distinção.

Neste trabalho, focamos em modelos baseados em *Transformer* (Vaswani et al., 2017), que apresentam resultados estado-da-arte em várias tarefas de classificação sentencial. Além disso, estudos prévios demonstram que esses modelos são sensíveis a propriedades linguísticas que observamos distintivas entre nossos domínios discursivos,[13] como: variações na estrutura sintática, dis-

[13]Dada a natureza dos modelos de aprendizado de máquina, não é possível afirmar, sem uma análise direcionada, se sua distinção dos domínios se baseia efetivamente nas características linguísticas que investigamos. No entanto, a capacidade desses modelos de desempenhar ta-

tribuição de categorias gramaticais e classes morfossintáticas, reconhecimento de entidades nomeadas e diferenças semânticas (Şapcı et al., 2024; Jung et al., 2022; Zhang et al., 2022; Wu et al., 2021).

Resultados citados na Seção 3, como (Rönnqvist et al., 2021), também evidenciam a eficácia dessa família de modelos na classificação de variação linguística em diferentes dimensões — incluindo em nível de registro, no qual, conforme discutimos na Seção 2, inserimos os domínios discursivos.

Testamos, como base para o desenvolvimento do classificador, quatro modelos baseados em *Transformer* — *bert-base* e *bert-large* (Souza et al., 2020), *albertina-100m* e *albertina-900m* (Rodrigues et al., 2023) — comparando-os com dois *baselines*: *Naive Bayes* e *Support Vector Machine*, que exploram somente a distribuição dos itens lexicais contidos nos exemplos.

Para melhor aferir a capacidade de classificação de domínios desses modelos, nesse conjunto de dados em específico, utilizamos a partição *test* do *carol-domain-sents*, dividindo-a em cinquenta partes e avaliando independentemente cada modelo em todas elas, de forma a obter cinquenta avaliações pareadas.

Reportamos o valor médio e respectivo desvio-padrão da medida F1 micro ao longo desses cinquenta subconjuntos, computados por meio da biblioteca *scikit-learn* (Pedregosa et al., 2011). A medida F1 micro consiste em computar o total global de verdadeiros positivos, falsos negativos, e falsos positivos para então calcular a medida F1 com esses valores. Essa é a única medida que reportamos, pois ela é numericamente equivalente a acurácia, precisão micro e cobertura micro quando calculadas em um *dataset* balanceado, que é o caso. Os resultados obtidos para cada modelo são apresentados na Tabela 14.

Para aferir a significância da diferença observada entre o valores das métricas obtidos para cada modelo, utilizamos, inspirados por Villanueva Llerena & Mauá (2017), o teste estatístico de Benavoli et al. (2014, 2017), um teste alinhado à abordagem bayesiana, que computa as probabilidades de um classificador M_1 ter uma métrica de performance maior, equivalente ou menor que outro classificador M_2, dados os resultados empíricos de performance pareada em diversos conjuntos. O teste recebe como parâmetro uma região de equivalência prática (*rope*) que

refas relacionadas a tais características, aliada à sua capacidade de detecção de propriedades relevantes, justifica considerá-los como candidatos plausíveis e coerentes com as análises previamente apresentadas.

Modelos	F1 Micro
svm	0.725 (0.003)
naive-bayes	0.779 (0.003)
albertina-100m	0.876 (0.002)
albertina-900m	0.879 (0.002)
bert-base	0.879 (0.002)
bert-large	0.885 (0.002)

Tabela 14: Média (desvio padrão) da medida F1 Micro para cada modelo.

F1 Micro				
Modelo 1	$>$	$=$	$<$	Modelo 2
naive-bayes	1.00	0.00	0.00	svm
naive-bayes	0.00	0.00	1.00	bert-large
naive-bayes	0.00	0.00	1.00	bert-base
naive-bayes	0.00	0.00	1.00	albertina-100m
naive-bayes	0.00	0.00	1.00	albertina-900m
svm	0.00	0.00	1.00	bert-large
svm	0.00	0.00	1.00	bert-base
svm	0.00	0.00	1.00	albertina-100m
svm	0.00	0.00	1.00	albertina-900m
bert-large	1.00	0.00	0.00	bert-base
bert-large	1.00	0.00	0.00	albertina-100m
bert-large	1.00	0.00	0.00	albertina-900m
bert-base	1.00	0.00	0.00	albertina-100m
bert-base	0.02	0.97	0.00	albertina-900m
albertina-100m	0.00	0.00	1.00	albertina-900m

Tabela 15: Comparação dos modelos na métrica F1 Micro. Maior valor destacado.

define a faixa de desempenho dentro da qual consideramos os dois classificadores como praticamente equivalentes. Para esta avaliação, tomamos um intervalo de equivalência prática de $[-0.001, 0.001]$, que corresponde a **rope** $= 0.1\%$. As probabilidades resultantes da aplicação do teste são apresentadas na Tabela 15.

Todos os modelos avaliados apresentaram desempenho superior ao limiar de sorteio aleatório, demonstrando sua eficácia na tarefa. Os *baselines* — *Naive Bayes* e *SVM* — obtiveram, conforme esperado, os piores resultados, evidenciando a importância de considerar informações contextuais extra-lexicais na classificação de domínios. No entanto, seu desempenho ainda foi satisfatório, indicando que a diferenciação lexical entre os domínios é significativa, em concordância com nossos achados referentes à variação vocabular e semântica entre os domínios da amostra. Entre os modelos baseados em *Transformer*, a ordem de performance foi a seguinte: *bert-large* superou *bert-base*, que, por sua vez, apresentou desempenho equivalente ao *albertina-900m*, ambos superiores ao *albertina-100m*. Assim, o *bert-large*

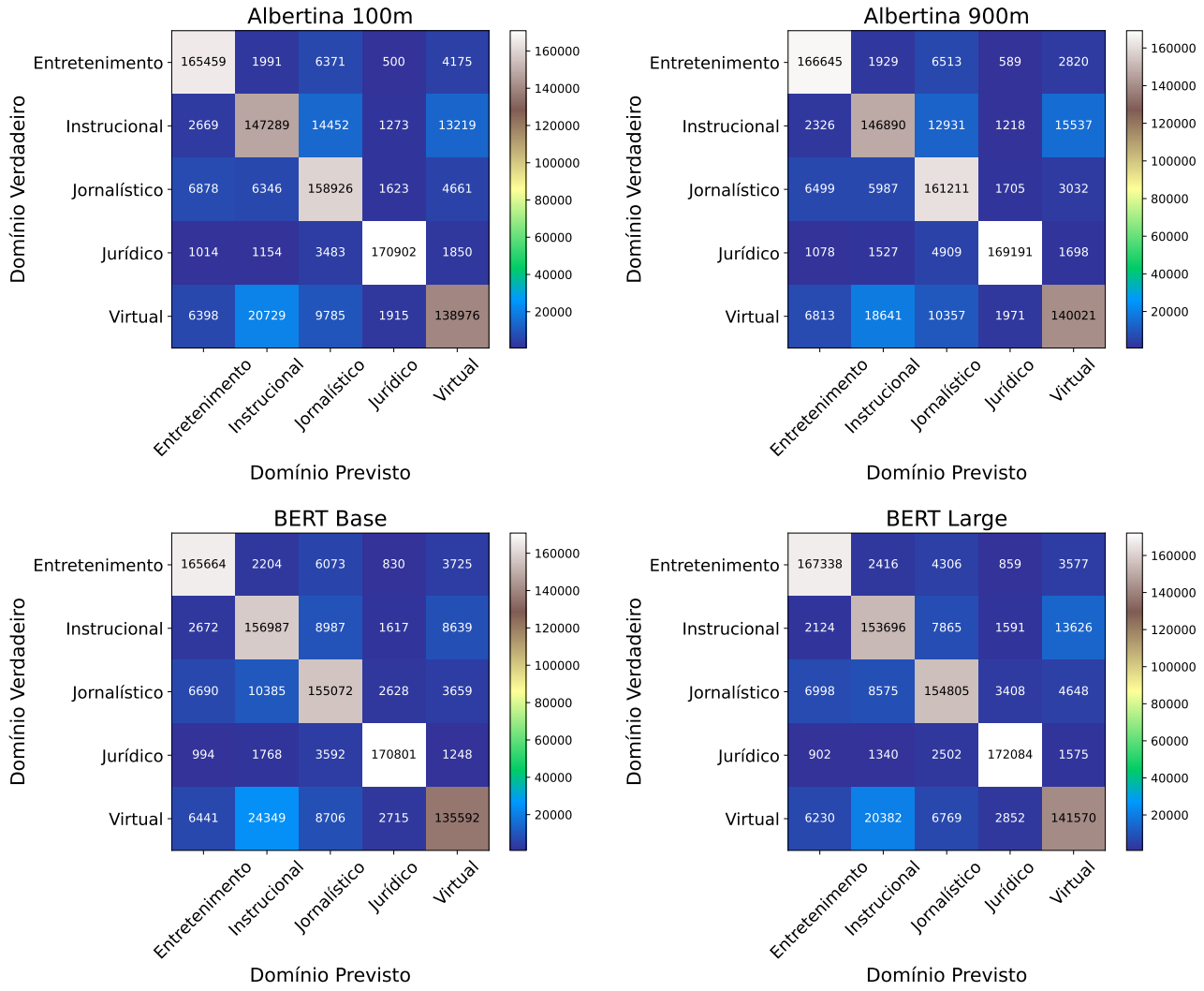


Figura 13: Matrizes de confusão para os modelos baseados em *Tranformers*.

se destaca como o melhor modelo entre os avaliados. Apesar disso, disponibilizamos todos os modelos, pois, juntos, representam um conjunto de recursos tecnológicos robustos, diversos e altamente eficazes para a diferenciação dos domínios estudados do português brasileiro.

Para oferecer uma visão mais detalhada do desempenho dos modelos, a Figura 13 apresenta as matrizes de confusão dos classificadores baseados em *Transformers*. Adicionalmente, a Tabela 17, nos apêndices, fornece alguns exemplos ilustrativos de erros de classificação cometidos por esses classificadores.

Nessas matrizes, a taxa de acerto é evidenciada pelos altos valores na diagonal principal, refletindo as métricas de performance relatadas. Observamos também que os modelos dessa família exibem um perfil de erro bastante consistente entre si.

A consistência entre as matrizes de confusão dos modelos baseados em *Transformer* pode indicar que seu perfil de desempenho esteja refletindo características dos dados e não dos modelos. Poderiam essas características estarem relacionadas à aparente escala de ordenação dos domínios que observamos anteriormente? Para averiguar essa possibilidade, apresentamos, na seção seguinte, uma análise exploratória da relação entre a confusão dos classificadores entre os diferentes domínios e suas posições na escala observada na nossa análise de discernibilidade.

7.1. Relação entre a Confusão dos Classificadores e a Similaridade dos Domínios

Observando-se as matrizes de confusão dos modelos baseados em *Transformer*, nota-se que o par de domínios posicionados nas extremidades da ordem observada na nossa análise de discernibilidade, *Jurídico* e *Entretenimento*, apresentam as menores taxas de confusão; já os pares *Instrucional-Jornalístico* e *Instrucional-Virtual*,

que estão no centro da ordenação observada, exibem as maiores taxas de confusão relativa.

Esta seção se propõe a investigar a consistência dessas associações. Para tal, apresentamos, a seguir, a estrutura metodológica que viabiliza essa análise.

Dada uma ordenação qualquer ϕ para os domínios, podemos definir a dissimilaridade $d_\phi(A, B)$ entre dois domínios A e B, segundo ϕ, como sendo o número de posições que separam A e B em ϕ. Dessa forma, podemos, construir uma matriz de dissimilaridade D_ϕ entre todos os pares de domínios, segundo ϕ, tal que $D_{\phi_{A,B}} = d_\phi(A, B)$.

Em contrapartida, podemos definir a distância de dois domínios A e B, segundo um classificador, como sendo o inverso da confusão observada entre esses dois domínios para esse classificador, sob o pressuposto de que quanto mais instâncias do domínio A um classificador classificar erroneamente como sendo um domínio B, maior a dificuldade do classificador distinguir A de B e, portanto, maior a proximidade dos dois domínios no seu espaço interno de classificação.

Para podermos interpretar essa distância localmente como uma dissimilaridade, precisamos que ela seja simétrica ($\forall x, y : d(x, y) = d(y, x)$) e possua reflexividade nula ($\forall x : d(x, x) = 0$). Podemos garantir isso, tomando uma versão simetrizada C_S da matriz de confusão no lugar da matriz de confusão C original, onde $C_{Si,j} = (C_{i,j} + C_{j,i})/2$ e tomando um nível de precisão p que garanta a diagonal nula. A matriz de dissimilaridade resultante desse processo é como a definida na equação 5, onde $trunc_p(\cdot)$ é a operação de truncar um número na p-ésima casa decimal.

$$D_{Ci,j} = trunc_p\left(\frac{2}{C_{i,j} + C_{j,i}}\right) \qquad (5)$$

Podemos aplicar o teste estatístico de Mantel (Mantel, 1967), que computa correlação entre matrizes de dissimilaridade, sobre D_ϕ e D_C, para averiguar o quanto a distância de dois domínios na ordenação ϕ em questão é preditiva da dissimilaridade dos domínios segundo os modelos.

Para o cálculo da correlação de Mantel, usamos a implementação da biblioteca `scikit-bio` (Rideout et al., 2023) e adotamos o Coeficiente de *Spearman* como medida de correlação básica entre os itens das matrizes. O teste retorna a correlação entre as duas matrizes de dissimilaridade comparadas.

Calculamos a correlação de Mantel entre a matriz de similaridade de confusão média dos modelos baseados em *Transformer* D_C e a dissimilaridade segundo a ordenação dos domínios D_ϕ, para todas as ordenações ϕ possíveis do nosso conjunto de cinco domínios discursivos. A Distribuição das correlações obtidas está representada no histograma da Figura 14.

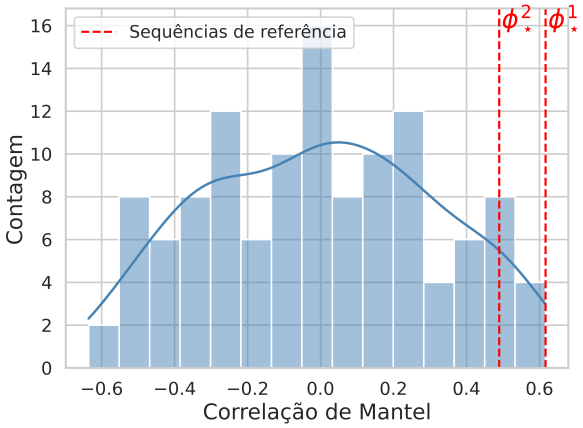


Figura 14: Distribuição da correlação de Mantel dentre a distância D_ϕ dos domínios segundo uma ordenação ϕ e a distância média de confusão dos domínios para os classificadores baseados em *Transformer* D_C, considerando-se todas as ordenações possíveis dos domínios do Carolina. Os valores correspondentes às ordenações $\phi^1_\star$ e $\phi^2_\star$ dos domínios, observadas na nossa análise de discernibilidade, estão marcadas pelas linhas tracejadas em vermelho.

Na nossa análise de discernibilidade dos domínios, apresentada na Seção 6, observamos duas ordenações ϕ bastante parecidas, que se repetiram sistematicamente nas distribuições de diversas variáveis — a ordenação $\phi^1_\star$ ={*Jurídico, Jornalístico, Instrucional, Fórum Virtual, Entretenimento*}, observada, principalmente, entre as propriedades numéricas globais das sentenças, e a ordenação $\phi^2_\star$ ={*Jurídico, Instrucional, Jornalístico, Fórum Virtual, Entretenimento*}, observada na ordem relativa dos domínios no que tange à taxa de aparição de várias etiquetas morfossintáticas (veja Seção 6.2).

As distâncias entre os domínios sob ambas as ordenações observadas, $\phi^1_\star$ e $\phi^2_\star$, apresentam altas correlações com a dissimilaridade de confusão dos classificadores, respectivamente $\rho^1_\star = 0.617$ e $\rho^2_\star = 0.489$ — valores representados pelas linhas tracejadas vermelhas na Figura 14. Se compararmos esses valores com as correlações das demais ordenações, vemos que ambas, $\phi^1_\star$ e $\phi^2_\star$, apresentam correlações entre os valores mais altos da distribuição, sendo a correlação associada a $\phi^1_\star$ a maior entre todos os valores observados e a correlação associada a $\phi^2_\star$ maior que 93.33% das correlações observadas.

Isso nos leva a concluir que existe uma correlação relevante entre a ordenação dos domínios nas propriedades estudadas na nossa análise de discernibilidade e sua dificuldade de distinção para fins de classificação automática. Essa observação sugere que essas propriedades ou (i) estão sendo usadas de alguma forma no processo de classificação dos domínios, ou (ii) que os classificadores estão acessando fenômenos subjacentes nos dados que estão relacionados, direta ou indiretamente, à distribuição das propriedades analisadas. Em especial, essa relação parece pender mais para o lado das propriedades numéricas globais da sentença associadas a $\phi^1_\star$, já que o valor de correlação observado nesse caso foi maior que no caso das demais propriedades ($\phi^2_\star$).

Dada a nossa análise da Seção 6.5, em que associamos a nossa escala de domínios à dimensão primária de variação que aparece em diferentes modelos multidimensionais, associada à dimensão funcional de oposição entre discurso oral-envolvido e literato-informacional, esses resultados parecem indicar que os classificadores baseados em *Transformers* estão conseguindo capturar, no seu treinamento, justamente essa dimensão primária de variação estatística entre os domínios, apelando para a distinção entre seus dois polos de discurso, como forma de diferenciá-los, sendo que, nos pontos em que essa diferenciação se torna mais difícil, também os modelos passam a falhar mais.

Além de fornecerem uma possível explicação para a eficácia e comportamento de classificação dos modelos, os resultados obtidos indicam também que uma análise prévia da discernibilidade e da natureza da variação entre classes, como a realizada neste trabalho, pode ser útil para prever a confusão dos modelos e, consequentemente, orientar o desenvolvimento de estratégias de mitigação, com potencial ganho de performance.

7.2. Classificação: Discussão e Comparação com a Literatura

Como mencionado na Seção 3, encontramos poucos trabalhos que se sobrepõem ao nosso em termos de língua-alvo, variedade intralinguística sendo classificada e tipo de modelo sendo utilizado para tal. Mesmo nesses casos, nosso trabalho ainda apresenta diferenças significativas, como o escopo de classificação — que aqui é sentencial, em comparação ao foco no nível de documento dado por outros trabalhos — além de estarmos usando um *dataset* novo, desenvolvido especificamente para esta pesquisa.

Dadas as dificuldades de uma comparação direta com resultados de contextos idênticos, tentamos construir um panorama geral da classificação de variedades e posicionar nossos resultados dentro desse panorama. Por isso, na Figura 15 apresentamos a performance para todos os trabalhos mencionados na Seção 3 que incluíram o desenvolvimento de classificadores automáticos.

Isso engloba resultados referentes a diferentes línguas e variedades, em sistemas de classificação e *datasets* distintos. Discriminamos a língua-alvo e o tipo de variedade intralinguística sendo classificada em cada ponto, segmentando os resultados em número de classes no sistema de classificação, representado no eixo horizontal.

O eixo vertical representa a performance obtida pelo classificador em termos de Medida F1 micro. Incluímos três tipos de resultados: (i) aqueles nos quais os autores explicitamente fornecem performance em F1 micro; (ii) aqueles nos quais os autores fornecem resultados em medida F1 de forma geral e nos quais pudemos deduzir ser F1 micro a partir do texto e (iii) aqueles nos quais os autores fornecem resultados em outras métricas, mas em arranjos experimentais passíveis de conversão para F1 Micro.

Dadas as diferenças nas configurações experimentais entre os resultados apresentados no gráfico da Figura 15, uma comparação quantitativa direta é impraticável. Contudo, em termos gerais, observamos três tendências principais: (i) existe uma considerável diversidade nos sistemas de classificação empregados pelos diferentes trabalhos, refletido no número de classes contemplados em cada caso; (ii) há uma tendência geral de queda do desempenho à medida que o tamanho do sistema de classificação aumenta; e (iii) modelos para a língua inglesa tendem a apresentar um desempenho superior aos de outras línguas, refletindo a quantidade e qualidade de recursos disponíveis.

No que tange ao nosso melhor resultado — representado no gráfico pelo marcador em formato de estrela — é possível notar que nosso classificador se encontra relativamente bem posicionado em comparação com os outros trabalhos mencionados. Essa constatação se mantém tanto na comparação com outros classificadores para a língua portuguesa quanto com aqueles desenvolvidos para variedades similares a registro, independentemente da língua. Ressaltamos que qualquer comparação direta é impossível nestas condições. Para tal, seria necessária a avaliação dos modelos sobre um mesmo conjunto de dados, bem como um sistema de tradução entre os sistemas de classificação e os escopos utilizados.

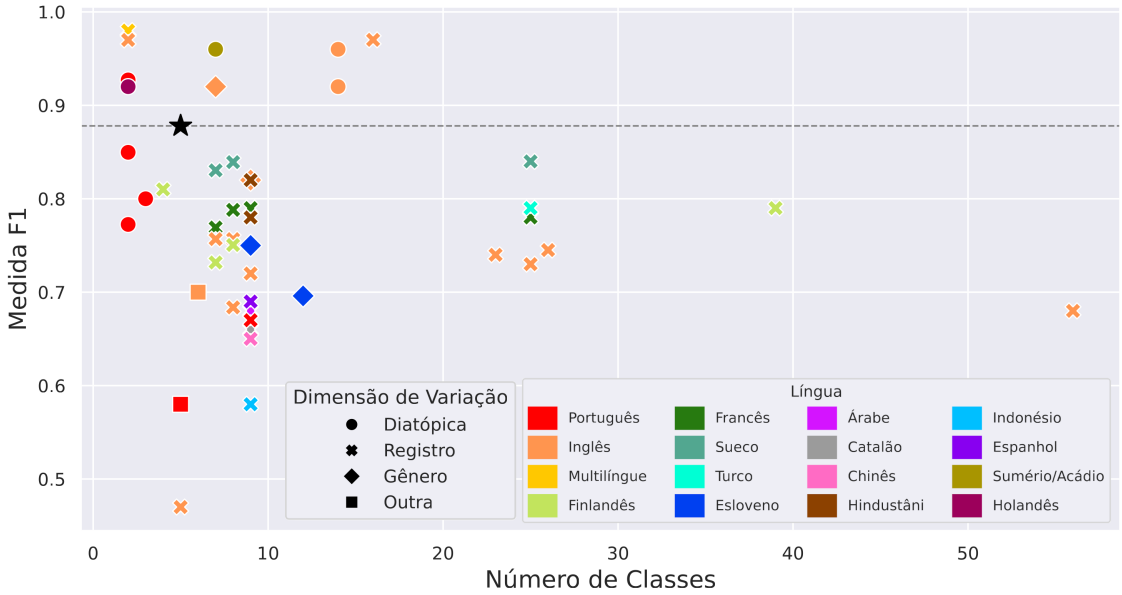


Figura 15: Panorama da performance de classificação para variedades intralinguísticas, incluindo os trabalhos mencionados na Seção 3, comparados ao nosso melhor resultado (marcador de estrela).

Esperamos, entretanto, que a política de construção e a estratégia de distribuição do nosso conjunto de dados — limitadas a documentos com licença permissiva, visando distribuição aberta para fins de pesquisa — estimulem, em um futuro próximo, o desenvolvimento de uma gama mais ampla de resultados sobre este mesmo *dataset*, viabilizando assim uma comparação mais abrangente e rigorosa.

8. Conclusões

Neste trabalho, nos debruçamos sobre a variação de domínios discursivos no português brasileiro em nível sentencial. Na primeira etapa examinamos a diferenciabilidade dos domínios na nossa amostra e caracterizamos sua variação. Na segunda parte, após concluir pela clara diferenciabilidade dos domínios no escopo sentencial, desenvolvemos classificadores de domínios nesse escopo. Construímos um *dataset* aberto para essa tarefa e treinamos quatro modelos baseados em *Transformers* e dois modelos clássicos, todos publicamente disponíveis.

Na nossa análise de discernibilidade e variação, identificamos diversas características distintivas em termos de domínio, o que nos permite retomar as perguntas iniciais, da Seção 1:

Q1 **Os domínios discursivos são computacionalmente discerníveis nesse escopo?** Sim. Os domínios avaliados são bastante discerníveis em nossa amostra;

Q2 **Quais propriedades os diferenciam?** Propriedades como grau de duplicação, comprimento de frases e palavras, distribuição de categorias gramaticais (*part-of-speech tags*) e tempo verbal são distintivas. Além disso, muitos pares domínios são relativamente distinguíveis nos espaços de *embeddings* e em termos de sobreposição vocabular.

Q3 **Quais problemas e tarefas de PLN podem se beneficiar da diferenciação de domínios nesse escopo?** Nossos resultados indicam que tarefas como etiquetagem gramatical, *tokenização*, segmentação, reconhecimento de entidades nomeadas e geração de *embeddings* podem se beneficiar de especialização em domínio.

Além disso, observamos um padrão sistemático de ordenação entre os domínios aparente em várias das propriedades numéricas analisadas e com cortes correspondentes em várias propriedades categóricas. Observamos padrões relacionados na distribuição vocabular e semântica dos domínios. Numa análise comparativa com modelos multidimensionais, concluímos que a ordenação reflete a dimensão de variação primária desses modelos, que envolve a distinção funcional entre discursos oralizados e envolvidos em oposição a discursos literários e mais informacionalmente densos.

Na segunda parte, os experimentos de classificação realizados confirmaram a viabilidade da tarefa, com o *bert-large* apresentando os melhores resultados. Realizamos também uma comparação entre a confusão entre os domínios por

parte dos classificadores baseados em *Transformers* e a distância dos mesmos domínios na ordenação observada na etapa anterior, atestando uma forte concordância entre ambos, o que sugere que esses classificadores baseiam-se significativamente na principal dimensão de variação dos domínios (a escala oral-envolvido vs. literato-informacional) para realizar a classificação.

Diversas direções para pesquisas futuras são possíveis. Destacamos: (i) o desenvolvimento de modelos de PLN especializados em domínios, (ii) uma exploração mais aprofundada da deduplicação de textos entre domínios, (iii) um estudo detalhado da relação entre complexidade textual e diferenças linguísticas observadas entre os domínios, em especial através de um desenvolvimento de um modelo de análise multidimensional sobre o *corpus*, adicionando a nossa análise um conjunto mais amplo e detalhado de *features*, (iv) a exploração do uso dos classificadores sentenciais para a construção de classificadores em nível de documento, possivelmente usando técnicas de agregação, e (v) o desenvolvimento de classificadores completamente baseados em *features* distintivas para fim de comparação. Além disso, seria interessante repetir os experimentos garantindo que sentenças do mesmo documento não sejam divididas entre conjuntos durante o treinamento, visando aumentar a robustez e a confiabilidade dos resultados.

Agradecimentos

Este trabalho foi realizado no Centro de Inteligência Artificial (C4AI-USP), com apoio da Universidade de São Paulo, da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) (processo #2019/07665-4) e da IBM Corporation. Este trabalho foi parcialmente apoiado pelo Ministério da Ciência, Tecnologia e Inovação, com recursos da Lei N. 8.248, de 23 de Outubro de 1991, no escopo PPI-SOFTEX, coordenada pela Softex e publicada como Residência em TIC 13, DOU 01245.010222/2022-44. Marcelo Finger contou com apoio parcial da FAPESP (processos #2015/21880-4, #2014/12236-1) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (bolsa PQ 302963/2022-7). Vanessa Monte contou com apoio do edital PrInt-Capes n^o 58/2022. Felipe Ribas Serras e Mariana Lourenço Sturzeneker foram apoiados pela IBM através de bolsa FUSP (Fundação de Apoio à Universidade de São Paulo) (Projeto 3541). Felipe R. Serras foi apoiado por uma bolsa PPI-SOFTEX manejada pela FUSP (Projeto 3970). Este trabalho foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e usou recursos do "Centro Nacional de Processamento de Alto Desempenho em São Paulo (CENAPAD-SP)". Os autores gostariam de agradecer a ajuda e amizade de Gabriela Lachi.

Referências

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta & Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. Em *25th International Conference on Knowledge Discovery & Data Mining (KDD)*, 2623–2631. 10.1145/3292500.3330701

Ansel, Jason, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu & Soumith Chintala. 2024. PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. Em *29th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 929–947. 10.1145/3620665.3640366

Bakhtin, Mikhail. 2016. *Os gêneros do discurso*. São Paulo: Editora 34

Bakker, Dik. 1998. Flexibility and consistency in word order patterns in the languages of europe. Em Anna Siewierska (ed.), *Constituent Order in the Languages of Europe*, 383–420. De Gruyter Mouton. 10.1515/9783110812206.383

Beltagy, Iz, Kyle Lo & Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. arXiv [cs.CL]. 10.48550/arXiv.1903.10676

Benavoli, Alessio, Giorgio Corani, Janez Demšar & Marco Zaffalon. 2017. Time for a change:

a tutorial for comparing multiple classifiers through bayesian analysis. *Journal of Machine Learning Research* 18(77). 1–36.

Benavoli, Alessio, Francesca Mangili, Giorgio. Corani, Marco Zaffalon & Fabrizio Ruggeri. 2014. A Bayesian Wilcoxon signed-rank test based on the dirichlet process. Em *31st International Conference on Machine Learning (ICML)*, II–1026—-II–1034.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62(2). 384–414. 10.2307/414678

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. 10.1017/cbo9780511621024

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27(1). 3–44. 10.1515/ling.1989.27.1.3

Biber, Douglas. 1994. An analytical framework for register studies. Em *Sociolinguistic Perspectives on Register*, 31–56. Oxford University Press. 10.1093/oso/9780195083644.003.0003

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press. 10.1017/cbo9780511519871

Biber, Douglas. 2001. Dimensions of variation among eighteenth-century speech-based and written registers. Em Douglas Biber & Susan Conrad (eds.), *Variation in English*, 200–214. London: Taylor & Francis

Biber, Douglas. 2006. Register: Overview. Em Keith Brown (ed.), *Encyclopedia of Language & Linguistics*, 476–482. Oxford: Elsevier 2nd edn. 10.1016/B0-08-044854-2/04306-6

Biber, Douglas & Susan Conrad. 2009a. *Historical evolution of registers, genres, and styles* 222–256. Cambridge: Cambridge University Press. 10.1017/CBO9780511814358.006

Biber, Douglas & Susan Conrad. 2009b. *Interpersonal spoken registers* 85–108. Cambridge: Cambridge University Press. 10.1017/CBO9780511814358.004

Biber, Douglas & Susan Conrad. 2009c. *Register, genre, and style*. Cambridge: Cambridge University Press. 10.1017/CBO9780511814358

Biber, Douglas & Susan Conrad. 2019. *Written registers, genres, and styles* 111–142. Cambridge: Cambridge University Press. 10.1017/9781108686136.005

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1(1). 1–37. 10.3366/cor.2006.1.1.1

Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. Cambridge: Cambridge University Press. 10.1017/9781316388228

Biber, Douglas & Edward Finegan. 1989. Drift and the evolution of English style: A history of three genres. *Language* 65(3). 487–517. 10.2307/415220

Biber, Douglas & Edward Finegan. 1992. The linguistic evolution of five written and speech-based English genres from the 17th to the 20th centuries. Em Matti Rissanen, Ossi Ihalainen, Terttu Nevalainen & Irma Taavitsainen (eds.), *History of Englishes*, 688–704. De Gruyter Mouton. 10.1515/9783110877007.688

Biber, Douglas & Edward Finegan. 2001. Diachronic relations among speech-based and written registers in English. Em Douglas Biber & Susan Conrad (eds.), *Variation in English*, 66–83. London: Taylor & Francis

Brown, Penelope & Colin Fraser. 1979. Speech as a marker of situation. Em Klaus R. Scherer & Howard Giles (eds.), *Social Markers in Speech*, 33–62. Cambridge: Cambridge University Press

Carapinha, Conceição. 2018. A linguagem jurídica. contributos para uma caracterização dos códigos legais. *Redis: Revista de Estudos do Discurso* 7. 91–119.

Castro, Dayvid, Ellen Souza & Adriano L.I. De Oliveira. 2016. Discriminating between Brazilian and European Portuguese national varieties on Twitter texts. Em *5th Brazilian Conference on Intelligent Systems (BRACIS)*, 265–270. 10.1109/BRACIS.2016.056

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

de Colla Furquim, Luis Otávio & Vera Lúcia Strube de Lima. 2012. Clustering and categorization of Brazilian Portuguese legal documents. Em *Computational Processing of the Portuguese Language (PROPOR)*, 272–283

Conrad, Susan. 2001. Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. Em Douglas Biber & Susan Conrad (eds.), *Variation in English*, 94–107. London: Taylor & Francis

Cortes, Corinna & Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3). 273–297. 10.1007/BF00994018

Costa, Pablo Botton, Matheus Camasmie Pavan, Wesley Ramos Santos, Samuel Caetano Silva & Ivandré Paraboni. 2023. BERTabaporu: Assessing a genre-specific language model for Portuguese NLP. Em *14th International Conference on Recent Advances in Natural Language Processing (RANLP)*, 217–223.

Costa, Sérgio Roberto. 2018. *Dicionário de gêneros textuais*. Autêntica Editora.

Crespo, Maria Clara R. Morales, Maria Lina de S. 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, Marcelo Finger, Maria Clara P. de Sousa, Cristiane Namiuti & Vanessa Martins do Monte. 2023. Carolina: a general corpus of contemporary Brazilian Portuguese with provenance, typology and versioning information. arXiv [cs.CL]. 10.48550/arXiv.2303.16098

Croft, William. 2002. *Typology and universals*. Cambridge: Cambridge University Press 2nd edn. 10.1017/CBO9780511840579

Crystal, David & Derek Davy. 1969. *Investigating English style*. London: Longman. 10.4324/9781315538419

de Faria Marques, Carolina Godoi. 2023. *Análise multidimensional dos textos legais federais brasileiros*: Universidade Federal de Minas Gerais (UFMG). Tese de Mestrado.

Delfino, Maria Claudia Nunes. 2021. Análise multidimensional: Os números na linguística. *Cadernos de Linguística* 2(4). e474. 10.25189/2675-4916.2021.v2.n4.id474

Dunn, Jonathan. 2019. Modeling global syntactic variation in English using dialect classification. Em *6th Workshop on NLP for Similar Languages, Varieties and Dialects*, 42–53. 10.18653/v1/W19-1405

Duranti, Alessandro. 1985. Sociocultural dimensions of discourse. Em Teun A. van Dijk (ed.), *Handbook of Discourse Analysis*, 193–230. London: Academic Press

Ehret, Katharina. 2021. An information-theoretic view on language complexity and register variation: Compressing naturalistic corpus data. *Corpus Linguistics and Linguistic Theory* 17(2). 383–410. 10.1515/cllt-2018-0033

Ehret, Katharina & Benedikt Szmrecsanyi. 2016. An information-theoretic approach to assess linguistic complexity. Em Raffaela Baechler & Guido Seiler (eds.), *Complexity, Isolation, and Variation*, 71–94. De Gruyter. 10.1515/9783110348965-004

Ehret, Katharina & Benedikt Szmrecsanyi. 2019. Compressing learner language: An information-theoretic measure of complexity in sla production data. *Second Language Research* 35(1). 23–45. 10.1177/0267658316669559

Ellis, Jeffrey & Jean Ure. 1969. Language varieties: Register. Em *Encyclopedia of Linguistics, Information and Control*, 251–259. London: Pergamon Press

Eskelinen, Anni, Amanda Myntti, Erik Henriksson, Sampo Pyysalo & Veronika Laippala. 2024. Building question-answer data using web register identification. Em *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2595–2611.

Ferguson, Charles A. 1994. Dialect, register, and genre: Working assumptions about conventionalization. Em Douglas Biber & Edward Finegan (eds.), *Sociolinguistic Perspectives On Register*, 15–30. Oxford University Press. 10.1093/oso/9780195083644.003.0002

Fonseca, Erick Rocha, Leandro Borges dos Santos, Marcelo Criscuolo & Sandra Maria Aluísio. 2016. Visão geral da avaliação de similaridade semântica e inferência textual. *Linguamática* 8(2). 3–13.

Ghadessy, Mohsen (ed.). 1988. *Registers of written English: Situational factors and linguistic features*. London: Pinter Publishers

Goswami, Koustava, Rajdeep Sarkar, Bharathi Raja Chakravarthi, Theodorus Fransen & John P. McCrae. 2020. Unsupervised deep language and dialect identification for short texts. Em *28th International Conference on Computational Linguistics (COLING)*, 1606–1617. 10.18653/v1/2020.coling-main.141

Gozdz-Roszkowski, Stanislaw. 2011. *Patterns of linguistic variation in american legal English*. Berlin, Germany: Peter Lang Verlag. 10.3726/978-3-653-00659-9

Gregory, Michael. 1967. Aspects of varieties differentiation. *Journal of Linguistics* 3(2). 177–198. 10.1017/S0022226700016601

Grillo, Sheila V. de Camargo. 2006. Esfera e campo. Em Beth Brait (ed.), *Bakhtin: Outros Conceitos-Chave*, Contexto

Gu, Yu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao & Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* 3(1). 10.1145/3458754

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Rodrigues & Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *XI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 122–131.

He, Pengcheng, Xiaodong Liu, Jianfeng Gao & Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv [cs.CL/cs.LG]. 10.48550/arXiv.2006.03654

Henriksson, Erik, Amanda Myntti, Saara Hellström, Anni Eskelinen, Selcen Erten-Johansson & Veronika Laippala. 2024. Automatic register identification for the open web using multilingual deep learning. arXiv [cs.CL]. 10.48550/arXiv.2406.19892

Hymes, Dell. 1974. *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press

Jung, Jeesu, Sangkeun Jung, Hyein Seo, Hyuk Namgung & Sungryeol Kim. 2022. Sequence alignment ensemble with a single neural network for sequence labeling. *IEEE Access* 10. 73562–73570. 10.1109/ACCESS.2022.3188107

Juola, Patrick. 2008. Assessing linguistic complexity. Em *Language Complexity*, 89–108. John Benjamins.

Jurafsky, Dan & James H. Martin. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall

Katinskaya, Anisya & Serge Sharoff. 2015. Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres. Em *5th Workshop on Balto-Slavic Natural Language Processing (BSNLP)*, 65–74.

Kauffmann, Carlos Henrique. 2005. *O corpus do jornal: variação lingüística, gêneros e dimensões da imprensa diária escrita*: Pontifícia Universidade Católica de São Paulo (PUC-SP). Tese de Mestrado.

Kauffmann, Carlos Henrique. 2020. *Linguística de corpus e estilo: análises multidimensional e canônica na ficção de Machado de Assis*: Pontifícia Universidade Católica de São Paulo (PUC-SP). Tese de Doutoramento.

Kauffmann, Carlos Henrique. 2022. Cognição e variação linguística de gêneros/registros jornalísticos: um estudo baseado em corpus. *Domínios de Lingu@gem* 16(4). 1266—-1291. 10.14393/DL52-v16n4a2022-3

Kessler, Brett, Geoffrey Nunberg & Hinrich Schuetze. 1997. Automatic detection of text genre. Em *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL/EACL)*, 32–38. 10.3115/976909.979622

Komesu, Fabiana Cristina. 2005. Blogs e as práticas de escrita sobre si na internet. Em Luiz Antônio Marcuschi & Antônio Carlos Xavier (eds.), *Hipertexto e gêneros digitais: novas formas de construção do sentido*, 110–119. Lucerna

Kutuzov, Andrey, Elizaveta Kuzmenko & Anna Marakasova. 2016. Exploration of register-dependent lexical semantics using word embeddings. Em *Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, 26–34.

Kuzman, Taja, Igor Mozetič & Nikola Ljubešić. 2023a. Automatic genre identification for robust enrichment of massive text collections: Investigation of classification methods in the era of large language models. *Machine Learning and Knowledge Extraction* 5(3). 1149–1175. 10.3390/make5030059

Kuzman, Taja, Peter Rupnik & Nikola Ljubešić. 2022. The GINCO training dataset for web genre identification of documents out in the wild. Em *13th Language Resources and Evaluation Conference (LREC)*, 1584–1594.

Kuzman, Taja, Peter Rupnik & Nikola Ljubešić. 2023b. Get to know your parallel data: Performing English variety and genre classification over MaCoCu corpora. Em *10th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 91–103. 10.18653/v1/2023.vardial-1.9

Kyröläinen, Aki-Juhani & Veronika Laippala. 2023. Predictive keywords: Using machine learning to explain document characteristics. *Frontiers in Artificial Intelligence* 5. 975729. 10.3389/frai.2022.975729

Laippala, Veronika, Jesse Egbert, Douglas Biber & Aki-Juhani Kyröläinen. 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation* 55(3). 757–788. 10.1007/s10579-020-09519-z

Laippala, Veronika, Roosa Kyllönen, Jesse Egbert, Douglas Biber & Sampo Pyysalo. 2019. Toward multilingual identification of online registers. Em *22nd Nordic Conference on Computational Linguistics*, 292–297.

Laippala, Veronika, Juhani Luotolahti, Aki-Juhani Kyröläinen, Tapio Salakoski & Filip Ginter. 2017. Creating register sub-corpora for the Finnish internet parsebank. Em *21st Nordic Conference on Computational Linguistics*, 152–161.

Laippala, Veronika, Samuel Rönnqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi & Sampo Pyysalo. 2020. From web crawl to clean register-annotated corpora. Em *12th Web as Corpus Workshop (WAC)*, 14–22.

Laippala, Veronika, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert & Sampo Pyysalo. 2023. Register identification from the unrestricted open web using the corpus of online registers of English. *Language Resources and Evaluation* 57(3). 1045–1079. 10.1007/s10579-022-09624-1

Laippala, Veronika, Anna Salmela, Samuel Rönnqvist, Alham Fikri Aji, Li-Hsin Chang, Asma Dhifallah, Larissa Goulart, Henna Kortelainen, Marc Pàmies, Deise Prina Dutra, Valtteri Skantsi, Lintang Sutawika & Sampo Pyysalo. 2022. Towards better structured and less noisy web data: Oscar with register annotations. Em *8th Workshop on Noisy User-Generated Text (w-NUT)*, 215–221.

Leal, Sidney Evaldo & Sandra Maria Aluísio. 2024. Complexidade textual e suas tarefas relacionadas. Em Helena M. Caseli & Maria G. V. Nunes (eds.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, BPLN.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2024. NILC-metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation* 58. 73–110. 10.1007/s10579-023-09693-w

Leal, Sidney Evaldo, Katerina Lukasova, Maria Teresa Carthery-Goulart & Sandra Maria Aluísio. 2022. RastrOS project: Natural language processing contributions to the development of an eye-tracking corpus with predictability norms for Brazilian Portuguese. *Language Resources and Evaluation* 56. 1333–1372. 10.1007/s10579-022-09609-0

van der Lee, Chris & Antal van den Bosch. 2017. Exploring lexical and syntactic features for language variety identification. Em *4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, 190–199. 10.18653/v1/W17-1224

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So & Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4). 1234–1240. 10.1093/bioinformatics/btz682

Lefer, Marie-Aude & Svetlana Vogeleer (eds.). 2016. *Genre and register-related discourse features in contrast*. Amsterdam: John Benjamins. 10.1075/bct.87

Loshchilov, Ilya & Frank Hutter. 2019. Decoupled weight decay regularization. arXiv [cs.LG/cs.NE/math.OC]. 10.48550/arXiv.1711.05101

Mantel, Nathan. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27(2_Part_1). 209–220

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv [cs.CL]. 10.48550/arXiv.1301.3781

Monte-Serrat, Dionéia, Mateus Machado & Evandro Ruiz. 2021. A machine learning approach to literary genre classification on Portuguese texts: circumventing NLP's standard varieties. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, 255–264. 10.5753/stil.2021.17805

Myntti, Amanda, Liina Repo, Elian Freyermuth, Antti Kanner, Veronika Laippala &

Erik Henriksson. 2024. Intersecting register and genre: Understanding the contents of web-crawled corpora. Em *4th International Conference on Natural Language Processing for Digital Humanities (NLP4DH)*, 386–397. 10.18653/v1/2024.nlp4dh-1.38

Nauege, João Muteteca. 2022. A língua de especialidade: um olhar sobre o português jurídico, tendências e desafios em Angola. *NJINGA e SEPÉ* 2(1). 247—256.

Ortiz Suárez, Pedro Javier, Benoît Sagot & Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Em *Workshop on Challenges in the Management of Large Corpora (CMLC)*, 9–16. 10.14618/ids-pub-9021

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(85). 2825–2830.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. GloVe: Global vectors for word representation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. 10.3115/v1/D14-1162

Pinto, Marcia Veirano. 2013. *A linguagem dos filmes norte-americanos ao longo dos anos: uma abordagem multidimensional*: Pontifícia Universidade Católica de São Paulo (PUC-SP). Tese de Doutoramento.

Pinto, Marcia Veirano. 2014. Dimensions of variation in north american movies. Em Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis, 25 years on*, 109–148. John Benjamins. 10.1075/scl.60.04vei

Polo, Felipe Maia, 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 & Renato Vicente. 2021. LegalNLP – natural language processing methods for the Brazilian legal language. arXiv [cs.CL]. 10.48550/arXiv.2110.15709

Pomikálek, Jan. 2011. *Removing boilerplate and duplicate content from web corpora*: Masaryk University. Tese de Doutoramento.

Repo, Liina, Brett Hashimoto & Veronika Laippala. 2023. In search of founding era registers: automatic modeling of registers from the corpus of founding era American English. *Digital Scholarship in the Humanities* 38(4). 1659–1677. 10.1093/llc/fqad049

Repo, Liina, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Miika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo & Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. Em *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 183–191. 10.18653/v1/2021.eacl-srw.24

Rideout, Jai Ram, Greg Caporaso, Evan Bolyen, Daniel McDonald, Yoshiki Vázquez Baeza, Jorge Cañardo Alastuey, Anders Pitman, Jamie Morton, Jose Navas, Kestrel Gorlick, Justine Debelius, Zech Xu, llcooljohn, adamrp, Joshua Shorenstein, Laurent Luce, Will Van Treuren, charudatta navare, Antonio Gonzalez, Colin J. Brislawn, Weronika Patena, Karen Schwarzberg, teravest, Jens Reeder, shiffer1, Igor Sfiligoi, nbresnick, Qiyun Zhu, Dr. K. D. Murray & Karan Sharma. 2023. biocore/scikit-bio. v.0.5.9. 10.5281/zenodo.8209901

Rodrigues, João, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso & Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer Albertina PT-*. Em *22nd EPIA Conference on Artificial Intelligence*, 441–453. 10.1007/978-3-031-49008-8_35

Rönnqvist, Samuel, Aki-Juhani Kyröläinen, Amanda Myntti, Filip Ginter & Veronika Laippala. 2022. Explaining classes through stable word attributions. Em *Findings of the ACL*, 1063–1074. 10.18653/v1/2022.findings-acl.85

Rönnqvist, Samuel, Valtteri Skantsi, Miika Oinonen & Veronika Laippala. 2021. Multilingual and zero-shot is closing in on monolingual web register classification. Em *23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 157–165.

Rousseeuw, Peter J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20. 53–65. 10.1016/0377-0427(87)90125-7

Sadeniemi, Markus, Kimmo Kettunen, Tiina Lindh-Knuutila & Timo Honkela. 2008.

Complexity of european union languages: A comparative approach. *Journal of Quantitative Linguistics* 15(2). 185–211. 10.1080/09296170801961843

Saeed, Elaf A., Ammar D. Jasim & Munther A. Abdul Malik. 2024. Cuneiform text dialect identification using machine learning algorithms and natural language processing. *Iraqi Journal of Information and Communication Technology* 7(2). 26–40. 10.31987/ijict.7.2.265

Şapcı, Ali Osman Berk, Hasan Kemik, Reyyan Yeniterzi & Oznur Tastan. 2024. Focusing on potential named entities during active label acquisition. *Natural Language Engineering* 30(3). 602—-624. 10.1017/S1351324923000165

Sardinha, Tony Berber. 2014. 25 years later: Comparing internet and pre-internet registers. Em Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis, 25 years on*, 81–108. John Benjamins. 10.1075/scl.60.03ber

Sardinha, Tony Berber. 2017. Text types in Brazilian Portuguese: a multidimensional perspective. *Corpora* 12(3). 483–515. 10.3366/cor.2017.0129

Sardinha, Tony Berber & Shannon Fitzsimmons-Doolan. 2025. *Lexical multidimensional analysis: Identifying discourses and ideologies*. Cambridge: Cambridge University Press

Sardinha, Tony Berber, Carlos Kauffmann & Cristina Mayer Acunzo. 2014a. Dimensions of register variation in Brazilian Portuguese. Em Tony Berber Sardinha & Marcia Veirano Pinto (eds.), *Multi-Dimensional Analysis, 25 Years On*, 35–80. John Benjamins. 10.1075/scl.60.02ber

Sardinha, Tony Berber, Carlos Kauffmann & Cristina Mayer Acunzo. 2014b. A multidimensional analysis of register variation in brazilian portuguese. *Corpora* 9(2). 239–271. 10.3366/cor.2014.0059.

Sardinha, Tony Berber & Marcia Veirano Pinto. 2017. American television and off-screen registers: a corpus-based comparison. *Corpora* 12(1). 85–114. 10.3366/cor.2017.0110

Sarti, Gabriele, Dominique Brunato & Felice Dell'Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. Em *Workshop on Cognitive Modeling and Computational Linguistics*, 48–60. 10.18653/v1/2021.cmcl-1.5

Serras, Felipe & Marcelo Finger. 2021. verBERT: Automating brazilian case law document multi-label categorization using BERT. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 237–246. 10.5753/stil.2021.17803

Sharoff, Serge. 2007. Classifying web corpora into domain and genre using automatic feature identification. *Cahiers du Cental* 5.

Skantsi, Valtteri & Veronika Laippala. 2023. Analyzing the unrestricted web: The Finnish corpus of online registers. *Nordic Journal of Linguistics* 1. 1–31. 10.1017/S0332586523000021

Sousa, Hugo, Rúben Almeida, Purificação Silvano, Inês Cantante, Ricardo Campos & Alípio Jorge. 2025. Enhancing Portuguese variety identification with cross-domain approaches. arXiv [cs.CL]. 10.48550/arXiv.2502.14394

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 403—-417. 10.1007/978-3-030-61377-8_28

Sparck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation* 28(1). 11–21.

Sturzeneker, Mariana, Maria Clara Crespo, Maria Lina Rocha, Marcelo Finger, Maria Clara Paixão de Sousa, Vanessa Martins do Monte & Cristiane Namiuti. 2022. Carolina's methodology: building a large corpus with provenance and typology information. Em *Digital Humanities and Natural Language Processing*, 53–58.

TEI Consortium. 2021. TEI P5: Guidelines for Electronic Text Encoding and Interchange. Last updated on 9th April 2021, revision 609a109b1.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. Em *Advances in Neural Information Processing Systems*,

Villanueva Llerena, Julissa Giuliana & Denis Deratani Mauá. 2017. *Multi-label classification based on sum-product networks*: Universidade de São Paulo. Tese de Mestrado. 10.11606/D.45.2017.tde-08122017-100124

Watanabe, Shuhei. 2025. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv [cs.LG/cs.AI]. 10.48550/arXiv.2304.11127

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 38–45. 10.18653/v1/2020.emnlp-demos.6

Wu, Shijie, Ryan Cotterell & Mans Hulden. 2021. Applying the transformer to character-level transduction. Em *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1901–1907. 10.18653/v1/2021.eacl-main.163

Zampieri, Marcos & Binyan Gebre. 2012. Automatic identification of language varieties: The case of Portuguese. Em *Conference on Natural Language Processing*, 233–237.

Zampieri, Marcos, Shervin Malmasi & Mark Dras. 2016a. Modeling language change in historical corpora: The case of Portuguese. Em *10th International Conference on Language Resources and Evaluation (LREC)*, 4098–4104.

Zampieri, Marcos, Shervin Malmasi, Octavia-Maria Sulea & Liviu P. Dinu. 2016b. A computational approach to the study of Portuguese newspapers published in Macau. Em *Natural Language Processing meets Journalism*, 47–51.

Zhang, Yu, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu & Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. Em *29th International Conference on Computational Linguistics (COLING)*, 4212–4227.

Zhou, Xujuan, Xiaohui Tao, Jianming Yong & Zhenyu Yang. 2013. Sentiment analysis on tweets for social events. Em *17th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 557–562. 10.1109/CSCWD.2013.6581022

Zuppardi, Maria Carolina. 2020. *Collocation dimensions in academic English*: Pontifícia Universidade Católica de São Paulo. Tese de Doutoramento

Apêndices

Tag	Classe	Descrição	Exemplos
SCONJ	Conjunção subordinativa	Palavra gramatical usada para introduzir orações subordinadas, estabelecendo dependência sintática.	"Ela chorou **quando** ouviu a notícia."; "Eu saí mais cedo **porque** estava cansado."
VERB	Verbo	Palavra que expressa ação, processo, estado ou ocorrência, e pode flexionar-se em tempo, modo, aspecto, pessoa e número.	"Ela **estuda** todos os dias."; "João **quebrou** o vaso."
PROPN	Nome Próprio	Subcategoria de substantivo que designa entidades únicas, como pessoas, lugares ou organizações.	"**Lucas** chegou cedo na escola."; "**Pedro** e **Ana** estudam para a prova juntos."
PRON	Pronome	Palavra que substitui ou acompanha um substantivo, referindo-se a participantes do discurso ou a elementos do contexto.	"**Ela** gosta de ler livros."; "**Isso** é muito interessante."
CCONJ	Conjunção coordenativa	Palavra ou expressão que liga duas orações ou termos independentes.	"Estudei bastante, **mas** não consegui tirar uma boa nota."; "Fui ao mercado, **e** comprei frutas."
ADV	Advérbio	Palavra que modifica verbos, adjetivos, outros advérbios ou frases inteiras, expressando circunstâncias como tempo, modo, lugar, intensidade etc.	"Ela correu **rapidamente** para o ônibus."; "Maria falou **muito** durante a reunião."
AUX	Verbo auxiliar	Verbo usado em conjunto com o verbo principal para formar tempos compostos, passivas ou perífrases verbais.	"Ela **está** estudando para a prova."; "Nós **vamos** viajar amanhã."
ADJ	Adjetivo	Palavra que qualifica ou caracteriza um substantivo, atribuindo-lhe uma propriedade ou estado.	"O cachorro **grande** latiu durante a noite."; "A casa **antiga** precisa de reformas."
DET	Determinante	Palavra que acompanha um substantivo, modificando-o e especificando seu sentido.	"**Este** livro é muito interessante."; "**A** menina brinca no parque."
NOUN	Substantivo comum	Palavra que designa seres, objetos, conceitos ou entidades de forma geral, podendo funcionar como núcleo do sujeito ou do objeto.	"O **carro** está estacionado na rua."; "A **criança** brinca no jardim."

Tag	Classe	Descrição	Exemplos
ADP	Preposição	Palavra invariável que estabelece uma relação entre dois termos de uma frase, subordinando um ao outro. Essa relação pode indicar diversas conexões, como posse, lugar, tempo, causa, entre outras.	"O livro está **sobre** a mesa."; "Ela foi **para** a escola cedo hoje."
INTJ	Interjeição	Palavra ou expressão que comunica emoções, reações ou chamamentos de forma independente da estrutura sintática.	"**Nossa**! Que dia incrível!"; "**Ai**! Eu me machuquei."
NUM	Numeral	Palavra variável que indica um número exato ou a posição que tal coisa ocupa numa série.	"Tenho **três** livros na minha mochila."; "Ele terminou em **primeiro** lugar na competição."
X	Outro / indefinido	Categoria usada para palavras que não se encaixam em nenhuma classe tradicional ou que não puderam ser analisadas.	"O termo **blorf** parece estranho e não pertence a nenhuma classe gramatical conhecida."; "No jogo, apareceu a palavra **zzzorp** que ninguém entendeu."
PUNCT	Pontuação	Sinais gráficos usados para estruturar e organizar o texto, indicando pausas, entonações e relações sintáticas.	"Vamos ao parque amanhã**!**"; "Eu gosto de frutas — especialmente maçãs, laranjas e bananas — porque são saudáveis**.**"
SYM	Símbolo	Qualquer caractere ou grupo de caracteres com função simbólica, como operadores matemáticos, hashtags, sinais monetários etc.	"Na fórmula, usamos o símbolo + para somar e o sinal **@** para marcar alguém na rede social."

Tabela 16: Descrições e exemplos das etiquetas de partes do discurso (*Part-of-Speech* — PoS) usadas neste trabalho.

Domínio verdadeiro	Resposta dos modelos	Texto
Jornalístico	Instrucional	A mistura de rock, country e R&B trouxeram muitos fãs.
	Entretenimento	Não dá boas entrevistas.
	Virtual	Rayane e Ygor não são atores profissionais, e os personagens têm os nomes dos atores.
	Jurídico	Caberá ao ministro vice-presidente, no exercício da Presidência do STF, decidir casos urgentes que forem encaminhados ao Tribunal.
Instrucional	Jornalístico	Os dados nos países em desenvolvimento são pouco claros.
	Entretenimento	Como é sua superfície?
	Virtual	As imagens devem ser salvas com qualidade suficiente para futura edição e visualização.
	Jurídico	Segundo a Constituição, somente a lei pode limitar a liberdade de expressão e estabelecer limites para sua expressão.
Entretenimento	Jornalístico	Tem gente treinando os músculos da criatividade.
	Instrucional	As pernas estão dobradas e jogadas para trás e os braços para frente.
	Virtual	Se devo ser a única a propor padrões exemplares a serem seguidos, que assim seja.
	Jurídico	Com o devido respeito, Presidente, cabe ao Senado decidir isso, não a nós.
Virtual	Entretenimento	Choro digitando isto e lembrando da minha rainha.
	Jornalístico	Antes de achar que há uma conspiração universal para nos derrubar, temos que esgotar as possibilidades de o problema estar dentro de casa.
	Instrucional	A falta de ar geralmente se desenvolve vários dias após os sintomas iniciais.
	Jurídico	Descobri que nossos Deputados Federais, em sua maioria, são péssimos.
Jurídico	Entretenimento	Ele está no último do círculo do Inferno de Dante.
	Jornalístico	Os brasileiros não suportam mais falsos protecionismos cujo único resultado é o atraso, a ignomínia de um povo.
	Instrucional	Em seguida, colocar a fita de papel microperfurada.
	Virtual	Se os anexos têm um tamanho muito grande, o que devo fazer?

Tabela 17: Exemplos de casos onde todos os modelos baseados em *transformer* erraram a predição do domínio. Apesar de esses serem apenas uma ínfima parcela dos exemplos de confusão dos modelos é possível observar diversos fenômenos de indução ao erro, seja em termos de estrutura das sentenças, assuntos tratados ou elementos lexicais característicos.