UNIVERSIDADE FEDERAL DE MINAS GERAIS
Faculdade de Letras
Programa de Pós-graduação em Estudos Linguísticos

Carolina Godoi de Faria Marques

ANÁLISE MULTIDIMENSIONAL DOS TEXTOS LEGAIS FEDERAIS BRASILEIROS

Belo Horizonte
2023

Carolina Godoi de Faria Marques

ANÁLISE MULTIDIMENSIONAL DOS TEXTOS LEGAIS FEDERAIS BRASILEIROS

Versão final

Dissertação apresentada ao Programa de Pós-Graduação em Estudos Linguísticos da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Linguística Teórica e Descritiva

Área de Concentração: Linguística Teórica e Descritiva
Linha de Pesquisa: Linguística de Corpus

Orientador: Profa. Dra. Lúcia de Almeida Ferrari

Belo Horizonte

2023

Marques, Carolina Godoi de Faria.
M357a Análise multidimensional dos textos legais federais brasileiros [manuscrito] / Carolina Godoi de Faria Marques. – 2023.

1 recurso online (208 f. : il., grafs., color., tabs., p&b.) : pdf.

Orientadora: Lúcia de Almeida Ferrari.

Área de concentração: Linguística Teórica e Descritiva.

Linha de Pesquisa: Linguística de Corpus.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Faculdade de Letras.

Bibliografia: f. 137-143.

Apêndices: f. 144-191.

Anexos: f. 192-208.

Exigências do sistema: Adobe Acrobat Reader.

1. Linguística de corpus – Teses. 2. Língua portuguesa – Variação – Teses. 3. Direito – Linguagem – Teses. I. Ferrari, Lúcia de Almeida. II. Universidade Federal de Minas Gerais. Faculdade de Letras. III. Título.

CDD : 410

Ficha catalográfica elaborada pelo Bibliotecário Israel José da Silva – CRB/6-2128

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ESTUDOS LINGUÍSTICOS

FOLHA DE APROVAÇÃO

ANÁLISE MULTIDIMENSIONAL DOS TEXTOS LEGAIS FEDERAIS BRASILEIROS

CAROLINA GODOI DE FARIA MARQUES

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTUDOS LINGUÍSTICOS, como requisito para obtenção do grau de Mestre em ESTUDOS LINGUÍSTICOS, área de concentração LINGUÍSTICA TEÓRICA E DESCRITIVA, linha de pesquisa Estudos Linguísticos Baseados em Corpora.

Aprovada em 30 de janeiro de 2023, pela banca constituída pelos membros:

Prof(a). Lucia de Almeida Ferrari - Orientadora

UFMG

Prof(a). Deise Prina Dutra

UFMG

Prof(a). Carlos Henrique Kauffmann

PUC-SP

Belo Horizonte, 30 de janeiro de 2023.

Documento assinado eletronicamente por **Lucia de Almeida Ferrari**, **Professora do Magistério Superior**, em 30/01/2023, às 16:19, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Deise Prina Dutra**, **Professora do Magistério Superior**, em 30/01/2023, às 16:22, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Carlos Henrique Kauffmann**, **Usuário Externo**, em 31/01/2023, às 14:12, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

AGRADECIMENTOS

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa (nº 88887.626989/2021-00) que permitiu minha dedicação integral aos estudos e à pesquisa.

Agradeço a minha orientadora Profa. Dra. Lucia Almeida Ferrari por todo o trabalho e assistência durante esses últimos 2 anos.

Agradeço ao Dr. Carlos Henrique Kauffmann pela etiquetação e pós processamento do corpus, assim como pela sua disponibilidade e ajuda durante a realização da análise multidimensional.

Agradeço à Prof. Dra. Deise Prina Dutra pelos seus esforços para a etiquetação do corpus.

Agradeço aos membros da banca pela disponibilidade e análise criteriosa deste trabalho.

Resumo

Nas últimas décadas, houve uma crescente produção científica na Linguística em interface com o Direito, voltada à linguagem jurídica. A linguagem jurídica é uma linguagem especializada, com características lexicais e gramaticais que a caracterizam como tal (DAHLMAN, 2006; RICHARD, 2011; GOŹDŹ-ROSZKOWSKI e POTRANDOLFO, 2015) o que permite classificá-la como um registro, segundo a perspectiva de Biber e Conrad (2009). Além disso, ela compreende uma ampla gama de sub-registros, tais como a linguagem utilizada nos tribunais, em artigos acadêmicos e nas legislações, cada uma com suas especificidades (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Por fim, cada idioma, cada tradição jurídica e cada país tem suas particularidades jurídicas. Estudos anteriores examinaram essa linguagem sob diferentes ângulos, incluindo terminologia, tradução, análise do discurso e variação linguística. Verifica-se ainda um grande número de corpora jurídicos/legais (PORTANDOLFO, 2012; GIAMPIERI, 2018). No português brasileiro, a literatura sobre a linguagem jurídica se volta para aspectos linguísticos específicos, em especial nas áreas da terminologia, tradução e análise do discurso, não havendo análises aprofundadas nem um corpus exclusivamente jurídico/legal brasileiro (FERRARI e CUNHA, 2022). Além disso, apesar do crescente interesse nessa linguagem, ainda são poucos os estudos sobre a sua variação. Visando preencher esta lacuna, partiu-se da perspectiva de registro para investigar a variação linguística dos textos legais federais brasileiros. Para tanto, utilizou-se a metodologia da Linguística de Corpus e a abordagem da Análise Multidimensional (AMD) (BIBER, 1988). A AMD é uma abordagem empírico-metodológica baseada em corpus aplicável a qualquer língua. Ela permite o estudo aprofundado dos registros através da descrição e caracterização de sua variação em dimensões via análise estatística de um corpus e interpretação funcional dos resultados (BIBER, 1988; BIBER e CONRAD, 2009). Existem dois tipos de AMD: completa e aditiva. A primeira, mais detalhada, permite a identificação de dimensões em um corpus. Já a segunda, mais global, aplica as dimensões identificadas em uma AMD completa a um determinado corpus (BERBER SARDINHA, 2013). No presente estudo realizamos uma AMD aditiva do corpus *LEX-BR-Ius* (FERRARI e MARQUES, em compilação) adotando como base o estudo de Berber Sardinha, Kauffmann e Acunzo (2014) visando mapear os textos legais em todas as dimensões de variação nele identificadas. Nosso corpus é escrito, monolíngue, sincrônico, anotado em *Part of speech* (POS) e lema, com marcação textual em *Modest* XML (HARDIE, 2014) e amostra textos legais federais brasileiros completos em vigor na data de extração (2022), balanceados por frequência de uso segundo os resultados da busca no portal JusBrasil. Concluiu-se que os textos legais são um registro segundo a perspectiva de Biber e Conrad (2009), sendo caracterizados pelo uso de verbos no futuro, verbos modais, conjunções coordenadas, orações subordinadas, construções passivas sem agente, substantivos, adjetivos, preposições, artigos definidos, nominalizações, riqueza lexical e densidade informacional. Os resultados revelaram ainda que o processo de criação de cada espécie normativa é distinto e que cada texto legal se distingue pelo seu contexto de comunicação, propósito comunicacional e assunto. Com esta pesquisa visa-se contribuir com os estudos sobre linguagem jurídica e variação linguística.

Palavras-chave: Português brasileiro. Textos legais federais brasileiros. Análise multidimensional. Registro. Variação.

Abstract

In the past decades, there has been a growing scientific production in Linguistics in interface with Law, focused on legal language. Legal language is a highly specialized language, with lexical and grammatical características that characterize it as such (DAHLMAN, 2006; RICHARD, 2011; GOŹDŹ-ROSZKOWSKI and POTRANDOLFO, 2015) what allow us to classify it as a register according to Biber and Conrad's perspective (2009). Moreover, it comprises a wide range of sub-registers, such as the language used in courts, in academic articles, and legislations, each with its specificities (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Finally, each language, each legal tradition, and each country has its own legal particularities. Previous studies have examined this language from different angles, including terminology, translation, discourse analysis, and linguistic variation. There is also a large number of legal corpora (PORTANDOLFO, 2012; GIAMPIERI, 2018). In Brazilian Portuguese, the literature on legal language addresses specific linguistic aspects, with most studies being in the areas of terminology, translation and discourse analysis, and there are no in-depth analyses or an exclusively Brazilian legal corpus (FERRARI and CUNHA, 2022). Despite its growing popularity, there are still few studies on legal language's variation. In order to fill this gap, we adopted the register perspective to investigate the linguistic variation of Brazilian Federal Statutory Legal Texts in force. To achieve our goals, we used the Corpus Linguistics methodology, the Multidimensional Analysis (MD) approach (BIBER, 1989), and the *LEX-BR-Ius* corpus (FERRARI and MARQUES, in preparation). The MD is a corpus-based empirical-methodological approach applicable to any language. This methodology allows for an in-depth study of registers by describing and characterizing their variation in dimensions via statistical analysis of a corpus and functional interpretation of the results (BIBER, 1988; BIBER and CONRAD, 2009). There are two types of MD: complete and additive. The first, more detailed, allows the identification of dimensions in a corpus. The second, more global, applies the dimensions identified in a full MD to a specific corpus (BERBER SARDINHA, 2013). In the present study we performed an additive MD of the LEX-BR-Ius corpus (FERRARI and MARQUES, in compilation) adopting the study by Berber Sardinha, Kauffmann, and Acunzo (2014) as a basis aiming to map legal texts into the dimensions of variation identified by it. The *LEX-BR-Ius* is a written, monolingual, synchronous, part-of-speech (POS) and lemma annotated, textually tagged with Modest XML (HARDIE, 2014) corpus of complete Brazilian Federal Statutory Legal Texts in force on the compilation date (2022), balanced by frequency of use according to the search results on the JusBrasil portal. We concluded that legal texts are a register according to Biber and Conrad's perspective (2009). They are characterized by the use of verbs in the future, modal verbs, coordinated conjunctions, subordinate clauses, passive constructions without agent, nouns, adjectives, prepositions, definite articles, nominalizations, lexical richness, and informational density. The results also show that the situational characteristics of each normative species are distinct with respect to the process followed for its creation, while each legal text is further distinguished by its communication context,

communication purpose, and subject matter. This research aims to contribute to studies on legal language and linguistic variation.

Keywords: Brazilian Portuguese. Brazilian Federal Statutory Legal Texts. Multidimensional Analysis. Register. Variation.

LISTA DE ILUSTRAÇÕES

LISTA DE GRÁFICOS

LISTA DE TABELAS

LSTA DE SIGLAS E ABREVIATURAS

AMD Análise Multidimensional

CR/88 Constituição da República Federativa do Brasil

CBVR Corpus Brasileiro de Variação de Registro

PB português brasileiro

POS *part of speech*

SUMÁRIO

1 INTRODUÇÃO

1.1 Contexto

O direito, assim como a linguagem, está presente na coletividade humana desde a sua origem, sem ele não seríamos capazes de viver em sociedade. Foi o estabelecimento de regras que possibilitou a convivência do ser humano em grandes grupos e o estabelecimento da sociedade. Cada sociedade tem suas leis e sua tradição jurídica, mas todas as sociedades modernas compartilham características jurídicas comuns, o que possibilita as interações humanas no mundo globalizado. Atualmente temos dois grandes sistemas jurídicos no mundo, o "*common law*", fruto da tradição jurídica anglo-saxônica, e o "*civil law*", fruto da tradição jurídica românica (GONÇALVES, 2018; LENZA, 2020). Dentro desses sistemas temos as chamadas grandes áreas ou ramos ou ainda, sistemas do direito, comuns à todas as tradições jurídicas, e a partir deles surgem subáreas a eles relacionados, sendo possível o seu desdobramento em outras menores para se adaptar à realidade da sociedade em questão (LENZA, 2020; SILVA, 2020). Além disso, cabe ressaltar que, como toda tradição jurídica, toda língua e todo país tem suas especificidades jurídicas, estudar o direito tanto da ótica jurídica quanto daquela linguística implica uma série de desafios.

Dentre esses desafios temos a linguagem produzida e utilizada nesse campo: a linguagem jurídica. A linguagem jurídica é uma linguagem altamente especializada, tendo assim características lexicais e gramaticais próprias (DAHLMAN, 2006; RICHARD, 2011; GOŹDŹ-ROSZKOWSKI e POTRANDOLFO, 2015). Essas características nos levam a considerar essa linguagem como um registro segundo a perspectiva proposta por Biber e Conrad (2009), ou seja, uma variedade da língua com características situacionais, linguísticas e funcionais próprias.

A linguagem jurídica é uma linguagem rica e plural que engloba, entre outros, a linguagem utilizada nos tribunais, artigos acadêmicos de direito, peças processuais e legislações, cada um com suas particularidades (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Dentre as suas diversas variantes, o que nos interessa são os textos legais. Os textos legais, ou seja, as normas jurídicas, apresentam termos técnicos,

construções gramaticais, padrões lexicais e estrutura próprios sendo por nós hipotizado se tratar de um sub registro da linguagem jurídica.

Em vista da grande quantidade de textos e da complexidade das características situacionais e linguísticas que envolvem os textos legais, assim como as particularidades de cada língua, país, sistema jurídico, abrangência das legislações e tempo disponível para a pesquisa, para possibilitar a identificação e descrição fidedigna das características a ele inerentes decidiu-se delimitar o estudo aos textos legais federais brasileiros em vigência[1], tendo sido amostrados com base na sua frequência de uso (vide 3.1). Para analisá-los foi utilizada a abordagem da análise multidimensional (AMD) (BIBER, 1988). Desenvolvida por Douglas Biber, ela possibilita a descrição e a caracterização da variação de registros de quaisquer línguas ou variedades em dimensões de variação. Para tanto, submete-se um corpus do (s) registro (s) estudado (s) a uma série de análises estatísticas e, subsequentemente, seus resultados são analisados funcionalmente proporcionando um retrato aprofundado e fidedigno de determinado registro. (BIBER, 1988; BIBER e CONRAD, 2009; BERBER SARDINHA, 2010).

É possível realizar a AMD de duas formas: AMD completa e AMD aditiva. Estas se diferenciam pela abrangência da análise e procedimentos metodológicos a serem seguidos, mas compartilham a mesma base teórica e objetivo, qual seja investigar a variação linguística dos registros. Enquanto a primeira oferece uma descrição detalhada dos registros analisados e revela como a sua variação se dá a partir da identificação de dimensões de variação, a segunda oferece um panorama geral da variação nos registros investigados através do seu mapeamento nas dimensões reveladas por AMD completa previamente realizada, incorporando-os aos registros deste estudo e complementando as dimensões nele identificadas. (BIBER e CONRAD, 2009; BERBER SARDINHA, 2013; BERBER SARDINHA, *et al,* 2019).

Para a realização da AMD, independentemente do tipo (completa ou aditiva), faz-se necessário um grande volume de palavras e diversidade de textos (BIBER, 1993). A linguística de corpus é uma metodologia que, com auxílio da computação, permite compilar e analisar um grande volume de amostras representativas da linguagem para

[1] Delimitou-se a pesquisa aos textos legais federais brasileiros vigentes no ano de 2022 – ano em que as seções do corpus utilizado nesta pesquisa foram compiladas.

os mais diversos estudos linguísticos (BERBER SARDINHA, 2004). Diante disso, utilizou-se essa metodologia para compilar e analisar o *LEX-BR-Ius* corpus (FERRARI e MARQUES, em compilação), um corpus composto de textos legais federais brasileiros em vigência na época da compilação.

Na presente pesquisa optou-se pela realização de um estudo comparativo entre os textos legais e aqueles do Corpus Brasileiro de Variação de Registro (CBVR) (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014). Para isso, as seções: Constituição, Códigos e Estatutos do *LEX-BR-Ius* (FERRARI e MARQUES, em compilação) foram submetidas à uma análise multidimensional aditiva usando como base para a comparação todas as dimensões de variação do português brasileiro identificadas por Berber Sardinha, Kauffmann e Acunzo (2014).

1.2 Problema

Diante do exposto, pretende-se responder às seguintes perguntas de pesquisa:

a) Como se dá a variação linguística nos textos legais federais brasileiros?
b) Os textos legais federais brasileiros podem ser considerados um registro?
c) Como os textos legais brasileiros se encaixam nas dimensões de variação do português brasileiro identificadas de Berber Sardinha, Kauffmann e Acunzo (2014)?

1.3 Objetivos

Para responder às perguntas de pesquisa estabeleceu-se como objetivo geral: investigar a variação linguística dos textos legais federais brasileiros em vigência no ano de 2022. Para que sejam possíveis tais análises, os objetivos específicos são:

a) Estudar a variação linguística dos textos legais federais brasileiros;
b) Verificar se os textos legais federais brasileiros podem ser classificados como um registro;
c) Comparar os textos legais federais brasileiros com os registros encontrados no Corpus Brasileiro de Variação de Registro (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014), em todas as dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014);

1.4 Justificativa

A produção científica na linguística em interface com o direito tem crescido exponencialmente desde os anos 90 (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Como fruto dessa multidisciplinariedade surgem estudos sobre a linguagem jurídica, analisada sob diversos aspectos. Entre eles temos um número significativo de pesquisas em várias áreas da linguística, incluindo terminologia, tradução, análise do discurso e variação linguística (GOŹDŹ-ROSZKOWSKI, 2012; FERRARI e CUNHA, 2022). Temos ainda um grande e crescente número de corpora especializados de linguagem jurídica: monolíngues e bilíngues paralelos ou comparados (PONTRANDOLFO, 2012; GIAMPIERI, 2018).

No português brasileiro, a literatura sobre a linguagem jurídica se volta para aspectos linguísticos específicos, sendo a maior parte dos estudos nas áreas da tradução e da análise do discurso (FERRARI e CUNHA, 2022). Destacam-se os trabalhos em análise do discurso jurídico do pioneiro Grupo de Pesquisa Linguagem e Direito e à Associação de Linguagem & Direito (ALIDI) e em terminologia pelo projeto TermiSul[2], que, com auxílio da linguística de corpus, constrói dicionários e glossários multilíngues de direito, entre outros. Em relação aos corpora não existe um corpus de português brasileiro compilado objetivando unicamente o estudo da linguagem jurídica brasileira. Temos subcorpora de corpora voltados para o estudo do português brasileiro em geral que abarcam uma ou mais variedades da linguagem jurídica (ex.: C-ORAL-BRASIL, que amostra a linguagem oral utilizada nos tribunais e o Corpus Brasileiro de Variação de Registro (CBVR), que amostra textos legais), para o estudo da tradução (ex.: COMET, que amostra contratos), entre outros (FERRARI e CUNHA, 2022; FERRARI e MARQUES, 2022).

Apesar do crescente interesse nesse tipo de pesquisa, são poucos os estudos voltados para a variação dessa linguagem (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Investigações prévias retornaram trabalhos e corpora relacionados à língua inglesa ou entre o inglês e outro par linguístico, mas nenhum corpus brasileiro exclusivamente jurídico/legal nem estudos voltados para a variação linguística dos textos

[2] http://www.ufrgs.br/termisul/

legais brasileiros. Diante da ausência de estudos sobre esse recorte específico optou-se por aprofundar-se nos estudos sobre linguagem jurídica, e, mais especificamente sobre a linguagem utilizada nos textos legais, sob a perspectiva do registro (BIBER e CONRAD, 2009) para investigar os textos legais federais brasileiros vigentes buscando compreender sua variação linguística.

Como nosso objetivo é estudar a linguagem utilizada nos textos legais, faz-se necessário um corpus representativo dessa variedade. Entretanto não existia um corpus que atendesse à nossa pesquisa, o que há são plataformas de busca de legislação[3] e a Base de Normas Jurídicas Federais (MARTIM; LIMA; ARAÚJO, 2018). Os primeiros são bancos de dados contendo os textos legais editados no país: elas são alimentadas constantemente pelas legislações atualizadas e disponibilizadas gratuitamente através de mecanismos de busca. Já a segunda é um conjunto de *datasets* coletado por cientistas da computação do Senado que contêm as legislações federais vigentes entre 4 de outubro de 1946 e 12 de abril de 2017.

Banco de dados são diferentes de corpora. A finalidade dos primeiros é armazenar informação, não tendo assim uma finalidade linguística, mas apenas de agregação de dados, enquanto a dos segundos é serem representativos de uma linguagem. Diferentemente dos corpora, na construção de um banco de dados não são levados em consideração questões como amostragem, representatividade ou balanceamento. Eles são coleções de textos sobre determinado objeto, no caso das plataformas em questão, as legislações. Não há amostra e sim todo o conjunto de textos legais criados no país. Além disso, os textos legais se diferenciam por espécie, assunto regulado e tem extensão muito distintas, não havendo balanceamento. Logo, essas plataformas não contemplam nossos objetivos, por isso, optamos por compilar o *LEX-BR-Ius* corpus (FERRARI e MARQUES, em compilação) para a nossa pesquisa.

[3]Destaca-se: http://www4.planalto.gov.br/legislacao/, https://www.jusbrasil.com.br/home, https://www.camara.leg.br/legislacao, https://normas.leg.br/busca e https://www.congressonacional.leg.br/.

1.5 Organização

A corroborar o exposto acima, o presente trabalho é dividido em cinco capítulos. O primeiro capítulo é destinado à introdução, na qual é contextualizada a pesquisa, são apresentadas as perguntas e objetivos que a guiaram e sua justificativa. Já o segundo capítulo é voltado para a revisão bibliográfica. Nele são introduzidos os conceitos, teorias e metodologias que guiaram a realização desta pesquisa, quais sejam: textos legais, registro, linguística de corpus e analise multidimensional. O terceiro capítulo, por sua vez, se ocupa da metodologia utilizada para a realização da pesquisa. Nele apresentamos, descrevemos e justificamos as escolhas, os métodos, os passos seguidos e as ferramentas utilizadas na compilação do corpus *LEX-BR-Ius* (FERRARI e MARQUES, em compilação) e na realização da Análise Multidimensional desse corpus. No quarto capítulo apresentamos e discutimos os resultados obtidos. Por fim, no quinto capítulo concluímos a pesquisa, retomando seus pontos principais e resultados relevantes, assim como seus futuros desdobramentos.

2. REVISÃO BIBLIOGRÁFICA

2.1 Os textos legais brasileiros

Em linhas gerais, os textos legais, popularmente chamados de leis, são as normas estabelecidas por determinada sociedade de forma a regular e proteger seja o ente sociedade que aqueles que a integram. Elas têm valor legal, são compulsórias e visam possibilitar a organização da sociedade e seus membros, assim como a sua convivência e proporcionar segurança jurídica. As normas podem ser extremamente abrangentes e detalhadas ou genéricas a depender do sistema adotado e prever desde aspectos procedimentais e administrativos, passando por princípios, direitos e deveres até proibições, sanções e penas.

No Brasil, os textos legais são as normas editadas pelo poder legislativo. Também ditas legislações, elas são classificadas em espécies normativas, cada uma com finalidade, criação e aplicação diversas segundo disposição da Constituição Federal. Elas são criadas através do processo legislativo[4] seguindo as regras procedimentais previstas na Constituição pelos atores por ela legitimados (Senador, Vereador, Presidente, etc.). Trata-se de um processo complexo com variações e particularidades referentes a cada espécie normativa a ser discutida. A sua não observância implica vício (formal ou material) e consequente inconstitucionalidade da norma. Para impedir isso e fiscalizar tal processo a constituição estabelece ainda um rigoroso sistema de controle de constitucionalidade que implica, entre outros uma verificação prévia/preventiva realizada seja pelo legislativo que pelo executivo nas fases de elaboração da norma e um controle repressivo/ posterior após a constituição da norma. Conforme nos ensina Lenza (2020), o processo legislativo envolve três grandes fases:

a) Fase de iniciativa: o processo é iniciado com a proposição do projeto de lei por parte da iniciativa geral, concorrente, privativa, popular, conjunta, parlamentar ou extraparlamentar, ou seja, todos aqueles a quem a Constituição legitima;

[4]Devido à complexidade e particularidades do processo legislativo aqui trazemos apenas uma visão geral para maiores detalhes ver: Silva (2020) e Lenza (2020).

b) Fase constitutiva: procede-se à análise, discussão, revisão e votação, do projeto de lei no âmbito legislativo por parte dos parlamentares, primeiramente pelas comissões e depois pelo Plenário (deliberação parlamentar). Se aprovado, o projeto segue para apreciação do chefe do executivo que pode decidir pela sanção (expressa ou tácita), ou seja, anuir o projeto, ou veto, discordar do projeto por considerá-lo inconstitucional (veto jurídico) ou contrário ao interesse público (veto político). O veto pode ser parcial (apenas de algumas partes do texto) ou total e deve sempre ser motivado. No caso de sanção seguimos para a próxima fase do processo, caso contrário o projeto é enviado ao legislativo para nova apreciação;

c) Fase Complementar: é a última fase do processo quando se há a promulgação e a publicação. A promulgação é o ato, geralmente realizado pelo chefe do executivo, que transforma o projeto de lei em lei atestando sua validade e, em outras palavras, trazendo-a à existência. Após a promulgação se dá a publicação do texto legal no Diário Oficial. Seu objetivo é fazer saber a todos a existência e conteúdo da nova norma e estabelecer quando ela deverá passar a ser cumprida. A partir da publicação não se pode alegar o não cumprimento da lei por desconhecimento.

Em relação ao conteúdo e redação das leis temos que estas devem ser gerais, abstratas e codificadas. Isso porque o Brasil adota o *Civil Law* como o sistema jurídico. Nesse sistema, fruto da tradição jurídica romano-germânica, as normas são criadas a partir de discussões teóricas e buscam abranger o maior número de cenários e questões imaginárias possíveis. O resultado dessas discussões é formalizado por escrito em forma de códigos, leis, etc. que são promulgados pelo chefe de estado adquirindo status legal e passando a produzir efeitos. A base do Estado e o que regulamenta todas as normas é a Constituição. Ela é hierarquicamente superior às demais normas e determina suas espécies e matérias a serem regulamentadas. (SILVA, 2020; LENZA, 2020)

A Constituição também estabelece e regula a forma de Estado, no caso do Brasil, o federalismo. O Brasil é uma Federação, portanto, o poder é distribuído entre os entes federados (União, estados membros, distrito federal e municípios), dotados de autonomia recíproca. Essa divisão de poderes entre os entes não é equânime: a União agrega a maior parte do poder e competências sendo ainda hierarquicamente superior aos demais entes (federalismo centrípeto). Apesar disso a autonomia entre eles permanece o que

permite, por exemplo, que cada ente elabore suas próprias leis, eleja seus representantes e crie e arrecade seus próprios impostos, desde que respeitados os limites e competências estabelecidas pela Constituição. Estas podem ser exclusivas, ou seja, apenas aquele ente é autorizado a realizar aquela ação, ou compartilhadas, que podem ser comuns aos entes federados (art. 23, CRF/88) ou concorrentes (art. 24, CRF/88), caracterizando nosso federalismo como misto: dual e de cooperação. (SILVA, 2020; LENZA, 2020)

Conforme exposto acima, o federalismo permite que os entes se auto organizem, podendo, entre outros criar leis. Isso significa que os textos legais podem ser federais, estaduais ou municipais, sendo os federais hierarquicamente superiores às demais e os estaduais superiores às municipais. Os textos legais são ainda classificados em espécies normativas, cada uma com função e regras próprias previstas pela Constituição. A nossa Carta Magna, em seu artigo 59, estabelece quais são as espécies normativas que podem ser editadas, são elas:

> Art. 59. O processo legislativo compreende a elaboração de:
> I emendas à Constituição;
> II leis complementares;
> III leis ordinárias;
> IV leis delegadas;
> V medidas provisórias;
> VI decretos legislativos;
> VII resoluções.
> Parágrafo único. Lei complementar disporá sobre a elaboração, redação, alteração e consolidação das leis. (BRASIL, 1988)

Como pode ser observado, podem ser editadas 7 espécies normativas, necessariamente elaboradas por meio do processo legislativo: emendas à constituição, leis complementares, leis ordinárias, leis delegadas, medidas provisórias, decretos legislativos e resoluções. A seguir detalhamos as espécies normativas às quais pertencem os textos legais brasileiros.

2.1.1 Constituição

Considerada lei maior, regula o estado, a divisão de poderes, as regras do ordenamento jurídico, os princípios fundamentais, etc. Por regular os fundamentos do

Estado, é única, só podendo existir uma Constituição por vez. Para que seja editada uma nova Constituição deve-se atender a uma série de requisitos e regras previstas no texto Constitucional, seguindo um procedimento especial e não o processo legislativo destinado à criação das demais normas. Por esse motivo, apesar de ser uma espécie normativa a Constituição não se encontra no rol do artigo 59 da CR/88 (BRASIL, 1988).

2.1.2 Emendas à Constituição

São normas que alteram o texto da Constituição. Ressaltamos que certas partes da Constituição não podem ser alteradas, conforme estabelecido pelo texto constitucional. Entre elas temos as cláusulas pétreas, artigos que regulam a forma federativa de Estado; o voto direto, secreto, universal e periódico; a separação dos Poderes; os direitos e garantias individuais, entre outros (LENZA, 2020).

Um exemplo seria a Emenda Constitucional Nº 112, de 27 de outubro de 2021 que altera o art. 159 da Constituição. Trata-se de um artigo que regula distribuição de recursos da União ao Fundo de Participação dos Municípios. A emenda trouxe uma nova redação a uma parte desse artigo e acrescentou uma nova determinação:

> Art. 1º O art. 159 da Constituição Federal passa a vigorar com a seguinte redação:
> "Art. 159.
> .....
> I - do produto da arrecadação dos impostos sobre renda e proventos de qualquer natureza e sobre produtos industrializados, 50% (cinquenta por cento), da seguinte forma:
> .....
> f) 1% (um por cento) ao Fundo de Participação dos Municípios, que será entregue no primeiro decêndio do mês de setembro de cada ano;
> ....." (NR)
> Art. 2º Para os fins do disposto na alínea "f" do inciso I do caput do art. 159 da Constituição Federal, a União entregará ao Fundo de Participação dos Municípios, do produto da arrecadação dos impostos sobre renda e proventos de qualquer natureza e sobre produtos industrializados, 0,25% (vinte e cinco centésimos por cento), 0,5% (cinco décimos por cento) e 1% (um por cento), respectivamente, em cada um dos 2 (dois) primeiros exercícios, no terceiro exercício e a partir do quarto exercício em que esta Emenda Constitucional gerar efeitos financeiros. (BRASIL, 2021)

2.1.3 Leis complementares

São normas cuja criação e finalidade está prevista taxativamente, ou seja, de forma explícita, na Constituição (LENZA, 2020). Como exemplo podemos citar a criação, incorporação, fusão, desmembramento e transformação de Estados e municípios prevista nos parágrafos 2º, 3º 3 4º do art. 18, CR/88:

> Art. 18. A organização político-administrativa da República Federativa do Brasil compreende a União, os Estados, o Distrito Federal e os Municípios, todos autônomos, nos termos desta Constituição.
> § 1º Brasília é a Capital Federal.
> § 2º Os Territórios Federais integram a União, e sua criação, transformação em Estado ou reintegração ao Estado de origem serão reguladas em lei complementar.
> § 3º Os Estados podem incorporar-se entre si, subdividir-se ou desmembrar-se para se anexarem a outros, ou formarem novos Estados ou Territórios Federais, mediante aprovação da população diretamente interessada, através de plebiscito, e do Congresso Nacional, por lei complementar.
> § 4º A criação, a incorporação, a fusão e o desmembramento de Municípios, far-se-ão por lei estadual, dentro do período determinado por Lei Complementar Federal, e dependerão de consulta prévia, mediante plebiscito, às populações dos Municípios envolvidos, após divulgação dos Estudos de Viabilidade Municipal, apresentados e publicados na forma da lei. (BRASIL, 1988; grifo nosso)

2.1.4 Leis ordinárias:

São as chamadas normas de competência residual, isso porque regulam tudo o que não for objeto por lei complementar, decreto legislativo ou resolução (LENZA, 2020). Um exemplo seria a Lei Nº 11.101, de 9 de fevereiro de 2005, que regula a falência e recuperação judicial e extrajudicial.

2.1.5 Leis delegadas

São normas editadas pelo Presidente da República para as quais ele não está no rol de legitimados para tal. Por isso, para editá-las ele deve primeiramente solicitar ao Congresso Nacional que lhe delegue poderes para tal. Caso o Congresso aprove a solicitação ele delimitará o assunto a ser legislado e as regras a serem seguidas pelo

presidente para tal (LENZA, 2020). Como exemplo podemos citar a Lei Delegada Nº 13, de 27 de agosto de 1992 que instituiu a Gratificação de Atividade para servidores do Poder Executivo.

2.1.6 Medidas provisórias

São normas editadas pelo Presidente da República por iniciativa própria, ou seja, não precisa solicitar aprovação, sob a justificativa de relevância e urgência. Estão regulamentadas no art. 62 da Constituição e diferentemente das outras espécies normativas, têm prazo de validade determinado, após publicadas tem força de lei e produzem efeitos por 60 dias, prorrogáveis por mais 60 dias. O legislativo discute a Medida Provisória enquanto ela está vigente e decide se a converte em lei ou se a rejeita, perdendo a eficácia desde a sua edição (efeito *ex tunc*) (LENZA, 2020).

Como exemplo de Medida provisória que foi convertida em lei temos a Medida Provisória N º 884, de 14 de junho de 2019 que altera a Lei nº 12.651, de 25 de maio de 2012 que regulava a proteção da vegetação nativa e foi convertida na Lei 13.887, de 17 de outubro de 2019. Já como exemplo de Medida provisória rejeitada podemos citar a Medida Provisória Nº 1.068, de 6 de setembro de 2021 que alterava as leis nº 12.965, de 23 de abril de 2014 e nº 9.610, de 19 de fevereiro de 1998, para tratar sobre o uso de redes sociais. Enquanto essa Medida Provisória vigorou, as alterações eram válidas, produzindo efeito em todo o território nacional. A partir do momento em que ela foi rejeitada, as leis alteradas voltaram a vigorar com o texto anterior à modificação como se essa nunca houvesse existido.

2.1.7 Decretos legislativos

Trata-se de normas que regulam matérias de competência exclusiva do Congresso Nacional, podendo, portanto, serem editadas apenas pelo próprio Congresso. O processo legislativo dessa espécie normativa é um pouco diferente do das demais, sendo sua promulgação e publicação realizada pelo Presidente do Senado Federal ao invés do Presidente da República. Este não tem participação no processo de criação dos

decretos legislativos, que como o próprio nome diz estão enquadrados no âmbito do legislativo. (LENZA, 2020)

Os decretos legislativos devem tratar das matérias elencadas nos artigos 49 e 62, § 3.º da Constituição, a seguir reproduzidos.

> Art. 49. É da competência exclusiva do Congresso Nacional:
> I - resolver definitivamente sobre tratados, acordos ou atos internacionais que acarretem encargos ou compromissos gravosos ao patrimônio nacional;
> II - autorizar o Presidente da República a declarar guerra, a celebrar a paz, a permitir que forças estrangeiras transitem pelo território nacional ou nele permaneçam temporariamente, ressalvados os casos previstos em lei complementar;
> III - autorizar o Presidente e o Vice-Presidente da República a se ausentarem do País, quando a ausência exceder a quinze dias;
> IV - aprovar o estado de defesa e a intervenção federal, autorizar o estado de sítio, ou suspender qualquer uma dessas medidas;
> V - sustar os atos normativos do Poder Executivo que exorbitem do poder regulamentar ou dos limites de delegação legislativa;
> VI - mudar temporariamente sua sede;
> VII - fixar idêntico subsídio para os Deputados Federais e os Senadores, observado o que dispõem os arts. 37, XI, 39, § 4º, 150, II, 153, III, e 153, § 2º, I;
> VIII - fixar os subsídios do Presidente e do Vice-Presidente da República e dos Ministros de Estado, observado o que dispõem os arts. 37, XI, 39, § 4º, 150, II, 153, III, e 153, § 2º, I
> IX - julgar anualmente as contas prestadas pelo Presidente da República e apreciar os relatórios sobre a execução dos planos de governo;
> X - fiscalizar e controlar, diretamente, ou por qualquer de suas Casas, os atos do Poder Executivo, incluídos os da administração indireta;
> XI - zelar pela preservação de sua competência legislativa em face da atribuição normativa dos outros Poderes;
> XII - apreciar os atos de concessão e renovação de concessão de emissoras de rádio e televisão;
> XIII - escolher dois terços dos membros do Tribunal de Contas da União;
> XIV - aprovar iniciativas do Poder Executivo referentes a atividades nucleares;
> XV - autorizar referendo e convocar plebiscito;
> XVI - autorizar, em terras indígenas, a exploração e o aproveitamento de recursos hídricos e a pesquisa e lavra de riquezas minerais;
> XVII - aprovar, previamente, a alienação ou concessão de terras públicas com área superior a dois mil e quinhentos hectares.
> XVIII - decretar o estado de calamidade pública de âmbito nacional previsto nos arts. 167-B, 167-C, 167-D, 167-E, 167-F e 167-G desta Constituição. (BRASIL, 1988)

> Art.62 § 3º As medidas provisórias, ressalvado o disposto nos §§ 11 e 12 perderão eficácia, desde a edição, se não forem convertidas em lei no prazo de sessenta dias, prorrogável, nos termos do § 7º, uma vez por igual período, devendo o Congresso Nacional disciplinar, por decreto legislativo, as relações jurídicas delas decorrentes. (BRASIL, 1988)

2.1.8 Resoluções

As resoluções são uma espécie normativa que regulamenta as matérias de competência privativa da Câmara dos Deputados e do Senado Federal, previstas nos art. 51 e 52 da Constituição e nos Regimentos internos dessas casas. Assim como os decretos legislativos, seguem um procedimento especial do qual o Presidente da República não participa, sendo a promulgação e a publicação realizada pelo presidente da própria casa legislativa (Câmara ou Senado) à qual compete regular determinada matéria. (LENZA, 2020)

Para fins exemplificativos reproduzimos a seguir o rol de matérias que competem à Câmara dos Deputados:

> Art. 51. Compete privativamente à Câmara dos Deputados:
> I - autorizar, por dois terços de seus membros, a instauração de processo contra o Presidente e o Vice-Presidente da República e os Ministros de Estado;
> II - proceder à tomada de contas do Presidente da República, quando não apresentadas ao Congresso Nacional dentro de sessenta dias após a abertura da sessão legislativa;
> III - elaborar seu regimento interno;
> IV - dispor sobre sua organização, funcionamento, polícia, criação, transformação ou extinção dos cargos, empregos e funções de seus serviços, e a iniciativa de lei para fixação da respectiva remuneração, observados os parâmetros estabelecidos na lei de diretrizes orçamentárias;
> V - eleger membros do Conselho da República, nos termos do art. 89, VII. (BRASIL, 1988)

Já ao Senado compete privativamente:

> Art. 52. Compete privativamente ao Senado Federal:
> I - processar e julgar o Presidente e o Vice-Presidente da República nos crimes de responsabilidade, bem como os Ministros de Estado e os Comandantes da Marinha, do Exército e da Aeronáutica nos crimes da mesma natureza conexos com aqueles;
> II processar e julgar os Ministros do Supremo Tribunal Federal, os membros do Conselho Nacional de Justiça e do Conselho Nacional do Ministério Público, o Procurador-Geral da República e o Advogado-Geral da União nos crimes de responsabilidade;
> III - aprovar previamente, por voto secreto, após argüição pública, a escolha de:
> a) Magistrados, nos casos estabelecidos nesta Constituição;

> b) Ministros do Tribunal de Contas da União indicados pelo Presidente da República;
> c) Governador de Território;
> d) Presidente e diretores do banco central;
> e) Procurador-Geral da República;
> f) titulares de outros cargos que a lei determinar;
> IV - aprovar previamente, por voto secreto, após argüição em sessão secreta, a escolha dos chefes de missão diplomática de caráter permanente;
> V - autorizar operações externas de natureza financeira, de interesse da União, dos Estados, do Distrito Federal, dos Territórios e dos Municípios;
> VI - fixar, por proposta do Presidente da República, limites globais para o montante da dívida consolidada da União, dos Estados, do Distrito Federal e dos Municípios;
> VII - dispor sobre limites globais e condições para as operações de crédito externo e interno da União, dos Estados, do Distrito Federal e dos Municípios, de suas autarquias e demais entidades controladas pelo Poder Público federal;
> VIII - dispor sobre limites e condições para a concessão de garantia da União em operações de crédito externo e interno;
> IX - estabelecer limites globais e condições para o montante da dívida mobiliária dos Estados, do Distrito Federal e dos Municípios;
> X - suspender a execução, no todo ou em parte, de lei declarada inconstitucional por decisão definitiva do Supremo Tribunal Federal;
> XI - aprovar, por maioria absoluta e por voto secreto, a exoneração, de ofício, do Procurador-Geral da República antes do término de seu mandato;
> XII - elaborar seu regimento interno;
> XIII - dispor sobre sua organização, funcionamento, polícia, criação, transformação ou extinção dos cargos, empregos e funções de seus serviços, e a iniciativa de lei para fixação da respectiva remuneração, observados os parâmetros estabelecidos na lei de diretrizes orçamentárias;
> XIV - eleger membros do Conselho da República, nos termos do art. 89, VII.
> XV - avaliar periodicamente a funcionalidade do Sistema Tributário Nacional, em sua estrutura e seus componentes, e o desempenho das administrações tributárias da União, dos Estados e do Distrito Federal e dos Municípios.
> Parágrafo único. Nos casos previstos nos incisos I e II, funcionará como Presidente o do Supremo Tribunal Federal, limitando-se a condenação, que somente será proferida por dois terços dos votos do Senado Federal, à perda do cargo, com inabilitação, por oito anos, para o exercício de função pública, sem prejuízo das demais sanções judiciais cabíveis. (BRASIL, 1988)

2.2 Os textos legais enquanto registro

Conforme exposto anteriormente, cada língua e cada tradição jurídica apresenta suas particularidades seja na criação que na aplicação de suas normas. O mesmo diz respeito à linguagem utilizada nos textos legais: a linguagem jurídica. Conforme Dahlman

(2006), Richard (2011), Goźdź-Roszkowski e Potrandolfo (2015), entre outros autores demonstram, a linguagem utilizada nesses textos é altamente especializada, tendo características léxico-gramaticais próprias. Em outras palavras, a linguagem jurídica é uma variedade da língua marcada por vários traços linguísticos próprios, usada em contextos específicos (DAHLMAN, 2006; RICHARD, 2011; GOŹDŹ-ROSZKOWSKI, 2012; CHOVANEC, 2013; GOŹDŹ-ROSZKOWSKI e POTRANDOLFO, 2015).

Além de terem características linguísticas específicas as linguagens especializadas estão geralmente atreladas a áreas do conhecimento ou domínios profissionais. No caso, a linguagem jurídica é a linguagem utilizada no âmbito do direito, entretanto, ela é extremamente variada, nos levando a falar de "linguagens jurídicas", "registros jurídicos" (CARAPINHA, 2018) ou ainda "gêneros jurídicos" (GOŹDŹ-ROSZKOWSKI, 2011) que variam dependendo do contexto de uso. Como vimos, a linguagem jurídica é usada em diversos contextos: nos tribunais, nas peças processuais, nos contratos, nos artigos científicos, nas legislações, etc. (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018). Em cada um desses contextos suas características divergem, indo contra a noção erroneamente difundia de que se trate de uma linguagem uniforme e homogênea.

A esse respeito, Carapinha (2018) afirma: "A linguagem jurídica não é, pois, monolítica; ela concretiza-se numa pluralidade de textos e de discursos com características muito distintas, que são usados por interlocutores diversos em situações comunicativas específicas." (CARAPINHA, 2018, p. 95). Já Goźdź-Roszkowski (2012), ressalta que "linguagem jurídica" é na realidade um termo guarda-chuva utilizado para denominar uma ampla gama de tipos textuais criados e usados em diferentes contextos no mundo jurídico, tendo um alto grau de variabilidade não só em seu interior, mas também que se diferencia a depender do sistema legal, da língua e do país em que vem utilizada. Logo, trata-se de uma linguagem extremamente plural, heterogênea e complexa cujos termos técnicos e construções gramaticais variam a depender do contexto (GOŹDŹ-ROSZKOWSKI, 2012; CARAPINHA, 2018).

Dentro desse *continuum* de variação optou-se por analisar a linguagem jurídica utilizada nos textos legais. Conforme exposto anteriormente, os textos legais são as normas que regulam a sociedade. No Brasil, elas são criadas pelo poder legislativo e

codificadas, ou seja, escritas e publicadas no Diário Oficial para poder obter eficácia, validade e aplicabilidade. Sob o ponto de vista linguístico, elas podem ser classificadas e estudadas sob diferentes óticas. Nesta pesquisa optamos por estudá-las sob a perspectiva do registro de Biber e Conrad (2009).

Para tanto, antes de introduzir a perspectiva adotada, faz-se necessário diferenciar as noções de gênero e registro. Essas noções são muitas vezes confundidas e algumas vezes consideradas sinônimos. Na literatura encontramos diversas definições, muitas das quais conflitantes. A seguir, para fins ilustrativos, destacamos algumas propostas de gênero e registro existentes na literatura.

Em relação à noção de gênero, ressaltamos a proposta da corrente Bakthiniana. Para eles gênero seria uma categoria de texto relativamente estável criada no âmbito de uma comunidade e por ela reconhecida visando a comunicação, sendo considerado uma "prática discursiva/ sócio-discursiva" (CARAPINHA, 2018). Para Swales (1990), por sua vez, gênero é um tipo de evento comunicativo, cuja finalidade é reconhecida pelos membros daquela comunidade de fala. Ele se realiza através da fala ou da escrita em circunstâncias específicas de produção e sua recepção depende do seu reconhecimento pelos falantes e que estes compartilharem um mesmo propósito comunicativo. Para que seja possível esse reconhecimento e a concretização da finalidade do gênero, ou seja, possibilitar a comunicação, é necessário que o texto tenha características prototípicas (estrutura, convenções, estruturas retóricas terminologias específicas, etc.) amplamente difundidas na comunidade de forma a serem facilmente reconhecidas como pertencentes a determinado gênero.

Já em relação à noção de registro, constata-se que muitas das propostas compartilham a importância dada ao contexto de uso de determinada variedade na análise linguística diferenciando-se pela forma com esta é realizada. Dentre elas destacamos a proposta de Haliday (1985), idealizador da teoria sistêmico funcional, que considera registro como variedade da língua, escolhida pelo falante com base na situação de uso. Para tanto, parte-se do princípio que a linguagem é um instrumento de interação utilizado pelos seres humanos para se comunicarem. Trata-se de um sistema de possibilidades do qual o falante se vale para transmitir determinada mensagem em determinado contexto. (HALIDAY, 1985)

Segundo essa proposta, basicamente, é o contexto social, ou seja, a situação de uso real na qual se encontra o falante, que direciona suas escolhas linguísticas, levando-o a preterir determinada variedade a outra naquele contexto específico. Para fazer essa escolha o falante leva em consideração: o campo, o meio e o modo (CARAPINHA, 2018). Segundo Carapinha (2018) estes "[...] dizem respeito (de forma muito simplificada) ao tópico, à relação social estabelecida entre os interlocutores e instaurada pelo texto ou pelo discurso, e ao meio de comunicação adotado, respetivamente." (CARAPINHA, 2018, p. 95). A linguagem produzida é então analisada funcionalmente a partir das metafunções da língua (ideacional, interpessoal e textual).

Feita a diferenciação entre gênero e registro, trazemos a perspectiva de registro de Biber e Conrad (2009), por nós adotada nesta pesquisa. Segundo eles, registro é: "[...] uma variedade associada com uma situação de uso específica (incluindo propósitos comunicativos próprios)" (BIBER e CONRAD, 2009, p. 6, tradução nossa)[5]. Em outras palavras, um registro é uma variedade da língua - uma categoria de textos que compartilham determinadas características situacionais ou sociais - utilizada em contextos específicos na sociedade. Um registro pode ser mais ou menos especializado e abranger variedades mais amplas como escrita a acadêmica que abrange uma grande quantidade de registros, tais como: artigo acadêmico, revisão de literatura, carta de intenção, resenha, etc. ou mais restritas como a resenha que engloba apenas esse registro específico. (BERBER SARDINHA, 2013)

Um registro é composto por características/itens/traços linguísticos, situacionais e funcionais próprios que nos permitem descrever qualquer variedade da língua em sua integralidade. Para Biber (1988, 1995), Biber e Conrad (2009), Berber Sardinha (2010) e Delfino (2021), as características linguísticas são as especificidades lexicais e gramaticais de um dado registro. Para fins de exemplificação, a seguir trazemos rapidamente alguns estudos que identificaram traços linguísticos característicos dos textos legais em inglês, italiano e português.

Na língua inglesa, Chovanec (2013), em sua pesquisa sobre a gramática nos textos do âmbito jurídico, aponta como características dos textos legais: construções impessoais; passivas elípticas; nominalizações; sintagmas nominais complexos e longos,

[5] *[...] a variety associated with a particular situation of use (including particular communicative purposes).*

cujas sentenças são conectadas usando parataxis ou hipotaxis; particípios presente e passados usados como modificadores de adjetivos para condensar a estrutura sintática da sentença. Ele também traz como particularidades desses textos a alta ocorrência de adjetivos, o uso de verbos modais, sentenças condicionais, estratégias coesivas específicas (ex.: explicitações, repetições e de advérbios do tipo *herein.* Além disso, foi constatada a pouca ocorrência de pronominalizações nesses textos, possivelmente para manter a consistência e evitar ambiguidades e interpretações diversas da originalmente pretendida, assim como o uso de uma estrutura lógica nas sentenças e a alta densidade lexical, provavelmente ligadas à hipótese anterior.

Em relação ao léxico jurídico/legal na língua inglesa, Richard (2018), em sua pesquisa sobre o léxico enquanto característica da linguagem jurídica, também destaca a alta densidade lexical. Ela especifica que o léxico especializado desse registro é marcado pelo empréstimo de outras línguas, em especial do latim, de outros sistemas jurídicos de *Common Law* e de outras áreas do conhecimento. Além disso, ela destaca o uso de léxico fonte-orientado e recipiente-orientado, polissemia, hiponímia e *weasel words*.

Já na língua italiana, Dhalman (2006), em seu estudo sobre as especificidades morfossintáticas dos textos legais, administrativos, jurisprudência e doutrina nessa língua, traz como particularidades destes o uso de: expressões fixas; estratégias de síntese (uso de ênclise com infinitivo do verbo modal, uso do infinitivo e do gerúndio na forma implícita, uso de abreviações e siglas, substituição de orações relativas por adjetivos, substituição de orações objetivas ou interrogativas indiretas por substantivos derivados de adjetivos); anteposição (topicalização, do adjetivo ao substantivo e do complemento de agente ao argumento do verbo em orações subordinadas implícitas); imperfeito narrativo; abstrações; nominalizações.

Por fim, na língua portuguesa, na variedade de Portugal, ressaltamos o estudo de Carapinha (2018) que investigou a linguagem jurídica nos códigos legais portugueses. Ela identificou como características desses textos a presença de lexemas específicos desse registro, léxico de domínio comum que adquire um significado distinto nesse registro, expressões vagas e genéricas, expressões em latim ou dele derivadas, assim como do grego e polissemia. Ela destaca ainda o uso de nominalizações, prefixação

(com/n-; sub- e im/n-), alta ocorrência de substantivos e adjetivos, expressões nominais frutos de composição, explicitações e conceptualizações, topicalizações e preferência pela terceira pessoa e uso do presente indicativo. Sobre a estrutura sintática a pesquisadora destaca a alta ocorrência de: sintagmas longos e complexos, impessoalidade, passivas, em especial elípticas, construções pronominais, orações reduzidas de particípio com valor temporal, orações reduzidas de gerúndio com valor condicional/temporal, orações subordinadas adverbiais temporais, orações condicionais de tipo hipotético e frases declarativas. Por fim ela ressalta o alto grau de estruturação e organização dos textos, que apresentam inúmeras subdivisões o que reflete em uma descontinuidade na leitura do texto.

Conforme exposto acima, em si tratando de textos legais a literatura indica que, ainda que seus traços linguísticos variem a depender da língua analisada, constatou-se que alguns itens linguísticos típicos da linguagem jurídica, tais como a nominalização e a riqueza lexical, estão presentes em todas as línguas analisadas, podendo ser considerados possíveis universais.

Além dos traços linguísticos, os registros também se distinguem segundo suas características situacionais. Estas descrevem o contexto da produção e ocorrência do registro. Para analisá-las, parte-se da experiência pessoal do pesquisador, da observação do registro, de informações fornecidas por informantes especialistas naquele registro, pesquisas anteriores sobre aquele registro ou ainda, análise de textos do registro pretendido. (BIBER e CONRAD, 2009)

Devido às muitas características situacionais existentes, Biber e Conrad (2009) sugerem alguns itens[6] para guiar o pesquisador na sua análise. Conforme essa proposta, a descrição é feita a partir da identificação dos participantes e sua relação, canal de comunicação, circunstâncias da produção, contexto, propósito comunicacional e assunto. Os participantes são: quem produz o texto (*adressor*: emissor), quem é o público-alvo daquele texto (*addressee*: receptor) e quem observa, mas não participa dessa comunicação (*on-lookers:* observadores). Quanto a relação entre os participantes avalia-se a interação ou ausência dessa entre eles, seu papel social, se há uma relação pessoal entre eles e se eles compartilham algum conhecimento. O canal, por sua vez, trata-se da

[6] Aqui trazemos apenas a um panorama, para maiores detalhes ver Biber e Conrad (2009) p. 40 a 46.

forma pela qual a comunicação ocorre: escrita, falada, sinalizada, etc. e qual o meio utilizado, ex.: impresso, eletrônico, etc. (BIBER e CONRAD, 2009)

Temos ainda as circunstâncias de produção, ou seja, o contexto no qual o texto foi produzido, por exemplo, um texto espontâneo, produzido naquele momento e já dirigido ao receptor ou um texto planejado, revisado e editado antes de chegar ao receptor. Já o contexto de comunicação diz respeito ao tempo e espaço em que o texto foi produzido e recebido. É possível que os participantes compartilhem o mesmo tempo e espaço, como em uma transmissão ao vivo na televisão (contemporâneo), mas também que esses não sejam compartilhados, como é o caso de textos escritos séculos atrás e lidos hoje (histórico). Também é importante analisar o propósito comunicativo, ou seja, porque aquele texto foi produzido. Este pode ser descrito por finalidade geral e. específica, factualidade e a expressão de um posicionamento. Por fim, temos o assunto do texto e sua abrangência: genérico ou específico. (BIBER e CONRAD, 2009)

Para fins de exemplificação, trazemos as características situacionais dos textos legais em geral identificadas a partir da experiência pessoal da pesquisadora, pesquisa bibliográfica e análise de textos legais seguindo Biber e Conrad (2009):

Tabela 1 – Características situacionais dos textos legais

PARTICIPANTES	Emissor	Institucional
	Receptor	Inúmeros
	Observadores	Sim
RELAÇÃO ENTRE OS PARTICIPANTES	Interatividade	Pouca
	Papel social	Representante e representado
	Relação pessoal	Não
	Conhecimento compartilhado	Geral e especializado
CANAL	Modo	Escrito
	Meio específico	Impresso e digital
CIRCUNSTÂNCIAS DE PRODUÇÃO		Revisado e editado
CONTEXTO DE COMUNICAÇÃO	O tempo e local de comunicação é compartilhado?	Sim e não
	Local da comunicação	Público
	Tempo	Histórico e contemporâneo
PROPÓSITO COMUNICACIONAL	Finalidade geral	Informar, descrever e explicar
	Finalidade específica	Enumerar as normas que regem a sociedade
	Factualidade	Factual
	Expressão de posicionamento	Não
ASSUNTO	Assunto geral	Direito
	Assunto específico	Sim
	Status social do receptor	Cidadão brasileiro e equiparados e instituições

Fonte: autora (2022), baseado em Biber e Conrad (2009)

Os textos legais têm como participantes um emissor institucional: o poder legislativo, seja ele federal, estadual ou municipal, que é exercido pelos parlamentares. Como receptor temos um número indeterminado de pessoas e instituições que estão sujeitas àquela norma. É possível haver observadores, uma vez que os textos legais são editados em um local público, em seções públicas e, no caso dos textos federais, são também transmitidos ao vivo na televisão aberta, sendo possível a qualquer pessoa assistir à sua produção.

Quanto à relação entre os participantes, a interação entre eles é pequena ou nula. Não há interação direta entre os parlamentares, ou seja, quem está por trás da instituição e ativamente edita os textos, e as demais pessoas. Entretanto, é possível ao cidadão comum participar da produção dos textos legais através de plebiscitos, discussões públicas ou ainda, em casos específicos previstos em Constituição, propor projeto de lei. Além disso, após a promulgação da lei, é possível aos sujeitos de direito ajuizar processo questionando a constitucionalidade de determinado texto legal.

Quanto aos papéis sociais do emissor e do receptor temos uma relação de representante-representado. Os membros do poder legislativo se encontram nessa posição porque foram eleitos pelo povo para representá-los no Congresso Nacional, por exemplo. Já o receptor são todos aqueles que eles estão representando: eles mesmos, já que também eles estão sujeitos às leis, os cidadãos brasileiros e aqueles a eles equiparados assim como as instituições. De forma geral não há relação pessoal entre os participantes e, apesar de compartilharem um conhecimento genérico, o emissor tem um conhecimento especializado que é compartilhado apenas por uma parcela dos receptores, como outros legisladores, advogados e membros do poder judiciário.

Já em relação ao canal seu modo é escrito e seu meio é tanto eletrônico, já que os textos legais são publicados e disponibilizados gratuitamente na internet em sites institucionais e privados, quanto impresso, uma vez que as editoras vendem os textos legais agrupados em *Vade mecums,* assim como a Constituição e os Códigos federais. Em relação às circunstâncias de produção, conforme explicitadas na seção anterior, os textos legais são produzidos por meio do processo legislativo, passando por várias revisões e edições tanto durante sua produção quanto após sua promulgação com modificações na sua redação.

Quando analisamos o contexto de comunicação, observamos que o tempo e local de comunicação podem ou não serem compartilhados pelos participantes. Na maior parte do tempo este não é compartilhado, mas existem alguns casos em que estes são compartilhados. O lugar da comunicação é público, qual seja a sede do poder legislativo, logo é possível assistir presencialmente às seções de discussão dos projetos de lei, e em alguns casos específicos, participar ativamente desse processo. Nesses casos o local de comunicação e seu tempo são compartilhados e contemporâneos. Entretanto, muitos

textos legais já foram produzidos, alguns datam de décadas ou até mesmo séculos atrás e continuam válidos. Como, por exemplo, o Código Penal que data de 1940, ou ainda, o Código Comercial que foi promulgado em 1850. Temos ainda outros textos que, apesar de não produzirem mais efeitos ainda são estudados seja por acadêmicos do direito que de outras áreas por seu valor histórico e informacional, como a Constituição de 1946 ou ainda a primeira Constituição brasileira que data de 1824. Nesses casos nem o local nem o tempo são compartilhados.

Ao analisarmos o propósito comunicacional, por sua vez, a finalidade dos textos legais de forma geral é informar as normas que regem a sociedade. Mais especificamente sua finalidade é informar, descrever, conceituar e até mesmo explicar as disposições em seu interior. Além disso, se queremos aprofundar mais, cada texto tem uma finalidade específica, sendo necessário analisá-los individualmente, por exemplo, o Estatuto do idoso se aplica especificamente às pessoas com 60 anos ou mais, trazendo seus direitos e deveres, assim como as sanções destinadas àqueles que os violam. Por serem normas genéricas, criadas para serem as mais objetivas possíveis e ao mesmo tempo impositivas, são textos factuais e não passíveis de posicionamento por parte dos seus autores.

Por fim, o assunto geral é o direito, mas como dito, cada lei tem suas especificidades e assuntos específicos tratados em seu interior. Por ser um texto destinado aos cidadãos brasileiros e equiparados, assim como as instituições é impossível identificar quantos e quais exatamente são os receptores, tornando a individualização e o detalhamento do seu status social impossível.

Além das características linguísticas e situacionais temos também as características funcionais. Estas descrevem a relação entre as duas primeiras, o que nos permite identificar as características próprias do registro, assim como descrevê-lo e analisá-lo. Para identificar essas características faz-se necessário um corpus do registro desejado anotado morfossintaticamente que deve ser submetido a uma análise quantitativa e qualitativa e comparado com outros registros de forma a revelar as características que lhe são próprias. (BIBER, CONRAD, 2009)

Por fim, cabe ressaltar que podemos analisar uma variedade da língua seja partindo da perspectiva de gênero que daquela de registro. Essas formas de análise,

conforme Biber e Conrad (2009), são complementares e se diferenciam em relação aos textos necessários para a realização da análise, as características linguísticas analisadas, a distribuição dessas características e a interpretação dos resultados.

Para a realização de uma análise sob a perspectiva do gênero são necessários textos na íntegra para se poder observar a organização estrutural e retórica do texto, assim com as características linguísticas prototípicas de determinado texto que geralmente ocorrem uma única vez. Quanto às características linguísticas observadas e sua distribuição, conforme já adiantamos, essas ocorrem poucas ou apenas uma vez ao longo do texto, são observadas: a estrutura do texto (ex.: eventuais partes ou divisões, formatação, etc.), organização retórica e convenções e expressões próprias do gênero. Por fim, a interpretação é voltada para a relação entre as características encontradas e as convenções de determinado gênero, ou seja, se elas confirmam determinado "formato" típico daquele gênero ou se fogem à norma. Por exemplo, se estamos estudando o gênero carta, entre as suas muitas características, o endereçamento é uma parte crucial da estrutura do texto e ocorre apenas uma vez. Se usássemos apenas uma parte da carta, como só o seu corpo, essa característica não seria observada, impedindo a caracterização correta do gênero. Se usássemos a carta na integra e ainda assim o endereçamento não constasse, ao interpretar os resultados diríamos que aquele texto foge ao gênero.

Para a realização da análise sob a perspectiva do registro, por sua vez, os textos não precisam ser integrais, apenas partes seriam suficientes para caracterizar determinado registro. Entretanto, por questão de melhor e maior representatividade da variação, textos na íntegra geralmente são preferidos quando se realiza esse tipo de análise. Além disso, recomenda-se comparar o registro estudado com outro (s) registros para que seja possível identificar e descrever as características específicas do registro estudado. Quanto às características linguísticas, todas as características léxico-gramaticais podem ser analisadas, focando no léxico e nas características linguísticas presentes em determinada amostra da variedade estudada. Em relação à distribuição, essa é variável, geralmente as características têm alta ocorrência e se distribuem ao logo de toda a amostra, mas também pode haver características com poucas ocorrências e em partes específicas do texto, apenas a análise revelará isso. Sua variação pode ser

representada através de um *continuum*, o qual não é possível de ser estabelecido em uma análise de gênero, uma vez que as diferenças entre os textos encontradas nesta análise são geralmente pequenas. Quanto à interpretação, como o registro engloba tanto as características linguísticas quanto aquelas situacionais, parte-se do pressuposto que elas servem um propósito comunicativo e interpreta-se os resultados funcionalmente.

Diante do exposto, hipotizamos que os textos legais sejam um registro segundo a perspectiva de Biber e Conrad (2009). Para tanto, pretende-se analisá-los sob essa perspectiva com auxílio da abordagem da análise multidimensional (BIBER, 1988), visando testar nossa hipótese.

2.3 A Linguística de Corpus

Para realizar uma análise linguística, em especial sob a perspectiva do registro, faz-se necessário uma diversidade e grande quantidade de textos que sejam representativos do registro analisado de forma a obter dados linguísticos significativos. Um bom meio para se obter um grande número dados linguísticos é a Linguística de Corpus. A Linguística de Corpus[7] é uma metodologia que permite a coleta e análise de um grande volume de amostras representativas da linguagem em uso *("real life language"),* seja sincrônica ou diacrônica, escrita ou falada, a serem usadas nos mais diversos estudos linguísticos empíricos (MCENERY e WILSON, 2001; MCENERY e HARDIE, 2012)

Segundo Berber Sardinha (2004):

> A Lingüística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjuntos de dados lingüísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade lingüística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador (BERBER SARDINHA, 2004, p 3)

Um corpus é, portanto, uma amostra significativa de textos autênticos que podem ser processados por computador, que sejam representativos da linguagem ou variedade

[7] Aqui apresentamos uma breve reflexão sobre a Linguística de corpus, seus usos, questões e desdobramentos. Para maiores aprofundamentos destacamos: Biber (1993); McEnery e Wilson (2001), Sardinha (2004); Sinclair (2005); McEnery e Hardie (2012); Stefanowitsch (2020).

pretendida, reunidos seguindo critérios específicos, podendo ser escritos, orais, transcrições e/ou multimodais. Suas principais características são: representatividade, autenticidade, amostragem, balanceamento, comparabilidade e computabilidade. (BERBER SARDINHA, 2004; STEFANOWITSCH, 2020)

A finalidade de todo corpus é ser representativo da linguagem ou variedade pretendida. Para termos um corpus verdadeiramente representativo e livre de vieses a sua arquitetura deve ser cuidadosamente pensada. Vários autores apontam para a importância da representatividade, mas não se tem um consenso do que seja a representatividade nem de como atingi-la. De forma ampla, um corpus representativo é aquele cuja amostra reflete a população. Para isso, ele deve retratar o uso e a distribuição dos fenômenos linguísticos quantitativa e qualitativamente proporcionais à língua em uso. Dessa forma é possível estabelecer generalizações a partir do corpus que se apliquem à língua como um todo. Dito isso, não há critérios objetivos para determinar a representatividade. Isso porque o corpus é uma amostra da língua, uma população crescente cujas dimensões são desconhecidas, sendo difícil de mensurar e, consequentemente de estabelecer o tamanho ideal da amostra. (BIBER, 1993; BERBER SARDINHA, 2004; SINCLAIR, 2005; MCENERY e HARDIE, 2012, STEFANOWITSCH, 2020)

O que temos são várias propostas que podem guiar o pesquisador no momento da arquitetura. Ressaltamos aquela defendida por Biber (1993) que, através de uma série de cálculos estatísticos mede qual deveria ser, quantitativa e qualitativamente, a composição interna do corpus para que esse seja representativo da linguagem pretendida. Pela complexidade dessa proposta, muitos adotam a extensão como critério. Estabelece-se uma quantidade de textos, palavras e registros que seja suficiente para representar a linguagem pretendida: geralmente o critério é o maior número de textos possível. Isso porque, como a língua é um sistema probabilístico, quanto maior a amostra, maior a probabilidade de abranger uma grande quantidade de traços linguísticos e, portanto, ser representativo da população. Entretanto, só a quantidade não é suficiente para se ter representatividade. (BIBER, 1993; BERBER SARDINHA, 2004; SINCLAIR, 2005; MCENERY e HARDIE, 2012, STEFANOWITSCH, 2020)

Faz-se necessário levar outros critérios em consideração, quais sejam: definir a população e a composição e método de amostragem. Estes aspectos devem ser considerados na elaboração da arquitetura: primeiro é necessário conhecer a população que se busca representar, seus tipos de texto, a ocorrência destes, assim como suas as características linguísticas e distribuição. A partir dessas informações se pensa na amostragem, quais textos devem compor o corpus, como e onde serão coletados, sua proporção no corpus e quantidade, de forma a ter-se uma amostra fidedigna à população. Logo, a quantidade de textos é apenas uma parte do processo que deve ser considerada durante a definição da amostra, são necessários outros elementos para se ter a representatividade. (BIBER, 1993; BERBER SARDINHA, 2004; SINCLAIR, 2005; MCENERY e HARDIE, 2012, STEFANOWITSCH, 2020)

As demais características dos corpora estão ligadas à representatividade. A autenticidade diz respeito à origem dos textos: eles devem ser textos reais de língua em uso, ou seja, produzidos em uma comunidade de fala com o propósito de comunicação e não para análise linguística. A amostragem se refere ao fato de que esses textos devem ser amostras dessa variedade, ou seja, uma parte do conjunto de textos existentes da linguagem em questão. O tamanho da amostra e os critérios utilizados para escolher os textos que a compõe variam de corpus para corpus, entretanto é importante que a amostra se equipare quantitativamente e qualitativamente à variedade pretendida, de forma a manter a representatividade. (BIBER, 1993; BERBER SARDINHA, 2004; SINCLAIR, 2005; MCENERY e HARDIE, 2012; STEFANOWITSCH, 2020)

Já o balanceamento visa equilibrar o corpus. Assim como a representatividade, seu conceito é problemático, não havendo um consenso entre os pesquisadores. Entretanto, todos concordam sobre sua importância na compilação de um corpus. Para realizá-lo são estabelecidos uma série de critérios para que o corpus seja equilibrado, não seja enviesado e, ao mesmo tempo, seja representativo. Ele deve ser feito de forma que os textos que compõem o corpus tenham um número de palavras aproximado e, no caso de subcorpora, a quantidade de textos devem refletir a incidência da variedade na língua, devendo sua distribuição ser proporcional a ela. (BIBER, 1993; BERBER SARDINHA, 2004; SINCLAIR, 2005; MCENERY e HARDIE, 2012; STEFANOWITSCH, 2020)

A comparabilidade se refere à possibilidade de comparar seja o corpus que as análises dele advindas com outros corpora. Por fim, a computabilidade se refere ao fato de que o corpus deve ser feito para ser lido por computador: é uma parte mais técnica à qual o compilador deve se atentar, devendo considerar o formato, codificação e extensão utilizada nos textos. Como pode ser visto, planejar e compilar um corpus que observe todos esses elementos é uma tarefa muito trabalhosa. Entretanto, conforme ressaltam Berber Sardinha (2004), Sinclair (2004) e McEnery e Hardie (2012) é extremamente importante levar essas características em consideração na arquitetura, de forma a compilar um corpus o mais próximo possível do ideal.

Quanto aos tipos, são vários os corpora que podem ser compilados[8]. Há corpora escritos e corpora orais, esses podem ser monolíngues, bilíngues ou multilíngues. No caso dos corpora bi e multilíngues eles podem ser paralelos ou comparáveis. Corpora paralelos são compostos por amostras de duas ou mais línguas na qual temos o texto original e sua tradução em disposição vertical, um ao lado do outro, com suas linhas alinhadas. Um representante desse tipo é o *European Parliament Proceedings Parallel Corpus* (Europarl)[9], um corpus paralelo multilíngue composto pelos documentos oficiais do Parlamento Europeu entre 1996 e 2011 em todas as 21 línguas oficiais da União Européia alinhados com o respectivo texto em inglês. Já os corpora comparáveis são compostos por dois (bilíngue) ou mais (multilíngue), subcorpora monolíngues, representativos de uma mesma amostra de linguagem em mais de uma língua, seguindo exatamente os mesmos critérios visando manter a representatividade, o equilíbrio e a comparabilidade dos corpora. Destacamos o *The International Comparable Corpus* (ICC)[10] um corpus comparável multilíngue que reúne amostras escritas e faladas em checo, finlandês, francês, alemão, irlandês, italiano, norueguês, polaco, eslovaco, sueco e chinês. (BERBER SARDINHA, 2004; MCENERY e HARDIE, 2012; STEFANOWITSCH, 2020).

[8] Aqui apresentamos apenas alguns aspectos dos corpora, para mais detalhes sobre os aqui apresentados e mais tipos classificações, bem como os critérios para compilação para cada tipo de corpus ver: Biber (1993), McEnery e Wilson (2001), Berber Sardinha (2004), Sinclair (2005), McEnery e Hardie (2012) e Stefanowitsch (2020).
[9] Disponível em: https://www.statmt.org/europarl/. Acesso em 20 set. 2022.
[10] Disponível em: https://korpus.cz/icc. Acesso em 20 set. 2022.

Em relação ao arco temporal amostrado temos corpora sincrônicos e diacrônicos. Os corpora sincrônicos são compostos por amostras produzidas em um período específico de tempo, visando representar a linguagem escrita e/ou falada de uma ou mais variedades em um recorte temporal específico (MCENERY; HARDIE, 2012). Como exemplo podemos citar o *British National Corpus* (BNC)[11], um marco na história da Linguística de Corpus por ser o primeiro a ter 100 milhões de palavras, que visa representar o inglês britânico escrito e falado do fim do século XX. Em oposição, os corpora diacrônicos são compostos por amostras produzidos em períodos de tempo distintos (MCENERY e HARDIE, 2012), um exemplo desse tipo de corpus é o *Colonia Corpus of Historical Portuguese*[12] que amostra textos escritos em português europeu e português brasileiro produzidos entre os séculos XVI e XX, totalizando 5,1 milhões de *tokens* (ZAMPIERI; BECKER, 2013).

Temos ainda corpora de amostragem (*sample corpus*) que representam uma variedade específica da língua em um período de tempo específico e são estáticos/fechados, ou seja, não são atualizados. Um exemplo é o corpus CHAVE[13] que amostra textos dos jornais Público e Folha de São Paulo publicados entre 1994 e 1995. Em contraste temos os corpora monitores, que amostram uma língua como um todo em diferentes arcos temporais, sendo assim compostos de textos representativos de diferentes variedades da língua pretendida e dinâmicos/abertos, e são expandidos ao longo do tempo. Aqui citamos o *Corpus of Contemporary American English* (COCA)[14], um corpus representativo do inglês americano com amostras escritas, faladas e transcrições, composto por 8 subcorpora: falado, ficção, revistas, jornais, textos acadêmicos, TV e legendas de filmes, blogs e sites (*spoken, fiction, popular magazines, newspapers, academic texts, and TV and Movies subtitles, blogs e other web pages)* totalizando aproximadamente 1 bilhão de palavras. O corpus é constantemente alimentado com novos dados, permitindo assim análises em diacronia. (BERBER SARDINHA, 2004; MCENERY e HARDIE, 2012)

[11] Disponível em: https://www.english-corpora.org/bnc/. Acesso em 20 set. 2022.
[12] Disponível em: http://corporavm.uni-koeln.de/colonia/. Acesso em 20 set. 2022.
[13] https://www.linguateca.pt/acesso/corpus.php
[14] https://www.english-corpora.org/coca/

Há ainda os corpora de aprendizes, compostos de textos escritos na segunda língua dos autores, como o *International Corpus of Learner English*[15]. Temos ainda os corpora especializados, geralmente de menor extensão, amostram textos de uma área específica do conhecimento ou pertencentes a um registro específico, por exemplo o *Business Letters*[16] que amostra cartas escritas em inglês na área de negócios.

Os corpora podem também ser anotados ou não anotados[17]. Corpora anotados são aqueles em que, no próprio texto, tenham acrescentado alguma informação fruto da análise linguística do corpus. Essas podem ser de tipo morfossíntático, informacional, prosódico, semântico, etc. Uma anotação muito usada é a de *Parts of Speach* (POS), na qual as palavras do corpus são associadas à sua respectiva classe gramatical. Geralmente essa atribuição é feita utilizando programas específicos, chamados *taggers*, *parsers* ou etiquetadores, que são programas ou scripts treinados para línguas específicas. Os corpora não anotados por sua vez são aqueles que comportam apenas as amostras, ou seja, o texto sem nenhum comentário/informação de tipo textual ou linguístico foi inserido em seu corpo *(raw text*). (BERBER SARDINHA, 2004; MCENERY e HARDIE, 2012; GRIES, 2017)

Por fim, cabe ressaltar que os corpora podem apresentar marcação textual[18], também chamada de *mark-up*, ou *mark-up* textual. Nela são acrescentadas informações ao texto limpo sobre o próprio texto ou sobre suas características, sendo ainda possível utilizá-la para armazenar seus metadados. Entre os sistemas mais utilizados para a marcação de corpora citamos o XML (*Extensible Markup Language*). Ele permite, a partir do uso de etiquetas delimitadas por parênteses angulares, acrescentar ou isolar informações em determinado documento de texto, possibilitando ainda organizá-lo hierarquicamente. (BURNARD, 2005; GRIES, 2017; MCENERY E XIAO, 2005; E HARDIE, 2014).

As etiquetas XML podem delimitar seja anotações linguísticas que marcações textuais. Esse sistema muitas vezes é implementado seguindo as diretrizes do TEI (*Text*

[15] https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html
[16] http://www.someya-net.com/concordancer/
[17] Aqui trazemos apenas uma noção geral do que seria a anotação linguística, para maiores detalhes ver: Gries (2017), Leech (2005), McEnery e Xiao (2005).
[18] Para maiores detalhes ver: Burnard (2005), Gries (2017), McEnery e Xiao (2005) e Hardie (2014).

Encoding Initiative[19]): um consórcio mundial que propõe uma série de regras e etiquetas voltadas para as humanidades na tentativa de padronizar a marcação dos textos nessas áreas, sendo hoje o padrão adotado mundialmente, inclusive para a construção de corpora. O famoso corpus BNC, por exemplo é marcado seguindo as diretrizes do TEI. Voltado para compilação de corpus, além do o TEI temos o XCES: o *Corpus Encoding System* que também se vale do XML, inspirado no primeiro, segundo seus criadores tem apenas as partes do TEI que foram julgadas uteis para a compilação de corpus. Temos também algumas propostas mais simples que as anteriores adotadas com menor frequência, como o TEI Lite standard, desenhado para atender a maior parte das necessidades dos usuários do TEI, mas sendo menos complexo e de fácil utilização. (BURNARD, 2005; GRIES, 2017; MCENERY E XIAO, 2005; E HARDIE, 2014).

Devido à sua complexidade, alto grau de detalhamento e muitas vezes requererem algum grau de conhecimento informático, muitas dessas diretrizes são de difícil aplicação para quem compila corpus, especialmente pesquisadores que trabalham sozinhos ou em pequenas equipes. Para satisfazer as necessidades desse grupo específico Hardie propôs o Modest XML (HARDIE, 2014) que traz as vantagens do XML completo sem precisar aplicá-lo na sua integridade. Uma proposta simples e fácil de ser aplicada voltada para quem quer marcar um corpus. Seguindo algumas regras, podemos marcar os textos utilizando seja as etiquetas standard quanto criar nossas próprias etiquetas voltadas para as necessidades da nossa pesquisa. Essa é a proposta que decidimos adotar em nosso corpus como será explanado nas seções seguintes.

No Brasil a compilação e disponibilização de corpora, conforme ressalta Mello (2012), é recente e geralmente realizada por pesquisadores das áreas de processamento de língua/linguagem natural (PLN), linguística computacional e lexicografia. Tem-se observado, contudo, um aumento no número de grupos de pesquisa e projetos que coletam e analisam corpora nacionais, além de pesquisas e eventos científicos na área da Linguística de Corpus. Em relação aos corpora compilados no Brasil, estes são, na sua maioria, escritos, mas nos últimos anos houve esforços para a criação de corpora orais. Têm-se assim corpora escritos e orais, sincrônicos e diacrônicos, monitores e de

[19] Disponível em: https://tei-c.org/. Acesso em 20 set. 2022.

amostragem, mono bi e multilíngues paralelos e comparáveis. (BERBER SARDINHA, 2004; MELLO, 2012).

Dentre os diferentes trabalhos, destaca-se o Corpus Brasileiro[20] compilado por Berber Sardinha e sua equipe, com o objetivo de representar o português brasileiro contemporâneo. Com mais de 1 bilhão de palavras, é um corpus escrito, sincrônico, etiquetado morfossintaticamente composto por amostras de vários registros, disponibilizado on-line. No âmbito dos estudos históricos, ressaltamos o Corpus Histórico do Português *Tycho Brahe*[21], compilado por Charlotte Galves, Aroldo Leal de Andrade e Pablo Faria, junto à UNICAMP em 2017 com a finalidade de representar o português escrito entre 1380 e 1978. Trata-se de um corpus escrito, de amostragem, diacrônico, com anotação sintática e morfológica, composto por 88 textos literários escritos em português por autores portugueses e brasileiros somando 3.544.628 palavras. O corpus está disponível online para pesquisa e também para download.

Já no âmbito dos estudos da tradução, da terminologia e do ensino há o Corpus Multilíngue para Ensino e Tradução (COMET)[22], desenvolvido pelo Departamento de Letras Modernas da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo Paulo e coordenado pela Profa. Dra. Stella Esther Ortweiler Tagnin. Trata-se de um corpus escrito sincrônico, de amostragem, multilíngue, comparável, composto por três subcorpora: Corpus Técnico-Científico (CorTec), Corpus de Tradução (CorTrad), e Corpus Multilíngue de Aprendizes (CoMAprend). O primeiro é composto por textos técnico científicos escritos em português e em inglês com assuntos comuns dividido em 20 subcorpora de acordo com a área do conhecimento, inclusive da área jurídica. O segundo é composto por traduções de textos em diversas línguas, anotado morfossintaticamente e semanticamente, sendo subdividido em três subcorpora: literário, técnico científico e jornalístico. O último é um corpus de aprendizes composto por textos escritos por estudantes de alemão, espanhol, francês, inglês e italiano. Todos os corpora estão disponíveis online para busca em plataforma própria.

[20]Disponível em: http://corpusbrasileiro.pucsp.br/cb/. Acesso em 20 set. 2022.
[21]Disponível em: http://www.tycho.iel.unicamp.br/corpus/index.html. Acesso em 20 set. 2022.
[22]Disponível em: https://comet.fflch.usp.br/projeto#comaprend. Acesso em 20 set. 2022.

Como exemplo de corpus de aprendizes temos também o C*orpus* do Inglês para Fins Acadêmicos (CorIFA)[23] (DUTRA, *et al.,* 2022), do Grupo de Estudos de Corpora Especializados e de Aprendizes (GECEA) da Universidade Federal de Minas Gerais. Trata-se de um corpus de textos escritos por aprendizes de inglês de vários níveis de proficiência da universidade amostrando diferentes registros acadêmicos. O mesmo grupo compila ainda corpora especializados, tais como o *Corpus of Articles in Chemistry* (CorAChem) e o *Corpus of Articles in Applied Linguistics* (CorAAL) que amostram, respectivamente artigos de química e de linguística aplicada escritos em inglês que seguem a estrutura IMRD(C) e publicados em revistas de alto impacto.

E por fim, como representante dos corpora orais ressaltamos o C-ORAL-BRASIL[24], corpus representativo da fala espontânea do português brasileiro, compilado por Tommaso Raso e Heliana Mello e braço brasileiro do projeto C-ORAL-ROM[25]. É um corpus oral, comparável, sendo disponibilizado o áudio e a transcrição alinhada segmentada em enunciados e unidades tonais. É composto por textos de fala formal e informal, dividido em subcorpora de acordo com o contexto de produção (Espontâneo Informal, Espontâneo Formal, Mídia e Telefônico), qual seja de domínio público ou privado, estes subdivididos nas seções: monólogos (1/3), diálogos (1/3) e conversações (1/3). Sua finalidade é representar as variações diafásica e diastrática do português brasileiro, com foco na variedade mineira. (RASO e MELLO, 2012, MELLO, 2012, FERRARI e BOSSAGLIA, 2020).

Em relação à linguagem jurídica identificou-se diversos estudos em várias áreas da linguística que utilizam a linguística de corpus para compilar e/ou analisar corpora já existentes que englobem esse grande registro. Conforme demonstrado por Pontrandolfo (2012) e Giampieri (2018) e retomado por Ferrari e Marques (2022) tem-se um grande e crescente número de corpora que abrangem a linguagem jurídica em seus mais variados aspectos em todo o mundo.

[23]Disponível em: https://sites.google.com/site/corpusifa/ e https://ola.unito.it/. Acesso em 20 set. 2022.
[24]Disponível em: http://www.c-oral-brasil.org/. Acesso em 20 set. 2022.
[25]Disponível em: http://www.elda.org/en/proj/coralrom.html. Acesso em 20 set. 2022.

Para fins de exemplificação[26], trazemos o *Cambridge Corpus of Legal English,* um subcorpus do *Cambridge English Corpus*[27], monolíngue, síncrono com aproximadamente 20 milhões de palavras, composto por livros, artigos e notícias de jornal relacionados ao mundo jurídico. Infelizmente esse corpus não está aberto à comunidade científica, sendo seu aceso restrito aos pesquisadores da universidade de Cambridge. Nessa linha temos também o *American Law Corpus,* corpus monolíngue, sincrônico, com aproximadamente 5 milhões de palavras, dividido em 7 subcorpora: artigos acadêmicos, memoriais, contratos, legislação, opiniões, artigos de jornais e livros didáticos (GOŹDŹ-ROSZKOWSKI, 2012). Compilado por Goźdź-Roszkowski, da Universidade de Łódz, com o objetivo de estudar a fraseologia jurídica.

Como representante dos corpora diacrônicos citamos o corpus *Proceedings of the Old Bailey (London's Central Criminal Court*)[28], um corpus monolíngue, diacrônico, com aproximadamente 127 milhões de palavras que amostra julgamentos criminais da corte de Londres entre 1674 e 1913. Destacamos também o *Corpus Juridisch Nederlands*[29], um corpus escrito monolíngue, diacrônico que amostra textos legais editados entre 1814 a 1989 na Holanda.

Dentre os corpora bilíngues trazemos o *Bononia Legal Corpus* (BoLC), um escrito, sincrônico, bilíngue (inglês-italiano) e comparável, de textos representativos das áreas jurídica, legislativa e administrativa em ambas as línguas. Estas foram escolhidas para representar a linguagem jurídica nos sistemas jurídicos*: civil law* e *common law*. Sua compilação teve início em 1997 e a previsão de duração do projeto foi de 4 anos (1997 a 2000). Em relação ao tamanho, foi estabelecido como mínimo de 10 milhões de palavras para cada subcorpus. (ROSSINI FAVRETTI, TAMBURINI, MARTELLI, 2001; 2007). Voltado para textos legais temos o *Reference corpus of Estonian: Legislation*[30]*,* um corpus escrito, sincrônico, paralelo de aproximadamente 1.8 milhões de palavras que

[26] Aqui trazemos apenas alguns dos muitos corpora jurídicos/legais existentes. Para maiores detalhes e mais corpora consultar Portandolfo (2012), Giampieri (2018), Ferrari e Marques (2022) e o site: https://legal-linguistics.net/data-collections/, um site que fornece uma lista de corpora jurídicos e uma rápida descrição de cada um.

[27]Disponível em: https://www.cambridge.es/en/about-us/cambridge-english-corpus. Acesso em 20 set. 2022.

[28]Disponível em: https://www.oldbaileyonline.org/. Acesso em 20 set. 2022.

[29] Disponível em: http://hdl.handle.net/10032/tm-a2-u2. Acesso em: 20 set. 2022.

[30] Disponível em: https://www.cl.ut.ee/korpused/segakorpus/seadused/. Acesso em: 20 set. 2022.

amostra leis da União Europeia em inglês e sua respectiva tradução para o estoniano. No Brasil temos o Corpus Técnico-Científico (CorTec), subcorpus do já citado Projeto CoMET (Corpus Multilíngue para Ensino e Tradução), trata-se de um corpus escrito, bilíngue (português-inglês), paralelo que amostra artigos científicos e textos técnicos de várias áreas do conhecimento, entre elas o direito.

Já no âmbito dos corpora multilíngue destacamos *o European Union Case Law* Corpus (EUCLCORP) (TRKLJA, MCAULIFFE, 2018), um corpus multilíngue, síncrono, comparável de jurisprudências do Tribunal de Justiça da União Europeia e de 8 tribunais nacionais no âmbito da União Europeia em compilação. Os textos datam de 1952 aos dias atuais, amostrando 23 línguas da União Europeia, sendo considerado o maior corpus multilíngue jurídico já criado (TRKLJA, MCAULIFFE, 2018). Por fim, neste âmbito citamos também o JUD-GENTT, um corpus multilíngue, comparável e paralelo que está sendo compilado na *Universidad Jaume I*, para pesquisas no âmbito da tradução. Amostra textos produzidos nos julgamentos criminais na Inglaterra, Espanha, Alemanha e França (PORTANDOLFO, 2012).

Assim como os corpora, as pesquisas sobre esse registro são variadas, dentre elas destaca-se as pesquisas nos âmbitos da terminologia (ex.: GOŹDŹ-ROSZKOWSKI e WITCZAK-PLISIECKA (eds), 2011; RICHARD, 2018); tradução (ex.: FANEGO e RODRÍGUEZ-PUENTE (eds) 2019), variação linguística (ex.: GOZDA-ROSZKOWSKI, 2011), análise do discurso (GOŹDŹ-ROSZKOWSKI, 2021) e sintaxe (ex.: CHOVANEC, 2013; DAHLMAN, 2006). No Brasil, conforme demonstrado por Ferrari e Cunha (2022) predomina-se, na linguística, pesquisas voltadas para a análise do discurso (ex.: BITTAR, 2009; SVOBODOVÁ, 2017), terminologia (ex.: MACIEL, 2001; ROCHA, 2017; TEIXEIRA et al 2019) e tradução (ex.: CARVALHO, 2006, 2014). Tem-se ainda a produção de dicionários (ex.: CASTRO, 2013; DINIZ, 1998; GUIMARÃES, 2013; KRIEGER *et al*, 2008; SANTOS, 2001), glossários (ex.: KRIEGER *et al*, 2004, 2006) e manuais de redação (ex.: AQUINO; DOUGLAS, 2017; BRASIL, 2004, 2006, 2018; PETRI, 2017; DAMIÃO; HENRIQUES, 2020).

Dentre os vários grupos e projetos de pesquisa existentes no Brasil destacamos os trabalhos em análise do discurso jurídico do pioneiro grupo de pesquisa Linguagem e Direito, junto Faculdade de Direito da Universidade Católica de Pernambuco

(UNICAP/CNPq), liderado pela professora Virgínia Colares. Trata-se de um grupo multidisciplinar que desenvolve pesquisas nas áreas de análise e crítica do discurso, hermenêutica e direito, com alguns trabalhos recentes sobre linguística forense. Também junto à Universidade Católica de Pernambuco foi criada a Associação de Linguagem & Direito (ALIDI)[31], que promove e divulga eventos, projetos e estudos sobre discurso jurídico e linguística forense no Brasil. Contribuindo ainda com a edição de publicações na área.

Já no âmbito da terminologia e tradução citamos o projeto TermiSul[32] (Projeto Terminológico Cone Sul) da Universidade Federal do Rio Grande do Sul. Idealizado pela Profa. Dra Maria da Graça Krieger em 1991, o projeto se volta para pesquisas nas áreas da terminologia e terminografia em linguagens especializadas nas línguas: português, alemão, espanhol, francês, inglês, italiano e russo. Dentre as linguagens especializadas abarcadas pelo projeto temos a linguagem jurídica, cujo estudo, com auxílio da linguística de corpus, resultou em dicionários e glossários multilíngues de direito, várias publicações e a criação de bases de dados legais, entre elas a Base de Dados Terminológica de CLEs da Linguagem Legal (BDT Cles LEGIS)[33], uma base de dados online com as combinatórias léxicas especializadas (CLEs) da legislação ambiental em português, alemão, espanhol, francês, inglês e italiano.

A Linguística de corpus também fornece ferramentas e métodos para a análise dos corpora. Entre eles destaca-se a Análise Multidimensional, uma abordagem metodológica empírica baseada em corpora proposta por Douglas Biber que será abordada na próxima seção. Essa metodologia permite descrever e caracterizar qualquer registro em termos de dimensões a partir de análises estatísticas de um corpus e da interpretação funcional dos resultados.

2.4 A análise multidimensional

A Análise Multidimensional (AMD) é uma metodologia da Linguística de Corpus desenvolvida por Douglas Biber (1988). Essa abordagem metodológica foi inicialmente

[31] Disponível em: http://www.alidi.com.br/. Acesso em: 08 set. 2022.
[32] Disponível em: http://www.ufrgs.br/termisul/. Acesso em: 08 set. 2022.
[33]Disponível em: http://www.ufrgs.br/termisul/cles/. Acesso em: 08 set. 2022.

criada para analisar a variação dos registros escritos e orais de língua inglesa, sendo primeiramente apresentada pelo pesquisador em 1985 e sistematizada em Biber (1988). Conforme Biber (1988), Biber e Conrad (2009), Berber Sardinha (2010) e Delfino (2021), a Análise Multidimensional permite analisar e descrever a variação linguística dos registros, podendo ser aplicada a qualquer língua e variedade. Para tanto, é necessária uma amostra de linguagem autêntica, composta por uma quantidade significativa de textos, que seja representativa do (s) registro (s) pesquisado (s). A amostra não pode ser enviesada, deve permitir a comparação com outras, e, para que seja possível o processamento dos dados e realização das análises, deve ser lida por computador. Para atender a esses critérios, faz-se assim necessário a utilização de um corpus.

É importante ressaltar que não só a arquitetura do corpus, mas também seu tamanho em número de palavras deve ser cuidadosamente considerado na realização da AMD. Isso porque faz-se necessário não só um corpus representativo e equilibrado da variedade estudada, mas também que tenha um número de palavras total que permita seu processamento, sendo desejável um corpus de tamanho médio. A esse respeito alerta Berber Sardinha (2013): “[...] um corpus gigantesco tornaria inviável a pesquisa em Análise Multidimensional, uma vez que ela depende da anotação morfossintática de todos os textos, da contagem das características linguísticas e de seu processamento estatístico.” (BERBER SARDINHA, 2013, p. 62-63).

A AMD segue a perspectiva do registro proposta por Biber e Conrad (2009). Seus objetivos, segundo Biber (1988), Biber e Conrad (2009), Berber Sardinha (2010) e Delfino (2021) são: (a) identificar os padrões de co-ocorrência léxico-gramaticais de uma língua e (b) comparar os registros a partir desses padrões. Para identificar os padrões de co-ocorrência léxico-gramaticais são utilizadas técnicas estatísticas e seus resultados são interpretados qualitativamente em termos funcionais para estabelecer-se as dimensões de variação.

A AMD parte da hipótese que:

> [...] fortes padrões de co-ocorrência de características linguísticas marcam as dimensões funcionais subjacentes. As características não co-ocorrem aleatoriamente nos textos. Se certas características co-ocorrerem

> consistentemente, então é razoável procurar uma influência funcional subjacente que encoraje seu uso. (BIBER, 1988, p. 13, tradução nossa)[34]

Logo, a recorrência de determinadas *características* linguísticas em padrões específicos em um corpus refletiria características situacionais e funcionais próprias de determinado registro. Os padrões de co-ocorrência léxico-gramaticais típicos de um registro, quando interpretados funcionalmente, passam a ser chamados de dimensões de variação. Em um único registro existem múltiplas dimensões e, a partir delas, é possível estabelecer um *continuum* de variação de um registro, suas características linguísticas próprias e determinar sua proximidade ou distância com outros registros. Elas nos permitem assim classificar, descrever e comparar os registros, assim como identificar empiricamente quantas dimensões existem em determinado registro, quais funções são independentes e quais estão associadas no interior de cada dimensão e sua importância para a caracterização do registro. (BIBER, 1988; BIBER e CONRAD, 2009; BERBER SARDINHA, 2010)

2.4.1 Metodologia da AMD

Ao trabalhar com a AMD, conforme nos ensinam Berber Sardinha (2010), Biber (1988 e 1995), Brezina (2018) e Delfino (2021) podemos optar por realizar a "AMD completa" *(Full MD)* ou a "aplicação de dimensões" também chamada de AMD aditiva. Ambas partem da mesma base teórica e proporcionam o estudo da variação dos registros analisados. Enquanto a AMD completa fornece um retrato mais detalhado, não só dos registros, mas também dos textos que compõe cada registro e identifica as dimensões de variação, a segunda oferece um retrato mais geral dos registros em análise e possibilita a sua comparação com o estudo anterior ao adicionar seus registros às dimensões de AMD completa realizada anteriormente. Isso não significa que a AMD aditiva não forneça uma riqueza de informações linguísticas dos textos analisados, esta é apenas menos detalha que a primeira.

---

[34] […] strong co-occurrence patterns of linguistic características mark underlying functional dimensions. Características do not randomly co-occur in texts. If certain características consistently co-occur, then it is reasonable to look for an underlying functional influence that encourages their use.

Cada tipo tem suas particularidades, mas um não exclui o outro. Eles se complementam e se enriquecem: enquanto a AMD completa traz um retrato detalhado ela está limitada pelo número de registros que consegue abarcar, a AMD aditiva a complementa ao adicionar às suas dimensões os registros por ela analisados sem ser necessário realizar todos os procedimentos metodológicos da AMD completa. Dessa forma a AMD completa aumenta em escopo, se tornando mais abrangente. A AMD aditiva, por sua vez, é enriquecida pelas referências da AMD completa, sendo considerada um bom ponto de partida para quem quer adentrar na análise multidimensional.

A AMD completa segue todos os passos propostos por Biber (1988), apresentados na próxima seção e permite. Já na AMD aditiva alguns deles etapas são desconsideradas e novas etapas acrescentadas. A primeira identifica as dimensões de variação de determinados registros e a segunda se vale dessas dimensões para a análise do corpus escolhido. Não necessariamente o pesquisador precisa utilizar todas as dimensões do estudo-base na AMD aditiva, é possível escolher a quais das dimensões do estudo original seu corpus será adicionado de acordo com os objetivos da sua pesquisa. Temos desde estudos que utilizam apenas uma das dimensões do estudo original até aqueles que se valem de todas elas

A AMD aditiva permite incorporar novos registros às dimensões de uma AMD completa, enriquecendo-a. Essa expansão de registros é uma das maiores vantagens desse tipo de abordagem, que se destaca ainda pelo menor tempo e conhecimentos estatísticos para a sua realização quando comparada àquela completa. Assim como a AMD completa, esse tipo de análise pode ser realizado tanto em pesquisas sincrônicas quanto diacrônicas e aplicada a quaisquer línguas e variedades e complementada por análises qualitativas. Diferentemente da AMD completa, a AMD aditiva não tem a etapa da análise fatorial, não sendo possível identificar novas dimensões.

É importante ressaltar que para a realização da AMD aditiva o corpus ao qual serão aplicadas as dimensões deve ter sido preferencialmente coletado, etiquetado e analisado da mesma forma que o corpus do estudo que identificou as dimensões. Além disso, a língua do (s) registro (s) em análise deve ser a mesma daquela do estudo original. Isso porque, esse tipo de pesquisa pode apresentar resultados não confiáveis ou

incompletos caso a língua e ou os procedimentos metodológicos na compilação e análise do corpus sejam diversos daqueles do estudo original.

A seguir detalhamos as etapas metodológicas de cada uma dessa formas de realização da AMD.

2.4.1.1 AMD Completa

Os procedimentos metodológicos da AMD completa propostos por Biber (1988) e retomados por Biber e Conrad (2009), Berber Sardinha (2010), Brezina (2018) e Delfino (2021) foram por nós resumidos em dez etapas, quais sejam: (1) Estudo prévio, (2) Escolha e preparação do corpus, (3) Determinação das variáveis linguísticas relevantes, (4) Contagem das variáveis linguísticas no corpus, (5) Normalização das ocorrências das variáveis linguísticas, (6) Análise fatorial, (7) Cálculo dos Z-escores, (8) Cálculo dos escores de dimensão de cada texto, (9) Cálculo da média dos escores de dimensão e (10) Interpretação funcional dos fatores em dimensões.

A seguir detalhamos cada uma delas:

(1) Estudo prévio: é preciso realizar um estudo prévio sobre o (s) registro (s) que se pretende abarcar na AMD de forma a investigar suas características linguísticas e situacionais.

(2) Escolha e preparação do corpus: é necessário selecionar dentre os corpora existentes um que atenda à pesquisa ou ainda compilar um corpus que seja representativo da variedade que se pretende analisar. Cabe ressaltar que o corpus deve ser etiquetado morfossintaticamente (*POS tagging*), de forma a permitir a identificação e análise das variáveis linguísticas. Além disso, para a possibilitar a realização da análise fatorial o corpus deve ser representativo, equilibrado e balanceado, devendo o pesquisador dar especial atenção ao número de textos por seção e palavras por textos que devem ser semelhantes e, ao mesmo tempo, ao número total de textos que, segundo Biber (1988) Biber e Conrad (2009), Berber Sardinha (2010), Brezina (2018) e Delfino (2021), idealmente deve ser ao menos cinco vezes o número de variáveis linguísticas. Nem sempre é possível alcançar esse ideal e o uso de um número menor de textos

também possibilita a realização da análise, mas é preciso levá-lo em consideração quando da escolha/compilação do corpus para a pesquisa.

Quanto à etiquetação é recomendado usar etiquetadores próprios para cada língua que apresentem uma alta taxa de precisão. Para a língua inglesa, por exemplo, é recomendado usar o *Bibber Tagger* (Biber, 1988) para etiquetar o corpus. Já na língua portuguesa brasileira recomenda-se o uso do etiquetador PALAVRAS (BICK, 2001, 2014).

(3) Determinação das variáveis linguísticas relevantes: deve-se investigar na literatura quais características linguísticas serão relevantes para a pesquisa, ou seja, aquelas que possivelmente permitirão identificar e distinguir os registros do corpus. Não existe uma lista ou número fixo de variáveis para a realização da AMD, este depende da língua e dos objetivos da pesquisa.

Cabe ressaltar que, devido às especificidades de cada língua e de cada pesquisa, algumas variáveis que são consideradas relevantes por um pesquisador podem não ser para outro. Biber (1988), por exemplo, identificou 67 variáveis para o estudo da variação escrita e oral na língua inglesa. Já Xiao (2009) identificou 141 variáveis em seu estudo sobre as variedades do inglês (ex. americano, britânico, etc.). Berber Sardinha, Kauffmann e Acunzo (2014), por sua vez, determinaram 190 variáveis (vide ANEXO A) para estudar as dimensões de variação do português brasileiro.

(4) Contagem das variáveis linguísticas no corpus: a ocorrência das variáveis linguísticas em cada texto deve ser contabilizada. Para tanto são utilizados *software* específicos. Para seu estudo sobre a variação na língua inglesa Biber (1988) desenvolveu o *TagCount* que etiqueta o corpus segundo as 67 variáveis por ele identificadas como relevantes para a língua inglesa. Para o estudo do português brasileiro, Berber Sardinha, Kauffmann e Acunzo (2014), por sua vez, se valeram do "PALAVRAS *Tag count*" (BERBER SARDINHA, 2013) um *software*, por eles desenvolvidos, para fazer contagem das características linguísticas.

(5) Normalização das ocorrências das variáveis linguísticas: normaliza-se por 1000 palavras a ocorrência das variáveis linguísticas em cada texto para manter a comparabilidade e possibilitar a realização das análises. Para tanto divide-se a frequência

da característica no texto pelo número de palavras do texto e multiplica-se o resultado por mil:

$$Freq.normalizada = \left(\frac{freq.\,da\,variável\,no\,texto}{\text{número de palavras do texto}}\right) \text{ x } 1000$$

(6) Análise fatorial: trata-se de uma série de cálculos estatísticos complexos realizados automaticamente por *softwares* especializados (ex.: IBM SPSS Statistics 23, R). Com ela investiga-se a correlação entre as variáveis, agrupando aquelas correlacionadas e excluindo-se aquelas que não apresentam correlação, reduzindo assim, o número de variáveis a serem analisadas. Dessa forma, os dados do corpus, quais sejam, as ocorrências das características linguísticas em cada texto normalizadas por 1.000, são analisados de forma a calcular a frequência das variáveis linguísticas em cada texto. O resultado é a identificação de padrões de co-ocorrência de variáveis nos textos, chamados de fatores. As variáveis que compõe os fatores estão relacionadas entre si, sendo elas positivas e negativas em distribuição complementar, ou seja, se uma variável positiva tem alta frequência a negativa terá baixa frequência no texto e vice-versa. (BIBER, 1988; BIBER e CONARD, 2009; BREZINA, 2018)

Antes de realizar a análise fatorial propriamente dita é preciso verificar se há correlação entre as variáveis, caso contrário não será possível combiná-las em fatores. Para isso, alguns testes estatísticos podem ser realizados e as variáveis que não passam nos testes são excluídas da análise.

Brezina (2018), por exemplo, sugere o Teste de Correlação de Pearson. Aplicável a variáveis contínuas, esse teste estabelece a correlação entre as variáveis em termos de coeficiente de correlação (r) [35]. Esse coeficiente vai de -1 a 1, sendo que os valores negativos indicam que a correlação é negativa, zero indica que não há correlação e, os números acima de zero, que a correlação é positiva. Quanto mais próximo de 1 é o valor, maior a correlação entre as variáveis. Para isso calcula-se o grau de covariância[36] dos dados que indica numericamente o quanto as variáveis estão correlacionadas ou não e seu resultado é dividido pela distância estatísticas das variáveis analisadas, obtida

[35] *r = covariance /SD1 x SD2* (BREZINA, 2018, p.142)
[36] *covariance = sum of multiplied distances from mean1 and mean2 /total no: of cases* – 1 (BREZINA, 2018, p.142)

através do seu desvio padrão. Além disso, para assegurar a confiabilidade dos resultados desse teste permitindo a generalização dessa correlação Brezina (2018) sugere que este seja complementado pelo cálculo do *p-value* ou o do intervalo de confiança.

Já Delfino (2021) aconselha a realização do Teste de Bartlett para determinar a existência correlação entre as variáveis. Esse teste determina se as variáveis analisadas têm a mesma variância ou não. Ou seja, se todas as variáveis estão a mesma distância da média ou não. Para tanto, estabelece-se uma hipótese nula na qual todas as variáveis têm a mesma variância e a hipótese contrária: existe mais de um valor de variância na população. Se o nível de significância for inferior a 0,05 a hipótese nula é rejeitada e a correlação é confirmada.

Verificada a correlação entre as variáveis, damos início à análise fatorial. Delfino (2021) aconselha a realização de duas análises fatoriais: não rotacionada e rotacionada. A primeira análise resulta em um gráfico de escarpa e uma tabela da variação compartilhada. No gráfico os fatores são representados pelos seus eixos (ex.: fator 1 = eixo x e fator 2 = eixo y) e as variáveis representadas pelos pontos distribuídos ao longo do plano do gráfico. Já a tabela de variação compartilhada traduz as informações do gráfico em dados numéricos. A análise conjunta desses dados possibilita a determinação do número de fatores. (BREZINA, 2018; DELFINO, 2021)

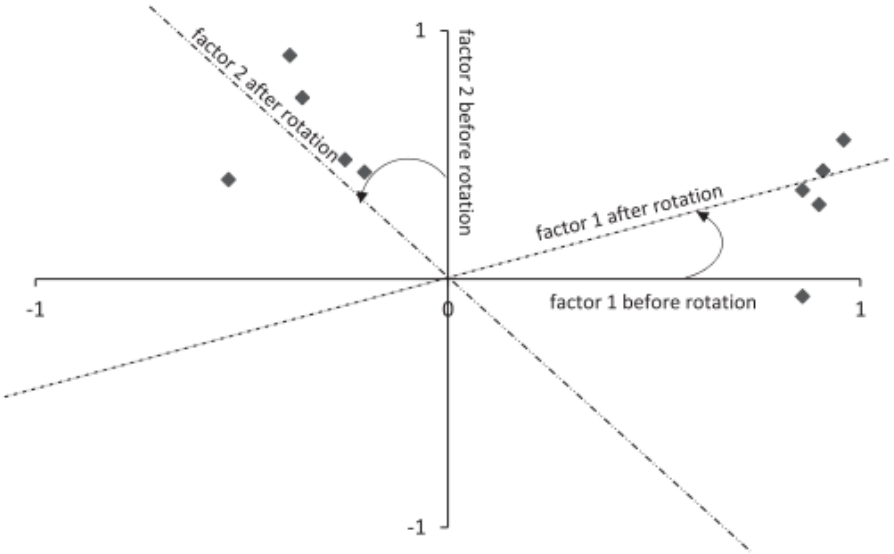
De forma a otimizar os resultados aconselha-se realizar uma segunda análise fatorial, dessa vez rotacionada. Nela os resultados da primeira análise são submetidos a rotação, girando-se os fatores do gráfico de escarpa, ou seja, sues eixos, sem alterar a posição das variáveis nele distribuídas. Segundo Biber (1988), quando rotacionamos os eixos as variáveis mais relevantes, consideradas assim representativas daquele registro, tendem a se enquadrar em apenas um fator que será então por elas caracterizado e as demais são excluídas.

Existem alguns tipos de rotação possíveis, cada tipo produz um resultado distinto, na rotação gira-se os eixos do gráfico em determinado ângulo a fim de abarcar apenas as variáveis mais relevantes, filtrando ainda mais os resultados. Brezina (2018) e Delfino (2021) aconselham o uso da rotação *promax* (rotação oblíqua), que parte do princípio de que há relação entre os fatores e é amplamente utilizada quando aplicamos a analise fatorial em corpora. Já Friginal e Hardy (2014) aconselham *varimax* (rotação

perpendicular), o padrão quando lidamos com analises não linguísticas e considera que pode não haver relação entre os fatores. Ambas são voltadas para um grande número de dados e visam reduzir o número de variáveis, selecionando apenas as mais relevantes.

Para melhor visualização do processo de rotação reproduzimos a imagem retirada de Brezina (2018) que ilustra uma rotação *promax*, ou seja, que gira os eixos do gráfico obliquamente em relação ao seu ponto de encontro. As variáveis abrangidas dentro desses eixos após a rotação são consideradas de maior significância, sendo mantidas e as não abrangidas são excluídas da análise.

Figura 1 – Rotação Promax



Fonte: Brezina (2018, p.166)

O próximo passo é decidir quais fatores extrair, para auxiliar nessa decisão existem algumas técnicas estatísticas. Brezina (2018) e Friginal e Hardy (2014) sugerem a produção de um *screen plot*, um gráfico que indica quais fatores devemos extrair ao dispor os fatores ao longo do eixo X e os *eigenvalues*[37] ao longo do eixo Y. A partir do gráfico aplica-se um critério de seleção. Aqui também existe mais de um critério possível.

[37] *Eigenvalue is a measure of how much variation in the data a factor explains – the larger the value the better. Mathematically, eigenvalue is a sum of squared factor loadings for all variables.* (BREZINA, 2018, p. 116)

Brezina (2018) traz: (a) os fatores que têm o *eigenvalue* maior de 1, (b) os fatores acima do ponto de inflexão do *screen plot,* (c) os fatores que estão acima da linha obtida pela Análise Paralela (PA)[38].

(7) Cálculo dos Z-escores: os resultados da análise fatorial devem ser padronizados para equilibrar as diferentes frequências das características individuais de modo que tanto as de alta quanto as de baixa frequência tenham um status semelhante no cálculo do escore das dimensões. Para tanto, calcula-se o z-escore de cada variável. Esse cálculo é feito subtraindo a frequência normalizada da variável da sua frequência média e o resultado dividido pelo seu desvio padrão. Seguindo a seguinte fórmula:

$$Z\ escore = \frac{freq.normalizada\ da\ variável - \ freq\ média\ da\ variável}{desvio\ padrão\ da\ variável}$$

Ao realizar essa operação as variáveis passam a terem uma base comum: uma média de 0 e um desvio padrão igual a 1. Com isso é possível identificar se elas têm uma frequência alta ou baixa em determinado texto sem que essa influencie no cálculo do escore de dimensão. Os z-escores positivos refletem as variáveis de maior frequência no texto e aqueles negativos as de menor frequência. Por exemplo, se temos uma variável com um Z-escore positivo igual a 3,1, isso significa que se trata de uma variável com alta frequência, estando 3,1 desvios padrão acima da média. Já uma variável negativa com, por exemplo, -2 de z-escore, tem baixa ocorrência naquele texto, estando 2 desvios padrão abaixo da média.

(8) Cálculo dos escores de dimensão de cada texto: utilizado para determinar o quanto determinado texto pontua em cada dimensão. Para tanto, em cada texto, subtrai-se, da soma dos Z-escores das variáveis resultantes positivas na extração de fatores, a soma dos Z-escores das variáveis negativas da extração de fatores:

$$Escore\ de\ dim. = \ (\sum dos\ z\ escores\ das\ variáveis\ positivos) - (\sum dos\ z\ escores\ das\ varáveis\ negativas)$$

[38] *[…] used to establish a baseline of eigenvalues obtained in computer simulation (Monte Carlo) with unrelated (random) variables. The factors extracted need to have eigenvalues clearly above this baseline.* (BREZINA, 2018, p. 166)

Se o texto pontua positivamente significa que tem alta frequência das variáveis que caracterizam aquela dimensão. Se pontua negativamente tem baixa frequência daquelas variáveis.

(9) Cálculo da média dos escores de dimensão: realizado para identificar como os registros se encaixam em cada dimensão. Para calculá-lo soma-se os escores de cada texto, calculados na etapa anterior, e divide-se seu resultado pelo número total de textos do registro:

$$Média\ do\ escore\ de\ dim. = \frac{\sum dos\ escores\ de\ dimensão\ de\ cada\ texto\ do\ registro}{número\ de\ textos\ do\ registro}$$

Quanto mais positivo pontua o registro mais prototípico daquela dimensão, quanto mais negativo menos prototípico.

(10) Interpretação funcional dos fatores em dimensões: interpretam-se funcionalmente os resultados como dimensões de variação. Cada dimensão traduz uma escala, com um polo positivo e outro negativo na qual todos os registros analisados são dispostos de acordo com a média dos seus escores de dimensão. O polo positivo agrupa os registros com maior ocorrência das variáveis que caracterizam a dimensão, quanto maior o valor da média de dimensão daquele registro mais prototípico daquela dimensão ele é. Já o polo negativo agrupa os registros com as variáveis de menor ocorrência na dimensão, os registros que pontuam os maiores valores negativos são os menos prototípicos daquela região. Dessa forma busca-se explicar os padrões quantitativos obtidos em termos funcionais e relacioná-los às características situacionais dos registros. Quanto mais distantes estão os registros na escala, mais distintos são e vice-versa.

2.4.1.2 AMD Aditiva

Para a realização da AMD aditiva, por sua vez, baseados em Berber Sardinha (2013) e Berber Sardinha *et al* (2019), trazemos 11 etapas a serem observadas, são elas: (1) Estudo prévio, (2) Escolha do estudo que servirá como base da AMD aditiva, (3) Escolha das dimensões a serem analisadas, (4) Identificação das características linguísticas das dimensões escolhidas, (5) Escolha e preparação do corpus, (6)

Contagem das variáveis linguísticas no corpus, (7) Normalização das ocorrências das variáveis linguísticas, (8) Cálculo dos Z-escores, (9) Cálculo dos escores de dimensão de cada texto, (10) Cálculo da média dos escores de dimensão dos novos registros e (11) Comparação dos novos registros com aqueles do estudo-base.

A seguir detalhamos cada uma delas:

(1) Estudo prévio: assim como na AMD completa, também na AMD aditiva faz-se necessária a realização de um estudo preliminar sobre o (s) registro (s) que se pretende estudar de forma a investigar suas características linguísticas e situacionais.

(2) Escolha do estudo que servirá como base da AMD aditiva: é necessário selecionar dentre as pesquisas de AMD completa uma que se enquadre nos objetivos da pesquisa pretendida. As dimensões utilizadas no estudo escolhido serão utilizadas como base para a adição dos novos registros. É importante que o estudo selecionado descreva detalhadamente os procedimentos metodológicos seguidos, os softwares utilizados, as variáveis que compõe cada dimensão, suas ocorrências, assim como, a média e o desvio padrão das variáveis linguísticas por dimensão para possibilitar o cálculo dos os escores de dimensão.

(3) Escolha das dimensões a serem analisadas: escolhe-se dentre as dimensões identificadas no estudo base aquelas que sejam pertinentes para a análise aditiva pretendida. É possível utilizar todas ou ainda, apenas uma, tudo depende dos objetivos da pesquisa.

(4) Identificação das características linguísticas das dimensões escolhidas: seleciona-se apenas as características linguísticas presentes na (s) dimensão (s) escolhida (s) para o estudo. Berber Sardinha *et al* (2019) alertam que quando determinada variável consta em mais de uma dimensão deve-se considerar que ela pertence à dimensão na qual tem uma maior ocorrência, nas demais ela geralmente será reportada entre parênteses no estudo original e deve ser ignorada.

(5) Escolha e preparação do corpus: assim como na AMD completa, deve-se selecionar um corpus para a realização da pesquisa e, no caso de não existir um corpus que atenda aos objetivos da pesquisa, deve-se compilar um. O corpus deve ser etiquetado morfossintaticamente, preferencialmente com o mesmo *tagger* que foi utilizado no estudo base.

(6) Contagem das variáveis linguísticas no corpus: contabiliza-se a ocorrência das variáveis linguísticas das selecionadas na etapa 4 em cada texto do corpus de estudo, preferencialmente utilizando o mesmo programa do estudo original.

(7) Normalização das ocorrências das variáveis linguísticas: também na AMD aditiva normaliza-se a ocorrência das variáveis linguísticas por 1000 palavras visando possibilitar a comparabilidade entre elas. Para tanto utiliza-se a fórmula apresentada na seção anterior, qual seja:

$$Freq.normalizada = \left(\frac{freq.da\ variável\ no\ texto}{\text{número de palavras do texto}}\right) \text{ x } 1000$$

(8) Cálculo dos Z-escores: conforme exposto na seção anterior, para calcular o escore das dimensões é necessário que as frequências das variáveis sejam niveladas para que tenham um mesmo peso, impedindo que sua frequência influencie no cálculo dos escores de dimensão. Para isso, realizamos o cálculo do Z-escore de cada variável. Entretanto, como na AMD aditiva não se realiza a análise fatorial o cálculo dos Z-escores muda um pouco quando comparado com aquele da AMD completa. Para possibilitar a aplicação das dimensões do estudo base ao corpus, na análise aditiva, utilizamos a frequência média e do desvio padrão de cada variável do estudo original para o cálculo do Z-escore das variáveis do corpus em análise. Para tanto subtrai-se a frequência normalizada da variável da frequência média da variável no estudo-base e divide-se esse valor pelo desvio padrão da variável no estudo-base, como pode ser observado na seguinte fórmula:

$$Z\ escore = \frac{freq.normalizada\ da\ variável - \ freq\ média\ da\ variável\ do\ estudo\ original}{desvio\ padrão\ da\ variável\ do\ estudo\ original}$$

(9) Cálculo dos escores de dimensão de cada texto: utilizado para determinar a pontuação de cada texto do corpus em análise em cada dimensão analisada do estudo base:

> Para calcular os escores do texto, somar as frequências padronizadas (Z-escores) de todas as características linguísticas do polo positivo das dimensões do estudo base. Em seguida, somar as características linguísticas do polo negativo (se existir) e subtrair a soma do polo negativo da soma do polo positivo. (BIBER, 1988, p. 95, tradução nossa)[39]

Em outras palavras, em cada dimensão, em cada texto do corpus em análise soma-se os Z-escores calculados na etapa anterior cujas variáveis tenham tido uma pontuação positiva no estudo original e dessa soma subtrai-se a soma dos Z-escores das variáveis que no estudo original tiveram uma pontuação negativa. No caso de dimensões com apenas um polo no estudo base apenas se soma os valores dos Z-escores. Para tanto, propõe-se a seguinte fórmula:

$$Escore\ de\ dim. =$$
$$(\sum dos\ z\ escores\ das\ variáveis\ positivas\ ) - (\sum dos\ z\ escores\ das\ variáveis\ negativas)$$

(10) Cálculo da média dos escores de dimensão dos novos registros: calcula-se a média dos escores de dimensão de cada registro do corpus em análise para possibilitar sua comparação com aqueles do estudo base em termos de dimensões. Para isso soma-se os escores de cada texto, calculados na etapa anterior, e divide-se seu resultado pelo número total de textos do registro, conforme a seguinte fórmula:

$$Média\ do\ escore\ de\ dim. = \frac{\sum dos\ escores\ de\ dimensão\ de\ cada\ texto\ do\ registro}{número\ de\ textos\ do\ registro}$$

(11) Comparação dos novos registros com aqueles do estudo-base: compara-se os registros do corpus em análise entre eles e com aqueles do estudo base. Para tanto adiciona-se a média dos escores de dimensão de cada registro à dimensão correspondente do estudo anterior, analisa-se as diferenças e semelhanças entre eles e

[39] *To compute the text scores, sum up the standardized frequencies (Z-scores) of all the linguistic characteristics loading on the positive pole of the dimensions for the reference study. Then add the linguistic características loading on the negative pole (if it exists) and subtract the summation of the negative pole from the summation of the positive pole.*

exemplifica-se as características linguísticas definidoras de cada dimensão identificadas nos novos registros com extratos de textos do corpus utilizado.

2.4.2 Modelos

Desde a apresentação da Análise Multidimensional em Biber (1988) essa metodologia tem sido aplicada em estudos linguísticos nas mais diversas línguas para o estudo dos mais variados registros. Durante as últimas décadas a AMD não ficou estagnada: ela se desenvolveu, passou por modificações e foram propostos novos modelos[40]. Conforme nos ensina Delfino (2021), apesar de ainda manterem grande similaridade com a proposta de Biber (1988), conhecida como clássica, funcional ou ainda gramatical, foram propostos modelos de AMD lexical. Enquanto a Análise Multidimensional proposta por Biber (1988) se baseia na interpretação funcional dos fatores, sendo observadas as *características* léxico-gramaticais dos registros, os modelos lexicais se baseiam na interpretação lexical dos fatores, que passam a ser de cunho exclusivamente lexical. Segundo Delfino (2021), existem hoje quatro modelos de AMD, sendo um funcional/gramatical (Biber, 1988) e três modelos lexicais: (a) temática, (b) colocacional e (c) semântica.

A AMD lexical temática foi desenvolvida por Berber Sardinha em 2014 com o objetivo de identificar as dimensões de variação lexical relacionada à representação da identidade nacional americana e brasileira em diacronia a partir de textos escritos em inglês nos séculos XIX, XX e XXI disponíveis no *Google Books*. Nesse estudo o pesquisador seguiu os mesmos passos da AMD clássica, alterando apenas as variáveis e a forma de etiquetar o corpus, definindo como variáveis apenas os substantivos, verbos, advérbios, adjetivos e expressões multipalavras e etiquetando o corpus por lema utilizando o *Tree Tagger*. Isso permitiu a identificação dos campos temáticos onde havia co-ocorrência lexical significante e a subsequente seleção de fatores lexicais que foram então interpretados em dimensões de variação lexical. (DELFINO, 2021)

O segundo modelo de AMD lexical trazido por Delfino (2021), chamado de AMD colocacional, também teve como idealizador Berber Sardinha. Apresentado em 2017,

[40]Aqui trazemos um panorama dos modelos propostos. Para mais informações ver Delfino (2021).

essa modalidade analisa as colocações sob a perspectiva do registro. Assim como o modelo anterior, seguem-se os mesmos passos de Biber (1988), sendo alterado apenas o foco: as palavras. Logo, ao invés de analisar as características léxico-gramatical em textos, contando-os e calculando seus escores, foram analisadas as palavras, medida a força de atração entre elas e calculados os escores visando identificar os colocados.

O modelo mais recente de AMD lexical foi proposto por Berber Sardinha e Delfino em 2021 com o objetivo de estudar a variação semântica no corpus de letras de músicas pop americanas por eles compilado. Da mesma forma que os demais modelos, foram seguidos os passos de Biber (1988). Aqui também a diferença entre os modelos se encontra na etiquetagem do corpus e nos critérios de determinação das variáveis. O corpus foi etiquetado em campos semânticos com o etiquetador UCREL[41] e as variáveis são campos semânticos determinados pelos pesquisadores. (DELFINO, 2021)

2.4.5 Análise Multidimensional do Português brasileiro

No português brasileiro a Análise Multidimensional mais abrangente já realizada é a já citada pesquisa de Berber Sardinha, Kauffmann e Acunzo (2014)[42]. Nela os pesquisadores analisaram o Corpus Brasileiro de Variação de Registro (CBVR), um corpus por eles compilado para essa pesquisa que abrange 48 registros (36 escritos e 12 falados) da língua portuguesa brasileira, totalizando 5.644.006 palavras. Trata-se de um corpus sincrônico, escrito e falado, anotado em *POS* e características lexicais usando o software PALAVRAS (BICK, 2000, 2014) e balanceado por número de textos por registro (20 textos cada). Segundo os autores, dentre os corpora já utilizados para AMD ao redor do mundo o CBVR é o corpus que abrange a maior quantidade de registros já compilado, tendo também um dos maiores números de palavras e textos em relação aos demais corpora usados nessa comparação.

O CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) é composto pelos seguintes registros:

1. *Academic articles*

[41] Disponível em: http://ucrel-api.lancaster.ac.uk/usas/tagger.html. Acesso em: 20 set. 2022.
[42] Aqui trazemos um panorama dessa pesquisa, para maiores detalhes ver: Berber Sardinha, Kauffmann e Acunzo (2014).

2. *Agreements*
3. *Blogs*
4. *BUSINESS CONFERENCE CALLS*
5. *Business letters*
6. *Campaign plans*
7. *Church liturgy*
8. *Comics*
9. *CONGRESSIONAL DEBATES*
10. *CONVERSATION*
11. *Editorials*
12. *Emails – Personal*
13. *Encyclopedia entries*
14. *Essays*
15. *. Facebook*
16. *Game instructions*
17. *General fiction*
18. *Government bids*
19. *Horoscope*
20. *INTERVIEWS – SOCIOLINGUISTIC*
21. *INTERVIEWS – PRESS*
22. *INTERVIEWS TV*
23. *Jokes*
24. *Legislation*
25. *Magazine celebrity*
26. *Magazine news*
27. *Medicine/drug labels*
28. *Minutes*
29. *Newspaper reportage*
30. *Non-fiction books*
31. *POLITICAL SPEECHES*
32. *Prep. school texts*

33. *Product labels*
34. *RADIO BROADCASTS*
35. *Recipes*
36. *Short stories*
37. *SOAP OPERAS*
38. *SONGS*
39. *TEXTBOOK DIALOGS*
40. *Textbook texts*
41. *Textbooks*
42. *Theses*
43. *TV NEWS*
44. *Twitter*
45. *User's/owner's manuals*
46. *Websites*
47. *Written exams*
48. *Youth fiction*

Fonte: Berber Sardinha, Kauffmann e Acunzo (2014, p. 38 e 39)

Os registros em caixa alta são aqueles orais e os demais os escritos. Destaca-se a ampla abrangência do corpus que amostra desde registros tradicionais como notícias de jornal e obras literárias até registros da internet como *Facebook* e *Twitter*. O corpus também se destaca pela inclusão de registros orais, pouco explorados no Brasil. Dentre seus registros, destacamos ainda o registro *Legislation* (legislação), que amostra de forma genérica os textos legais. Nessa seção temos a Constituição, Códigos, Leis Ordinárias, Decretos, Medidas Provisórias e um Ato institucional, totalizando 20 textos e 125 mil palavras (2,2% do total de palavras do corpus).

Em relação à análise multidimensional propriamente dita, no estudo foi realizada uma AMD completa seguindo a metodologia proposta por Biber (1988). O primeiro passo foi compilar e anotar o CBVR em POS, usando o PALAVRAS (BICK, 2000, 2014), a anotação foi então revisada manualmente por amostragem e quando encontrados erros estes foram corrigidos. Em seguida, as 190 variáveis por eles identificadas (vide ANEXO

A) em pesquisa bibliográfica prévia como relevantes para a variação do registro e que possibilitavam sua identificação automática foram então contadas no corpus com auxílio do pós processador por eles desenvolvido para esse propósito: “PALAVRAS Tag count” (BERBER SARDINHA, 2013). Para tanto, eles: “alimentaram o corpus anotado e os padrões de busca avançada por eles elaborado (baseado nas etiquetas individualmente e combinadas) e revisaram cuidadosamente os resultados” (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014, p. 37, tradução nossa)[43]. As ocorrências dessas variáveis foram então normalizadas por 1.000.

Feito isso, os dados foram submetidos a testes estatísticos para determinar se havia correlação entre variáveis, quais sejam o teste Kaiser-Meyer-Olkin e o teste de Esfericidade de Bartlett. Ambos os testes indicaram uma correlação significativa entre as variáveis. O próximo passo foi realizar a análise fatorial. Foram feitas duas análises fatoriais no software SPSS 20: uma não rotacionada, com *principal axis* como método de extração, sendo desconsideradas as variáveis com valor inferior a .2, e uma rotacionada com rotação promax, sendo extraídas apenas as características com valor igual ou maior a + ou - .3. No total foram extraídas 93 (vide ANEXO B) das 190 variáveis iniciais na análise fatorial. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Na sequência esses valores foram padronizados pelo cálculo dos Z-escore e utilizados para calcular os escores de dimensões de cada texto e a média do escore das dimensões para cada registro (vide ANEXO C para a estatística descritiva desse estudo). De posse desses resultados os autores realizaram *o F-test* (ANOVA) e o teste R2. O primeiro é utilizado para verificar se existem diferenças significativas entre os registros em todas as dimensões, já o segundo permite prever quantas dimensões serão identificadas a partir do cálculo da porcentagem da variação dos escores de dimensão. Ambos os testes sugeriram uma diferença significativa entre os registros. Por fim, os resultados foram interpretados funcionalmente em termos de dimensões de variação. Essas são representadas em escalas ao longo das quais os registros vêm distribuídos de acordo com os sues escores de dimensão calculados na etapa precedente. A escala tem dois polos: um negativo e outro positivo. Os registros com maior pontuação no polo

[43] *[…] fed the tagged corpus and our detailed hand-crafted search patterns (based on individual tags and tag combinations) and whose output we carefully revised for accuracy.*

positivo são os registros mais prototípicos de determinada dimensão, enquanto os com as maiores pontuações negativas são os menos prototípicos. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

No total Berber Sardinha, Kauffmann e Acunzo (2014) identificaram 6 dimensões de variação no português brasileiro: (*1) Oral versus literate discourse* (discurso oral vs. discurso letrado)*, (2) Argumentation* (argumentação), *(3) Involved versus informational production* (produção envolvida vs. produção informacional), *(4) Directive discourse* (discurso diretivo), *(5) Future versus past time orientation (*discurso orientado para o futuro vs. discurso orientado para o passado) e *(6) Reported discourse* (discurso indireto) (vide ANEXO D). Segundo os pesquisadores todas essas dimensões têm correspondência em maior ou menor nível com aquelas identificadas em outras línguas ou variedades

A dimensão 1: *Oral versus literate discourse* reúne 49 características linguísticas, com 35 no polo positivo e 14 no polo negativo (vide ANEXO D). Trata-se da dimensão com maior variação e correlação entre os fatores. Essa dimensão é considerada pelos autores, após observação dos resultados de AMD realizadas em outros idiomas, como potencialmente universal e reflete a diferença entre o discurso oral e aquele letrado. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

No polo positivo prevalecem os verbos, pronomes e advérbios. Os pronomes e verbos de primeira pessoa, por exemplo que realizam a função de trazer o foco para o emissor, enquanto aqueles de segunda pessoa se voltam para o receptor. Nesse polo figuram a grande maioria dos registros orais e predominam aqueles voltados para o diálogo, ressaltando a forte presença de interação entre os participantes. Os registros com maior pontuação positiva e consequentemente mais prototípicos dessa dimensão são: canções, diálogos de livros didáticos e novelas. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Já no polo negativo predominam os nomes e as preposições cuja função de condensar informações nos textos, aumenta a densidade informacional. Essa é realizada pelas nominalizações, particípios passados, substantivos abstratos e compostos, adjetivos atributivos e preposições. Nesse polo predominam os textos escritos informativos e com maior grau de especialização, sendo os mais negativos e, logo, os

menos prototípicos: bulas, licitações e leis. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

A dimensão 2, *Argumentation,* por sua vez, reúne 19 características linguísticas, todas no polo positivo (vide ANEXO D). Predominam as orações não finitas, infinitivas e relativas/adjetivas introduzidas por pronomes relativos. Estes realizam a função de estruturar uma argumentação, convencer o interlocutor e justificar determinada afirmação. Os registros estão distribuídos de forma que os mais prototípicos envolvem alto grau de argumentação, como discursos políticos e entrevistas de jornal, mas temos também o horóscopo, o registro identificado como mais prototípico dessa dimensão e que à primeira vista não seria considerado um registro argumentativo. Os menos prototípicos, por sua vez, apresentam pouca ou nenhuma argumentação como receitas, leis e sites. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Já a dimensão 3: *Involved versus informational production* reúne 16 características linguísticas, sendo 13 no polo positivo e 3 no polo negativo (vide ANEXO D). Verifica-se assim uma prevalência de características positivas, ou seja, características com alta frequência nos textos. Entre elas destacamos a alta ocorrência de perguntas, em especial as *tag questions* e as perguntas de sim ou não, assim como dos marcadores de discurso e contrações. A predominância dessas características aponta para um alto grau de interação entre os falantes que essa se dá em um contexto mais informal. Essa tendência é refletida nos registros prototípicos: entrevistas sociolinguísticas, as conversações e as novelas. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Já o polo negativo reúne poucas características, quais sejam: relação *type-token*, adjetivos atributivos de lugar e pronomes possessivos. Destas apenas a primeira, que reflete a variedade do léxico, tem uma carga significativa nessa dimensão, já que as demais tiveram pontuações maiores em outras dimensões, não sendo consideradas na análise dessa dimensão. Temos assim neste polo textos com uma ampla gama lexical o que permitiria a transmissão de informações de forma mais condensada e específica, sendo rótulos de produtos, sites e bulas de remédio os registros com maior pontuação negativa. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Essa dimensão, assim como a 1, é uma das com maior correlação entre os fatores e é provável que seja universal. Nos demais estudos de AMD essa dimensão e a

dimensão 1 são consideradas como uma única dimensão. Apesar de se assemelhar à primeira dimensão, os pesquisadores argumentam que no português brasileiro as dimensões 1 e 3 são dimensões distintas. Essa é mais especializada que aquela, representando de forma mais direta e precisa a interação entre os participantes e a carga informacional dos registros. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Segundo eles:

> Em resumo, consideramos que subjacente a ambas as Dimensões 1 e 3 está a mesma distinção básica entre "discurso estereotipado oral- isto é, conversação" - e "discurso estereotipado escrito - isto é, exposição informativa" (Biber 1995, p. 238). Entretanto, argumentamos que o contexto de produção dos registros letrados/informacionais diferencia as duas dimensões em nosso caso. Assim, embora os termos letrado e informativo tenham sido confundidos em pesquisas anteriores, em nosso estudo atual sentimos a necessidade de separá-los e atribuir-lhes significados particulares a fim de refletir as diferentes circunstâncias de produção que operam sobre os registros. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014, p. 63, tradução nossa)[44]

A dimensão 4: *Directive discourse,* assim como a 2 tem apenas um polo. Ela reúne 7 características linguísticas no polo positivo (vide ANEXO D). Ela é marcada principalmente pela presença de verbos no presente do subjuntivo e imperativo. Estes geralmente exercem a função de dar instruções ou ordens para a execução de tarefas visando atingir determinado objetivo prático. Devido à natureza dessas características o registro mais marcado é receitas, seguido por manuais e instruções de jogos. Já entre aqueles menos marcados, ou seja, aquelas menos prototípicos, ressaltam-se as bulas, que apesar do senso comum, apresentaram baixa ocorrência das características que refletem diretrizes. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

A dimensão 5, intitulada *Future versus past time orientation,* reúne 15 características linguísticas, 9 positivas e 6 negativas (vide ANEXO D). No polo positivo destaca-se a presença de verbos no futuro. Já no polo negativo predominam os verbos no passado. Esse contraste entre futuro e passado pode ser observado nos registros com maior pontuação em cada polo: licitações governamentais, instruções de jogos e

---

[44] *In short, we consider that underlying both Dimensions 1 and 3 is the same basic distinction between 'stereotypically spoken discourse – that is, conversation' – and 'stereotypically written discourse – that is, informational exposition' (Biber 1995, p. 238). However, we argue that the context of production of the literate/informational registers sets the two dimensions apart in our case. Hence, although the terms literate and informational have been conflated in previous research, in our current study we felt the need to separate them and attach particular meanings to them in order to reflect the different production circumstances operating on the registers.*

horóscopo no polo positivo *versus* livros não ficcionais, histórias curtas/contos e ficção infanto-juvenil. Observa-se que os registros no polo negativo são em sua maioria narrativos, enquanto aqueles do polo positivo são mais informativos e voltados para a previsão de acontecimentos futuros. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Essa dimensão, ao contrário das demais, não tem correspondente nos outros estudos que aplicam AMD. Nestes verifica-se o contraste entre o presente e o passado sendo a dimensão geralmente intitulada como narrativa. Apesar de ter sido constatado a tendência à narração no polo negativo, devido à presença do passado, como essa se opõe, nessa pesquisa, ao uso do futuro os pesquisadores entenderam que não se tem uma dimensão narrativa clara. Essa diferença com os demais estudos pode ser justificada: pela arquitetura do corpus, cuja representação dos registros literários é inferior aos demais estudos analisados ou pela estrutura da língua portuguesa. Faz-se assim necessário a realização de mais estudos para verificar essas hipóteses. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Por fim, a dimensão 6, *Reported discourse,* reúne 13 características, 9 no polo positivo e 4 no polo negativo (vide ANEXO D). Entretanto, as características do polo negativo têm um valor superior em outros fatores, tendo os autores optado por ignorá-las na análise dessa dimensão. Foram consideradas então só as características do polo positivo. Este é marcado pelo uso de pronomes, em especial aqueles raros, cuja ocorrência é extremamente alta no registro mais prototípico dessa dimensão, qual seja os sermões religiosos. Trata-se de textos que usam formas e léxico de português arcaico e também aquelas pouco usadas no dia a dia, em especial pronomes como *vos* e *los*. Este é seguido por contos e ficção infanto-juvenil, o que reflete a tendência de discurso relatado/indireto. Como registros menos prototípicos temos as bulas, os rótulos e as receitas. (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014)

Para a realização da presente pesquisa adotamos o estudo de Berber Sardinha, Kauffmann e Acunzo (2014) acima descrito como estudo-base para a realização de uma análise multidimensional aditiva. Pretende-se adicionar nosso corpus às dimensões de variação eles identificadas de forma a verificar se: (a) os textos legais são um registro; (b) como os textos legais brasileiros se encaixam nas dimensões de variação de Berber

Sardinha, Kauffmann e Acunzo (2014); (c) quais são as diferenças e semelhanças entre os textos legais e os registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014).

3. METODOLOGIA

Para cumprir os objetivos pretendidos utilizou-se a metodologia da Linguística de Corpus e a abordagem da Análise Multidimensional (AMD) (BIBER, 1988) para compilar e analisar o corpus *LEX-BR-Ius* (FERRARI e MARQUES, em preparação), visando estudar sua variação linguística. Para tanto o corpus foi adicionado às dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014). Para a realização dessa pesquisa nos valemos dos *softwares*: “IBM SPSS Statistics 23” (IBM, 2015), "Microsoft Excel 2016“, "Notepad++ 7.8.9” (HO, 2020), “PALAVRAS” (BICK, 2000; 2014) e “PALAVRAS Tag Count” (BERBER SARDINHA, 2013). A seguir detalharemos o processo e os métodos utilizados na pesquisa.

3.1 O *LEX-BR-Ius*

O *LEX-BR-Ius*[45] (FERRARI e MARQUES, em preparação), é um dos corpora do projeto *BR-Ius* (FERRARI, em preparação) que está compilando um conjunto de corpora de linguagem jurídica. O *LEX-BR-Ius* (FERRARI e MARQUES, em preparação) é composto por textos legais federais brasileiros integrais em vigência na data da compilação, realizada ao longo do ano de 2022, e selecionados com base em sua frequência de uso, medida a partir do número de resultados obtidos na busca no site especializado Jusbrasil[46].

O corpus é dividido em 7 seções: Constituição, Emendas à Constituição, Códigos, Leis complementares, Leis ordinárias, Medidas provisórias e Estatutos. A decisão por estas seções foi baseada na divisão adotada pelo Planalto brasileiro no seu portal de legislação[47] e será explicada na próxima seção. Os textos legais que compõem o corpus

[45] Seu nome faz referência à sua amostra: textos legais brasileiros, significando “leis/legislação do direito brasileiro”. *Lex* e *Ius* são termos do latim, língua na qual surgiu o direito romano, base do direito brasileiro, significando respectivamente: lei e direito, já a sigla BR se refere ao Brasil.

[46] Trata-se de um site agregador de conteúdo relacionado ao direito brasileiro. Nele encontramos as legislações, jurisprudência, artigos, notícias, etc. Em suma, todos os textos de acesso livre disponíveis na internet que envolvam o mundo jurídico brasileiro. Sendo utilizado majoritariamente pelos operadores do direito, optamos por utilizá-lo para obter o número de menções online dos textos legais de forma a filtrar seu uso/relevância no país, sendo selecionados aqueles com 10.000 resultados ou mais. Disponível em: https://www.jusbrasil.com.br/home. Acesso em: 20 set. 2022.

[47] Disponível em: http://www4.planalto.gov.br/legislacao/. Acesso em: 20 set. 2022.

foram retirados do site em questão, onde estão disponíveis gratuitamente e são atualizados diariamente todos os textos legais brasileiros, com exceção das resoluções e os decretos legislativos. Entendemos que o conjunto de textos legais selecionados atendem aos critérios estabelecidos por Biber (1993) para dar representatividade ao nosso corpus.

Trata-se, portanto, de um corpus escrito, monolíngue, síncrono, potencialmente comparável, anotado POS e lema, com marcação XML, seguindo a proposta *Modest XML* (HARDIE, 2014), de grandes dimensões. Pretende-se disponibilizar o corpus gratuitamente online em quatro versões: Limpa (*raw text),* Marcação XML, Anotada (POS e lema) e Completa (marcação XML e anotação), além de um cabeçalho com os metadados de cada texto em XML. A seguir descrevemos os passos da compilação do corpus.

Para auxiliar o processo de compilação foi criado um guia disponibilizado no APÊNDICE A, que contém também, de forma detalhada, todas as etiquetas por nós criadas.

3.1.1 Seleção dos textos

Para a compilação do corpus, o primeiro passo foi a seleção dos textos que o comporiam. Como já dito, trata-se de um corpus especializado de textos legais brasileiros. Diante disso estabeleceu-se como critérios para guiar a busca e seleção dos textos que comporiam o corpus do português: a) os textos devem ser textos legais federais; b) os textos devem estar em vigência; c) os textos devem ser representativos das espécies normativas existentes no país; d) os textos devem ser baixados na íntegra.

Como queremos compilar um corpus especializado é necessário que os textos que compõem o corpus sejam representativos da variedade desejada: textos legais. Para que o corpus seja equilibrado, é desejável que inclua uma variedade significativa e uma quantidade relevante de textos. É ainda desejável que os textos sejam integrais, pois apenas uma parte de um texto não é necessariamente representativa do todo (SINCLAIR, 2005). Essa observação é extremamente válida quando falamos de textos legais. Em um único texto legal vários assuntos são abordados, além disso, em distintas partes de um

mesmo texto as estruturas linguísticas e léxico utilizados mudam drasticamente. Logo, selecionar, ainda que arbitrariamente apenas uma parte do texto pode significar uma perda considerável de dados linguísticos. Por fim, devido ao tempo e recursos disponíveis para a compilação e subsequente pesquisa, assim como a abrangência e quantidade de legislações existentes no país decidiu-se delimitar o a textos legais federais brasileiros já que são aplicáveis a todo o território nacional e em número relativamente pequeno quando comparado ao total de textos legais brasileiros.

Quanto ao critério da vigência esclarecemos que textos legais são editados e modificados diariamente. Um texto vigente hoje pode não o ser amanhã, ou ainda vigorar com modificações. Alguns textos, como por exemplo a Constituição Federal, datam de mais de 30 anos, entretanto seu texto é atualizado frequentemente. Trata-se de um estudo linguístico sincrônico, logo o que nos interessa é a linguagem utilizada hoje nos textos legais. Diante disso optou-se apenas por textos que estejam vigentes, ou seja, estão sendo aplicados e produzindo efeitos na data da compilação[48].

Por fim, a representatividade é chave quando compilamos um corpus. Como demonstrado nas seções anteriores temos 8 espécies normativas: Constituição, Emendas à constituição, Leis complementares, Leis ordinárias, Leis delegadas, Medidas Provisórias, Decretos Legislativos e Resoluções. Se queremos um corpus representativo de textos legais, espera-se que amostras de todas as espécies normativas estejam inclusas. Entretanto, diferentemente das demais espécies que se encontram disponibilizadas online gratuitamente em diversos sites, os Decretos Legislativos e as Resoluções, exclusivos do Congresso Nacional, são de difícil acesso, e, como as matérias por eles reguladas tem caráter mais administrativo ou executivo que jurídico, optou-se por não os incluir no corpus.

Também se optou por não incluir as leis delegadas, isso porque, apesar de previstas na Constituição, só foram editadas 2 leis delegadas desde o advento da constituição, ambas em 1992, ambas com assunto puramente administrativo, uma vez que instituíam gratificações para servidores públicos que não são mais aplicáveis, tendo

[48] Devido à alta mutabilidade dos textos legais optou-se por marcar, entre outros, todas as alterações passadas pelos textos legais até o momento da compilação. Esse recurso permite também a realização de pesquisas diacrônicas, sendo possível identificar e retraçar todas as modificações pelas quais cada texto passou.

seu pagamento ocorrido entre 1992 e 1993. Diante do exposto, o corpus é composto por textos legais pertencentes às seguintes espécies normativas: Constituição, Emendas à Constituição, Leis complementares, Leis ordinárias e Medidas Provisórias.

Estabelecidos esses critérios teve início o segundo passo: a busca pelos textos. Para tanto, realizou-se uma busca no site: Portal da Legislação, um site pertencente ao Planalto onde são disponibilizadas gratuitamente online todas as legislações, com exceção das resoluções e dos decretos legislativos, e também textos de caráter administrativo. O Portal adota uma classificação um pouco diferente da Constituição quando se refere aos textos legais. As categorias disponíveis no site mesclam as espécies normativas previstas na Constituição e categorias por eles cunhadas. São elas: Constituição, Códigos, Leis ordinárias, Leis Delegadas, Leis Complementares, Estatutos, Decretos, Decretos-Leis, Decretos não numerados, Mensagens de veto total, Medidas provisórias, Projetos de Lei, Pareceres da AGU, Propostas de Emenda à Constituição.

Conforme exposto anteriormente, na CR/88 não existe previsão de uma espécie normativa intitulada “código”. A denominação “código” é utilizada pela doutrina e pelos três poderes (legislativo, judiciário e executivo) para se referir aos textos legais que regulam as grandes áreas do direito, como direito civil e direito penal, estabelecendo ainda diretrizes gerais para a área a serem seguidas nas demais normas. Temos assim, por exemplo, o Código Penal que estabelece as normas basilares do direito penal no país. O termo “código”, na realidade, se refere ao assunto da norma e o termo geralmente consta na ementa do texto, como pode ser observado na ementa do Código Civil: “Institui o Código Civil. ” (BRASIL, 2002). Em comum, todos os textos referidos como Códigos são normas federais, ou seja, aplicáveis compulsoriamente em todo o território nacional, mas sua espécie normativa varia. O código pode ser, por exemplo uma Lei ordinária, como é o caso do Código Civil, mas também um Decreto-Lei, como é o Código Penal.

Assim como os códigos também os estatutos não são espécies normativas, são o assunto e a designação dada aos textos legais que regulam grupos específicos da população. Como exemplo trazemos o Estatuto do idoso, que estabelece normas específicas voltadas para as necessidades das pessoas com 60 anos ou mais.

Decretos, por sua vez, se referem aos textos de natureza administrativa editados pelo Presidente da República enquanto chefe do Executivo. Sua finalidade é

regulamentar os textos legais e a organização da administração pública. Assim como os decretos, os chamados os Decretos não Numerados são editados pelo chefe do Executivo e tem natureza administrativa. Eles regulam assuntos da administração pública como a abertura de crédito, não tendo caráter normativo.

Diferentemente dos casos acima, os Decretos-Leis são uma espécie normativa. Entretanto, eles não estão previstos na CR/88, mas sim naquelas anteriores a esta. Se trata de textos legais editados pelo Presidente da República. Ao criar a constituição atual a Assembleia Constituinte optou por extinguir essa espécie normativa, mas manter alguns dos Decretos-leis vigentes na época, como por exemplo o Decreto-Lei nº 2.848, de 7 de dezembro de 1940, popularmente conhecido como Código Penal.

Já a categoria Mensagens de veto total engloba os vetos do Presidente a projetos de lei, juntamente com suas razões para tal. Geralmente projetos são vetados por serem considerados inconstitucionais ou contrários ao interesse público. As mensagens de veto não são normas e sim um ato administrativo.

A categoria Projetos de Lei, por sua vez, engloba todos os projetos de lei: em tramitação e tramitados. Enquanto projetos eles não têm caráter normativo: a partir do momento em que termina a tramitação do projeto e ele é promulgado, ele se torna um texto legal propriamente dito.

Temos ainda os Pareceres da AGU, que são os entendimentos da Advocacia Geral da União sobre questões legislativas e jurídicas. Não têm caráter normativo, mas vinculam a Administração Pública a seguir aquele entendimento ao lidar com a questão por ele abordada.

Por fim, as Propostas de Emenda à Constituição, assim como os projetos de lei, são projetos/propostas de emendas que, ao fim do processo legislativo, caso aprovadas, se tornam Emendas à Constituição, adquirindo status normativo.

Diante disso, selecionamos todos os textos legais disponibilizados no site que atendiam aos nossos critérios. No total foram obtidos 15.546 textos pertencentes às seguintes categorias do site: Constituição, Emendas à Constituição, Códigos, Leis complementares, Leis ordinárias, Medidas provisórias e Estatutos.

A seguir apresentamos uma tabela com a relação de textos por seção do corpus e seu número aproximado de palavras, ou seja, antes da limpeza, em janeiro de 2022:

Tabela 2 - Relação de textos por seção

Seção	Nº textos	Nº de palavras aprox.
Constituição	1	115264
Emendas à Constituição	114	91240
Códigos	17	1465098
Estatutos	18	200568
Leis complementares	192	Em compilação
Leis ordinárias	13491	Em compilação
Medidas provisórias	1713	Em compilação
TOTAL	15546	1872170

Fonte: autora (2022)

Ressaltamos que, por se tratar de um corpus ainda em compilação, não temos os dados de todas as categorias selecionadas, apenas das seções já compiladas. Entretanto, como pode ser notado pela tabela, já com os dados atuais temos um número de textos e palavras por categoria altamente variável. Temos desde a constituição com um único texto até as leis ordinárias com 13.491 textos. O mesmo se aplica ao número de palavras por texto: a Emenda Constitucional 66 (BRASIL, 2010), por exemplo tem 43 palavras, já a Consolidação das Leis do Trabalho conta com 167.415 palavras (BRASIL, 1943). Diante dessa discrepância nos números constatou-se que utilizar o número de palavras para selecionar os textos resultaria na exclusão de muitos textos e enviesaria a amostra, tornando essa opção inviável. Decidiu-se então selecionar os textos com base em sua frequência de uso, medida a partir do número de resultados obtidos na busca pelo texto no site especializado Jusbrasil, um site agregador de conteúdo relacionado ao direito brasileiro, sendo selecionados aqueles com 10.000 resultados ou mais.

Aqui temos a tabela com a relação de textos por seção após a seleção e limpeza dos textos:

Tabela 3 – Relação de textos por seção após a seleção

Seção	Nº textos	Nº de palavras	Nº de palavras do menor texto	Nº de palavras do maior texto	Média
Constituição	1	97082	97082	97082	97082
Emendas à Constituição	48	50241	43	12015	1046.69
Códigos	13	716411	12232	167415	55108.54
Estatutos	11	125356	2370	35241	11396.00
Leis complementares	Em compilação	Em compilação	Em compilação	Em compilação	Em compilação
Leis ordinárias	Em compilação	Em compilação	Em compilação	Em compilação	Em compilação
Medidas provisórias	Em compilação	Em compilação	Em compilação	Em compilação	Em compilação
TOTAL	73	989090			

Fonte: autora (2022)

Como o corpus ainda está em compilação no momento temos os dados de 4 das 7 seções, contabilizando 73 textos e aproximadamente 1 milhão de palavras. Aqui é possível notar que apesar dos esforços ainda assim há uma grande diferença na composição de cada um deles, uma questão na qual ainda estamos trabalhando e à qual esperamos responder, pelo menos parcialmente, com o presente trabalho.

3.1.2 Download dos textos e organização

Os textos legais foram extraídos do site do Planalto brasileiro e convertidos em ".txt" no software "Notepad++ 7.8.9" (HO, 2020) com codificação UTF-8 e salvos em diretórios no Desktop, correspondentes à seção ao qual pertencem. Por motivos de organização e didáticos optou-se por adotar a classificação utilizada pelo Portal da Legislação como divisão para nossos subcorpora. Logo, ficou estabelecido que o *LEX-BR-Ius* (FERRARI e MARQUES, em preparação) seria composto por 7 seções:

Constituição, Emendas à Constituição, Códigos, Leis complementares, Leis ordinárias, Medidas provisórias e Estatutos.

A cada texto foi atribuído um código de identificação (ID) para facilitar a organização e identificação do mesmo (para mais detalhes vide ANEXO A). O código é formado pela inicial do subcorpus em maiúsculo, número da lei, *underline,* dia, mês e ano da publicação da lei. Exemplo: à Lei nº 11.101 de 9 de fevereiro de 2005 foi atribuída a ID: LO11.101_9.02.2005. LO é a abreviação estabelecida para leis ordinárias, 11.101 é o número da lei e 9.02.2005 é sua data de publicação em numeral.

O texto extraído do site do planalto foi nomeado com o código de identificação por nós atribuído e *underline* original para indicar que aquele texto está tal qual o disponível no site. Exemplo: LO11.101_9.02.2005_original. Esse arquivo se tornou nossa cópia de segurança. Como pretende-se disponibilizar seja uma versão *raw text* que uma em ".xml", o arquivo nomeado como original foi aberto no "Notepad++ 7.8.9" (HO, 2020) e copiado. Nesse mesmo programa criaram-se outros dois novos arquivos e colou-se o texto legal anteriormente copiado. Esses novos arquivos foram nomeados respectivamente: ID_limpo e ID_xml. O primeiro será a versão raw text do texto e foi salvo em ".txt" e o segundo a versão em .xml, sendo salvo em ".xml". Exemplo: LO11.101_9.02.2005_limpo e LO11.101_9.02.2005_xml.

Em seguida, utilizou-se o programa "Notepad++ 7.8.9" (HO, 2020), para a criação de um cabeçalho em ".xml" seguindo a proposta de Hardie (2014) com os metadados de cada texto que também foram colocados em uma tabela Excel para fins de controle. O cabeçalho contém as seguintes informações: ID, Norma; Ementa; Tipo de norma; Assunto; Área do direito; Presidente em exercício; Data da promulgação; Data de publicação; Início da vigência; Alterações; Número de artigos; Número de palavras; Fonte; Data da extração; Subcorpus; Pesquisador; Revisor. Essas informações foram extraídas do próprio texto legal e da tabela fornecida no site do Planalto com as informações de cada texto legal. Os cabeçalhos foram intitulados com o código respectivo do texto *underline* cabeçalho e salvos em formato XML. Ex.: LO11.101_9.02.2005_cabecalho.

Aqui, a título de exemplo temos o cabeçalho do Código florestal (BRASIL, 2012):

Figura 2 – Cabeçalho do Código Florestal

```
<cabecalho>
<texto id = "C12.651_25.05.2012"/>
<norma v = "LEI Nº 12.651 DE 25 DE MAIO DE 2012"/>
<ementa v = "Dispõe sobre a proteção da vegetação nativa; altera as Leis nºs 6.938, de 31 de agosto de 1981, 9.393, de 19 de dezembro de 1996, e 11.428, de 22 de dezembro de 2006; revoga as Leis nºs 4.771, de 15 de setembro de 1965, e 7.754, de 14 de abril de 1989, e a Medida Provisória nº 2.166-67, de 24 de agosto de 2001; e dá outras providências."/>
<tipo v = "código"/>
<assunto v = "CODIGO, PROTEÇÃO, VEGETAÇÃO, FLORESTA, ECOLOGIA, AREA DE PROTEÇÃO AMBIENTAL, MEIO AMBIENTE, RESERVA ECOLOGICA, ZONA COSTEIRA, ZONA RURAL, ZONA URBANA, CORRELAÇÃO, ATIVIDADE AGROPECUARIA. CRITERIOS, OBRIGATORIEDADE, RECUPERAÇÃO, FAIXA, TERRAS, PROXIMIDADE, CURSO D'AGUA."/>
<area v = "CODIGO FLORESTAL, POLITICA DO MEIO AMBIENTE."/>
<presidente v = "Dilma Rousseff"/>
<promulgacao v = "25 de Maio de 2012"/>
<publicacao v = "28 de Maio de 2012"/>
<vigencia v = "Esta Lei entra em vigor na data de sua publicação."/>
<alteracao v = "MPV 571, DE 25/05/2012: ACRESCE O ART. 1º-A; ALTERA O ART. 3º, 4º, 5º, 6º, 10; ACRESCE O CAPÍTULO III-A DO USO ECOLOGICAMENTE SUSTENTÁVEL DOS APICUNS E SALGADOS - ART. 11-A; ALTERA OS ARTS. 14, 15, 17, 29, 35, 36, 41, 58; ACRESCE OS ARTS. 61-A, 61-B, 61-C E 78-A.; LEI 12.727, DE 17/10/2012: ACRESCE ARTS. 1º-A, 11-A, 61-A, 61-B, 61-C, 78-A E ALTERA ARTS. 3º, 4º, 5º, 6º, 10, 12, 14, 15, 16, 17, 18, 29, 35, 36, 41, 42, 58, 59, 66 E 83 (VETADO); MPV 724, DE 04/05/2016: ALTERA ART. 82-A; LEI 13.295, DE 14/06/2015: ALTERA ARTS. 29 E 78-A; LEI 13.335, DE 14/09/2016: ALTERA ART. 59; MPV 759, DE 22/12/2016: ALTERA ARTS. 64 E 65; LEI 13.465, DE 11/07/2017: ALTERA ARTS. 64, E 65.; MPV 867, DE 26/12/2018: ALTERA ART. 59; MPV 884, DE 14/06/2019: ALTERA ART. 29; LEI 13.887, DE 17/10/2019: ALTERA ARTS. 29 E 59.; LEI 14.285, DE 29/12/2021: ALTERA ARTS.  3º E 4º."/>
<artigos v = "84"/>
<palavras v = "21393"/>
<fonte v = "http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm"/>
<extracao v = "15/06/2022"/>
<subcorpus v = "códigos"/>
<pesquisador v = "carolina marques"/>
<revisor v="lucia ferrari"/>
</cabecalho>
```

Fonte: autora (2022)

Para melhor organização, dentro do diretório do subcorpus foi destinada uma pasta para cada texto legal, nomeada com seu código de identificação e dentro dela foram salvos o arquivo com o texto original em “.docx”, para preservar a formatação original, incluindo itálicos e tachados, e “.txt”, o arquivo da versão limpa (*raw text)*, o arquivo da versão com marcação XML e o arquivo do cabeçalho.

3.1.3 Limpeza

Os arquivos das versões *raw text* e XML passaram então por um tratamento semiautomático visando sua limpeza. Esse processo foi realizado no software “Notepad++ 7.8.9” (HO, 2020) com auxílio de expressões regulares. Foram apagados os elementos extratextuais (ex.: tabelas, anexos, data de promulgação, nome da lei, etc.) e linhas em branco.

O processo de limpeza da versão *raw text* seguiu as seguintes etapas:

1. Abriu-se o texto em ".txt" no programa em questão;
2. Estabeleceu-se UTF-8 como codificação;
3. Apagou-se:

a) Brasão das Armas Nacionais da República Federativa do Brasil;
b) CÂMARA DOS DEPUTADOS;
c) Centro de Documentação e Informação;
d) Presidência da República;
e) Secretaria-Geral;
f) Subchefia para Assuntos Jurídicos;
g) Hiperlinks;
h) Nome da lei;
i) Ementa da lei;
j) Frase de abertura (ex.: O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:);
k) Divisões (capítulo, seção, título, etc.);
l) Modificações (ex.: VETADO; "Caput" do artigo com redação dada pela Lei nº 14.112, de 24/12/2020, publicada na Edição Extra B do DOU de 24/12/2020, em vigor 30 dias após a publicação);
m) Data da promulgação (ex: Brasília, 10 de janeiro de 2002; 181 o da Independência e 114 o da República.);
n) Nome do presidente e envolvidos na lei;
o) Observação sobre a publicação no DOU;
p) Linhas em branco;

4. Procedeu-se à revisão do arquivo limpo.

Já o processo de limpeza da versão XML seguiu as seguintes etapas:

1. Abriu-se o texto em ".xml" no programa em questão;
2. Estabeleceu-se UTF-8 como codificação;
3. Apagou-se:

a) Brasão das Armas Nacionais da República Federativa do Brasil;

b) CÂMARA DOS DEPUTADOS;
c) Centro de Documentação e Informação;
d) Presidência da República;
e) Secretaria-Geral;
f) Subchefia para Assuntos Jurídicos;
g) Hiperlinks;
h) Linhas em branco;
4. Procedeu-se à revisão do arquivo limpo.

3.1. 4 Balanceamento

Após a limpeza foi realizado uma primeira tentativa de balanceamento das seções do corpus. O objetivo do balanceamento é criar um corpus equilibrado, ou seja, com número total de palavras por texto e por seções aproximados, visando possibilitar a comparabilidade, ao mesmo tempo em que se mantém a representatividade da linguagem, sendo fidedigno à realidade. Entretanto, como realçado nas seções anteriores, constatamos uma discrepância muito grande no número de textos e palavras por seção e por texto.

Trata-se de um corpus ainda em compilação e pelos motivos acima expostos decidiu-se suspender o balanceamento até a conclusão da presente pesquisa, cujos resultados, espera-se, nos ajudarão nessa etapa. No momento estamos repensando nossa arquitetura e critérios de seleção e questionando se a escolha por dividir o corpus por seção seguindo as categorias estabelecidas pelo Planalto é válida ou não. O presente trabalho, pretende entre outros verificar, através da análise multidimensional das seções já compiladas do corpus, se há realmente diferenças linguísticas e situacionais significativas que justifiquem a divisão por nós adotada ou indiquem uma melhor configuração para nosso corpus.

3.1. 5 Mark up Modest XML

A versão XML do corpus passou por uma marcação textual XML seguindo a proposta de Hardie (2014). Para o projeto reputamos que uma marcação em ".xml" completa (*full XML*) seguindo ou não as diretrizes do TEI ou do XCES seriam demasiadamente detalhadas, por isso optamos por seguir a proposta de Hardie (2014). Trata-se uma proposta de marcação XML mais básica que se destaca pela sua simplicidade e flexibilidade. Criada pensando nas necessidades de quem compila corpora, permite adicionar ou isolar informações no texto por meio de etiquetas que são delimitadas por parênteses angulares (< e >). Essas etiquetas podem ser tanto aquelas já convencionadas do sistema XML, "*de facto standard tags*", quanto as criadas pelo próprio compilador segundo suas necessidades, ou seja, existe a possibilidade de personalização das etiquetas. Dessa forma a marcação pode ser adaptada aos mais diversos tipos de corpora e pesquisas linguísticas, mas ao mesmo tempo ser de fácil aplicação, não sendo preciso ter conhecimentos de computação para realizá-la. (HARDIE, 2014).

No nosso corpus a marcação foi feita com etiquetas por nós elaboradas em português brasileiro seguindo a nomenclatura do direito brasileiro. As etiquetas foram criadas objetivando identificar os textos, armazenar seus metadados em forma de cabeçalho, separar as seções dos textos normativos e seus artigos. No total foram criadas 34 etiquetas e, para evitar problemas de codificação e processamento, todas as etiquetas são compostas apenas por letras minúsculas e não possuem acentos ou sinais gráficos. Foi também criado um guia para auxiliar os pesquisadores no momento da marcação (vide APÊNDICE A).

A escolha por criar etiquetas em português brasileiro ao invés de em inglês, que é o padrão na marcação de corpus, se justifica pelas particularidades da área jurídica como um todo e também terminológicas, o que poderia causar problemas no entendimento do significado das etiquetas. Isso porque, a linguagem jurídica é altamente complexa, sendo a terminologia utilizada em cada país muito específica. Essas diferenças e problemas de transposição de significado de uma língua para a outra se tornam ainda mais evidentes quando lidamos com línguas de países que utilizam sistemas jurídicos distintos.

É exatamente esse o caso do português brasileiro jurídico e do inglês jurídico. Enquanto o inglês jurídico é utilizado no sistema jurídico de *common law,* no Brasil o sistema jurídico vigente é o *civil law* (LENZA, 2020; SOUZA, 2020). Como a própria base do ordenamento jurídico é distinto muitas das normas, princípios e conceitos adotados no direito brasileiro não existem ou são aplicados de forma distinta nos países que adotam o *common law,* o que torna o estabelecimento de equivalentes tradutórios uma tarefa complexa (LENZA, 2020; SOUZA, 2020; CASTRO 2013). A depender do caso e do contexto se sugere traduzir literalmente, em outros, a recomendação é recorrer a adaptações, aproximando o significado da realidade do público-alvo, ou ainda, a inclusão de explicações e clarificações no corpo do texto ou em notas de rodapé para que as diferenças entre os sistemas e línguas não prejudiquem a transmissão da mensagem (CASTRO, 2013).

Um exemplo dessa diferença de significados é o próprio termo “lei” que no Brasil é usado tanto como termo “guarda-chuva” para os textos legais como parte da denominação de algumas espécies normativas. Independente da acepção, no Brasil, lei é um texto legai editado pelo poder legislativo. A tradução literal para o inglês seria “*law*”, entretanto, esse termo não pode ser considerado um equivalente tradutório de lei. Apesar de ser uma palavra da língua inglesa utilizada amplamente no inglês jurídico, nos países de *common law,* a palavra “*law”* se refere tanto à lei criada pelo legislativo que aos precedentes vinculantes (*case laws)* estabelecidos pelo judiciário. Logo, temos uma diferença importante de uso e significado entre os termos “lei” e “*law”*, sendo necessário recorrer a adaptações para indicar em inglês jurídico que em português leis são criadas exclusivamente pelo legislativo e, ao contrário indicar em português que *law* se refere tanto a texto editados pelo legislativo quanto pelo judiciário através do estabelecimento de precedentes. (CASTRO, 2013)

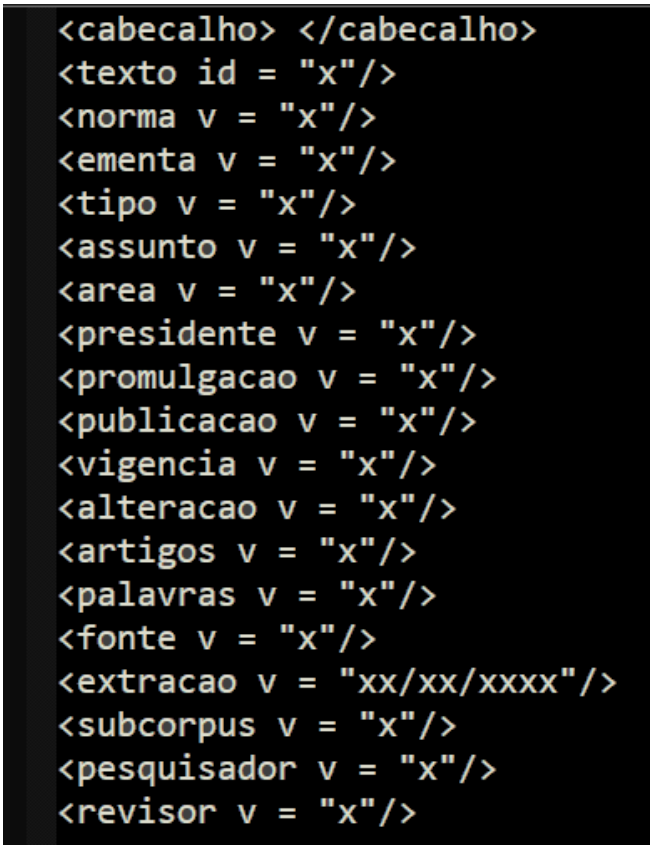
Outro exemplo que aponta uma diferença significativa entre esses sistemas é a palavra artigo. No Brasil, artigo é a unidade mínima do texto legal. É a menor divisão existente, sendo utilizada para introduzir e ordenar os assuntos abordados em determinada norma. Vem usado na sua forma abreviada seguida do número correspondente do trecho em ordem crescente e, na mesma linha, do trecho que introduziu, que geralmente ocupa poucas linhas. Ex.: “Art. 1º Toda pessoa é capaz de

direitos e deveres na ordem civil.” (BRASIL, 2002). Em inglês jurídico existe o termo “*article”*, mas também nesse caso seu significado não corresponde aquele brasileiro. “*Article”* é usado para subdividir partes do texto legal, muitas vezes tratando diversos temas e a parte delimitada de grandes extensões, podendo abranger várias páginas do texto legal. Dentro dos textos delimitados por esse termo encontramos ainda outras subdivisões menores. Por fim, na estrutura do texto legal essa palavra consta na íntegra, em letras maiúsculas e centralizada, estando separada do texto. (CASTRO, 2013)

A marcação foi realizada no “Notepad++ 7.8.9” (HO, 2020) com codificação UTF-8 e extensão XML.

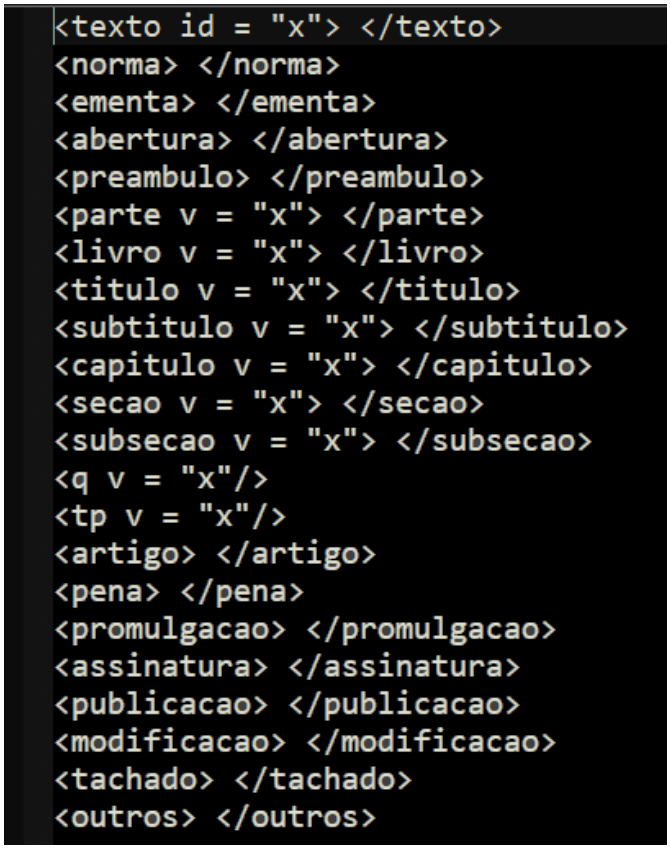
A seguir reproduzimos as etiquetas do cabeçalho e do texto:

Figura 3 – *Tagset* cabeçalho



Fonte: autora (2022)

Figura 4 – *Tagset* texto



Fonte: autora (2022)

Aqui temos um exemplo do texto Código Florestal (BRASIL, 2012) marcado com as etiquetas textuais.

Figura 5 – Código Florestal com marcação XML

```
<texto id = "C12.651_25.05.2012">
<norma> LEI Nº 12.651, DE 25 DE MAIO DE 2012.
</norma>
<ementa> Dispõe sobre a proteção da vegetação nativa; altera as Leis nºs 6.938, de 31 de agosto de 1981, 9.393, de 19 de dezembro de 1996, e 11.428, de 22 de dezembro de 2006; revoga as Leis nºs 4.771, de 15 de setembro de 1965, e 7.754, de 14 de abril de 1989, e a Medida Provisória nº 2.166-67, de 24 de agosto de 2001; e dá outras providências.
</ementa>
<abertura> A PRESIDENTA DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:
</abertura>
<capitulo v = "CAPÍTULO I DISPOSIÇÕES GERAIS">
<artigo> Art. 1º
<modificacao> (VETADO)
</modificacao>
</artigo>
<tachado> <artigo> Art. 1º-A. Esta Lei estabelece normas gerais com o fundamento central da proteção e uso sustentável das florestas e demais formas de vegetação nativa em harmonia com a promoção do desenvolvimento econômico, atendidos os seguintes princípios:
<modificacao> (Incluído pela Medida Provisória nº 571, de 2012)
</modificacao>
```

Fonte: autora (2022)

3.1.6 Anotação

A anotação do nosso corpus foi feita por Carlos Kauffmann a quem agradecemos. Para tanto, os textos limpos em ".txt" passaram por um processo de tokenização, anotação *Part of Speech* (POS) e lematização no *software* "PALAVRAS" (BICK, 2000; 2014), o mesmo utilizado na anotação do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014). Trata-se de um *parser* para a língua portuguesa desenvolvido por Eckhard Bick com altas taxas de confiabilidade e que vem sendo continuamente aprimorado ao longo dos últimos 20 anos. Segundo Bick (2014):

> O anotador PALAVRAS é um sistema de análise gramatical modular, apoiado em léxico e baseado em regras CG, que atribui etiquetas morfossintáticas, estruturas de dependência e outras informações gramaticais a todas as formas de palavras em português, permitindo a fácil transformação de sua saída em diferentes

> tradições linguísticas (por exemplo, árvores constituintes) ou formatos técnicos (por exemplo, Tiger xml). (BICK, 2014, p. 298, tradução nossa)[49]

Trata-se assim, de um software de última geração amplamente utilizado na anotação de corpora em português que anota automaticamente mais de 300 traços linguísticos (sintáticos, morfológicos, semânticos, etc.). Para tanto utiliza um sistema baseado em regras, contando com mais de 6.000 regras gramaticais aplicável a vários tipos de corpora (ex.: escrito, orais, histórico, etc.).

Feita a anotação com o PALAVRAS (BICK, 2000; 2014), nosso corpus foi processado pelo "PALAVRAS *Tag Count*" (BERBER SARDINHA, 2013) visando a contagem das características linguísticas anteriormente anotadas. Trata-se de um pós processador de textos em português brasileiro desenvolvido por Berber Sardinha e sua equipe para contar as etiquetas das 190 características linguísticas identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) como relevantes para o estudo da variação do português brasileiro (vide ANEXO A) a partir da Análise Multidimensional do Corpus Brasileiro de Variação de Registro (CBVR) (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014). Desde então, esse processador vem passando por atualizações e correções visando uma melhor performance e pode ser aplicado nos mais diversos corpora de português, nos sendo gentilmente disponibilizado pelos seus criadores para a realização desta pesquisa.

[49] *The PALAVRAS parser is a modular, lexicon-supported and CG rule-based system of grammatical analysis, that assigns morphosyntactic tags, dependency structures and other grammatical readings to all word-forms in running Portuguese text, allowing for easy, tag-based transformation of its output into different linguistic traditions (e.g., constituent trees) or technical formats (e.g., Tiger xml).*

Aqui temos um trecho do Código Florestal (BRASIL,2012) anotado e processado.

Figura 6 - Código Florestal anotado

```
<texto id = "C12.651.25.05.2012">
  <word id="1" form="Art." base="art." postag="N" morf="M S" extra="*" head="0"
  deprel="NPHR"/>
  <word id="2" form="1º" base="1º" postag="ADJ" morf="M S" extra="NUM-ord jh NER:prednum
  np-close" head="0" deprel="N<"/>
  <word id="3" form="Art." base="Art." postag="PROP" morf="M S" extra="official * np-long"
  head="0" deprel="N<"/>
  <word id="4" form="1º-A" base="1º-A" postag="NUM" morf="M/F P" extra="card np-long"
  head="0" deprel="N<"/>
  <word id="5" form="." base="." postag="pu" morf="--" extra="--" head="0" deprel="PU"/>
  <word id="6" form="Esta" base="este" postag="DET" morf="F S" extra="prop2 * dem" head="0"
  deprel=">N"/>
  <word id="7" form="Lei" base="lei" postag="N" morf="F S" extra="prop prop2 * conv" head="0"
  deprel="SUBJ>"/>
  <word id="8" form="estabelece" base="estabelecer" postag="V" morf="PR 3S IND VFIN"
  extra="vH fmc mv" head="0" deprel="FS-STA"/>
```

Fonte: autora (2022)

3.2 Análise Multidimensional aditiva do *LEX-BR-Ius*

Na presente pesquisa buscou-se inicialmente realizar uma Análise Multidimensional completa, conforme apresentada em 2.4.1.1, dos textos legais federais brasileiros em vigência, de forma a verificar se eles são um registro, se há diferença linguísticas significativas entre os tipos legais que permitam classificá-los como registros diversos, e descrever sua variação em dimensões. Entretanto, devido ao número total de textos do nosso corpus (73) e à discrepância entre o número de textos e de palavras por texto e por seções, a realização da análise multidimensional completa se tornou inviável. Isso porque, para a realização desse tipo de análise é fundamental ter um corpus balanceado e equilibrado com um grande número de textos (ocorrências), sendo recomendado que esse número supere em ao menos 5 vezes o de variáveis (BREZINA, 2018).

Pretendíamos reproduzir em nosso estudo os passos de Berber Sardinha, Kauffmann e Acunzo (2014) utilizando as variáveis por eles consideradas relevantes para o estudo das dimensões de variação do português brasileiro na sua pesquisa. Estas totalizam 190 variáveis (vide ANEXO A), logo, o número desejável de textos seria por volta de 900 textos. A diferença entre o número ideal de textos e o número de textos do nosso corpus é muito grande: quase 14 vezes maior que o de textos do nosso corpus.

Outro fator de peso que impossibilita a aplicação dessa abordagem ao nosso corpus tal como se encontra é a falta de equilíbrio entre e dentro de suas seções. Para que a comparação seja possível e fidedigna à realidade é necessário que haja um número aproximado de textos e também de palavras por subcorpus e de palavras por texto dentro de cada subcorpus. (BIBER, 1988; BIBER e CONRAD, 2009; BREZINA, 2018)

Para contornar esse problema pensou-se em realizar um recorte no corpus. Várias possibilidades foram levantadas. Primeiramente, pensou-se em aumentar o número de textos da seção Emendas à Constituição, já que é a seção com menor número de palavras, de forma a resolver a diferença gritante entre o número de palavras total das seções. Entretanto, essa seção, já tem uma quantidade muito superior de textos (48) quando comparada com as demais (1, 13 e 11 textos). Além disso, o número de palavras por texto é, em média, 1047 palavras, muito inferior ao das demais seções. Para a realização da AMD completa, não só o número de palavras por seção, mas também a quantidade de textos por seção e por texto devem ser semelhantes para que o corpus seja equilibrado. Por fim, após aprovada, a maior parte do texto da emenda, é incorporado à Constituição, passando a fazer parte daquele texto. Logo, grande parte do texto de cada emenda já está contido na seção Constituição, gerando uma certa sobreposição entre as seções Emendas à Constituição e Constituição. Por tanto, essa opção foi descartada.

Pensou-se então em excluir a seção Emendas à Constituição da análise e analisar só a Constituição, os Códigos e Estatutos, mas ainda assim teríamos uma diferença significativa na composição das seções. Decidiu-se então, além de excluir essa seção, estabelecer um número fixo de textos e palavras por seção restante e retirar trechos de cada texto que as compõe. Apesar de obtermos êxito na execução dessa ideia, atingindo um corpus equilibrado, essa composição vai contra uma das decisões metodológicas que tomamos no início da compilação do corpus, qual seja, a de manter os textos na íntegra ao invés de utilizar trechos, já que selecionar, ainda que arbitrariamente apenas uma parte do texto pode significar uma perda considerável de dados. Logo, também esta opção foi descartada.

Para manter os textos na integra e possibilitar a comparação entre os subcorpora teríamos que ter um número aproximado de textos e palavras por texto. Pensamos então

em realizar um recorte, excluindo a seção Emendas à Constituição pelos motivos expostos acima e incluir apenas os textos de 10 mil a 40 mil palavras das seções Códigos e Estatutos. Dessa forma obtemos 5 textos em cada seção, com um total de palavras próximo a 100 mil palavras em cada texto, o que possibilita compará-las com a Constituição, que também tem mais ou menos esse número de palavras. Esse novo recorte possibilitou alcançar o equilíbrio do corpus, entretanto temos ainda outro problema: a quantidade de textos. Dessa forma alcançamos o equilíbrio, mas o problema da quantidade de textos foi acentuado. Nesse recorte temos um total de 11 textos (1 Constituição, 5 Códigos e 5 Estatutos), com cerca 100 mil palavras por seção, mas o ideal para a realização da AMD completa é de cerca 900 textos. Logo, apesar do nosso recorte ter um número de palavras e textos semelhante por serem poucos os textos que o compõe não seria possível caracterizar o registro. Por isso descartamos também essa ideia.

Por fim pensou-se em manter a separação em seções, considerando como texto cada um dos artigos dos textos legais ao invés do texto legal na íntegra. Para tanto, foi feita a segmentação de cada texto legal em ".txt" com configuração UTF-8 por artigo, utilizando um script. Foi então realizada a revisão manual desses arquivos, uma vez que, notou-se alguns problemas na segmentação[50]. Também com essa nova configuração tem-se uma discrepância nos números por seção e, por causa do pequeno número de palavras por texto (artigo), constatou-se que não seria possível caracterizar adequadamente a variação, já que são poucas as ocorrências das variáveis linguísticas por texto. Devido ao pouco tempo disponível para a conclusão da pesquisa, e, apesar de promissora, essa proposta também foi descartada.

Diante do exposto, optou-se por alterar um pouco nossos objetivos originais, realizando uma AMD aditiva a partir das dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) de forma a comparar nosso corpus e suas seções

[50] Quais sejam: alguns artigos não foram segmentados corretamente, tendo uma parte ficado em um arquivo e a outra em outro arquivo. Observou-se também que alguns arquivos continham mais de um artigo. Também se notaram erros na segmentação dos artigos com título, característica recorrente no Código Penal. Nestes o título do artigo foi considerado parte do artigo anterior, sendo incluído no arquivo deste e o artigo ao qual dá título foi isolado em outro arquivo. Por fim, alguns artigos que trazem modificações na redação de artigos de outras leis foram segmentados de forma que o enunciado do artigo foi salvo em um arquivo e a nova redação em outro arquivo.

globalmente com registros encontrados no CBVR (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014). Para tanto, nos inspiramos no estudo de Berber Sardinha (2013) que realizou uma análise aditiva dos registros da internet às dimensões identificadas por Biber (1988).

Conforme exposto na seção 2.4.1.2 a AMD aditiva proporciona o estudo da variação dos registros a partir do mapeamento dos registros estudados nas dimensões identificadas em estudo de AMD completo anteriormente realizado. A AMD aditiva enriquece a AMD completa à qual se adicionam os novos registros, sendo também por esta enriquecida pela comparação dos seus registros com aqueles do estudo original. Entre as suas vantagens está o menor tempo de realização e menor dificuldade estatística, não sendo necessária a realização da análise fatorial e, portanto, não é necessário haver um rigor tão grande na arquitetura do corpus, em especial no que diz respeito ao equilíbrio e número de textos. Isso permite aplicar essa abordagem ao nosso corpus em sua atual configuração apesar das suas distorções, o que não seria possível na AMD completa. Além disso, essa abordagem se destaca por sua flexibilidade, que permite ao pesquisador escolher quais dimensões serão utilizadas e obtenção de resultados confiáveis e detalhados que permitem traçar um panorama dos registros estudados. (BERBER SARDINHA, 2013; BERBER SARDINHA, *et al.*, 2019)

Tomadas essas decisões metodológicas, deu-se início à AMD aditiva. Para a sua realização seguimos os passos descritos na seção 2.4.1.2, reproduzidos a seguir, e utilizamos o software Microsoft Excel 2016[51].

(1) Estudo prévio: foi realizado um estudo prévio na literatura sobre linguagem jurídica em geral e também sobre aquela utilizada nos textos legais em inglês, italiano e português e sobre uma amostra de textos legais federais brasileiros para investigar suas características linguísticas e situacionais (vide seção 2.2). Após examinar as características situacionais de nossos textos, concluímos que essas se distinguem a depender da espécie normativa e também por texto em relação ao seu contexto de comunicação e assunto;

[51]Devido à dimensão das planilhas produzidas ao longo da análise não é possível disponibilizá-las em anexo. Para ter acesso a essas planilhas favor acessar o link: https://docs.google.com/spreadsheets/d/1TFRAcRokwgbF2Gu9BfzOYNZsV2amH9wE/edit?usp=share_link&ouid=113578582801313013593&rtpof=true&sd=true

(2) Escolha do estudo que servirá como base da AMD aditiva: escolheu-se como estudo-base, ao qual adicionamos nosso corpus, o estudo de Berber Sardinha, Kauffmann e Acunzo (2014) que determinou as dimensões do português brasileiro a partir da Análise Multidimensional do Corpus Brasileiro de Variação de Registro (CBVR) (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014), descrito anteriormente (vide seção 2.4.5). Uma vez que pretendemos trabalhar com a língua portuguesa brasileira, este estudo se mostrou ideal, já que se trata da AMD completa mais abrangente realizada sobre essa variedade;

(3) Escolha das dimensões a serem analisadas: optou-se por aplicar nosso corpus a todas as dimensões identificadas pelo estudo, quais sejam: (*1) Oral versus literate discourse (discurso letrado), (2) Argumentation, (3) Involved versus informational production, (4) Directive discourse, (5) Future versus past time orientation e (6) Reported discourse,* de forma a obter um maior grau de detalhamento e identificar mais diferenças entre os registros;

(4) Identificação das características linguísticas das dimensões escolhidas: selecionamos apenas as características linguísticas presentes nas dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D) disponibilizadas no estudo separadas por dimensão e por polo entre as páginas 44 e 56. No total são 93 características, sendo 49 na primeira dimensão, 19 na segunda, 16 na terceira, 7 na quarta, 15 na quinta e 13 na sexta (vide ANEXO D). Como algumas variáveis são carregadas em mais de uma dimensão, para a análise aditiva consideramos que elas só compõem a dimensão na qual tem o maior valor numérico.

Por exemplo, a variável *Pronouns: Possessive* (pronomes:possesivos) ocorre seja na dimensão 1 que na 6 (vide ANEXO D), sendo que na primeira pontua .605 e na segunda .424. Isso significa que os pronomes possessivos têm uma frequência 0.065 desvios padrão acima da média na primeira dimensão e 0.424 desvios padrão na sexta dimensão. Em outras palavras, essa variável apresenta alta frequência em ambas as dimensões, mas sua frequência na dimensão 1 supera aquela da 6. Logo, ao analisar as dimensões, considera-se que essa variável pertença à dimensão 1, onde apresenta um maior valor, e ignora-se sua presença na dimensão 6.

(5) Escolha e preparação do corpus: como dito anteriormente, foi compilado o *LEX-BR-Ius* (FERRARI, MARQUES, em compilação). O corpus foi etiquetado morfossintacticamente[52] com o mesmo etiquetador do estudo-base, qual seja o PALAVRAS (BICK, 2001, 2014) conforme exposto na seção anterior. Além disso, decidiu-se por utilizar apenas as seções Constituição, Códigos e Estatutos, desconsiderando a seção Emendas à Constituição, já que grande parte do seu texto consta já na Constituição, o que poderia gerar certo viés na pesquisa.

A seguir apresentamos a configuração do corpus ao qual aplicamos as dimensões:

Tabela 4 – Seções utilizadas para a AMD aditiva

Seção	Nº textos	Nº de palavras
Constituição	1	97082
Códigos	13	716411
Estatutos	11	125356
TOTAL	**25**	**938849**

Fonte: autora (2022)

(6) Contagem das variáveis linguísticas no corpus[53]: utilizou-se o "PALAVRAS Tag count" (BERBER SARDINHA, 2013) desenvolvido por Berber Sardinha e sua equipe para a realização da contagem das variáveis no nosso corpus. Trata-se do mesmo pós processador utilizado na realização da AMD completa na qual estamos nos baseamos. Após a contagem foi extraída uma planilha em ".csv" com as ocorrências de cada variável em cada texto do corpus. Nesta planilha selecionamos apenas os dados que constavam das dimensões identificadas no estudo base.

É importante ressaltar que das 93 variáveis do estudo base (vide ANEXO B) verificamos que seis delas não foram contabilizadas no nosso corpus, quais sejam: *QU questions* (perguntas QU – dimensão 1), *Average word length* (média do tamanho da palavra - dimensão 1), *Tag questions* (perguntas anotadas - dimensão 3), *Questions: Yes or No question* (perguntas de sim ou não - dimensão 3), *Type-token ratio* (relação type-

[52] A etiquetação foi feita por Carlos Kauffmann, a quem agradecemos.
[53] Também a contagem foi realizada por Carlos Kauffmann com auxílio do pós processador em questão. Novamente gostaríamos de agradecê-lo pelo suporte.

token - dimensão 3) *e Modals*: Haver que/haver de (modais - dimensão 6). Contatamos os autores, mas eles não souberam explicar por que essas variáveis não foram contadas no nosso corpus. Hipotizamos que, com as sucessivas atualizações do programa, essas variáveis podem ter sido excluídas, ou ainda que houve algum erro durante a execução do programa em nosso corpus o que fez com que essas variáveis não fossem geradas.

(7) Normalização das ocorrências das variáveis linguísticas: normalizamos as ocorrências das variáveis selecionadas em cada texto por mil palavras e as separamos por dimensão, visando possibilitar a comparação dos dados. Para isso utilizamos a seguinte fórmula:

$$Freq.normalizada = \left(\frac{freq.da\ variável\ no\ texto}{\text{número de palavras do texto}}\right) \text{x } 1000$$

Para fins exemplificativos reproduzimos a seguir uma parte da planilha na qual normalizamos as ocorrências das variáveis da dimensão 6: *Reported discourse* identificada por Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D) no nosso corpus:

Tabela 5 - Ocorrência normalizada da variável cjfinal por texto

ID	Seção	cjfinal		
		original	nº palavras texto	norm.
C10.406_10.01	cod	3	101629	0.029519133
C12.651_25.05	cod	0	21386	0
C13.105_16.03	cod	8	91469	0.087461326

Fonte: autora (2022)

Na tabela acima, realizada no Microsoft Excel (2016), normalizamos a ocorrência da variável cjfinal: *conjunction subordinating final* (conjunção subordinativa final/de finilidade) que compõe a dimensão 6 de Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D). Nela temos, em ordem, a ID atribuída ao texto, a seção do corpus à qual ele pertence, a frequência real da variável no texto, o número de palavras do texto e o valor normalizado da variável.

Para calcular o valor normalizado da variável cjfinal no texto C10.406_10.01, por exemplo, aplicamos a fórmula acima apresentada dividindo o valor de ocorrências reais da variável (3) pelo número de palavras do texto (101.629) e multiplicamos o resultado por mil, obtendo sua frequência normalizada: 0,029519133:

$$Freq.normalizada\ cjfinal = \left(\frac{3}{101629}\right) \text{ x } 1000$$

$$Freq.normalizada\ cjfinal = (0.000029521933) \text{ x } 1000$$

$$Freq.normalizada\ cjfinal = \mathbf{0,029519133}$$

O mesmo procedimento foi aplicado a todos os textos do corpus.

(8) Cálculo dos Z-escores: visando nivelar a frequência das variáveis para que elas tenham um mesmo peso no cálculo dos escores de dimensão calculamos o Z-escore de cada variável. Esse passo é fundamental para impedir que os escores de dimensão sejam influenciados pela alta ou baixa frequência de determinada variável. Esse cálculo foi feito utilizando a frequência normalizada da variável calculada na etapa precedente e a frequência média e o desvio padrão da variável no estudo base disponibilizadas pelos autores no anexo do estudo entre as páginas 69 e 74 (vide ANEXO C).

Ressaltamos que não conseguimos individualizar na estatística descritiva do estudo original (vide ANEXO C) os dados correspondentes à variável: *Que clause controlled by adjective (stance)* (oração “que” controlada por adjetivo (posicionamento)) identificada na dimensão 2 desse estudo (vide ANEXO D). Isso porque na tabela disponibilizada pelos autores (vide ANEXO C) essa variável se repete duas vezes, nas posições 113 e 114 dessa tabela com valores distintos: “113 *Que clause cntrld. by adjective* .00 5.38 .45 .76; 114 *Que clause cntrld. by adjective* .00 7.03 1.24 1.28” (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014, p. 72)

Contatamos os autores, que também não conseguiram individualizar qual das duas opções (113 e 114) corresponderia aos dados da variável em questão. Logo optou-se por excluir essa variável da nossa análise.

O cálculo dos Z-escores foi feito em uma planilha do Microsoft Excel 2016 utilizando a fórmula:

$$Z\ escore = \frac{freq.normalizada\ da\ variável - \ freq\ média\ da\ \text{variável}\ do\ estudo\ original}{desvio\ padrão\ da\ variável\ do\ estudo\ original}$$

Para fins exemplificativos reproduzimos uma parte da planilha na qual calculamos o Z-escore da variável *Subordinating (final) clause (cjfinal*) da dimensão 6 (vide ANEXO D):

Tabela 6 – Z-escore da variável cjfinal da dimensão 6

ID	Seção	cjfinal			
		norm	média CBVR	DP CBVR	Z-escore
C10.406_10.01	cod	0.029519	0.01	0.08	0.243989
C12.651_25.05	cod	0	0.01	0.08	-0.125
C13.105_16.03	cod	0.087461	0.01	0.08	0.968267

Fonte: autora (2022)

Na tabela acima temos, em ordem, o código atribuído ao texto, a seção do corpus à qual ele pertence, a frequência normalizada da variável cjfinal em cada texto, a frequência média dessa variável no estudo original (0.01), seu desvio padrão (0.08) no estudo original (0,08) e o Z-escore. Para calcular o Z-escore dessa variável aplicamos a fórmula acima apresentada subtraindo sua frequência normalizada da frequência média do estudo original e dividimos pelo seu desvio padrão no estudo original.

No texto C10.406_10.01, por exemplo, o cálculo foi realizado da seguinte forma:

$$Z\ escore\ cjfinal = \frac{0.029519 - \ 0.01}{0.08}$$

$$Z\ escore\ cjfinal = \frac{0.019519}{0.08}$$

$$Z\ escore\ cjfinal = \mathbf{0.243989}$$

Isso significa que a variável cjfinal no texto C10.406_10.01 é aprox. 0,24 desvios padrão maior que a média do estudo original.

(9) Cálculo dos escores de dimensão de cada texto: calculamos os escores de dimensão de cada texto do nosso corpus visando obter sua pontuação na escala de cada dimensão identificada por Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D). Para tanto, somamos os Z-escores, calculados na etapa precedente, de todas as variáveis localizadas no polo positivo da dimensão no estudo original (vide ANEXO D) e, nas dimensões com polo negativo, subtraímos desse resultado a soma dos Z-escores das variáveis localizadas no polo negativo da mesma dimensão no estudo original (vide ANEXO D).

Para as dimensões que possuem apenas um polo utilizamos a fórmula:

$$Escore\ de\ dim. = \sum dos\ z\ escores$$

A seguir reproduzimos uma parte da planilha na qual calculamos o escore de dimensão dos nossos textos na dimensão 6: *Reported discourse* (vide ANEXO D):

Tabela 7 – Escores de dimensão por texto da dimensão 6

ID	Seção	cjfinal	objprnrare	prn3obl	vb2	vbpubl	Escore de dimensão
C10.406_10.01	cod	0.243989	1.734199	1.892805	-0.27233	0.167703	3.766366636
C12.651_25.05	cod	-0.125	-0.44562	-0.82386	-0.27607	-0.40337	-2.073923734
C13.105_16.03	cod	0.968267	0.747552	0.964749	-0.19125	1.092019	3.581340568

Fonte: autora (2022)

A dimensão 6 de Berber Sardinha, Kauffmann e Acunzo (2014) é composta pelas variáveis: cjfinal (*Subordinating (final) clause:* oração subordinativa final/de finalidade), objprnrare (*rare object pronouns:* pronome objeto raro), prn3obl *(pronoun third person oblique:* pronome oblíquo de terceira pessoa*),* vb2 *(verb second person*: verbo na segunda pessoa*)* e vbpub *(verb public:* verbo público*),* todas no polo positivo da dimensão (vide ANEXO D). Para calcular como cada texto do nosso corpus pontua nessa dimensão

recorremos ao cálculo do escore de dimensão. Para isso, utilizamos o z-escore de cada variável em cada texto, calculados na etapa anterior, e reproduzidos na coluna nomeada com o nome da variável na tabela acima. Como nessa dimensão, no estudo original, todas as variáveis estão no polo positivo, para calcular o escore de dimensão apenas somamos os valores de todos os z-escores de determinado texto.

Por exemplo, no texto C10.406_10.01, na dimensão 6 o cálculo realizado foi:

$$Escore\ de\ dim.\mathrm{C10.406\_10.01} = (\mathrm{cjfinal} + \mathrm{objprnrare} + \mathrm{prn3obl} + \mathrm{vbpubl} + \mathrm{vb2})$$

$$Escore\ de\ dim.\mathrm{C10.406\_10.01} = (0.243989 + 1.734199 + 1.892805 - 0.27233 + 0.167703)$$

$$Escore\ de\ dim.\mathrm{C10.406\_10.01} = \mathbf{3.766366636}$$

Já para as dimensões com dois polos a fórmula aplicada foi:

$$Escore\ de\ dim. = (\sum dos\ z\ escores\ das\ variáveis\ positivas\ ) - (\sum dos\ z\ escores\ das\ variáveis\ negativas)$$

Esse é o caso da dimensão 5: nela temos tanto o polo positivo quanto o negativo (vide anexo D). No polo positivo do estudo original temos as variáveis: *Verbs: Future subjunctive mood (vbsubfut)* (verbos: futuro do subjuntivo)*, Conjunctions: Coordinating (ou) (cjou)* (conjunções: coordenada ou)*, Verbs: Future present tense (vbfutpres)* (verbos: futuro do presente), *Modals: Dever (mddever)* (modais: dever), *Modals: Poder (mdpoder)* (modais: poder), *Subordinating (conditional) clause (cjcond)* (oração (condicional subordinativa)*, Adverbs: Likelihood (advlikl)* (advérbios: probabilidade), *Conjunctions: Coordinating (phrasal) (cjcoorphr)* (conjunção coordenativa frasal). Já no polo negativo do estudo original temos: *Nouns: Place (nplac)* (substantivos: lugar), *Verbs: Past subjunctive mood (vbsubpast)* (verbos: passado subjuntivo), *Adjectives: Affiliative (adjaffi)* (adjetivos: afiliativo), *Verbs: Imperfect (vbimpf)* (verbos: imperfeito), *Verbs: Past indicative tense (vbpast)* (verbos: passado indicativo).

Nesse caso, o cálculo do escore de dimensão foi:

$$Escore\ de\ dim. = (vbsubfut + cjou + vbfutpres + mddever + mdpoder + cjcond + advlikl + cjcoorphr) - (nplac + vbsubpast + adjaffi + vbimpf + vbpast)$$

Para fins exemplificativos reproduzimos uma parte da planilha na qual calculamos o escore de dimensão dos nossos textos na dimensão 5: *Future versus past time orientation* (vide ANEXO D):

Tabela 8 – Escores de dimensão por texto da dimensão 5

ID	seção	adjaffi	advlikl	cjcond	cjcoorphr	cjou	mddever	mdpoder	nplac	vbfutpres	vbimpf	vbpast	vbsubfut	vbsubpast	Escore de dimensão
C10.406_10.01	cod	-0.454	0.621	2.729	4.568	3.941	0.999	3.547	-0.379	1.975	-0.595	-0.945	4.603	-0.445	**25.809**
C12.651_25.05	cod	-0.041	0.635	2.729	10.418	2.886	2.557	0.717	2.053	1.151	-0.676	-1.096	-0.091	-0.610	**21.377**
C13.105_16.03	cod	-0.323	1.976	2.729	3.9217	3.981	1.619	2.531	-0.464	3.379	-0.670	-0.999	4.003	-0.591	**27.192**

Fonte: autora (2022)

Para calcular o escore de dimensão do texto C10.406_10.01, na dimensão 5 o cálculo realizado foi:

$$Escore\ de\ dim. = (vbsubfut + cjou + vbfutpres + mddever + mdpoder + cjcond + advlikl + cjcoorphr) - (nplac + vbsubpast + adjaffi + vbimpf + vbpast)$$

$$Escore\ de\ dim. = (4.603871 + 3.941662 + 1.975289 + 0.999334 + 3.54774 + 2.729997 + 0.621847 + 4.568742) - (-0.37942 + (-0.94583) + (-0.45442) + (-0.59567) + (-0.94583))$$

$$Escore\ de\ dim. = 22.98848 - (-2.82114)$$

$$Escore\ de\ dim. = \mathbf{25.80962587}$$

(10) Cálculo da média dos escores de dimensão dos novos registros: no nosso caso temos apenas um registro: os textos legais. Calcular a média dos escores de dimensão desse registro permite aplicar as dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014) ao nosso corpus possibilitando verificar como os textos legais se encaixam nas dimensões do estudo original.

É importante ressaltar que optamos por calcular, além da média dos escores de dimensão do registro também a média dos escores de dimensão por seção do corpus de forma identificar possíveis diferenças entre as seções. Como nas demais etapas, nos valemos do Microsoft Excel 2016 para realizar os cálculos.

Para o cálculo do escore dimensão para o registro textos legais utilizamos a seguinte fórmula:

$$Média\ do\ escore\ de\ dim. \ = \frac{\sum dos\ escores\ de\ dimensão\ de\ cada\ texto}{número\ de\ textos\ do\ corpus}$$

Para o cálculo do escore dimensão por seção do corpus utilizamos a seguinte fórmula:

$$Média\ do\ escore\ de\ dim.\ por\ seção = \frac{\sum dos\ escores\ de\ dimensão\ de\ cada\ seção}{número\ de\ textos\ por\ seção}$$

A seguir reproduzimos a planilha na qual calculamos as médias de dimensão da seção Códigos na dimensão 6 Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D).

Tabela 9 – Média de dimensão da seção Códigos na dimensão 6

ID	Seção	Escore de dimensão
C10.406_10.01	cod	3.766366636
C12.651_25.05	cod	-2.073923734
C13.105_16.03	cod	3.581340568
C2.848_7.12	cod	1.418173419
C227_28.02	cod	-0.595714209
C24.643_10.07	cod	0.889157949
C3.689_03.10	cod	3.591258323
C4.737_15.07	cod	0.918833507
C5.172_25.10	cod	1.212190843
C5.452_01.05	cod	0.053959643
C556_25.06	cod	2.521037069

C8.078_11.09	cod	0.446398329
C9.503_23.09	cod	-1.188670815
	Total	14.54040753
	Média	**1.118492887**

Fonte: autora (2022)

Na tabela temos a relação de todos os textos da seção Códigos do nosso corpus e seus respectivos escores de dimensão calculados na etapa anterior. Para calcular a média de dimensão somamos os escores de dimensão de todos os textos e em seguida dividimos o resultado pelo número de textos da seção:

$$Média\ do\ escore\ de\ dim.\ da\ seção\ Códigos = 3.766366636 + (-2.073923734) + 3.581340568 + 1.418173419 + (-0.595714209) + 0.889157949 + 3.591258323 + 0.918833507 + 1.212190843 + 0.053959643 + 2.521037069 + 0.446398329 + (-1.188670815)/13$$

$$Média\ do\ escore\ de\ dim.\ da\ seção\ Códigos = \frac{14.54040753}{13}$$

$$Média\ do\ escore\ de\ dim.\ da\ seção\ Códigos = \mathbf{1.118492887}$$

(11) Comparação dos novos registros com aqueles do estudo-base: inseriu-se as médias de dimensão do nosso corpus como um todo e por seção na planilha com as médias de dimensão do estudo original e gerou-se um gráfico de barras para melhor visualização dos dados. Em seguida, comparou-se o nosso corpus como um todo com as suas seções e ambos com os registros do estudo original em cada uma das seis dimensões nele identificadas (vide ANEXO D) visando determinar as diferenças e semelhanças entre eles. Na próxima seção apresentaremos os resultados referentes a cada dimensão analisada com exemplos dos textos prototípicos do nosso corpus em cada dimensão para ilustrá-los.

É importante comentar que no artigo onde foi publicado o estudo original as médias de dimensão são ilustradas em gráficos de barras por dimensão, mas eles não contêm a média de dimensão referentes a cada registro. Nos basearmos neles para realizar a adição do nosso registro iria contra o rigor metodológico necessário para a realização da

análise e os resultados seriam, no máximo, aproximativos. Diante disso, para possibilitar a realização da nossa análise, contatamos os autores do estudo e pedimos acesso à planilha com as médias de dimensão do estudo original, entretanto, por circunstâncias várias, não foi possível obter acesso à planilha. Para contornar esse problema, Carlos Kauffmann, um dos autores do estudo, gentilmente nos disponibilizou a planilha com as contagens absolutas do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) e a partir dela recriamos, com auxílio dos softwares "IBM SPSS Statistics 23" (IBM, 2015) e Microsoft Excel 2016, os passos do estudo original, visando gerar a tabela com as médias de dimensão desse estudo.

É importante ressaltar ainda que, na planilha que nos foi fornecida, não constava o registro *User's/owner's manuals* (manual do usuário/proprietário), nos tendo sido esclarecido que este foi adicionado em uma etapa subsequente do estudo, e que não seria possível ter acesso à planilha na qual ele foi incluído. Logo, esse registro não consta na nossa análise, tendo sido nosso corpus comparado com 47 dos 48 registros do estudo base. Além disso, conforme exposto na etapa 6 desta pesquisa, constatamos que, por algum possível problema técnico, apesar de usarmos o mesmo etiquetador e pós processador do estudo original, seis variáveis desse estudo não foram contabilizadas no nosso corpus. São elas: *QU questions* (qsqu – dimensão 1), *Average word length* (wl - dimensão 1), *Tag questions* (qsttag - dimensão 3), *Questions: Yes or No question* (qsnyn- dimensão 3), *Type-token ratio* (ttr - dimensão 3) *e Modals*: Haver que/haver de (mdhaver - dimensão 6). Algumas dessas variáveis não teriam muito peso no nosso corpus, como por exemplo *Questions: Yes or No question*, já que, nos textos legais não existem perguntas de sim ou não. Outras, como a variável *Type-token ratio* seriam de extrema importância na análise para verificar a diversidade lexical dos textos legais, apontada em estudos anteriores como uma das principais características da linguagem jurídica. Por fim, na etapa 8 da nossa pesquisa, não conseguimos determinar qual seria a média e o desvio padrão no estudo original da variável: *Que clause controlled by adjective (stance)* (adjque), identificada na dimensão 2 desse estudo (vide ANEXO D)

Diante destes problemas, optamos por desconsiderar essas variáveis na análise, excluindo-as seja da planilha do nosso corpus que daquela do estudo original. Essa decisão se apoia no fato de que, como essas variáveis não foram computadas no nosso

corpus e algumas delas trariam resultados significativos para a caracterização dos textos legais, para manter os parâmetros dos corpora os mais próximos possíveis e possibilitar uma comparação mais fidedigna, elas deveriam ser desconsideradas. Por isso, alertamos que os resultados, ainda que baseados no estudo original, são diferentes dele, inclusive aqueles referentes à distribuição e pontuação dos registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) nas dimensões, já que um registro (*User's/owner's manuals)* e 7 variáveis foram desconsiderados na análise.

A recriação dos dados do estudo base, ao invés de utilizar a planilha com as médias de dimensão originais, assim como a ausência de sete variáveis, compromete a precisão e detalhamento do estudo. Estamos cientes que realizar a análise dessa forma não é ideal, mas diante dos imprevistos e da impossibilidade de obter a planilha original, essa foi a única forma por nós encontrada de dar prosseguimento à pesquisa. Esperamos em um futuro próximo, caso tenhamos acesso à planilha com as médias de dimensão do estudo original, realizar uma nova tentativa de etiquetar o corpus incluindo todas as variáveis do estudo original para podermos então refazer a pesquisa e ter resultados mais fidedignos e precisos.

(12) Cálculo da significância estatística da variação: para verificar se os textos legais são um registro, confirmando nossa hipótese inicial e respondendo nossa primeira pergunta de pesquisa, recorreu-se, inspirados em Berber Sardinha (2013) e Berber Sardinha, Kauffmann e Acunzo (2014), aos testes estatísticos F (ANOVA) e R^2. Esses testes permitem mensurar a variação das dimensões em relação a cada registro. Aplicam-se ambos os testes porque estes se complementam. Enquanto o ANOVA permite identificar se há diferenças significativas entre os registros nas dimensões identificadas com base na sua média de dimensão ele não identifica onde está essa diferença, já o teste R^2 mensura em porcentagem o quanto dessa variação pode ser explicada pelos textos de determinado registro. (BERBER SARDINHA, 2013; BREZINA, 2018)

Para a realização desses testes nos valemos dos softwares: Microsoft Excel 2016 e "IBM SPSS Statistics 23" (IBM, 2015). Primeiramente fizemos o upload dos dados do nosso corpus juntamente com aqueles do CVBR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) que se encontravam em uma planilha do Excel no software "IBM SPSS

Statistics 23” (IBM, 2015). Em seguida, utilizamos a função *Analyse: General linear model* e selecionamos a opção *Univariate*, carregando como variável dependente a dimensão e como variável fixa o registro. O output obtido foi, entre outros, a tabela *“Tests of Between Subject effects”,* que dispõe o valor do teste F (ANOVA), a significância estatística dos resultados e seu R^2. Esse processo foi repetido para cada dimensão. Os dados obtidos foram então copiados e organizados em uma única tabela no Microsoft Excel 2016 por dimensão para uma melhor visualização e serão apresentados na próxima seção.

4 RESULTADOS E DISCUSSÃO

4.1 Os corpora

Antes de trazer os resultados da AMD aditiva propriamente dita é importante fazer uma rápida comparação entre o corpus em análise – o *LEX-BR-Ius* (FERRARI e MARQUES, em compilação) e o corpus do estudo base – o Corpus Brasileiro de Variação e Registro (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014).

Primeiramente, cabe ressaltar que o CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) foi compilado com o objetivo de abranger a maior quantidade possível de registros do português brasileiro, sendo composto por 48 registros entre orais e escritos, totalizando cerca de 5.6 milhões de palavras e 960 textos. Já o *LEX-BR-Ius* (FERRARI e MARQUES, em preparação) foi compilado para ser um corpus representativo de textos legais federais brasileiros, composto assim por apenas um único registro: os textos legais, totalizando cerca de 1 milhão de palavras. Observa-se assim uma clara diferença seja de objetivos que de composição e abrangência dos corpora não só relativa ao número de registros abarcados, mas também entre os números de palavras desses corpora.

Em relação à arquitetura destacamos que enquanto no corpus do estudo base cada seção (48) corresponde a um registro que é amostrado por 20 textos considerados representativos desse registro, totalizando 960 textos, no nosso corpus, apesar de amostrarmos apenas um registro é dividido em 7 seções segundo a categoria do texto legal e não estabelecemos limites para a amostra de cada seção havendo grande variedade seja no número de palavras que de textos por seção.

Em relação às seções do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) cabe ressaltar que ele tem uma seção dedicada aos textos legais, qual seja *Legislation.* Esta seção do corpus é composta por 20 textos legais, com aproximadamente 125mil palavras, o que corresponde a 2,2% do total de palavras do corpus. Já nosso corpus, apesar de ainda estar em compilação totaliza no momento 73 textos com quase um milhão de palavras. No design dessa seção todos os textos legais foram considerados como pertencentes à mesma categoria intitulada: “*legislation*”. Nós,

por outro lado, optamos por separar nosso corpus por seção segundo a categoria do texto legal devido às diferenças procedimentais e situacionais de cada uma delas.

Nós buscamos amostrar todas as categorias de textos legais, incluindo grande parte delas no nosso corpus (Constituição, Códigos, Estatutos, Emendas à Constituição, Leis ordinárias, Leis complementares e Medidas Provisórias). No CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014), por sua vez, apenas algumas delas estão presentes, quais sejam: Constituição, Códigos, Leis Ordinárias, Decretos, Medidas Provisórias e Ato institucional, sendo que a maioria deles (15) promulgada e publicada em 2001, o que representaria um possível recorte temporal, e sua proporção não reflete aquela da realidade. Além disso, alguns textos já não estão mais em vigência, como é o caso do Ato complementar nº8. Nota-se ainda que não foi esclarecido de onde ou quando os textos foram extraídos, informações importantes quando se trabalha com textos legais já que seus textos passam por alterações frequentemente. Por isso, no nosso corpus, nos metadados informamos seja a data da extração que a fonte, utilizando sempre os textos em vigor na data da extração e retirando-os do Portal da Legislação do Planalto.

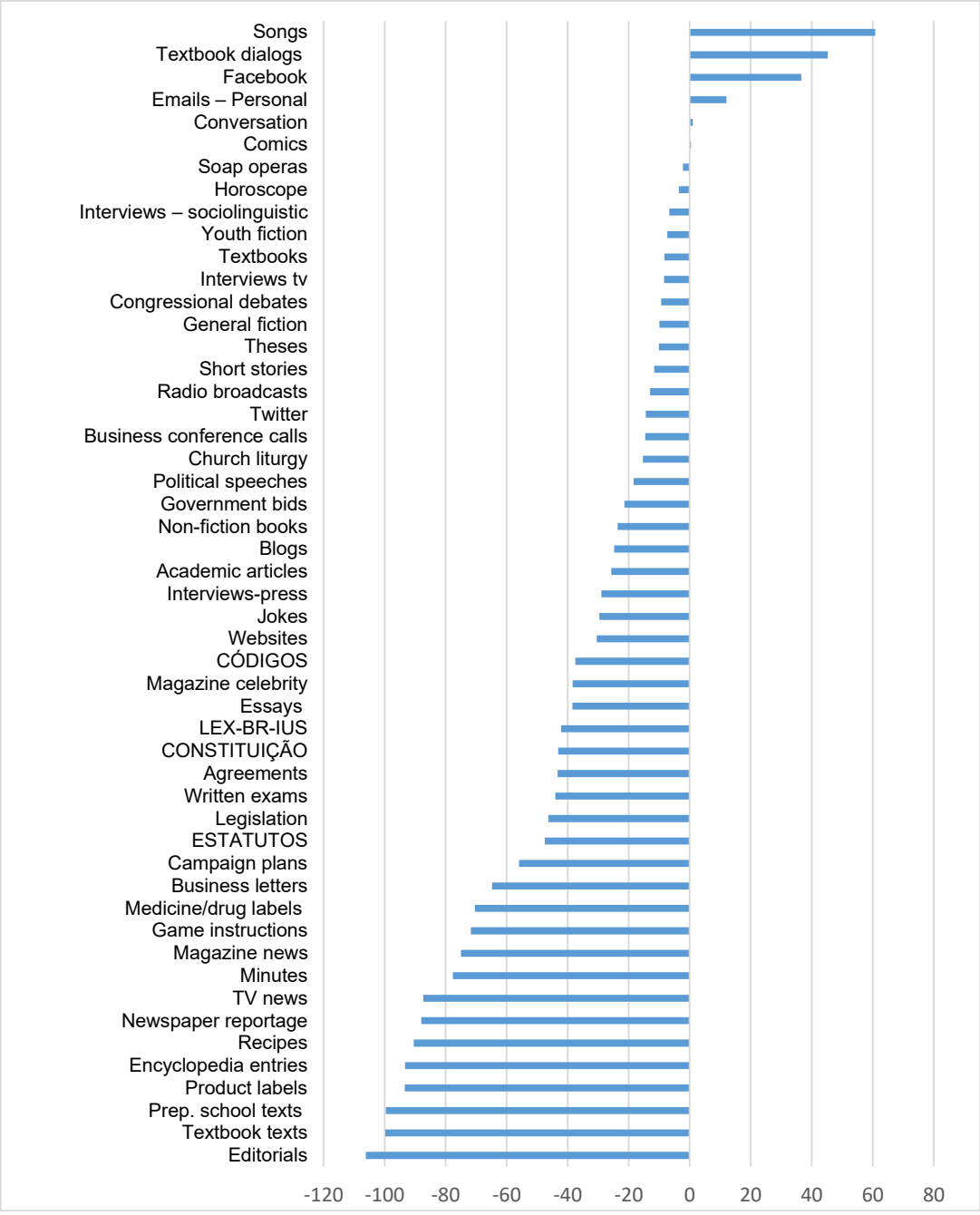
Por fim, quanto ao processamento dos textos, percebe-se que os arquivos da seção *Legislation* que nos foram gentilmente cedidos para a presente pesquisa não passaram por um processo de limpeza tão rigoroso e detalhado quanto aquele por nós realizado. No CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) foram mantidos o texto completo com todas as informações textuais e extratextuais, muitas das quais por nós consideradas, na compilação do nosso corpus, irrelevantes para a caracterização do registro, tais como: “Edição revista, atualizada e ampliada”, “Impresso no Brasil”, “Senado Federal” “Subsecretaria de Informações”, a assinatura dos parlamentares envolvidos no projeto de lei, números de página, o índice da Constituição e do Código de Proteção e Defesa do Consumidor, etc.

4.2 Dimensão 1: *Oral versus literate discourse*

A seguir apresentamos um gráfico de barras com as médias de dimensão do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) referente à dimensão de variação 1 de Berber Sardinha, Kauffmann e Acunzo (2014), intitulada *oral versus literate*

discourse (vide ANEXO D) ao qual adicionamos as médias de dimensão do nosso corpus como um todo e por seção para possibilitar a comparação entre os registros do estudo original e o nosso, conforme exposto anteriormente. Para ter acesso aos dados numéricos que basearam o gráfico vide APÊNDICE B.

Gráfico 1 – Textos legais adicionados à dimensão 1: *Oral versus literate discourse*



Fonte: autora (2022)

No gráfico, na lateral esquerda temos os registros analisados. Os nomes das seções do nosso corpus assim como do corpus estão em caixa alta, enquanto os nomes dos registros do estudo original estão em caixa baixa para facilitar sua diferenciação visual. Na parte inferior do gráfico temos a escala numérica referente aos escores dos registros nessa dimensão. Tomando o número 0 como ponto de referência, à sua esquerda temos o polo negativo da dimensão (*literate discourse*) e à direita o polo positivo (oral). Já as barras representam o valor do escore do registro ao qual se referem. Quanto mais positivo é esse valor mais prototípico o registro, ou seja, ele apresenta mais características do discurso oral. Quanto mais negativo, menos traços do discurso oral e mais características do discurso letrado *(literate discourse*).

Observa-se que dentre os registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) aquele com maior pontuação positiva é *Songs* (canções) (aprox.60.82). Isso significa que esse é o registro que melhor representa o discurso oral, ou seja, aquele mais prototípico dentre os registros analisados. Ele se caracteriza pela presença de verbos e pronomes na primeira e segunda pessoa, assim como o uso de verbos mentais e de ação ressaltando a interação entre os participantes e a alternância de foco entre emissor e receptor. Já no polo negativo o registro com maior pontuação é *Editorials* (editoriais*)* (aprox. -106.16). Sendo assim, o registro mais "letrado" da amostra. Esse polo se destaca pela alta frequência de nomes e preposições tendo uma alta densidade informacional e letramento.

Em relação ao nosso registro: textos legais, observamos que todos eles, seja as seções que o corpus como um todo se encontram no polo negativo da dimensão 1. Isso significa que nos textos legais predominam as características do discurso letrado. Sua pontuação aproximada, foi: Constituição (-43.11), Códigos (-37.48), Estatutos (-47.45) e *LEX-BR-Ius* (-42.09). Observa-se que a seção com maior pontuação negativa, ou seja, aquela mais "letrada" e, consequentemente com menor número de características do discurso oral é a seção Estatutos e aquele com maior pontuação negativa é a seção Códigos. Já o corpus como um todo se encontra em uma posição intermediária, entre elas e mais próximo da Constituição. Como pode ser visto, apesar de todas as seções conterem textos legais, há uma pequena diferença na variação entre elas e também entre

elas e o registro *legislation (-46.3),* que se aproxima mais da nossa seção Estatutos a ser esclarecida em estudos futuros.

O polo negativo dessa dimensão *(literate discourse*) onde se encontra o nosso corpus, é caracterizado pela presença de: *adjectives: attributive position* (adjetivos: posição atributiva), *nouns: compound* (substantivos: composto), *articles: definite* (artigos: definido), *prepositions: all* (preposições: todos), *nouns: abstract* (substantivos: abstrato)*, adjectives: topical* (adjetivos: tópico), *nominalization in subject position* (nominalização na posição de sujeito), *past participle* (particípio passado), *adjectives: relational* (adjetivos: relacional) *agentless passives* (passivas sem agente), *pronouns: relative* qual or cujo (pronomes: relativo qual ou cujo) e *reduced progressive clause* (oração progressiva reduzida). A seguir trazemos trechos do texto E12.288_20.07.2010 (Estatuto da Igualdade Racial), o texto do nosso corpus com maior pontuação negativa (-57.41) nessa dimensão, ou seja, aquele mais "letrado" para observar essas características em contexto:

(1) Art. 1o Esta Lei institui o Estatuto da Igualdade Racial, destinado a garantir à população negra a efetivação da igualdade de oportunidades, a defesa dos direitos étnicos individuais, coletivos e difusos e o combate à discriminação e às demais formas de intolerância étnica.

Parágrafo único. Para efeito deste Estatuto, considera-se:

I - discriminação racial ou étnico-racial: toda distinção, exclusão, restrição ou preferência baseada em raça, cor, descendência ou origem nacional ou étnica que tenha por objeto anular ou restringir o reconhecimento, gozo ou exercício, em igualdade de condições, de direitos humanos e liberdades fundamentais nos campos político, econômico, social, cultural ou em qualquer outro campo da vida pública ou privada; (BRASIL, 2010)

(2) Art. 4o A participação da população negra, em condição de igualdade de oportunidade, na vida econômica, social, política e cultural do País será promovida, prioritariamente, por meio de: (BRASIL, 2010)

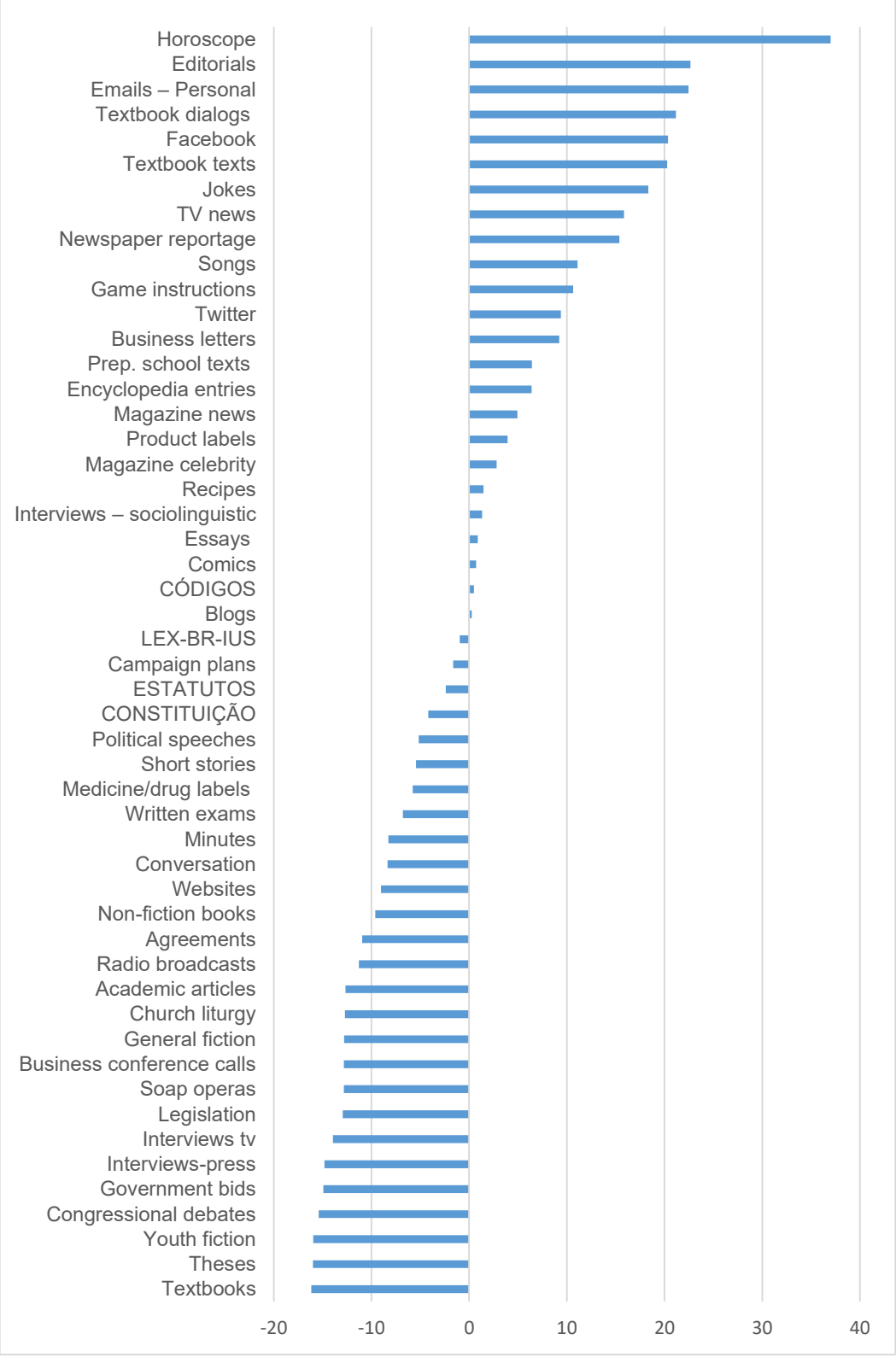
(3) Art. 47. É instituído o Sistema Nacional de Promoção da Igualdade Racial (Sinapir) como forma de organização e de articulação voltadas à implementação do conjunto de políticas e serviços destinados a superar as desigualdades étnicas existentes no País, prestados pelo poder público federal. (BRASIL, 2010)

Como pode ser observado em todos trechos reproduzidos acima temos uma alta incidência de artigos definidos (ex.: o, a, as), preposições (ex.: de, da, em) e adjetivos atributivos tais como “negra”, em “população negra” nos trechos 1 e 2, e “étnicos”, em “direitos étnicos individuais” no trecho 1º e “étnicas” em “desigualdades étnicas” no trecho 3. No trecho 1 temos ainda o uso de vários sintagmas nominais preposicionados como: “igualdade de oportunidades”, “igualdade de condições” e “formas de intolerância étnica”, o uso do particípio passado em “baseada” e substantivos abstratos, como “cor”. Já no trecho 2 temos novamente particípio passado (promovida) e a presença de nominalização na posição de sujeito (A participação). Por fim, no trecho 3 temos uma oração passiva não agentiva. Como pode ser observado nos trechos temos as principais características identificadas no polo negativo da dimensão em análise, confirmando a posição dos textos legais como textos letrados.

4.3 Dimensão 2: *Argumentation*

Em relação à dimensão de variação 2: *Argumentation* de Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D), após adicionarmos nossos dados àqueles do estudo original temos:

Gráfico 2 – Textos legais adicionados à dimensão 2: *Argumentation*



Fonte: autora (2022)

Nessa dimensão se distinguem os registros com maior grau de argumentação daqueles com pouca ou quase nenhuma argumentação. Observa-se que, dentre os registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014), aquele com maior pontuação positiva (aprox. 36.99) é, surpreendentemente, *Horoscope* (horóscopo). Os registros mais prototípicos se caracterizam pela presença orações não finitas, infinitivas e relativas/adjetivas introduzidas por pronomes relativos que exercem o papel de estruturar a argumentação visando convencer o interlocutor do que está sendo dito. Já no polo negativo o registro com maior pontuação é *Textbooks* (livros didáticos) (aprox. -16.14). Sendo assim, o registro com menor grau de argumentação, tendo baixa ocorrência das características mais frequentes dessa dimensão.

No nosso corpus, por sua vez, temos seja pontuações negativas que positivas: Constituição (-4.18), Códigos (0.48), Estatutos (-2.39) e *LEX-BR-Ius* (-0.97). Como pode ser observado a seção Códigos é a única que se encontra no polo positivo, mas as pontuações se todas as seções e do corpus são baixas e próximas ao zero, o que representaria um grau mediano abaixo de argumentação e todos os valores são próximos uns dos outros na escala, tendo uma menor variação entre as seções quando comparado com a dimensão anterior. A seção do corpus com maior pontuação positiva, ou seja, aquela mais "argumentativa" é a seção Códigos e aquela com maior pontuação negativa é a seção Constituição. Já o corpus como um todo se encontra no polo negativo, mas é aquele com menor pontuação negativa quando comparado com as demais seções. Ao comparamos com a seção *legislation* observamos que ela se localiza no polo negativo, sendo mais negativa (-12.93) que os textos do nosso corpus e, consequentemente, menos argumentativa.

Nossos textos seriam assim nem decisivamente argumentativos nem completamente destituídos de argumentação. Conclui-se que a argumentação não exerce um papel primordial nos textos legais. Por se tratar de textos normativos seu objetivo principal é informar sobre as normas que regem a sociedade. Essas normas são impositivas, logo devem ser seguidas independente de concordância por parte do receptor. Isso explicaria por que a argumentação não exerce um papel determinante nesses textos, já que, devido à sua natureza impositiva e informativa não há necessidade de se argumentar, no texto da norma, sua razão de existência ou porque essa deva ser

respeitada. A argumentação ocorre em uma etapa precedente da criação do texto legal, antes mesmo do início do processo legislativo, durante a escrita do projeto de lei.

Berber Sardinha, Kauffmann e Acunzo (2014) identificaram como características dessa dimensão: *que clause controlled by noun* (oração que controlada por substantivo), *pronouns: relative que* (pronomes: relativo que), *adverbs:* comparative (advérbios: comparativo)*, nouns: cognition* (substantivos: cognitivo), *que or infinitive clause controlled by noun (stance)* (que ou oração infinitiva controlada pro substantivo (posicionamento)), *infinitive clause controlled by adjective* (oração infinitiva controlada por adjetivo), *que clause controlled by preposition* (oração que controlada por preposição), *pronouns: demonstrative* (pronomes: demonstrativo)*, infinitive clause controlled by preposition* (oração infinitiva controlada por preposição), *infinitive clause controlled by ease or difficulty adjective* (oração infinitiva controlada por adjetivo facilitador ou dificultador), *adverbs: hedge* (advérbios: fronteira), *articles: indefinite* (artigos: indefinido), *verbs: future preterit tense* (verbos: futuro do pretérito) *e conjunctions: coordinating (adversative)* (conjunções: coordenativa (adversativa)).

A seguir trazemos alguns trechos do texto E9.474_22.07.1997 (Estatuto do refugiado), o texto do nosso corpus com maior pontuação positiva nessa dimensão (12.34), logo o que apresenta mais características argumentativa, para ilustrar algumas dessas características:

(1) Art. 1º Será reconhecido como refugiado todo indivíduo que:

I - devido a fundados temores de perseguição por motivos de raça, religião, nacionalidade, grupo social ou opiniões políticas encontre-se fora de seu país de nacionalidade e não possa ou não queira acolher-se à proteção de tal país;

II - não tendo nacionalidade e estando fora do país onde antes teve sua residência habitual, não possa ou não queira regressar a ele, em função das circunstâncias descritas no inciso anterior;

III - devido a grave e generalizada violação de direitos humanos, é obrigado a deixar seu país de nacionalidade para buscar refúgio em outro país. (BRASIL, 1997)

(2) Art. 4º O reconhecimento da condição de refugiado, nos termos das definições anteriores, sujeitará seu beneficiário ao preceituado nesta Lei, sem prejuízo do disposto

em instrumentos internacionais de que o Governo brasileiro seja parte, ratifique ou venha a aderir. (BRASIL, 1997)

(3) Art. 7º O estrangeiro que chegar ao território nacional poderá expressar sua vontade de solicitar reconhecimento como refugiado a qualquer autoridade migratória que se encontre na fronteira, a qual lhe proporcionará as informações necessárias quanto ao procedimento cabível.

§ 1º Em hipótese alguma será efetuada sua deportação para fronteira de território em que sua vida ou liberdade esteja ameaçada, em virtude de raça, religião, nacionalidade, grupo social ou opinião política. (BRASIL, 1997)

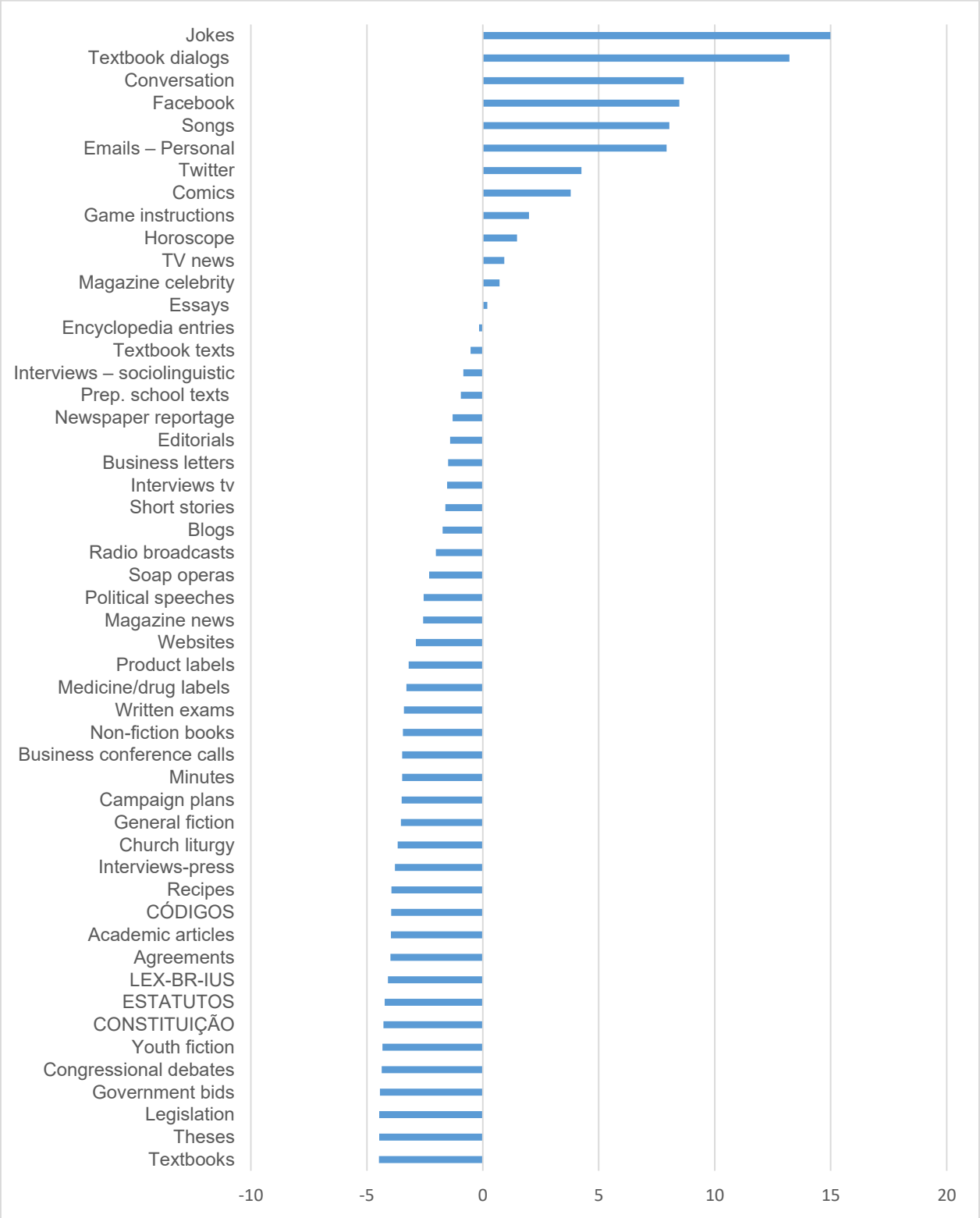
(4) Art. 39 II - a prova da falsidade dos fundamentos invocados para o reconhecimento da condição de refugiado ou a existência de fatos que, se fossem conhecidos quando do reconhecimento, teriam ensejado uma decisão negativa; (BRASIL, 1997)

No trecho 1 destacamos a presença de orações infinitivas controladas por preposição no inciso III. Já no trecho 2 ressaltamos a oração controlada por preposição: ”de que o governo seja parte”. No trecho 3 podemos observar o uso de substantivos cognitivos, tais como “vontade” e “hipótese”, o uso do pronome relativo “que” e de oração controlada por substantivo no § 1º. Por fim, no trecho 4 temos um exemplo de verbo no futuro do pretérito (teriam ensejado) e de artigo indefinido (uma).

4.4 Dimensão 3: *Involved versus informational production*

Já na dimensão de variação 3 *(Involved versus informational production*) de Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D) a distribuição dos registros é:

Gráfico 3 – Textos legais adicionados à dimensão 3: *Involved versus informational production*



Fonte: autora (2022)

Em relação aos registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) aquele com maior pontuação positiva é *Jokes* (piadas) (aprox. 14.98), sendo o registro com maior grau de interação entre os participantes. No polo negativo, por sua vez, predominando textos de caráter informativo. O registro com maior pontuação nesse polo é *Textbooks* (aprox. -4.48), marcado pela sua diversidade lexical que exerce um papel importante na transmissão de informações.

Quando observamos a pontuação dos textos legais nessa dimensão temos que todos se localizam no polo negativo, logo seriam textos com maior caráter informacional em detrimento da interatividade. A pontuação por seção e do corpus foi: Constituição (-4.28), Códigos (-3.95), Estatutos (-4.23) e *LEX-BR-Ius* (-4.09). Todos com pontuações bem aproximadas, sendo o mais negativo, logo o mais informacional a Constituição e o menos negativo os Códigos. Interessante notar que nessa dimensão a pontuação dos textos legais se aproxima daquela da seção *legislation* (-4.46) do estudo original.

As variáveis que caracterizam *involved production* são: *contractions* (contrações), discourse marker (marcadores de discurso), pronouns: *third person singular, in subject position* (pronomes: terceira pessoa singular na posição de sujeito), *pronouns: third person plural, in subject position* (pronomes: terceira pessoa plural na posição de sujeito), *conjunctions: coordinating (conclusive)* (conjunções: coordenativa (conclusiva)), *Adverbs: Place* (advérbios: lugar) e *Modals:* Ter que/ter de (Modais: ter que/ter de). O outro polo dessa dimensão *(informational production)* é caracterizado no estudo original apenas pela variável *Type-token ratio* (relação type-token), entretanto, conforme explicado na seção Metodologia, essa variável foi excluída do presente estudo por não ter sido contabilizada no nosso corpus devido a problemas técnicos. Logo consideramos na análise apenas as variáveis do polo positivo dessa dimensão acima expostas. Como todos os nossos textos pontuaram negativamente isso significa que neles constam poucas ou nenhuma dessas variáveis, a variável *Contractions* (contrações), por exemplo não ocorreu nenhuma vez no nosso corpus. Logo ilustrar essas características com nossos textos é impossível em alguns casos.

A seguir trazemos trechos de alguns textos do nosso corpus buscando exemplificar algumas dessas características:

(1) Art. 2º V - garantir aos índios a permanência voluntária no seu habitat, proporcionando-lhes ali recursos para seu desenvolvimento e progresso; (BRASIL, 1973)

(2) Art. 23. As ilhas ou ilhotas, que se formarem no álveo de uma corrente, pertencem ao domínio público, no caso das águas públicas, e ao domínio particular, no caso das águas comuns ou particulares.

§1º Se a corrente servir de divisa entre diversos proprietários e elas estiverem no meio da corrente, pertencem a todos esses proprietários, na proporção de suas testadas até a linha que dividir o álveo em duas partes iguais. (BRASIL 1934)

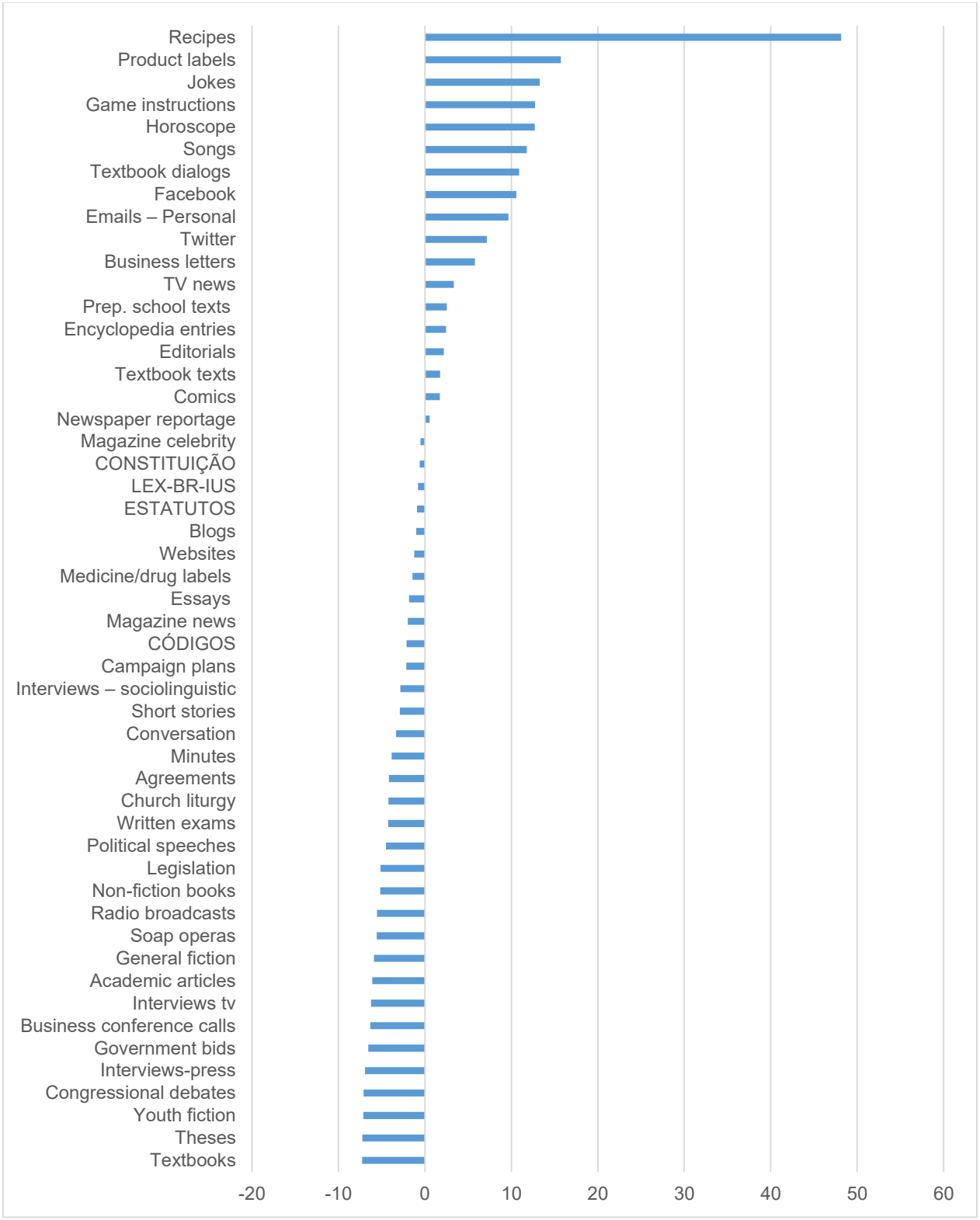
(3) Art. 97. Não poderá o dono do prédio abrir poço junto ao prédio do vizinho, sem guardar as distâncias necessárias ou tomar as precisas precauções para que ele não sofra prejuízo. (BRASIL 1934)

No trecho 1, retirado do texto E6.001_19.12.1973 (estatuto do índio), temos um advérbio de lugar: “ali” que se refere ao habitat dos índios. No trecho 2, retirado do texto C24.643_10.07.1934 (Código de águas) temos o pronome pessoal de terceira pessoa plural na posição de sujeito “elas” que se refere às “ilhas ou ilhotas” no caput do artigo. Já no trecho 3 temos um exemplo também do Código de águas com um pronome pessoal de terceira pessoa singular na posição de sujeito (ele).

4.5 Dimensão 4: *Directive discourse*

Na dimensão de variação 4 *(Directive discourse*) de Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D), por sua vez, nossos resultados podem ser resumidos no gráfico a seguir:

Gráfico 4 – Textos legais adicionados à dimensão 4: *Directive discourse*



Fonte: autora (2022)

Nessa dimensão o registro do estudo original com maior pontuação positiva é *Recipes* (receitas) (aprox. 48.16), marcado pela presença de verbos no presente do subjuntivo e imperativo que reforçam o aspecto instrucional/diretivo dessa dimensão. No polo negativo, temos os registros menos prototípicos com pouca ou nenhuma função de dar ordens ou instruções. O registro com maior pontuação negativa é *Textbooks* (aprox. -7.25).

Nosso registro se enquadra no polo negativo dessa dimensão, não sendo caracterizado pela presença de orientações para a execução de tarefas. A pontuação por seção e do corpus como um todo é: Constituição (-0.61), Códigos (-2.11), Estatutos (-0.92) e *LEX-BR-Ius* (-0.81). Também nessa dimensão as pontuações são próximas umas das outras, sendo mais negativa a seção Códigos e a menos negativa a Constituição. Quanto à seção *legislation*, também essa se localiza no polo negativo, mas sua pontuação é um pouco superior àquelas do nosso corpus (-5.15).

As características abrangidas nessa dimensão. Logo típicas de um discurso "diretivo" são: *verbs: present subjunctive mood* (verbos: presente do subjuntivo), *verbs: imperative mood* (verbos: imperativo), *nouns: concrete* (substantivos: concreto), *subject omission* (omissão de sujeito), *verbs: facilitation* (verbos: facilitação), *conjunctions: coordinating (clausal)* (conjunção coordenativa (oração)). A seguir trazemos trechos do texto C556_25.06.1850 (Código comercial) do nosso corpus com maior pontuação nessa dimensão (1.51), sendo aquele que mais apresenta, ainda que em baixa frequência, as características dessa dimensão, visando ilustrá-las:

(1) Art. 4 - Ninguém é reputado comerciante para efeito de gozar da proteção que este Código liberaliza em favor do comércio, sem que se tenha matriculado em algum dos Tribunais do Comércio do Império, e faça da mercancia profissão habitual (artigo nº 9). (BRASIL, 1850)

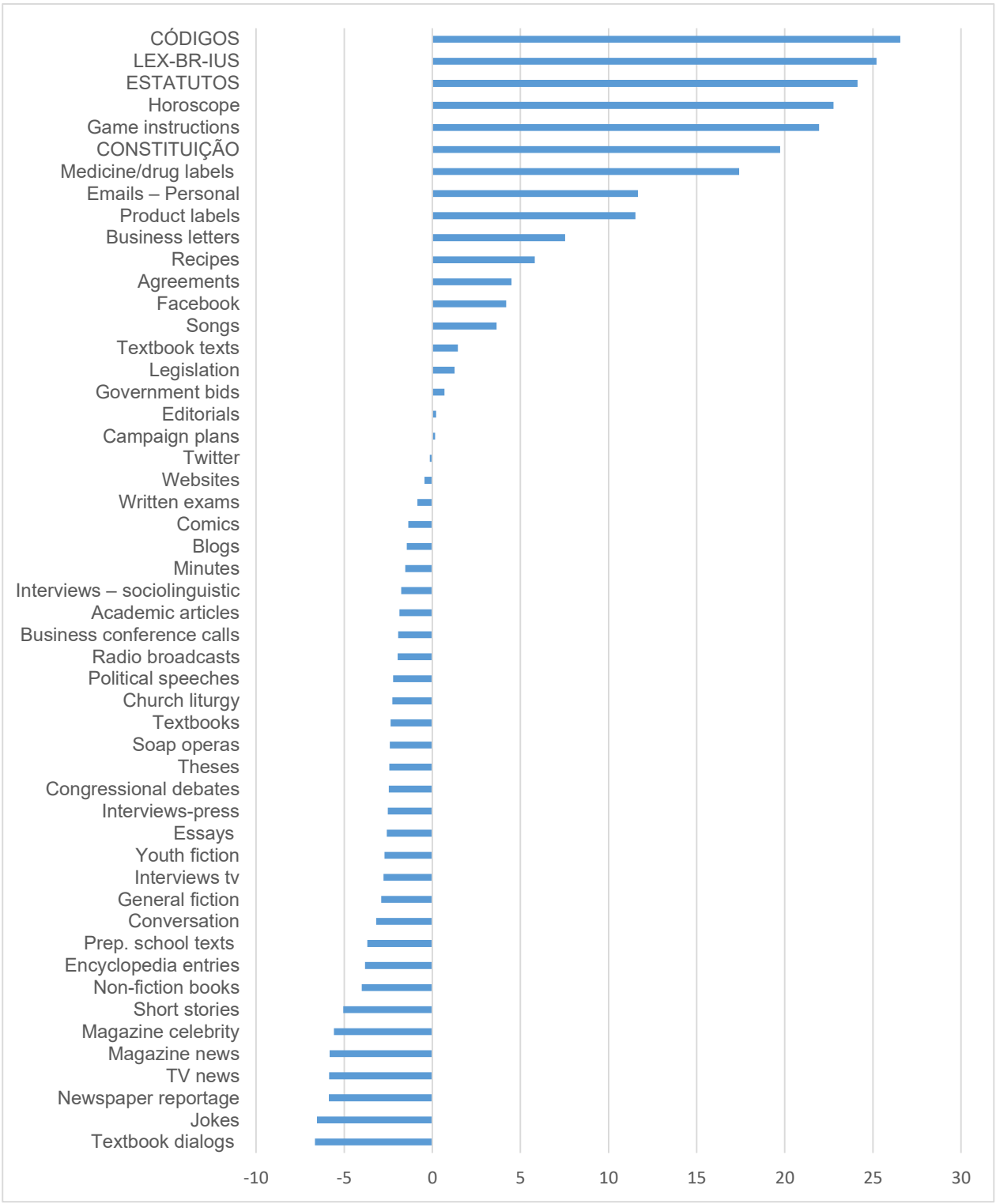
(2) *Art. 36 - Para ser corretor, requer-se ter mais de 25 (vinte e cinco) anos de idade, e ser domiciliado no lugar por mais de 1 (um) ano.* (BRASIL, 1850)

No trecho 1 temos verbos no presente do subjuntivo (faça e tenha), substantivos concretos como “artigo” e a conjunção coordenativa “e” Já no trecho 2 destacamos o verbo facilitador: “requerer”.

4.6 Dimensão 5: *Future versus past time orientation*

Na dimensão de variação 5: *Future versus past time orientation* de Berber Sardinha, Kauffmann e Acunzo (2014) (vide ANEXO D), os registros estão distribuídos conforme o gráfico a seguir:

Gráfico 5 – Textos legais adicionados à dimensão 5: *Future versus past time orientation*



Fonte: autora (2022)

Dentre os registros do CBVR (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014) aquele com maior pontuação positiva é *Horoscope* (aprox. 22.75), marcado pelo uso do futuro utilizado para fazer previsões. A ele se opõe àquele com maior pontuação negativa*: Textbook dialogs* (diálogos de livros didáticos) (aprox. -6.65), marcado pelo uso do passado.

Os textos legais, por sua vez, estão na dimensão positiva, sendo o registro com maior pontuação nessa dimensão, superando inclusive o registro mais prototípico do estudo original. Nosso corpus pontuou: Constituição (19.73), Códigos (26.54), Estatutos (24.12) e *LEX-BR-Ius* (25.21). Esse registro é assim marcado pela alta frequência de verbos no futuro e uso dos modais dever e poder, conjunções coordenadas e orações subordinadas. Essas características exercem a função de estabelecer como se dá a aplicação das normas e suas consequências em casos concretos aplicáveis após a promulgação da norma. É importante ressaltar que, nessa dimensão há grande diferença nos valores do nosso corpus e aquele da seção *legislation (1.25).* Apesar de ambos se encontrarem no polo positivo, esta última tem uma pontuação significativamente menor quando comparada com as nossas o que deverá ser investigado futuramente.

Como todos os nossos textos se enquadram no polo positivo dessa dimensão, sendo marcados pela orientação verso o futuro trazemos aqui as variáveis desse polo: *verbs: future subjunctive mood* (verbos: futuro do subjuntivo), *conjunctions: coordinating (ou)* (conjunções: coordenativa (ou)), *verbs: future present tense* (verbos: futuro do presente), *modals: dever* (modais: dever), *modals: poder* (modais: poder), *subordinating (conditional) clause* (oração subordinativa (condicional)), *adverbs: likelihood* (adjetivos: probabilidade) *e conjunctions: coordinating (phrasal)* (conjunções: coordenativa (frasal)).

A seguir ilustraremos essas variáveis em contexto em trechos do texto C8.078_11.09. 1990 (Código de defesa do consumidor) o texto com maior pontuação positiva do nosso corpus nessa dimensão (33.55):

(1) *Art. 2° Consumidor é toda pessoa física <u>ou</u> jurídica que adquire <u>ou</u> utiliza produto <u>ou</u> serviço como destinatário final.*

Parágrafo único. Equipara-se a consumidor a coletividade de pessoas, ainda que indetermináveis, que haja intervindo nas relações de consumo. (BRASIL, 1990)

(2) *Art. 8º § 2º O fornecedor deverá higienizar os equipamentos e utensílios utilizados no fornecimento de produtos ou serviços, ou colocados à disposição do consumidor, e informar, de maneira ostensiva e adequada, quando for o caso, sobre o risco de contaminação.* (BRASIL, 1990)

(3) *Art. 10. O fornecedor não poderá colocar no mercado de consumo produto ou serviço que sabe ou deveria saber apresentar alto grau de nocividade ou periculosidade à saúde ou segurança.* (BRASIL, 1990)

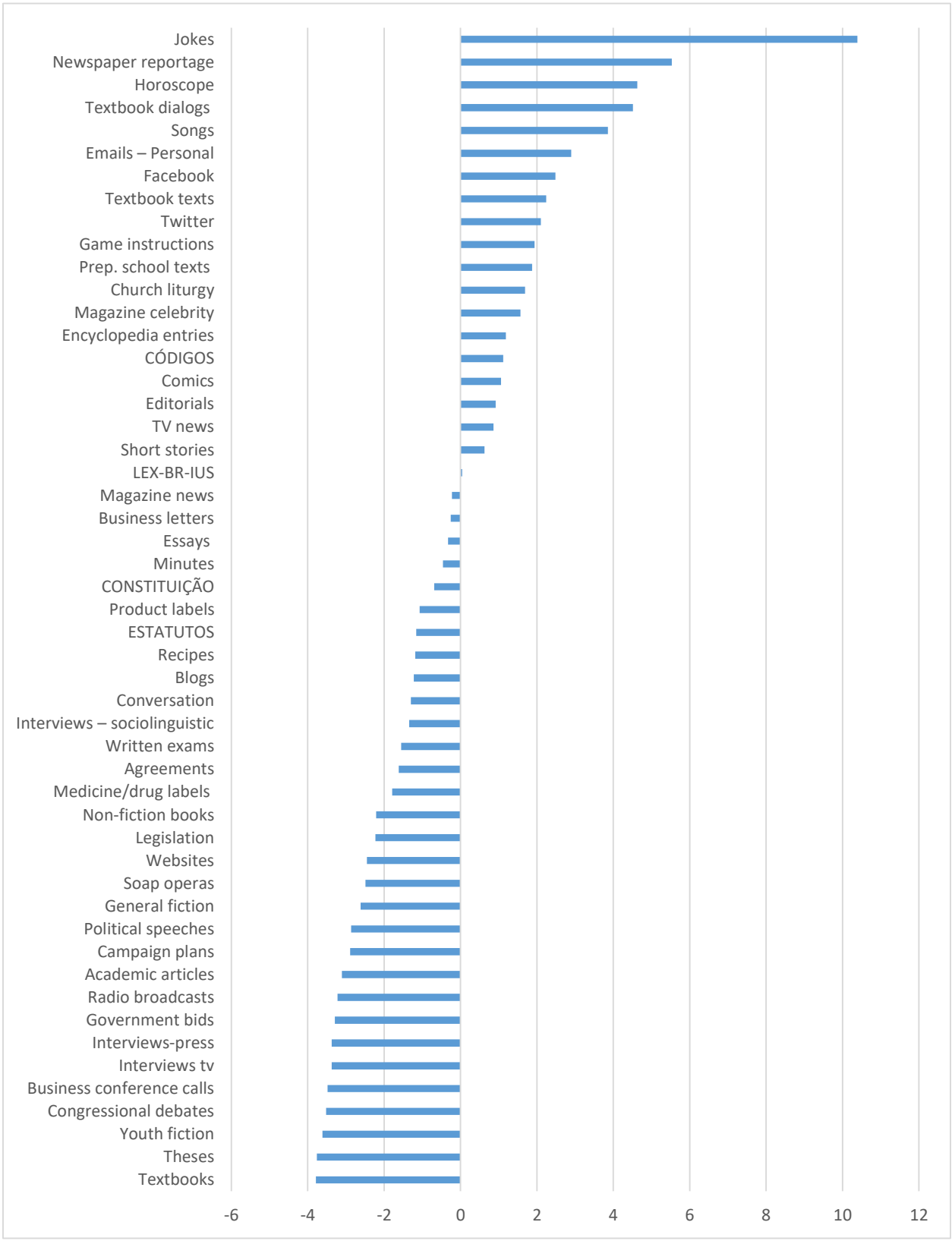
(4) *Art. 39 VIII - colocar, no mercado de consumo, qualquer produto ou serviço em desacordo com as normas expedidas pelos órgãos oficiais competentes ou, se normas específicas não existirem, pela Associação Brasileira de Normas Técnicas ou outra entidade credenciada pelo Conselho Nacional de Metrologia, Normalização e Qualidade Industrial (Conmetro);* (BRASIL, 1990)

No trecho 1 temos a conjunção coordenada “ou” e seu uso frasal. No trecho 2 destacamos a presença do verbo modal “dever” conjugado no futuro do presente e de verbo no futuro subjuntivo (for). Já no trecho 3 temos o verbo modal “poder”, também no futuro do presente. No trecho 4, por sua vez temos um exemplo de oração subordinativa condicional.

4.7 Dimensão 6: *Reported discourse*

Por fim, a última dimensão de variação do PB identificada por Berber Sardinha, Kauffmann e Acunzo (2014) é a dimensão 6: *Reported discourse.* A seguir reproduzimos o gráfico no qual adicionamos nosso corpus aos registros por eles analisados no estudo original:

Gráfico 6 – Textos legais adicionados à dimensão 6: *Reported discourse*



Fonte: autora (2022)

Essa dimensão é marcada pelo uso do discurso relatado/indireto, realizado através do uso de pronomes raros, possessivos e oblíquos, verbos na segunda pessoa e verbos públicos, assim como orações subordinadas e uso de léxico mais arcaico e/ou pouco utilizado no dia a dia. O registro do estudo original mais prototípico dessa dimensão, ou seja, aquele com maior pontuação positiva é *Jokes* (aprox. 10.39), já aquele mais negativo é *Textbooks* (aprox. -3.78).

Nessa dimensão nossos textos pontuaram em ambos os polos: Constituição (-0.68), Códigos (1.11), Estatutos (-1.16) e *LEX-BR-Ius* (0.04). Enquanto o corpus se localiza no polo positivo da dimensão, juntamente com a seção Código, as seções Constituição e Estatutos estão no polo negativo. A maior pontuação positiva é da seção Códigos, já a maior pontuação negativa é da seção Estatutos. Percebe-se assim que o discurso indireto não é uma característica marcante dos textos legais. Quanto à seção *legislativo,* ela se encontra no polo negativo (-2.2), sendo mais negativa que nossos textos.

São características dessa dimensão: *pronouns: rare in object position* (pronomes: raro em posição de objeto), *verbs: second person* (verbos: segunda pessoa), *subordinating (final) clause* (oração subordinativa final/de finalidade), *pronouns: third person, object position* (pronomes: terceira pessoa na posição de objeto) e *verbs: public* (verbos: público). A seguir trazemos um trecho do texto C10.406_10.01.2002 (Código Civil) do nosso corpus com maior pontuação nessa dimensão (10.84), logo o texto com mais incidência das características do discurso indireto:

(1) *Art. 12. Pode-se exigir que cesse a ameaça, ou a lesão, a direito da personalidade, e reclamar perdas e danos, sem prejuízo de outras sanções previstas em lei.*

Parágrafo único. Em se tratando de morto, terá legitimação para requerer a medida prevista neste artigo o cônjuge sobrevivente, ou qualquer parente em linha reta, ou colateral até o quarto grau. (BRASIL, 2002)

(2) *Art. 20. Salvo se autorizadas, ou se necessárias à administração da justiça ou à manutenção da ordem pública, a divulgação de escritos, a transmissão da palavra, ou*

a publicação, a exposição ou a utilização da imagem de uma pessoa poderão ser proibidas, a seu requerimento e sem prejuízo da indenização que couber, se lhe atingirem a honra, a boa fama ou a respeitabilidade, ou se se destinarem a fins comerciais. (BRASIL, 2002)

(3) *Art. 1322 Parágrafo único. Se nenhum dos condôminos tem benfeitorias na coisa comum e participam todos do condomínio em partes iguais, realizar-se-á licitação entre estranhos e, antes de adjudicada a coisa àquele que ofereceu maior lanço, proceder-se-á à licitação entre os condôminos, a fim de que a coisa seja adjudicada a quem afinal oferecer melhor lanço, preferindo, em condições iguais, o condômino ao estranho.* (BRASIL, 2002)

No trecho 1 destacamos o uso do pronome de terceira pessoa na posição de objeto “se” e dos verbos públicos “reclamar” e “requerer”. No trecho 2, por sua vez, ressaltamos o uso do pronome raro em posição de objeto “lhe”. Já no trecho 3 temos um exemplo de oração subordinativa final.

4.8 ANOVA e R^2

Para medir se as diferenças entre os registros são significativas, o quanto os textos analisados abarcam a variação projetada e determinar a capacidade das dimensões de discriminarem corretamente entre os diferentes grupos de textos recorreu-se aos testes estatísticos *F-test* (ANOVA) e R^2. Para tanto seguiu-se o processo apresentado na seção Metodologia para calcular a variação do nosso registro nas dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014).

Na tabela a seguir reproduzimos nossos resultados:

Tabela 10 – ANOVA e R²

Dimensão	F	p	R²
1	45.6569	.000	69%
2	39.917	.000	66.10%
3	29.0722	.000	58.40%
4	89.4327	.000	81.60%
5	25.2681	.000	54.80%
6	23.928	.000	53.40%

Fonte: autora (2022)

Na tabela, na primeira coluna temos a dimensão à qual correspondem os dados. Na segunda coluna temos os resultados do teste F (ANOVA) que, juntamente com a terceira coluna "p", indica a variação dos textos em cada dimensão e se esta é estatisticamente significante ou não. Esclarece-se que o "p" faz referimento ao *p-value*[54] um teste estatístico que permite estabelecer se o resultado é estatisticamente significante ou não, sendo considerados estatisticamente significantes os resultados cujo *p-value* é inferior a 0.05 (5%). Já na coluna R² temos, em porcentagem o quanto a variação é capturada pelo corpus. Segundo Berber Sardinha (2014), se o resultado do R ² é superior a 20% ele é relevante, significando que os textos analisados naquela dimensão abarcam uma variação significativa dos registros.

Por exemplo, na dimensão 1 o resultado do teste ANOVA interpretado em conjunto com o *p-value* indica que há uma variação estatisticamente significante entre os registros analisados em relação às suas médias de dimensão. Logo, todos os registros se

[54] *A p-value is a probability value (p stands for probability) and is one of the outcomes of a statistical test. P-value can be defined as the probability that the data would be at least as extreme as that observed if the null hypothesis were true. […] If the p-value is small enough, usually smaller than 0.05, i.e. 5%, we reject the null hypothesis and conclude that the observed difference is unlikely to be due to chance and therefore the result is statistically significant. This means that the difference observed in the corpus (sample) is likely to be a true difference in the population (all language use). If the p-value is equal to or is larger than 0.05 (or 5%) we conclude that there is not enough evidence in the corpus to reject the null hypothesis.* (BREZINA, 2018, p. 12-13)

distinguem entre si em um grau que permite, juntamente com a análise das dimensões e seus escores classificá-los como registros distintos. Já o R^2 indica que os textos analisados abrangem cerca de 69% da variação capturada pelos registros nessa dimensão, bem superior ao corte de 20% trazido por Berber Sardinha (2014), logo capturam um grau significante de variação.

A dimensão com maior variação entre os registos é a dimensão 4: *Directive discourse*, cujo F teste pontuou 89.43. Nessa dimensão os textos analisados capturam 81.6% da variação prevista pelo R^2. Já a dimensão com menor variação entre os registros é a dimensão 6: *Reported discourse,* que, apesar de ter um resultado estatisticamente significante, essa é de 23.92 e os textos abrangem 53.4% da variação prevista. Os resultados indicam que a diferenças entre os registros é significante em todas as dimensões analisadas e que elas são capazes de prever entre 53.4% e 81.6% das diferenças entre os registros em análise.

5 CONCLUSÃO

Conforme exposto ao longo deste trabalho a linguagem jurídica é um registro extremamente amplo cujas características mudam a depender da língua e do sistema jurídico adotado em determinado país. Apesar da sua riqueza e do crescente interesse nessa linguagem nos últimos anos, seja entre estudiosos do direito que da linguística, trata-se de um registro ainda pouco explorado. Dentre os muitos registros que compõe a linguagem jurídica, na presente pesquisa escolhemos estudar os textos legais. Por questões de tempo e abrangência, já que a quantidade de textos legais é enorme e novos textos são editados diariamente, e aqueles já existentes passam por mudanças frequentemente, optou-se por realizar um recorte que possibilitasse a realização da pesquisa e ao mesmo tempo caracterizasse essa variedade. Por tanto, optou-se por ter como objeto da pesquisa os textos legais federais brasileiros em vigência.

Para a realização dessa pesquisa partimos das seguintes perguntas de pesquisa: a) Como se dá a variação linguística nos textos legais federais brasileiros?; b) Os textos legais federais brasileiros podem ser considerados um registro?; c) Como os textos legais brasileiros se encaixam nas dimensões de variação do português brasileiro identificadas de Berber Sardinha, Kauffmann e Acunzo (2014)?. Para respondê-las estabelecemos como objetivo geral investigar a variação linguística dos textos legais federais brasileiros em vigência no ano de 2022 e como objetivos específicos: a) Estudar a variação linguística dos textos legais federais brasileiros;b) Verificar se os textos legais federais brasileiros podem ser classificados como um registro; c) Comparar os textos legais federais brasileiros com os registros encontrados no Corpus Brasileiro de Variação de Registro (BERBER SARDINHA, KAUFFMANN e ACUNZO, 2014), em todas as dimensões identificadas por Berber Sardinha, Kauffmann e Acunzo (2014).

Para tanto, adotamos a perspectiva de registro de Biber e Conrad (2009) e utilizamos a metodologia da linguística de corpus e a abordagem da Análise Multidimensional (BIBER,1988). Para dar início à pesquisa primeiramente, fez-se necessário, um corpus representativo dos textos legais federais brasileiros, entretanto, no momento da pesquisa não existia um corpus que satisfizesse nossos objetivos. Isso nos levou, com base na linguística de corpus, a compilar nosso próprio corpus: o *LEX-*

BR-lus (FERRARI, MARQUES, em compilação). Em seguida, devido a circunstâncias imprevistas e problemas com a composição do corpus, estabeleceu-se que suas seções: Constituição, Estatutos e Códigos seriam submetidas a uma análise multidimensional do tipo aditiva adotando como estudo base aquele de Berber Sardinha, Kauffmann e Acunzo (2014).

Para responder nossas perguntas e cumprir nossos objetivos realizamos uma análise multidimensional aditiva, aplicando ao nosso corpus todas as dimensões de variação identificadas por Berber Sardinha, Kauffmann e Acunzo (2014), quais sejam: *(1) Oral versus literate discourse, (2) Argumentation, (3) Involved versus informational production, (4) Directive discourse, (5) Future versus past time orientation e (6) Reported discourse.* Em seguida, contrastamos nosso corpus globalmente e por seções com os registros do Corpus Brasileiro de Variação e Registro (BERBER SARDINHA, KAUFFMANN, ACUNZO, 2014). Por fim, para complementar nossa análise os dados foram submetidos aos testes estatísticos: ANOVA e R^2.

Diante dos nossos resultados concluímos que os textos legais são um registro segundo a perspectiva de Biber e Conrad (2009) uma vez que nosso corpus apresenta cargas fatoriais únicas em cada dimensão, tendo sido ainda constatado através dos testes estatísticos uma variação e diferenças estatisticamente significativas entre os registros analisados em todas as dimensões do estudo original. Além disso, constatou-se que as diferentes seções do nosso corpus apresentam cargas fatoriais distintas entre si em todas as dimensões estando mais próximas ou mais distantes umas das outras a depender da dimensão analisada. Em relação às características situacionais temos que o processo de criação de cada espécie normativa é distinto. Além disso, cada texto legal, independente da espécie, apresenta um contexto de comunicação, propósito comunicacional e assunto próprios.

Contatou-se que os textos legais são marcados pelo discurso letrado, sendo predominantemente informativos, recorrendo frequentemente ao futuro para transmitir informações relativas às normas e suas aplicações. Essas características são observadas nos textos a partir da presença expressiva de verbos no futuro, verbos modais (dever e poder), conjunções coordenadas, orações subordinadas, construções passivas sem agente, substantivos, adjetivos, preposições, artigos definidos e

nominalizações que refletem uma grande riqueza lexical e densidade informacional com um discurso voltado para o futuro. Esses traços linguísticos correspondem a alguns traços observados em estudos prévios a respeito dessa variedade, inclusive em outras línguas, como a nominalização, a riqueza lexical e o uso de passivas, o que nos leva a hipnotizar que essas características seriam possíveis universais da linguagem jurídica. Mais pesquisas nesse sentido se fazem necessárias para confirmar essa hipótese.

Diante dos resultados obtidos hipotiza-se ainda que as seções do nosso corpus são sub registros do registro textos legais. Essa hipótese se baseia no fato de que foram constatadas diferenças significativas entre as características situacionais, linguísticas e carga fatorial das seções do corpus em todas as dimensões analisadas. Essas diferenças, juntamente com as características comuns compartilhadas por todas as seções apontam para a possibilidade de as seções do nosso corpus serem subregistros dos textos legais, sendo necessários mais estudos a esse respeito para confirma-la.

Durante a realização desta pesquisa nos deparamos com vários desafios e problemas metodológicos a começar pela própria arquitetura do corpus que, apesar de todo o cuidado e estudo empregado na sua concepção, ainda assim apresentou falhas quanto ao equilíbrio e balanceamento das seções. Estas nos impediram de realizar a análise multidimensional completa, nossa proposta inicial, e nos levaram a reformular nossos objetivos e perguntas de pesquisa. Para resolver as questões metodológicas que foram se apresentando durante a pesquisa, decisões tiveram que ser tomadas para que fosse possível levar a pesquisa a término. Apesar de termos obtido resultados interessantes e válidos para a caracterização dos textos legais que esperamos contribuam com o campo e inspirem pesquisas futuras, nem todas as decisões levaram a resultados satisfatórios, não tendo sido possível realizar uma pesquisa tão refinada quanto pretendíamos inicialmente.

Quanto aos desdobramentos dessa pesquisa, pretende-se refazê-la considerando todas as variáveis do estudo original, o que não foi possível no momento, pois, devido a problemas técnicos, nosso corpus não foi etiquetado com todas as variáveis do estudo original e não tivemos tempo hábil para refazer esse processo. Visamos ainda repensar a arquitetura do nosso corpus de forma a torná-lo mais equilibrado e balanceado. Após chegarmos a uma nova configuração do corpus e fizermos os ajustes necessários

daremos prosseguimento à sua compilação e, ao completá-la, disponibilizá-lo gratuitamente online, juntamente com sua descrição e o guia de anotação por nós elaborado (vide APÊNDICE A) contribuindo com mais pesquisas sobre a linguagem jurídica brasileira voltada para os textos legais.

Espera-se ainda submeter nosso corpus a uma AMD completa, visando responder aos questionamentos que surgiram durante a realização dessa pesquisa e refinar mais as análises realizadas para estudar mais detalhadamente esse registro, assim como comparar com os resultados da presente pesquisa para observar se eles se verificam ou não. A esse respeito, buscamos ainda identificar suas dimensões de variação a fim de verificar se estas coincidem com aquelas do PB ou se existem outras dimensões intrínsecas aos textos legais. Por fim, buscamos determinar se as seções do nosso corpus podem ser consideradas sub-registros dos textos legais.

REFERÊNCIAS

AQUINO, R.; DOUGLAS, W. **Manual de português e redação jurídica**. 6. ed. Niterói: Impetus, 2017.

BERBER SARDINHA, Tony. **Linguística de Corpus.** Barueri, SP: Manole, 2004.

BERBER SARDINHA, Tony. Variação entre registros da Internet. *In:* SHEPHERD, T. G.; SALIÉS, T. G. (Eds.). **Linguística da Internet**. São Paulo: Contexto, 2013. p. 55–85.

BERBER SARDINHA, Tony. **Pós-processador PT Tag Count**. 2013.

BERBER SARDINHA, Tony. A abordagem metodológica da Análise Multidimensional. **Gragoatá**. Niterói, n. 29, p. 107-125, 2. sem. 2010. Disponível em: https://periodicos.uff.br/gragoata/article/view/33077. Acesso em: 19 jan. 2022.

BERBER SARDINHA, Tony; PINTO, Marcia Veirano (eds.). **Multi-dimensional analysis**: 25 years on a tribute to Douglas Biber. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2014.

BERBER SARDINHA, Tony *et al.* Adding Registers to a Previous Multi-Dimensional Analysis. *In*: BERBER SARDINHA, T.; VEIRANO PINTO, M. (Eds.). **Multi-Dimensional Analysis: Research Methods and Current Issues**. London: Bloomsbury, 2019. p. 165–186.

BERBER SARDINHA, Tony; PINTO, Marcia Vierano (eds.). **Multi-Dimensional Analysis:** Research Methods and Current Issues. Londres: Bloomsbury Academic, 2019.

BERBER SARDINHA, Tony; KAUFFMANN, Carlos; ACUNZO, Cristina Mayer. Dimensions of register variation in Brazilian Portuguese. *In*: PINTO, Marcia Veirano (eds.). **Multi-dimensional analysis**: 25 years on a tribute to Douglas Biber. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2014.

BIBER, Douglas. **Variations across speech and writing.** Cambridge: CUP.1988.

BIBER, Douglas. Representativeness in Corpus Design. **Literary and Linguistic Computing**, v. 8, n. 4, Oxford: Oxford University Press, p. 243-257, 1993.

BIBER, Douglas. **Dimensions of register variation: A cross-linguistic perspective.** Cambridge: CUP.1995.

BIBER, D; CONRAD, S**. Register, genre, and style**. Cambridge: CUP. 2009.

BICK, E.. PALAVRAS, a constraint grammar-based parsing system for Portuguese. In T. SARDINHA, Berber, e FERREIRA, T. São Bento (Eds.), **Working with Portuguese corpora** (pp. 279–302). London: Bloomsbury, 2014.

BICK, E.. **The Parsing System "Palavras":** Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Dr.phil. thesis. Aarhus University. Aarhus, Denmark: Aarhus University Press. November 2000.

BITTAR, E. C. B. **Linguagem**. 4. ed. São Paulo: Saraiva, 2009.

BRASIL. **Constituição da República Federativa do Brasil**, de 10 de outubro de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 10 de set. 2020.

BRASIL. **Lei nº 12.651, de 25 de maio de 2012** (Código Florestal). Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm. Acesso em: 10 de set. 2020.

BRASIL. **Manual de Redação**. Brasília: Câmara dos Deputados, 2004.

BRASIL. **Manual de Redação da Presidência da República / Casa Civil**. 3. ed., rev., atual. e ampl. Brasília: Presidência da República, Subchefia de Assuntos Jurídicos, 2018.

BRASIL. **Manual de Redação Parlamentar e Legislativa**. Brasília: Senado Federal, Consultoria Legislativa, 2006.

BREZINA; Vaclav. **Statistics in Corpus Linguistics. A Practical Guide.** Cambridge: Cambridge University Press, 2018.

BURNARD, Lou. Metadata for corpus work. In: WYNNE, M (eds.). **Developing Linguistic Corpora**: a Guide to Good Practice. Oxford, 2005. Disponível em: https://users.ox.ac.uk/~martinw/dlc/chapter3.htm. Acesso em 30 set. 2022.

CAO, Y.; XIAO, R. A multi-dimensional contrastive study of English abstracts by native and non-native writers. **Corpora**, 8(2), 209–234, 2013.

CARAPINHA, Conceição. A linguagem jurídica. Contributos para uma caracterização dos Códigos Legais. **REDIS: Revista de Estudos do Discurso**, nº 7, 2018. Disponível em: https://ojs.letras.up.pt/index.php/re/article/view/6200. Acesso em: 10 set. 2021.

CARVALHO, L. **Inglês Jurídico Tradução e Terminologia**. São Paulo: Lexema, 2014.

CARVALHO, L. Os dicionários jurídicos bilíngües e o tradutor - dois binômios em Direito Contratual. **TradTerm**, v. 12, p. 309-347, 2006. DOI https://doi.org/10.11606/issn.2317-9511.tradterm.2006.46903. Acesso em: 10 set. 2021.

CASTRO, Marcílio Moreira de. **Dicionário de direito, economia e contabilidade**: português-inglês/ inglês-português. 4. ed. Rio de Janeiro: Forense, 2013.

CHOVANEC, Jan. **Grammar in the Law.** In: CHAPELLE, Carol A.(ed.). The Encyclopedia of Applied Linguistics. New Jersey: Blackwell Publishing Ltd, 2013. Disponível em: https://www.academia.edu/3389303/Grammar_in_the_Law. Acesso em: 6 set. 2020.

CONRAD, S.; BIBER, D. Multidimensional methodology and the dimensions of register variation in English. In: CONRAD, S.; BIBER, D. (eds.). **Variation in English: multidimensional studies**, 2001, p. 13 - 42. Harlow: Pearson Education.

DAHLMAN, Roberta Colonna. **Specialità del linguaggio giuridico italiano**. Stockholm University, 2006. Disponível em: https://www.researchgate.net/publication/27818945_Specialita_del_linguaggio_giuridico_italiano. Acesso em: 6 set. 2020.

DAMIÃO, R. T.; HENRIQUES, A. **Curso de português jurídico**. 14. ed. São Paulo: Atlas, 2020.

DELFINO, Maria Claudia Nunes. Análise Multidimensional: os números na linguística. **Cadernos de Linguística**. v. 2, n. 4, 2021. Disponível em: https://www.researchgate.net/publication/354864891_Analise_multidimensional_os_numeros_na_Linguistica. Acesso em: 19 jan. 2022.

DINIZ, M. H. **Dicionário jurídico**. São Paulo: Saraiva, 1998.

FANEGO, Teresa; RODRÍGUEZ-PUENTE, Paula (eds). **Corpus-based Research on Variation in English Legal Discourse**. Amsterdam/Philadelphia: John Benjamins Publishing Company, 2019.

FERRARI, L.A.; BOSSAGLIA, G. The C-ORAL-BRASIL project for Brazilian Portuguese spoken corpora. **Caplletra,** Valencia, n.69, p. 201-220, 2020. Disponível em: https://ojs.uv.es/index.php/caplletra/article/view/17269. Acesso em: 24 jan. 2022.

FERRARI, L.A.; CUNHA, E. L. T. P. Reflexões metodológicas sobre datasets e linguística de corpus: uma análise preliminar de dados legislativos. **Domínios de Lingu@gem,** [S. l.], v. 16, n. 4, p. 1571–1607, 2022. Disponível em: https://seer.ufu.br/index.php/dominiosdelinguagem/article/view/64146. Acesso em: 23 de set. 2022.

FERRARI, Lúcia de Almeida; MARQUES, Carolina Godoi de Faria. O LEX-BR-Ius: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras. **ANTARES**, v.14, n.34, 2022. Disponível em: http://www.ucs.br/etc/revistas/index.php/antares/article/view/11150/5328. Acesso em: 19 dez. 2022.

GIAMPIERI, Patrizia. **Online Parallel and Comparable Corpora for Legal Translations**. In: Altre modernità / Otras modernidades / Autres modernités / Other Modernities, N. 20, 237-252, Milano, 2018. Disponível em: <

https://www.researchgate.net/publication/329365608_Online_Parallel_and_Comparable_Corpora_for_Legal_Translations>. Acesso em: 6 set. 2020.

GRIES, Stefan Th.; BEREZ, Andrea L.. Linguistic Annotation in/for Corpus Linguistics. In: IDE, N.; PUSTEJOVSKY, J. (eds.). **Handbook of Linguistic Annotation**. Springer Science+Business Media Dordrecht, p. 379-409, 2017.

GONÇALVES, Carlos Roberto. **Direito civil parte geral**. São Paulo: Saraiva, 2018

GOTTI, Maurizio. The translation of legal texts: Interlinguistic and intralinguistic perspectives. ESP Today, 4(1), p. 5–21, 2016.

GOŹDŹ-ROSZKOWSKI, Stanisław. **Patterns of Linguistic Variation in American Legal English**: A Corpus-Based Study. Frankfurt am Main: Peter Lang, 2011.

GOŹDŹ-ROSZKOWSKI, Stanisław. Corpus Linguistics in Legal Discourse. **International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique**, 34, 1515-1540, 2021. Disponível em: https://link.springer.com/article/10.1007/s11196-021-09860-8#:~:text=Thus%2C%20Corpus%20Linguistics%20is%20becoming,most%20relevant%20and%20complete%20results. Acesso em: 6 set. 2020.

GOŹDŹ-ROSZKOWSKI, Stanisław. Frequent phraseology in contractual instruments: A corpus-based study. In: GOTTI, Maurizio e GIANNONI, Davide Simone (eds). **New Trends in Specialized Discourse Analysis**. Bern: Peter Lang, 2006, p. 147–161.

GOŹDŹ-ROSZKOWSKI, Stanisław. Legal Language. *In*: CHAPELLE, Carol A. (eds.). **The Encyclopedia of Applied Linguistics.** John Wiley e Sons, 2012. p. 3281-3287.

GOŹDŹ-ROSZKOWSKI, Stanisław; PONTRANDOLFO, Gianluca. Facing the facts: Evaluative patterns in English and Italian judicial language. In: BHATIA, Vijay; GARZONE, Giuliana; SALVI, Rita (eds). **Language and Law in Professional Discourse**. Newcastle upon Tyne: Cambridge Scholars, 2014, p. 10–28.

GOŹDŹ-ROSZKOWSKI, Stanisław; PONTRANDOLFO, Gianluca. **Legal Praseology Today: Corpus-based Applications Across Legal Languages and Genres**. In: International Journal of Specialized Comunication, v. XXXVII, 2015. Disponível em: https://www.academia.edu/18714805/Legal_Phraseology_Today_Corpus_based_Applications_Across_Legal_Languages_and_Genres. Acesso em: 6 set. 2020.

GOŹDŹ-ROSZKOWSKI, Stanisław; WITCZAK-PLISIECKA, Iwona (eds). **Special Issue on Legal Terminology: Approaches and Applications**. Research in Language, vol. 9.1, Łódź: Łódź University Press, 2011. Disponível em: https://www.academia.edu/5789565/Special_Issue_on_Legal_Terminology_Approaches_and_Applications. Acesso em: 6 set. 2020.

GUIMARÃES, D. T. **Dicionário técnico jurídico**. São Paulo: Rideel, 2013.

HALLIDAY, M. A. K. **An Introduction to Functional Grammar.** London: Edward Arnold, 1985.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. **ICAME Journal**, 38 (1), 73–103, 2014. Disponível em: https://doi.org/10.2478/icame-2014-0004. Acesso em: 20 jul. 2021.

HO, Dong. **Notepad++** (Version 7.8.9) [Computer Software]. 2020. Disponível em:https://notepad-plus-plus.org/downloads/v7.8.9/ **.** Acesso em: 5 de mar. 2020.

IBM. **IBM SPSS Statistics** (Version 23) [Computer Software]. IBM, 2015. Disponível em: https://www.ibm.com/support/pages/downloading-ibm-spss-statistics-23. Acesso em: 20 de jan. 2022.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; FINATTO, M. J. B. **Dicionário de Direito Ambiental**. 2. ed. Rio de Janeiro: Lexikon, 2008.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; FINATTO, M. J. B.; REUILLARD, P. C. R. **Glossário de Gestão Ambiental**. São Paulo: Disal Editora, 2006.

KRUGER, Alet; WALLMACH, Kim; MUNDAY, Jeremy (eds.). **Corpus-Based Translation Studies Research and Application**. New York: Continuum International Publishing Group, 2011.

LEECH, Geoffrey. Adding Linguistic Annotation. In: WYNNE, M (eds.). **Developing Linguistic Corpora:** a Guide to Good Practice. Oxford, 2005. Disponível em: https://users.ox.ac.uk/~martinw/dlc/chapter2.htm. Acesso em 30 set. 2022.

LENZA, Pedro. **Direito Constitucional esquematizado**. São Paulo: Saraiva, 2020.
MARTIM, H. de *et al*. Base de normas jurídicas brasileiras: uma iniciativa de open government data. **Perspectivas em Ciência da Informação**, v. 23, n. 4, p. 133, 2018.

MACIEL, A. M. B. **Para o reconhecimento da especificidade do termo jurídico**. 2001. 291 f. Tese. (Doutorado em Estudos da Linguagem) – Programa de Pós-Graduação em Letras. Universidade Federal do Rio Grande do Sul, 2001.

MCENERY, Tony; HARDIE, Andrew. **Corpus Linguistics: Method, Theory and Practice**. Cambridge: Cambridge University Press, 2012.

MCENERY, Tony; WILSON Andrew. **Corpus Linguistics**. Edinburgh: Endiburgh UP, second edition, 2001.

MCENERY, Tony; XIAO, Richard; TONO, Yukio. Unit 3 Corpus markup. In: MCENERY, Tony; XIAO, Richard; TONO, Yukio. **Corpus-based Language Studies***:* An Advanced Resource Book. Londres: Routledge Applied Linguistics Series, 2005. Disponível em:

https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A03.pdf. Acesso em 30 set. 2022.

MCENERY, Tony; XIAO, Richard; TONO, Yukio. Unit 4 Corpus annotation. In: MCENERY, Tony; XIAO, Richard; TONO, Yukio. **Corpus-based Language Studies**: An Advanced Resource Book. Londres: Routledge Applied Linguistics Series, 2005. Disponível em: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A04.pdf. Acesso em 30 set. 2022.

MELLO, Heliana. Os corpora orais e o c-oral-brasil. In: RASO, Tommaso; MELLO, Heliana (eds.). **C-ORAL-BRASIL I:** Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012, p. 31-54.

ONESTI, Cristina. Methodology for building a text-structure oriented legal corpus. **Comparative Legilinguistics**, p. 37- 48, 2011.

PETRI, M. J. C. **Manual de linguagem jurídica**. 3. ed. São Paulo: Saraiva, 2017.

PONTRANDOLFO, Gianluca. Legal Corpora: an overview**. Rivista Internazionale di Tecnica della Traduzione**, p. 121-136, Trieste, 2012. Disponível em: https://www.openstarts.units.it/bitstream/10077/9783/1/12Pontrandolfo.pdf**.** Acesso em: 6 set. 2020.

RASO, T. O corpus C-ORAL-BRASIL. In: RASO, Tommaso; MELLO, Heliana (eds.). **C-ORAL-BRASIL I:** Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012, p. 55-90.

RICHARD, Isabelle. Is legal lexis a characteristic of legal language? **Lexis** [Online], 11, 2018. Disponível em: https://www.researchgate.net/publication/324949333_Is_legal_lexis_a_characteristic_of_legal_language**.** Acesso em: 6 set. 2020.

ROCHA, Jean Michel Pimentel**. Fraseologia jurídico-comercial e proposta de um glossário de colocações especializadas trilíngue baseado em corpus**. Dissertação apresentada à Universidade Estadual Paulista, 2017. Disponível em: https://repositorio.unesp.br/handle/11449/149766. Acesso em: 6 set. 2020

ROSSINI FAVRETTI, R.; TAMBURINI, F.; MARTELLI, E.. Words from Bononia Legal Corpus. In: Text Corpora and Multilingual Lexicography (W.Teubert ed.), John Benjamins, 2007, pp. 11-30.

ROSSINI FAVRETTI, R.; TAMBURINI, F.; MARTELLI, E.. Words from Bononia Legal Corpus. International Journal of Corpus Linguistics, Vol. 6 (Special Issue), 2001, 13-34.

SANTOS, W. **Dicionário jurídico brasileiro**. Belo Horizonte: Del Rey, 2001.

SILVA, José Afonso da. **Curso de direito constitucional positivo.** Salvador: Juspodivm, 2020.

SINCLAIR, John. Corpus and Text. In: WYNNE, M (eds.). **Developing Linguistic Corpora**: a Guide to Good Practice. Oxford: Oxbow Books, 2005, p. 1-16. Disponível em: 6 set. 2020.

STEFANOWITSCH, Anatol. **Corpus Linguistics**: A Guide to the Methodology. Berlin: Language Science Press, 2020.

SVOBODOVÁ, I. Modalidade não epistêmica na linguagem jurídica: um estudo contrastivo. **Caligrama**, Belo Horizonte, v. 22, n. 2, p. 103-133, 2017. DOI http://dx.doi.org/10.17851/2238-3824.22.2.103-133

SWALES, John. **Genre Analysis: English in Academic and Research Settings**. Cambridge: CUP, 1990.

TEIXEIRA, W. R.; LIMA, J. A. O.; ARAUJO, L. C.; VIERO, D. M.; SANTANA, F. F.; HERINGER, F. R. A.; MARTIM, H.; VIEIRA FILHO, J. J. Exemplo de extração de definições em textos articulados de normas jurídicas com o apoio do processamento de linguagem natural. **Cadernos de Informação Jurídica**, v. 6, n. 1, p. 49-64, 2019. Disponível em: http://hdl.handle.net/20.500.11959/brapci/119039. Acesso em: 20 dez. 2021.

TIERSMA, P.. **Legal Language**. The University of Chicago Press, 1999.

TRKLJA, Aleksandar; MCAULIFFE, Karen. The European Union Case Law Corpus (EUCLCORP): A Multilingual Parallel and Comparative Corpus of EU Court Judgments (March 5, 2018). In: FRANK, A. U. *et al* (eds.). **Proceedings of the Second Workshop on Corpus-Based Research in the Humanities: CRH-2**. Gerastree Proceedings, Vol. 1, p. 217-226. Disponível em: https://ssrn.com/abstract=3134457. Acesso em: 06 nov. 2022.

ZAMPIERI, Marcos; BECKER, Martin. Colonia: Corpus of Historical Portuguese. 2013. Disponível em: https://www.marcoszampieri.com/papers/colonia2013.pdf. Acesso em: 09 jun. 2022.

APÊNDICE A – Guia de marcação do *LEX-BR-lus*

GUIA DE MARCAÇÃO DOS TEXTOS DO *LEX-BR-lus*

Belo Horizonte

2022

SUMÁRIO

1 MARCAÇÃO XML

O corpus *LEX-BR-Ius* (FERRARI e MARQUES, em preparação) possui uma marcação textual XML seguindo a proposta de XML modesto de Hardie (2014). Trata-se de uma proposta de marcação XML mais simples em relação ao XML completo (*full XML*) que é o sistema de marcação mais difundido globalmente. Assim como a marcação XML completa, trata-se de um sistema que permite adicionar ou isolar informações no texto ou seus metadados por meio de etiquetas que são delimitadas por parênteses angulares (< e >) (HARDIE, 2014).

Criada pensando nas necessidades de quem compila corpora essa proposta se destaca pela simplicidade e flexibilidade, não tendo a complexidade de regras e critérios quanto o XML completo. Permite tanto o uso das etiquetas já convencionadas desse sistema, “de facto standard tags”, quanto a criação de etiquetas próprias que atendam às necessidades do pesquisar, ou seja, a possibilidade de personalização das etiquetas. Dessa forma a marcação pode ser adaptada aos mais diversos tipos de corpora e pesquisas linguísticas, mas ao mesmo tempo ser de fácil aplicação, não sendo preciso ter conhecimentos de computação para realizá-la. (HARDIE, 2014).

Para tanto, segundo Hardie (2014), determinadas regras devem ser seguidas, quais sejam:

- Codificação UTF-8;
- Extensão .xml;
- As etiquetas vêm entre parênteses angulares, ex.: <etiqueta>;
- As etiquetas são sensíveis a maiúsculas e minúsculas, recomenda-se usar sempre letra minúscula e ignorar acentuação gráfica;

Obs.: As informações que vêm entre as aspas nas etiquetas são exceção da regra acima, nelas é permitido letras maiúsculas, minúsculas, acentuação, números e caracteres especiais;

- É recomendável que as etiquetas tenham apenas uma palavra e contenham apenas letras;

- Todas as etiquetas devem ser fechadas, para isso usar “/” entre o primeiro parêntese angular e a etiqueta para fechar no caso de etiquetas de região, ex.: <etiqueta> [texto] </etiqueta> e entre a etiqueta e o último parêntese angular no caso de etiquetas pontuais, ex.: <etiqueta/>;
- Deve haver uma etiqueta que delimite o conteúdo de todo o arquivo, ex.: <texto> [texto] </texto>;
- Dar um espaço entre a etiqueta e a palavra ou etiqueta anterior/seguinte, ex.:

Errado: <artigo>Art. 1º Não há crime sem lei anterior que o defina. Não há pena sem prévia cominação legal<modificacao> (Redação dada pela Lei nº 7.209, de 11.7.1984)</modificacao></artigo>

(BRASIL, 1940)

Certo: <artigo> Art. 1º Não há crime sem lei anterior que o defina. Não há pena sem prévia cominação legal. <modificacao> (Redação dada pela Lei nº 7.209, de 11.7.1984) </modificacao> </artigo>

(BRASIL, 1940)

- Todos os "espaços brancos" contam como um único espaço, ex.: <etiqueta> [texto] </etiqueta> é igual a <etiqueta> [texto] </etiqueta>;
- A quebra de linha é irrelevante, ex.:

<etiqueta> [texto] </etiqueta> é a mesma coisa de:

<etiqueta>

[texto]

</etiqueta>;

- As partes do texto devem ser marcadas com etiquetas de região, ex.:<etiqueta> [texto] </etiqueta>
- Os aspectos pontuais do texto devem ser marcados com etiquetas pontuais, ex.:<etiqueta/>
- As etiquetas devem estar perfeitamente aninhadas: as etiquetas que delimitam partes não devem se sobrepor, sempre fechar uma etiqueta antes de abrir outra, ex.: <parte> [texto] </parte> <parte> [texto] </parte>

- Etiquetas com atributos são utilizadas para especificar ou explicar algo sobre a etiqueta. Elas devem ser marcadas com a etiqueta, seu “atributo”, ou seja, o tipo de informação e o “valor”, que é a informação que queremos colocar na etiqueta. Elas vêm como: <etiqueta x= “y”> [texto] </etiqueta>, se marcarem partes do texto, e <etiqueta x= “y”/> se forem aspectos pontuais;
- Substituir os caracteres especiais caso apareçam no texto:
  - & por &
  - < por <
  - > por >
- Para verificar se está tudo certo com a marcação clicar no nome do arquivo com o botão direito do mouse, selecionar “abrir com” e escolher um navegador, ex.: “Internet Explorer”, “Google Chrome”, “Mozilla”, etc. Se estiver tudo certo o arquivo vai abrir no navegador, se não aparecerá uma mensagem de erro indicando qual é o erro e a linha e a coluna onde ele está.

No nosso corpus a marcação será feita com etiquetas por nós elaboradas para identificar os textos, armazenar seus metadados em forma de cabeçalho, separar as seções dos textos normativos e seus artigos. Nas próximas seções apresentamos as etiquetas que serão utilizadas.

2 ETIQUETAS

2.1. ID

Código de identificação do texto. Formado pelas iniciais da espécie normativa em maiúsculo, número, dia, mês e ano da publicação. Deve ser incluída em todos os arquivos (original, limpo *(raw text), xml e header*).

Nos arquivos: original, limpo e XML essa etiqueta delimitará todo o conteúdo do texto. Logo ela virá na primeira linha do arquivo no editor de texto e deve ser fechada na última linha do texto, sendo assim uma etiqueta de região que tem como atributo a ID do texto.

Etiqueta: <texto id = “INICIASNº_DATA”>

Exemplo:

<texto id = “LO11.101_9.02.2005”>

[texto]

</texto>

(BRASIL, 2005)

No arquivo do cabeçalho, essa etiqueta será uma etiqueta pontual que especifica qual é a ID do texto ao qual o cabeçalho se refere. Logo, sua sintaxe é diversa da etiqueta usada nos outros arquivos, tem a “/” no final e não deve ser fechada com outra etiqueta. Ela já está incluída no modelo de cabeçalho na posição correta e com a sintaxe certa, basta apenas preenchê-la. A etiqueta que delimitará todo o conteúdo do cabeçalho é a própria etiqueta de cabeçalho.

Etiqueta: <texto id = “INICIASNº_DATA”/>

Exemplo:

<cabecalho>

<texto id = “LO11.101_9.02.2005”/>

...

</cabecalho>

(BRASIL, 2005)

Iniciais dos textos legais a serem usadas na formação da ID:

- Constituição =CR
- Emendas à Constituição = EC
- Códigos = C
- Leis complementares= LC
- Leis ordinárias = LO
- Medidas provisórias = MP
- Decretos = D
- Estatutos = E

2.2 Etiquetas do cabeçalho

Os metadados do texto legal vão ser incluídos em forma de cabeçalho em arquivo separado do texto legal e também em tabela Excel. O cabeçalho vai conter as seguintes informações: ID, Norma; Ementa; Tipo de norma; Assunto; Área do direito; Presidente em exercício; Data da promulgação; Data de publicação; Início da vigência; Alterações; Número de artigos; Número de palavras; Fonte; Data da extração; Subcorpus; Pesquisador. Essas informações serão extraídas do próprio texto legal e da tabela fornecida pelo planalto com as informações de cada texto legal (para acessá-la basta abrir a norma no site e clicar no seu nome, isso abrirá a tabela).

Obs.: Quando a informação não estiver disponibilizada no texto legal nem na tabela colocar como atributo: "informação não fornecida"

Ex.: <area v = "informação não fornecida"/>

Etiqueta: <cabecalho>

Exemplo:

<cabecalho>

<texto id = “LO11.101_9.02.2005”/>

<norma v = “LEI Nº 11.101 , DE 9 DE FEVEREIRO DE 2005”/>

<ementa v = “Regula a recuperação judicial, a extrajudicial e a falência do empresário e da sociedade empresária”/>

<tipo v = “lei ordinária”/>

<assunto v = "dispositivos, normas, extinção, concordata, objetivo, projeto, recuperação, situação financeira, empresa. Criação, comitê, objetivo, competência, homologação, plano, prazo determinado, recuperação, situação financeira, empresa. Requerimento, falência, prerrogativa, juiz, descumprimento, prazo, recuperação, plano, situação financeira, empresa."/>

<area v = “direito comercial; sociedades comerciais”/>

<presidente v = “Luiz Inácio Lula da Silva”/>

<promulgacao v = “09/02/2005”/>

<publicacao v = “09/02/2005”/>

<vigencia v = “120 (cento e vinte) dias após sua publicação”/>

<alteracao v = "LEI 11.127, DE 28/06/2005: ACRESCE PARÁGRAFO 5º AO ART. 192; LEI 11.196, DE 21/11/2005: ALTERA O ART. 199; LEI 12.873, DE 24/10/2013: ACRESCE PAR. 2º, RENUMERANDO-SE O ATUAL PARÁGRAFO ÚNICO PARA 1º DO ART. 48; LCP 147, DE 07/08/2014: ALTERA ARTS. 24, 26, 41, 45, 48, 68, 71, 72 E 83; MPV 881, DE 30/04/2019: ACRESCE ART. 82-A; LEI 14.112, de 24/12/2020: Altera arts. 6º, 10, 14, 16, 22, 24, 36, 39, 48, 49, 50, 51, 52, 54, 56, 58, 59, 60, 61, 63, 66 67, 69, 73, 75, 83, 84, 86, 99, 104, 131, 141, 142, 143, 145, 156, 158, 159, 161, 163, 164, 168, 189, 191, E 196. Acresce arts. 6º-A,

6º-B, 6º-C, 7º-A, 20-A, 20-B, 20-C, 20-D, 45-A, 48-A, 50-A, 51-A, 56-A, 58-A, 60-A, 66-A, 69-A, 69-B, 69-C, 69-D, 69-E, 69-F, 69-G, 69-H, 69-I, 69-J, 69-K, 69-L, 70-A, 82-A, 114-A, 144-A, 159-A, 167-A, 167-B, 167-C, 167-D, 167-E, 167-F, 167-G, 167-H, 167-I, 167-J, 167-K, 167-L, 167-M, 167-N, 167-O, 157-P, 167-Q, 167-R, 167-S, 167-T, 167-U, 167-V, 167-2, 167-X, 167-Y, 189-A, 193-A. Revoga § 7º do art. 6º; incisos IV e V do caput, com as respectivas alíneas, e § 4º, todos do art. 83; inciso I do caput do art. 84; parágrafo único do art. 86; incisos II e III do caput e §§ 1º, 2º, 4º, 5º e 6º, todos do art. 142; §§ 2º e 3º do art. 145; incisos III e IV do caput do art. 158; art. 157; e § 2º do art. 159.”/>

<artigos v = “201”/>

<palavras v = “33.223”/>

<fonte v = “https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11101.htm” />

<extracao v = “14/07/2021”/>

<subcorpus v = "leis ordinárias"/>

<pesquisador v = “carolina marques”/>

</cabecalho>

(BRASIL, 2005)

2.2.1 Norma

Delimita o nome do tipo normativo.

Etiqueta: <norma v = “x”/>

Exemplo: <norma v = “LEI Nº 11.101, DE 9 DE FEVEREIRO DE 2005”/>

(BRASIL, 2005)

2.2.2 Ementa

Delimita o assunto ou pequena explicação sobre a norma.

Etiqueta: <ementa v = “x”/>

Exemplo:

<ementa v = “Regula a recuperação judicial, a extrajudicial e a falência do empresário e da sociedade empresária”/>

(BRASIL, 2005)

2.2.3 Tipo de norma

Espécie normativa, é a classificação pelo tipo de norma. Usaremos:

- constituição;
- emenda à constituição;
- código;
- lei ordinária;
- lei complementar;
- medida provisória;
- decreto;
- estatuto.

Etiqueta: <tipo v = “x”/>

Exemplo:

<tipo v = “lei ordinária”/>

(BRASIL, 2005)

2.2.4 Assunto

Assuntos e conteúdos abrangidos pelo texto legal. Disponibilizado na ficha técnica do texto no site do planalto (http://www4.planalto.gov.br/legislacao/).

Etiqueta: <assunto v = “x”/>

Exemplo:

<assunto v = "DISPOSITIVOS, NORMAS, EXTINÇÃO, CONCORDATA, OBJETIVO, PROJETO, RECUPERAÇÃO, SITUAÇÃO FINANCEIRA, EMPRESA. CRIAÇÃO, COMITÊ, OBJETIVO, COMPETÊNCIA, HOMOLOGAÇÃO, PLANO, PRAZO DETERMINADO, RECUPERAÇÃO, SITUAÇÃO FINANCEIRA, EMPRESA. REQUERIMENTO, FALÊNCIA, PRERROGATIVA, JUIZ, DESCUMPRIMENTO, PRAZO, RECUPERAÇÃO, PLANO, SITUAÇÃO FINANCEIRA, EMPRESA.”/>

(BRASIL, 2005)

2.2.5 Área do direito

Em qual (s) área (s) do direito o texto legal se enquadra segundo a ficha técnica do site do planalto em “classificação de direito”.

Etiqueta: <area v = “x”/>

Exemplo:

<area v = “direito comercial; sociedades comerciais”/>

(BRASIL, 2005)

2.2.6 Presidente em exercício

Nome do presidente em exercício na época da promulgação do texto legal. Essa informação está na tabela do site em “Chefe de Governo”

Etiqueta: < presidente v = “x”/>

Exemplo:

<presidente v = “Luiz Inácio Lula da Silva”/>

(BRASIL, 2005)

2.2.7 Data da promulgação

Data em que a lei foi promulgada/assinada. Esta informação está na tabela do site em “Data de assinatura”

Etiqueta: <promulgacao v = “x”/>

Exemplo:

<promulgação v = “09 de Fevereiro de 2005”/>

(BRASIL, 2005)

2.2.8 Data de publicação

Data em que a lei foi publicada no Diário Oficial da União.

Etiqueta: <publicacao v = “x”/>

Exemplo:

<publicacao v = “09 de Fevereiro de 2005”/>

(BRASIL, 2005)

2.2.9 Início da vigência

Data a partir da qual a lei se torna vigente, ou seja, em que entra em vigor. Essa informação geralmente consta no último artigo da lei. Outra forma de obtê-la é clicar no hiperlink “Vigência” logo no início da página da lei, alinhado à esquerda entre o título da lei e a frase de abertura.

Etiqueta: <vigencia v = “x”/>

Exemplo:

<vigencia v = “120 (cento e vinte) dias após sua publicacao”/>

(BRASIL, 2005)

2.2.10 Alterações

Normas que modificaram o texto legal e quais pontos foram modificados. Consultar ficha técnica da lei no site do planalto para obter essas informações. Caso não haja, colocar “não”.

Obs.: colocar todas as alterações em um único parágrafo e separá-las com ponto e vírgula.

Obs.: colocar apenas as alterações, caso haja uma seção de “correlações” esta não deve ser incluída.

Etiqueta: <alteracao v = “x”/>

Exemplo:

<alteracao v = "LEI 11.127, DE 28/06/2005: ACRESCE PARÁGRAFO 5º AO ART. 192; LEI 11.196, DE 21/11/2005: ALTERA O ART. 199; LEI 12.873, DE 24/10/2013: ACRESCE PAR. 2º, RENUMERANDO-SE O ATUAL PARÁGRAFO ÚNICO PARA 1º DO ART. 48; LCP 147, DE 07/08/2014: ALTERA ARTS. 24, 26, 41, 45, 48, 68, 71, 72 E 83; MPV 881, DE 30/04/2019: ACRESCE ART. 82-A; LEI

14.112, de 24/12/2020: Altera arts. 6º, 10, 14, 16, 22, 24, 36, 39, 48, 49, 50, 51, 52, 54, 56, 58, 59, 60, 61, 63, 66 67, 69, 73, 75, 83, 84, 86, 99, 104, 131, 141, 142, 143, 145, 156, 158, 159, 161, 163, 164, 168, 189, 191, E 196. Acresce arts. 6º-A, 6º-B, 6º-C, 7º-A, 20-A, 20-B, 20-C, 20-D, 45-A, 48-A, 50-A, 51-A, 56-A, 58-A, 60-A, 66-A, 69-A, 69-B, 69-C, 69-D, 69-E, 69-F, 69-G, 69-H, 69-I, 69-J, 69-K, 69-L, 70-A, 82-A, 114-A, 144-A, 159-A, 167-A, 167-B, 167-C, 167-D, 167-E, 167-F, 167-G, 167-H, 167-I, 167-J, 167-K, 167-L, 167-M, 167-N, 167-O, 157-P, 167-Q, 167-R, 167-S, 167-T, 167-U, 167-V, 167-2, 167-X, 167-Y, 189-A, 193-A. Revoga § 7º do art. 6º; incisos IV e V do caput, com as respectivas alíneas, e § 4º, todos do art. 83; inciso I do caput do art. 84; parágrafo único do art. 86; incisos II e III do caput e §§ 1º, 2º, 4º, 5º e 6º, todos do art. 142; §§ 2º e 3º do art. 145; incisos III e IV do caput do art. 158; art. 157; e § 2º do art. 159.”/>

(BRASIL, 2005)

2.11 Número de artigos

Quantidade de artigos presente na norma.

Etiqueta: <artigos v = “x”/>

Exemplo:

<artigos v = “201”/>

(BRASIL, 2005)

Obs.: caso na norma a numeração dos artigos recomece, colocar na etiqueta a soma dos números.

Exemplo.: No código comercial a contagem vai até o artigo 913 e depois recomeça do art 1º ao 30. No total temos 913 + 30 = 943, logo na etiqueta colocamos o resultado da soma: <artigos v = “943”/>

Obs.: Não colocar ponto para marcar casa decimal.

Exemplo: <artigos v =”2056”/>

(BRASIL, 2002)

2.2.12 Número de palavras

Número de palavras total do texto.

Obs.: não colocar ponto entre as casas dos números, ex.: 2.300 (errado) -> 2300 (certo).

Etiqueta: <palavras v = “x”/>

Exemplo: <palavras v = “33223”/>

(BRASIL, 2005)

2.2.13 Fonte

Site do qual foi extraído o texto normativo.

Etiqueta: <fonte v = “x”/>

Exemplo:

<fonte v = “https://www.planalto.gov.br/ccivil_03/_ato2004-2006/2005/lei/l11101.htm”/>

(BRASIL, 2005)

2.2.14 Data da extração

Data em que o texto normativo foi extraído do site.

Etiqueta: <extracao v = “xx/xx/xxxx”/>

Exemplo:

<extração v = “14/07/2021”/>

(BRASIL, 2005)

2.2.15 Subcorpus

Subcorpus do qual a norma faz parte. São eles:

- constituição;
- emendas à constituição;
- códigos;
- leis ordinárias;
- leis complementares;
- medidas provisórias;
- decretos;
- estatutos.

Etiqueta: <subcorpus v = “x”/>

Exemplo:

<subcorpus v = "leis ordinárias"/>

(BRASIL, 2005)

2.2.16 Pesquisador

Nome e último sobrenome do responsável pela anotação do texto. Tudo em letra minúscula.

Etiqueta: <pesquisador = “x”/>

Exemplo:

<pesquisador v ="carolina marques"/>

2.2.17 Revisor

Nome e último sobrenome do responsável pela revisão da anotação. Tudo em letra minúscula. Essa etiqueta deve ser acrescentada ao cabeçalho na hora da revisão.

Etiqueta: <revisor = “x”/>

Exemplo:

<revisor v ="carolina marques"/>

2.3 Etiquetas do texto

O texto vai ser delimitado pela etiqueta <texto id = “x”> (vide 3.1) e dentro dele usaremos as etiquetas para delimitar as partes do texto, os artigos e as modificações.

Etiqueta: <texto id = “x”>

Exemplo.:

<texto id = “LO11.101_9.02.2005”>

[texto]

</texto>

(BRASIL, 2005)

2.3.1 <norma>

Delimita o do título/nome do tipo normativo (tipo normativo, número e data da promulgação).

Etiqueta: <norma>

Exemplo:

<norma> LEI No 10.406, DE 10 DE JANEIRO DE 2002 </norma>

(BRASIL, 2002)

Exemplo:

<norma> DECRETO-LEI No 2.848, DE 7 DE DEZEMBRO DE 1940 </norma>

(BRASIL, 1940)

2.3.2 <ementa>

Delimita o assunto ou pequena explicação sobre a norma.

Etiqueta: <ementa>

Exemplo:

<ementa> Institui o Código Civil. </ementa>

(BRASIL, 2002)

Exemplo: <ementa> Código Penal. </ementa>

(BRASIL, 1940)

Exemplo:

<ementa> Regula a recuperação judicial, a extrajudicial e a falência do empresário e da sociedade empresária. </ementa>

(BRASIL, 2005)

2.3.3 <abertura>

Delimita a primeira frase da norma, esta anuncia a sua promulgação/decretação/sanção, etc. pelo Presidente da república.

Etiqueta: <abertura>

Exemplo:

<abertura> O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei: </abertura>

(BRASIL, 2002) e (BRASIL, 2005)

Exemplo:

<abertura> O PRESIDENTE DA REPÚBLICA, usando da atribuição que lhe confere o art. 180 da Constituição, decreta a seguinte Lei: </abertura>

(BRASIL, 1940)

2.3.4 <preambulo>

Explicação: introduz/apresenta a norma. Pode ter natureza puramente política ou funcional.

Preambulo: <preambulo>

Exemplo:

<preambulo> PREAMBULO

Nós, representantes do povo brasileiro, reunidos em Assembléia Nacional Constituinte para instituir um Estado Democrático, destinado a assegurar o exercício dos direitos sociais e individuais, a liberdade, a segurança, o bem-estar, o desenvolvimento, a igualdade e a justiça como valores supremos de uma sociedade fraterna, pluralista e sem preconceitos, fundada na harmonia social e comprometida, na ordem interna e internacional, com a solução pacífica das controvérsias, promulgamos, sob a proteção de Deus, a seguinte CONSTITUIÇÃO DA REPÚBLICA FEDERATIVA DO BRASIL.

</preambulo>

(BRASIL, 1988)

2.3.5 Partes e subdivisões dos textos legais

As subdivisões da lei (parte, livro, título, etc.) serão marcadas com etiquetas de partes com atributo. A etiqueta deve delimitar todo o conteúdo da subdivisão, sendo aberta no título da subdivisão e fechada após o último item incluído naquela subdivisão. Lembrando que existe uma hierarquia entre os vários tipos de subdivisão, devendo essa ser respeitada tanto na etiqueta de abertura quanto na etiqueta de fechamento.

O título das subdivisões será marcado como atributo da subdivisão, ele é a informação sobre a etiqueta que queremos ressaltar.

Atenção: o atributo vem sempre entre aspas. O atributo será antecedido do identificador “v” de valor, que corresponde ao nome da parte e o símbolo de igualdade (=). Tanto o atributo quanto seu respectivo identificador vêm só na etiqueta de abertura. A etiqueta de fechamento não tem atributo nem identificador.

Exemplo:

<titulo v = "TÍTULO II ÁGUAS PÚBLICAS EM RELAÇÃO AOS SEUS PROPRIETÁRIOS">

<capitulo v = "CAPÍTULO ÚNICO">

<artigo> Art. 29. As águas públicas de uso comum, bem como o seu álveo, pertencem:

I – A União:

a) quando marítimas;

b) quando situadas no Território do Acre, ou em qualquer outro território que a União venha a adquirir, enquanto o mesmo não se constituir em Estado, ou for incorporado a algum Estado;

c) quando servem de limites da República com as nações vizinhas ou se extendam a território estrangeiro;

d) quando situadas na zona de 100 kilometros contigua aos limites da República com estas nações;

e) quando sirvam de limites entre dois ou mais Estados;

f) quando percorram parte dos territórios de dois ou mais Estados.

II – Aos Estados:

a) quando sirvam de limites a dois ou mais Municípios;

b) quando percorram parte dos territórios de dois ou mais Municípios.

III – Aos Municípios:

a) quando, exclusivamente, situados em seus territórios, respeitadas as restrições que possam ser impostas pela legislação dos Estados.

§ 1º Fica limitado o domínio dos Estados e Municípios sobre quaisquer correntes, pela servidão que a União se confere, para o aproveitamento industrial das águas e da energia hidráulica, e para navegação;

§ 2º Fica, ainda, limitado o domínio dos Estados e Municípios pela competência que se confere a União para legislar, de acordo com os Estados, em socorro das zonas periodicamente assoladas pelas secas.

</artigo>

<artigo> Art. 30. Pertencem a União os terrenos de marinha e os acrescidos natural ou artificialmente, conforme a legislação especial sobre o assunto.

</artigo>

<artigo> Art. 31. Pertencem aos Estados os terrenos reservados as margens das correntes e lagos navegáveis, si, por algum título, não forem do domínio federal, municipal ou particular.

Parágrafo único. Esse domínio sofre idênticas limitações as de que trata o Art. 29.

</artigo>

</capitulo>

</titulo>

(BRASIL, 1934)

Obs.: Em alguns casos, entre o nome da parte (parte, título, capítulo, etc.) e seu título temos um aviso de modificação. Nesses casos vamos marcar com a etiqueta da parte correspondente apenas seu nome e título. A modificação será marcada separadamente logo abaixo dessa marcação com sua etiqueta própria e não dentro da etiqueta da parte.

Exemplo:

CAPÍTULO II-A

(Incluído pela Lei nº 10.467, de 11.6.2002)

DOS CRIMES PRATICADOS POR PARTICULAR CONTRA A ADMINISTRAÇÃO PÚBLICA ESTRANGEIRA

(BRASIL, 1940)

Errado: <capitulo v = "CAPÍTULO II-A(Incluído pela Lei nº 10.467, de 11.6.2002) DOS CRIMES PRATICADOS POR PARTICULAR CONTRA A ADMINISTRAÇÃO PÚBLICA ESTRANGEIRA">

Certo: <capitulo v = " CAPÍTULO II-A DOS CRIMES PRATICADOS POR PARTICULAR CONTRA A ADMINISTRAÇÃO PÚBLICA ESTRANGEIRA ">

<modificacao> (Incluído pela Lei nº 10.467, de 11.6.2002) </modificacao>

2.3.5.1 <parte>

Tipo de subdivisão dentro da norma.

Etiqueta: <parte v = “x”>

Exemplo:

<parte v = “Parte geral”>

[texto]

</parte>

(BRASIL, 2002)

2.3.5.2 <livro>

Tipo de subdivisão dentro da norma.

Etiqueta: <livro v = “x”>

Exemplo:

<livro v = “LIVRO I DAS PESSOAS”>

[texto]

</livro>

(BRASIL, 2002)

2.3.5.3 <titulo>

Tipo de subdivisão dentro da norma.

Etiqueta: <titulo v = “x”>

Exemplo:
<titulo v = “TITULO I DAS PESSOAS NATURAIS”>

[texto]

</titulo>

(BRASIL, 2002)

2.3.5.4 <subtitulo>

Tipo de subdivisão dentro da norma.

Etiqueta: <subtitulo v = “x”>

Exemplo:

<subtítulo v = “SUBTITULO II DA SOCIEDADE PERSONIFICADA”>

[texto]

</subtitulo>

(BRASIL, 2005)

2.3.5.5 <capitulo>

Tipo de subdivisão dentro da norma.

Etiqueta: <capitulo v = “x”>

Exemplo:

<capitulo v = “CAPITULO I Da Personalidade e da Capacidade”>

[texto]

</capitulo>

(BRASIL, 2002)

2.3.5.6 <secao>

Tipo de subdivisão dentro da norma.

Etiqueta: <secao v = “x”>

Exemplo:

<secao v = “Secao I Disposições Gerais”>

[texto]

</secao>

(BRASIL, 2005)

2.3.5.7 <subsecao>

Tipo de subdivisão dentro da norma.

Etiqueta: <subsecao v = “x”>

Exemplo:

<subsecao v = “Subsecao I Da Retrovenda”>

[texto]

</subsecao>

(BRASIL, 2002)

2.3.6 <q>

Em algumas normas, por exemplo código penal e código processual penal alguns artigos, incisos e até mesmo parágrafos têm títulos. Nesses casos vamos usar uma etiqueta pontual com esse título como atributo.

Etiqueta: <q v = “x”/>

Exemplo:

<q v= “Lugar do crime”/>

<modificacao> (Redação dada pela Lei nº 7.209, de 1984) </modificacao>

<artigo> Art. 6º Considera-se praticado o crime no lugar em que ocorreu a ação ou omissão, no todo ou em parte, bem como onde se produziu ou deveria produzir-se o resultado.

<modificacao> (Redação dada pela Lei nº 7.209, de 1984) </modificacao>

</artigo>

(BRASIL, 1940)

<artigo> Art. 12. Diz-se o crime:

<q v = " Crime consumado"/>

I - consumado, quando nele se reunem todos os elementos de sua definição legal;

<q v = "Tentativa"/>

II - tentado, quando, iniciada a execução, não se consuma, por circunstâncias alheias à vontade do agente.

<q v = " Pena da Tentativa"/>

Parágrafo único. Salvo disposição em contrário, pune-se a tentativa com a pena correspondente ao crime consumado, diminuida de um a dois terços.

</artigo>

(BRASIL, 1940)

2.3.7 <tp>

Tipo penal é o nome do crime ou contravenção penal. Vai ser uma etiqueta pontual com atributo, este será o nome do crime/contravenção penal.

Etiqueta: <tp v = “x”/>

Exemplo:

<tp v = “Difamação”/>

<artigo v = "Art. 139"> - Difamar alguém, imputando-lhe fato ofensivo à sua reputação:

<pena> Pena - detenção, de três meses a um ano, e multa.</pena>

<tp v = “Exceção da verdade”/>

Parágrafo único - A exceção da verdade somente se admite se o ofendido é funcionário público e a ofensa é relativa ao exercício de suas funções.

</artigo>

(BRASIL, 1940)

2.3.9 <artigo>

Texto do artigo.

Etiqueta: <artigo>

Exemplo:

<artigo> Art. 1o Toda pessoa é capaz de direitos e deveres na ordem civil. </artigo>

(BRASIL, 2002)

2.3.10 <pena>

Pena atribuída a determinado tipo penal.

Etiqueta: <pena>

Exemplo:

<pena> Pena – reclusão, de 2 (dois) a 4 (quatro) anos, e multa. </pena>

(BRASIL, 2005)

2.3. 11 <promulgacao>

Lugar e data da promulgação da lei. Promulgação é quando o presidente assina a lei, ela passa a existir, mas ainda não produz efeitos, não pode ser aplicada.

Etiqueta: <promulgacao>

Exemplo:

<promulgacao> Brasília, 10 de janeiro de 2002; 181 o da Independência e 114 o da República. -> </promulgacao>

(BRASIL, 2002)

2.3.12 <assinatura>

Nome do presidente e outras pessoas envolvidas na criação da norma, constam no fim da norma.

Etiqueta: <assinatura>

Exemplo:

<assinatura> LUIZ INÁCIO LULA DA SILVA

Márcio Thomaz Bastos

Antonio Palloci Filho

Ricardo José Ribeiro Berzoini

Luiz Fernando Furlan </assinatura>

(BRASIL, 2005)

2.3.13 <publicacao>

Observação sobre a publicação no Diário Oficial da União.

Etiqueta: <publicacao>

Exemplo:

<publicacao> Este texto não substitui o publicado no DOU de 31.12.1940 e retificado em 3.1.1941 </publicacao>

(BRASIL, 1940)

2.3.14 <modificacao>

Modificações no texto legal, podem ser: veto total ou parcial, apagamento total ou parcial ou acréscimo na redação de um dispositivo existente ou criação de novo dispositivo. São modificações: Revogado, Revogado pela lei x, Redação dada pela lei x, Incluído pela lei x, Vetado, etc.

Etiqueta: <modificacao>

Exemplo:

<modificacao> (“Caput” do artigo com redação dada pela Lei nº 14.112, de 24/12/2020, publicada na Edição Extra B do DOU de 24/12/2020, em vigor 30 dias após a publicacao) </modificacao>

(BRASIL, 2005)

Obs.: Caso haja algum sinal de pontuação depois da modificação ele deverá ser apagado seja na versão limpa que na versão XML.

Exemplo:

Art. 35. (VETADO).

<artigo> Art. 35. <modificacao> (VETADO) </modificacao> </artigo>

[CÓDIGO DE PROCESSO CIVIL]

Obs.: As modificações devem ser marcadas uma a uma e não como bloco, a etiqueta abre logo antes da modificação e fecha logo depois de cada modificação.

Exemplo:

<artigo> Art. 57. A exclusão do associado só é admissível havendo justa causa, assim reconhecida em procedimento que assegure direito de defesa e de recurso, nos termos previstos no estatuto. <modificacao> (Redação dada pela Lei nº 11.127, de 2005) </modificacao>

<tachado> Parágrafo único. Da decisão do órgão que, de conformidade com o estatuto, decretar a exclusão, caberá sempre recurso à assembléia geral </tachado>

Parágrafo único. <modificacao> (revogado) </modificacao> <modificacao> (Redação dada pela Lei nº 11.127, de 2005) </modificacao>

</artigo>

(BRASIL, 2002)

Obs.: a etiqueta de modificação deve estar dentro da etiqueta do artigo.

Exemplo:

<artigo> Art. 60. A convocação dos órgãos deliberativos far-se-á na forma do estatuto, garantido a 1/5 (um quinto) dos associados o direito de promovê-la. <modificacao> (Redação dada pela Lei nº 11.127, de 2005) </modificacao>

</artigo>

(BRASIL, 2002)

Obs.: algumas vezes a modificação não é delimitada por parênteses, por isso atenção na hora de marcar para não deixar passar batido essa modificação. Ela deve ser marcada da mesma forma que as demais.

Exemplo:

§2º Revogado.

§2º <modificacao> Revogado </modificacao>

2.3.15 <tachado>

Texto anterior à modificação. Vem tachado no texto da lei e antecede a nova redação, caso exista.

Etiqueta: <tachado>

Exemplo:

Original:

~~§ 4º Na recuperação judicial, a suspensão de que trata o caput deste artigo em hipótese nenhuma excederá o prazo improrrogável de 180 (cento e oitenta) dias contado do deferimento do processamento da recuperação, restabelecendo-se, após o decurso do prazo, o direito dos credores de iniciar ou continuar suas ações e execuções, independentemente de pronunciamento judicial.~~

§ 4º Na recuperação judicial, as suspensões e a proibição de que tratam os incisos I, II e III do caput deste artigo perdurarão pelo prazo de 180 (cento e oitenta) dias, contado do deferimento do processamento da recuperação, prorrogável por igual período, uma única vez, em caráter excepcional, desde que o devedor não haja concorrido com a superação do lapso temporal. (Redação dada pela Lei nº 14.112, de 2020)

Marcado:

<tachado> § 4º Na recuperação judicial, a suspensão de que trata o caput deste artigo em hipótese nenhuma excederá o prazo improrrogável de 180 (cento e oitenta) dias contado do deferimento do processamento da recuperação, restabelecendo-se, após o decurso do prazo, o direito dos credores de iniciar ou continuar suas ações e execuções, independentemente de pronunciamento judicial. </tachado>

§ 4º Na recuperação judicial, as suspensões e a proibição de que tratam os incisos I, II e III do caput deste artigo perdurarão pelo prazo de 180 (cento e oitenta) dias, contado do deferimento do processamento da recuperação, prorrogável por igual período, uma única vez, em caráter excepcional, desde que o devedor não haja concorrido com a superação do lapso temporal. <modificacao> (Redação dada pela Lei nº 14.112, de 2020) </modificacao>

(BRASIL, 2005)

Obs.: as etiquetas de parte, artigo, tipo penal, pena, etc. são obrigatórias, ainda que o texto esteja tachado. Nesse caso colocamos primeiro a etiqueta de tachado e depois a da parte correspondente.

Exemplo:

Original: ~~Art. 16 A lei que alterar o processo eleitoral só entrará em vigor um ano após sua promulgação~~

Marcado:

<tachado>

<artigo> Art. 16 A lei que alterar o processo eleitoral só entrará em vigor um ano após sua promulgação </artigo>

</tachado>

(BRASIL, 1988)

Obs.: No caso de modificações de textos tachados, independentemente de a modificação estar tachada ou não vamos incluir a etiqueta de modificação dentro da etiqueta de tachado.

Exemplo:

Original:

~~Art. 6º São direitos sociais a educação, a saúde, a alimentação, o trabalho, a moradia, o lazer, a segurança, a previdência social, a proteção à maternidade e à infância, a assistência aos desamparados, na forma desta Constituição.~~ ~~(Redação dada pela Emenda Constitucional nº 64, de 2010)~~

Marcado:

<tachado>

<artigo> Art. 6º São direitos sociais a educação, a saúde, o trabalho, a moradia, o lazer, a segurança, a previdência social, a proteção à maternidade e à infância, a assistência aos desamparados, na forma desta Constituição.

<modificacao> (Redação dada pela Emenda Constitucional nº 26, de 2000) </modificacao>

</artigo>

</tachado>

(BRASIL, 1988)

Original:

~~IV - de provimento, pelo Superior Tribunal de Justiça, de representação do Procurador-Geral da República, no caso de recusa à execução de lei federal~~. (Revogado pela Emenda Constitucional nº 45, de 2004)

Marcado:

<tachado> IV - de provimento, pelo Superior Tribunal de Justiça, de representação do Procurador-Geral da República, no caso de recusa à execução de lei federal.

<modificacao> (Redação dada pela Emenda Constitucional nº 25, de 2004) </modificacao>

</tachado>

(BRASIL, 1988)

Obs.: Caso toda a parte/livro/título/capítulo etc. tiver sido revogado/alterado por uma mesma lei usamos só uma etiqueta de tachado que colocamos antes de abrir a parte rachada e fechamos ela depois do final dessa parte.

Exemplo:

<tachado>

<parte v = "PARTE PRIMEIRA DO COMÉRCIO EM GERAL">

<modificacao> Parte revogada pela Lei 10.406, de 10.1.2002 </modificacao>

<titulo v = "TÍTULO I Dos Comerciantes">

<capitulo v = "Capítulo I Das Qualidades Necessárias para ser Comerciante">

<artigo> Art. 1 - Podem comerciar no Brasil:

1 - Todas as pessoas que, na conformidade das leis deste Império, se acharem na livre administração de suas pessoas e bens, e não forem expressamente proibida neste Código.

2 - Os menores legitimamente emancipados.

3 - Os filhos-famílias que tiverem mais de 18 (dezoito) anos de idade, com autorização dos pais, provada por escritura pública. O filho maior de 21 (vinte e um) anos, que for associado ao comércio do pai, e o que com sua aprovação, provada por escrito, levantar algum estabelecimento comercial, será reputado emancipado e maior para todos os efeitos legais nas negociações mercantis.

4 - As mulheres casadas maiores de 18 (dezoito) anos, com autorização de seus maridos para poderem comerciar em seu próprio nome, provada por escritura pública. As

que se acharem separadas da coabitação dos maridos por sentença de divórcio perpétuo, não precisam da sua autorização.

Os menores, os filhos-famílias e as mulheres casadas devem inscrever os títulos da sua habilitação civil, antes de principiarem a comerciar, no Registro do Comércio do respectivo distrito.

</artigo>

[…]

<artigo> Art. 456 - O tempo para a prescrição de obrigações mercantis contraídas, e direitos adquiridos anteriormente à promulgação do presente Código, será computado e regulado na conformidade das disposições nele contidas, começando a contar-se o prazo da data da mesma promulgação.

</artigo>

</titulo>

<modificacao> Parte revogada pela Lei 10.406, de 10.1.2002 </modificacao>

</parte>

</tachado>

(BRASIL, 1850)

A mesma lógica se aplica aos artigos. Se for um artigo inteiro revogado por uma mesma lei é só abrir a etiqueta antes da de artigo e fechar depois dela.

Exemplo:

<tachado> <artigo> Art. 3 o São absolutamente incapazes de exercer pessoalmente os atos da vida civil:

I - os menores de dezesseis anos;

II - os que, por enfermidade ou deficiência mental, não tiverem o necessário discernimento para a prática desses atos;

III - os que, mesmo por causa transitória, não puderem exprimir sua vontade. </artigo> </tachado>

(BRASIL, 2002)

2.3.16 <outros>

Em algumas leis vamos encontrar informações extratextuais, como observações e expedientes que não se enquadram nas nossas etiquetas. Como não fazem parte do texto da lei em si, ou seja, não são artigos essas devem ser apagadas na versão limpa e marcadas na versão XML para não se confundir com o texto. Nesses casos vamos usar a etiqueta <outros>.

Etiqueta: <outros>

Exemplo:

<outros> Carta de Lei, pela qual V. M. I. Manda executar o Decreto d'Assembléa Geral, que Houve por bem Sanccionar, sobre o Codigo Commercial do Imperio do Brasil, na fórma acima declarada.

Para Vossa Magestade Imperial Ver. <outros>

(BRASIL, 1850)

REFERÊNCIAS

BRASIL. **Constituição da República Federativa do Brasil**, de 10 de outubro de 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 10 de set. 2020.

BRASIL. **Decreto-lei no 2.848, de 7 de dezembro de 1940** (Código Penal). Disponível em: http://www.planalto.gov.br/CCIVIL_03/Decreto-Lei/Del2848.htm. Acesso em: 10 de set. 2021.

BRASIL. **Decreto nº 24.643 de 10 de julho de 1934** (Código de Águas). Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/D24643.htm. Acesso em: 10 de set. 2021.

BRASIL. **Lei nº 556, de 25 de junho de 1850** (Código Comercial). Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/LIM/LIM556.htm. Acesso em: 10 de set. 2021.

BRASIL. **Lei nº 10.406 de 10 de janeiro de 2002** (Código Civil). Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/2002/L10406.htm#art2044. Acesso em: 10 de set. 2021.

BRASIL. **Lei nº 11.101 de 9 de fevereiro de 2005**. Disponível em: http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2012/Lei/L12651.htm. Acesso em: 10 de set. 2020.

HARDIE, A. (2014). Modest XML for Corpora: Not a standard, but a suggestion. **ICAME Journal**, 38(1), 73–103. https://doi.org/10.2478/icame-2014-0004.

APÊNDICE B – *LEX-BR-Ius* e seções adicionados às dimensões identificadas por Berber Sardinha, Kauffmann, Acunzo (2014)

Dimensão 1: *Oral versus literate discourse*

Registro	D1
Editorials	-106.160184
Textbook texts	-99.74903669
Prep. school texts	-99.58144055
Product labels	-93.40033435
Encyclopedia entries	-93.34821167
Recipes	-90.40932769
Newspaper reportage	-88.014399
TV news	-87.33350623
Minutes	-77.65818976
Magazine news	-74.96375828
Game instructions	-71.73994181
Medicine/drug labels	-70.37609962
Business letters	-64.81267243
Campaign plans	-55.90334792
ESTATUTOS	-47.4500862
Legislation	-46.30433594
Written exams	-44.04253314
Agreements	-43.3129415
CONSTITUIÇÃO	-43.11512122
LEX-BR-IUS	-42.09524708
Essays	-38.42318443
Magazine celebrity	-38.35980061
CÓDIGOS	-37.48577751
Websites	-30.48515688
Jokes	-29.60943154
Interviews-press	-28.95587673
Academic articles	-25.71806617
Blogs	-24.77025628
Non-fiction books	-23.63186762
Government bids	-21.37441424
Political speeches	-18.34941296
Church liturgy	-15.31171893
Business conference calls	-14.59950854
Twitter	-14.40698304
Radio broadcasts	-12.93067851

Short stories	-11.57490082
Theses	-10.04760469
General fiction	-9.8700785
Congressional debates	-9.298093681
Interviews tv	-8.413056564
Textbooks	-8.247842546
Youth fiction	-7.375062473
Interviews – sociolinguistic	-6.671183158
Horoscope	-3.521922816
Soap operas	-2.152091365
Comics	0.487217556
Conversation	1.07952029
Emails – Personal	12.05439133
Facebook	36.57706773
Textbook dialogs	45.25191139
Songs	60.8248333

Dimensão 2: *Argumentation*

Registro	**D2**
Textbooks	-16.14839
Theses	-15.98824
Youth fiction	-15.97052
Congressional debates	-15.40989
Government bids	-14.93931
Interviews-press	-14.82598
Interviews tv	-13.9518
Legislation	-12.93474
Soap operas	-12.83517
Business conference calls	-12.81631
General fiction	-12.81306
Church liturgy	-12.72266
Academic articles	-12.67348
Radio broadcasts	-11.29161
Agreements	-10.95509
Non-fiction books	-9.606825
Websites	-9.018018
Conversation	-8.345286
Minutes	-8.26839
Written exams	-6.777247

Medicine/drug labels	-5.781976
Short stories	-5.44501
Political speeches	-5.166722
CONSTITUIÇÃO	-4.183141
ESTATUTOS	-2.399809
Campaign plans	-1.636197
LEX-BR-IUS	-0.971869
Blogs	0.2685317
CÓDIGOS	0.4834086
Comics	0.7018744
Essays	0.8804369
Interviews – sociolinguistic	1.3227011
Recipes	1.4578864
Magazine celebrity	2.8183573
Product labels	3.9286728
Magazine news	4.9396875
Encyclopedia entries	6.4038207
Prep. school texts	6.4360243
Business letters	9.2129281
Twitter	9.3858552
Game instructions	10.647759
Songs	11.091258
Newspaper reportage	15.38561
TV news	15.858464
Jokes	18.340132
Textbook texts	20.264803
Facebook	20.350391
Textbook dialogs	21.163203
Emails – Personal	22.462431
Editorials	22.656042
Horoscope	36.993927

Dimensão 3: *Involved versus informational production*

Registro	D3
Textbooks	-4.482415
Theses	-4.467495
Legislation	-4.467077
Government bids	-4.430409
Congressional debates	-4.359225

Youth fiction	-4.328465
CONSTITUIÇÃO	-4.281133
ESTATUTOS	-4.233418
LEX-BR-IUS	-4.0902
Agreements	-3.985682
Academic articles	-3.960152
CÓDIGOS	-3.954329
Recipes	-3.937806
Interviews-press	-3.789502
Church liturgy	-3.666426
General fiction	-3.530983
Campaign plans	-3.503113
Minutes	-3.474935
Business conference calls	-3.474467
Non-fiction books	-3.446897
Written exams	-3.399608
Medicine/drug labels	-3.299044
Product labels	-3.19234
Websites	-2.889013
Magazine news	-2.577485
Political speeches	-2.547265
Soap operas	-2.316101
Radio broadcasts	-2.019632
Blogs	-1.734238
Short stories	-1.61158
Interviews tv	-1.540446
Business letters	-1.497919
Editorials	-1.412562
Newspaper reportage	-1.298835
Prep. school texts	-0.948329
Interviews – sociolinguistic	-0.840396
Textbook texts	-0.530074
Encyclopedia entries	-0.161331
Essays	0.1887138
Magazine celebrity	0.7216497
TV news	0.9267367
Horoscope	1.474758
Game instructions	1.9944504
Comics	3.791822
Twitter	4.2487725
Emails – Personal	7.9234918

Songs	8.0439876
Facebook	8.4735576
Conversation	8.6617643
Textbook dialogs	13.217245
Jokes	14.984328

Dimensão 4: *Directive discourse*

Registro	**D4**
Textbooks	-7.257644099
Theses	-7.227496115
Youth fiction	-7.113217389
Congressional debates	-7.096617587
Interviews-press	-6.911954676
Government bids	-6.559171001
Business conference calls	-6.320970352
Interviews tv	-6.24617314
Academic articles	-6.087625418
General fiction	-5.885883392
Soap operas	-5.585640167
Radio broadcasts	-5.554832762
Non-fiction books	-5.182989791
Legislation	-5.150955325
Political speeches	-4.521128474
Written exams	-4.264058457
Church liturgy	-4.228133471
Agreements	-4.177045603
Minutes	-3.838516026
Conversation	-3.340098427
Short stories	-2.888736145
Interviews – sociolinguistic	-2.842408676
Campaign plans	-2.142334245
CÓDIGOS	-2.11665222
Magazine news	-1.989936868
Essays	-1.841514441
Medicine/drug labels	-1.44257328
Websites	-1.237234681
Blogs	-0.989243237
ESTATUTOS	-0.927348551
LEX-BR-IUS	-0.811094347

CONSTITUIÇÃO	-0.612297875
Magazine celebrity	-0.519380094
Newspaper reportage	0.54790333
Comics	1.726815103
Textbook texts	1.770480484
Editorials	2.197723433
Encyclopedia entries	2.445338615
Prep. school texts	2.535715263
TV news	3.350623993
Business letters	5.793695305
Twitter	7.154665462
Emails – Personal	9.672433701
Facebook	10.58950561
Textbook dialogs	10.89035374
Songs	11.77957446
Horoscope	12.71313415
Game instructions	12.75130285
Jokes	13.29051515
Product labels	15.73003736
Recipes	48.16755981

Dimensão 5: *Future versus past time orientation*

Registro	D5
Textbook dialogs	-6.6547352
Jokes	-6.5440362
Newspaper reportage	-5.8665034
TV news	-5.8621968
Magazine news	-5.8250606
Magazine celebrity	-5.5893923
Short stories	-5.0585635
Non-fiction books	-4.0097661
Encyclopedia entries	-3.8185798
Prep. school texts	-3.6936875
Conversation	-3.187847
General fiction	-2.8987335
Interviews tv	-2.772967
Youth fiction	-2.7141683
Essays	-2.5896889
Interviews-press	-2.5237381

Congressional debates	-2.4758206
Theses	-2.4445979
Soap operas	-2.4136037
Textbooks	-2.3759878
Church liturgy	-2.2686412
Political speeches	-2.220245
Radio broadcasts	-1.974283
Business conference calls	-1.9379753
Academic articles	-1.8640841
Interviews – sociolinguistic	-1.7731662
Minutes	-1.5369923
Blogs	-1.4528648
Comics	-1.3685281
Written exams	-0.850188
Websites	-0.4421717
Twitter	-0.1400972
Campaign plans	0.14953513
Editorials	0.21102445
Government bids	0.68321844
Legislation	1.25793477
Textbook texts	1.44914869
Songs	3.64493177
Facebook	4.19278418
Agreements	4.48536481
Recipes	5.80936255
Business letters	7.5369507
Product labels	11.5284022
Emails – Personal	11.6713017
Medicine/drug labels	17.4035635
CONSTITUIÇÃO	19.731713
Game instructions	21.9469569
Horoscope	22.7578488
ESTATUTOS	24.1251775
LEX-BR-IUS	25.21
CÓDIGOS	26.5477523

Dimensão 6: *Reported discourse*

Registro	D6
Textbooks	-3.787524
Theses	-3.756808
Youth fiction	-3.611026
Congressional debates	-3.519138
Business conference calls	-3.481716
Interviews tv	-3.373175
Interviews-press	-3.373175
Government bids	-3.290245
Radio broadcasts	-3.218218
Academic articles	-3.107522
Campaign plans	-2.887375
Political speeches	-2.861551
General fiction	-2.616992
Soap operas	-2.486086
Websites	-2.450234
Legislation	-2.229816
Non-fiction books	-2.205673
Medicine/drug labels	-1.790304
Agreements	-1.614834
Written exams	-1.555304
Interviews – sociolinguistic	-1.343259
Conversation	-1.297512
Blogs	-1.22542
Recipes	-1.186971
ESTATUTOS	-1.161055
Product labels	-1.067445
CONSTITUIÇÃO	-0.688823
Minutes	-0.459094
Essays	-0.325233
Business letters	-0.25314
Magazine news	-0.224528
LEX-BR-IUS	0.043199
Short stories	0.6294025
TV news	0.8634941
Editorials	0.9225393
Comics	1.0587218
CÓDIGOS	1.1184929

Encyclopedia entries	1.1860812
Magazine celebrity	1.5673482
Church liturgy	1.6902756
Prep. school texts	1.873758
Game instructions	1.940208
Twitter	2.1028999
Textbook texts	2.2434791
Facebook	2.4831643
Emails – Personal	2.900134
Songs	3.856172
Textbook dialogs	4.516347
Horoscope	4.6267949
Newspaper reportage	5.5297991
Jokes	10.391971

ANEXO A – Variáveis consideradas relevantes para a análise multidimensional português brasileiro

Sigla	Variável	Etiqueta
adj	*adjective_affiliative*	adjaffi
adj	*adjective_all*	adjall
adj	*adjective_attributive*	adjattr
adj	*adjective_color*	adjcolr
adj	*adjective_evaluative*	adjeval
adj	*adjective_except_evaluative*	adjexceval
adj	*adjective_postmodifying_attributive*	adjpost
adj	*adjective_predicative*	adjpred
adj	*adjective_premodifying_attributive*	adjpre
adj	*adjective_relational*	adjrela
adj	*adjective_size*	adjsize
adj	*adjective_superlative*	adjsup
adj	*adjective_time*	adjtime
adj	*adjective_topical*	adjtopi
adj	*stance_adjective_inf_clause_attitude*	adjattiinf
adj	*stance_adjective_inf_clause_factive*	adjfactinf
adj	*stance_adjective_inf_clause_likelihood*	adjliklinf
adj	*stance_adjective_que_clause_attitude*	adjattique
adj	*stance_adjective_que_clause_factive*	adjfactque
adj	*stance_adjective_que_clause_likelihood*	adjliklque
adj	*stance_adjective_que_or_inf_clause*	adjqueinfcl
adv	*adverb_all*	advall
adv	*adverb_amplifier*	advampl
adv	*adverb_attitudinal*	advatt
adv	*adverb_compound*	advcmpd
adv	*adverb_downtoner*	advdown
adv	*adverb_emphatic*	advemph
adv	*adverb_except_time_manner_place*	advother
adv	*adverb_factive*	advfact
adv	*adverb_hedge*	advhedg
adv	*adverb_intensity*	advints
adv	*adverb_likelihood*	advlikl
adv	*adverb_long*	advlong
adv	*adverb_mode*	advmanner
adv	*adverb_nao*	advnao
adv	*adverb_negative_except_nao*	advneg
adv	*adverb_nonfactual*	advnonf

adv	*adverb_place*	advpl
adv	*adverb_split_auxiliary*	advsplit
adv	*adverb_time*	advtime
adv	*comparative*	advcomp
adv	*discourse_markers_conjunctional_adverbs*	discmrkr
art	*article_def*	artdef
art	*article_indef*	artindef
cj	*conjunction_coordinating*	cjcoor
cj	*conjunction_coordinating_additive*	cjadd
cj	*conjunction_coordinating_adversative*	cjadv
cj	*conjunction_coordinating_clausal*	cjcoorcls
cj	*conjunction_coordinating_conclusive*	cjcncl
cj	*conjunction_coordinating_ou*	cjou
cj	*conjunction_coordinating_phrasal*	cjcoorphr
cj	*conjunction_subordinating*	cjsub
cj	*conjunction_subordinating_causal*	cjcaus
cj	*conjunction_subordinating_concessive*	cjcncsv
cj	*conjunction_subordinating_conditional*	cjcond
cj	*conjunction_subordinating_conformative*	cjcnfm
cj	*conjunction_subordinating_final*	cjfinal
cj	*conjunction_subordinating_proportional*	cjprop
cj	*conjunction_subordinating_temporal*	cjtemp
cl	*CU_complement_clause*	vbCU
cl	*agentless_passive*	clpassless
cl	*infinitive_clause_controlled_adjective*	clinfadj
cl	*infinitive_clause_controlled_adjective_affective*	clinfadjaff
cl	*infinitive_clause_controlled_adjective_certainty*	clinfadjcert
cl	*infinitive_clause_controlled_adjective_ease_difficulty*	clinfadjease
cl	*infinitive_clause_controlled_adjective_evaluation*	clinfadjeval
cl	*infinitive_clause_controlled_adjective_willingness*	clinfadjwill
cl	*infinitive_clause_controlled_preposition*	clinfprp
cl	*noun_subject_position*	clnsbjc
cl	*passive_postnominal_modifier*	clpostnom
cl	*passive_with_por*	clpasspor
cl	*que_adjective_complement_clause*	adjque
cl	*que_clause_controlled_adjective*	clqueeadj
cl	*que_clause_controlled_adjective_certainty*	clqueadjcert
cl	*que_clause_controlled_adjective_ease_difficulty*	clqueadjease
cl	*que_clause_controlled_adjective_evaluation*	clqueadjeval
cl	*que_clause_controlled_adverb*	clqueeadv
cl	*que_clause_controlled_preposition*	clqueeprp

cl	*que_clause_controlled_verb*	clqueevb
cl	*que_noun_complement_clause*	nounque
cl	*que_verb_complement_clause_indicative*	vbqueindic
cl	*que_verb_complement_clause_subjunctive*	vbquesubjc
cl	*se_passive*	clsepass
cl	*subordinate_clause_except_causal_concessive_conditional*	cjsubother
md	*modal_all*	mdall
md	*modal_conseguir*	mdconseguir
md	*modal_dever*	mddever
md	*modal_haver*	mdhaver
md	*modal_obligation*	mdoblig
md	*modal_parecer*	mdparecer
md	*modal_poder*	mdpoder
md	*modal_precisar*	mdprecisar
md	*modal_ter*	mdter
n	*nominalization*	nominlz
n	*nominalization_in_subject_position*	nominlzsubj
n	*noun_abstract*	nabst
n	*noun_all*	nall
n	*noun_animate*	nanim
n	*noun_cognition*	ncogn
n	*noun_compound*	ncomp
n	*noun_concrete*	nconc
n	*noun_group_institution*	ngrpi
n	*noun_place*	nplac
n	*noun_quantity*	nqtty
n	*noun_technical*	ntech
n	*proper_noun*	nprop
n	*stance_noun_inf_clause_attitude*	nattitinf
n	*stance_noun_inf_clause_factual*	nfactlinf
n	*stance_noun_inf_clause_likelihood*	nliklhinf
n	*stance_noun_inf_clause_non-factual*	nnonfcinf
n	*stance_noun_que_clause_attitude*	nattitque
n	*stance_noun_que_clause_factual*	nfactlque
n	*stance_noun_que_clause_likelihood*	nliklhque
n	*stance_noun_que_clause_non-factual*	nnonfcque
n	*stance_noun_que_or_inf_clause*	nqueinfcl
prn	*nominal_pronoun_subject_position*	prnnomsubj
prn	*possessive_pron*	prnposs
prn	*pronoun_demonstrative*	prndem
prn	*pronoun_first_person_plural_subject_position*	prn1plusubj

prn	*pronoun_first_person_plural_subject_position*	prn1plusubj
prn	*pronoun_first_person_singular_subject_position*	prn1sngsubj
prn	*pronoun_firstperson_oblique*	prn1obl
prn	*pronoun_second_person_plural_subject_position*	prn2plusubj
prn	*pronoun_second_person_singular_subject_position*	prn2sngsubj
prn	*pronoun_secondperson_oblique*	prn2obl
prn	*pronoun_secondperson_oblique*	prn2obl
prn	*pronoun_secondperson_oblique*	prn2obl
prn	*pronoun_third_person_plural_subject_position*	prn3plusubj
prn	*pronoun_third_person_singular_subject_position*	prn3sngsubj
prn	*pronoun_thirdperson_oblique*	prn3obl
prn	*quantifiers*	prnqtf
prn	*rare_object_pronouns*	objprnrare
prn	*relative_pronoun_preceded_by_preposition*	prnrelprep
prn	*relative_pronoun_qual_cujo*	prnqualcujo
prn	*relative_pronoun_que*	prnque
prp	*preposition_all*	prpall
qs	*QU_question*	qsqu
qs	*YN_question*	qsyn
qs	*question_tags*	qsttag
vb	*contractions*	contrac
vb	*focus_marker*	focusmkr
vb	*future*	vbfutpres
vb	*future_ir*	vbfutir
vb	*future_preterite*	vbfutpret
vb	*gerund*	vbgerall
vb	*imperative*	vbimp
vb	*imperfect*	vbimpf
vb	*indicative*	vbindic
vb	*infinitive*	vbinf
vb	*past*	vbpast
vb	*pastparticiple*	vbpastprt
vb	*perfect_aspect*	vbpfaspct
vb	*pluperfect*	vbplupf
vb	*present*	vbpres
vb	*progressive*	vbprog
vb	*progressive_infinitive*	vbproginf
vb	*progressive_phrase*	vbprogphr
vb	*ser_estar_main*	vbserestar
vb	*stance_verb_inf_clause_causative*	vbcausinf
vb	*stance_verb_inf_clause_cognitive*	vbcogninf

vb	*stance_verb_inf_clause_desire*	vbdesrinf
vb	*stance_verb_inf_clause_probability*	vbprobinf
vb	*stance_verb_inf_clause_speech*	vbspchinf
vb	*stance_verb_que_clause_causative*	vbcausque
vb	*stance_verb_que_clause_cognitive*	vbcognque
vb	*stance_verb_que_clause_desire*	vbdesrque
vb	*stance_verb_que_clause_probability*	vbprobque
vb	*stance_verb_que_clause_speech*	vbspchque
vb	*stance_verb_que_or_inf_clause*	vbqueinfcl
vb	*subject_drop*	subjdrop
vb	*subjunctive_future*	vbsubfut
vb	*subjunctive_past*	vbsubpast
vb	*subjunctive_present*	vbsubpres
vb	*verb_action*	vbact
vb	*verb_aspect*	vbaspct
vb	*verb_auxiliary*	vbaux
vb	*verb_communication*	vbcomm
vb	*verb_exist_relation*	vbexist
vb	*verb_facilitation*	vbfacil
vb	*verb_firstperson*	vb1
vb	*verb_mental*	vbment
vb	*verb_occurrence*	vbocc
vb	*verb_personal_infinitive*	vinfpers
vb	*verb_private*	vbpriv
vb	*verb_public*	vbpubl
vb	*verb_secondperson*	vb2
vb	*verb_suasive*	vbsua
vb	*verb_thirdperson*	vb3
vb	*verbs_all*	vball
vd	*clause_length*	clauselgth
vd	*type-token_ratio*	ttr
vd	*word_count*	wrcount
vd	*word_length*	wl

Fonte: Berber Sardinha, Kauffmann, Acunzo (2014)

ANEXO B – Variáveis identificadas após análise fatorial do CBVR

1 Adjectives: Affiliative
2 Adjectives: Attributive position
3 Adjectives: Attributive, pre-modifying
4 Adjectives: Evaluative
5 Adjectives: Predicative position
6 Adjectives: Relational
7 Adjectives: Topical
8 Adverbs: Não
9 Adverbs: Amplifier
10 Adverbs: Comparative
11 Adverbs: Emphatic
12 Adverbs: Hedge
13 Adverbs: Intensity
14 Adverbs: Likelihood
15 Adverbs: Manner
16 Adverbs: Negative, except não
17 Adverbs: Place
18 Adverbs: Time
19 Articles: Definite
20 Articles: Indefinite
21 Clause types: Infinitive clause controlled by adjective
22 Clause types: Infinitive clause controlled by ease or difficulty adjective
23 Clause types: Infinitive clause controlled by preposition
24 Clause types: Que clause controlled by adjective (stance)
25 Clause types: Que clause controlled by adverb
26 Clause types: Que clause controlled by noun
27 Clause types: Que clause controlled by preposition
28 Clause types: Que clause controlled by verb in indicative mood
29 Clause types: Que or infinitive clause controlled by noun (stance)
30 Clause types: Reduced progressive clause
31 Clause types: Subordinating (conditional)
32 Clause types: Subordinating (final)
33 Conjunctions: Coordinating (ou)
34 Conjunctions: Coordinating (adversative)
35 Conjunctions: Coordinating (clausal)
36 Conjunctions: Coordinating (conclusive)
37 Conjunctions: Coordinating (phrasal)
38 Modals: Dever
39 Modals: Haver que/haver de
40 Modals: Poder
41 Modals: Precisar
42 Modals: Ter que/ter de
43 Nouns: Abstract
44 Nouns: Cognition
45 Nouns: Compound

46 Nouns: Concrete
47 Nouns: Nominalization in subject position
48 Nouns: Place
49 Other características of verb or noun phrases: Agentless passive
50 Other: Contractions
51 Other: Discourse marker
52 Other: Subject omission
53 Prepositions: All
54 Pronouns: Demonstrative
55 Pronouns: First person singular, in subject position
56 Pronouns: First person, object position
57 Pronouns: Nominal in subject position
58 Pronouns: Possessive
59 Pronouns: Quantifier
60 Pronouns: Rare in object position
61 Pronouns: Relative qual or cujo
62 Pronouns: Relative que
63 Pronouns: Second person singular, in subject position
64 Pronouns: Second person, object position
65 Pronouns: Third person plural, in subject position
66 Pronouns: Third person singular, in subject position
67 Pronouns: Third person, object position
68 Questions: QU questions
69 Questions: Tag questions
70 Questions: Yes or No question
71 Verbs: Ir future
72 Verbs: Action
73 Verbs: Communication
74 Verbs: Facilitation
75 Verbs: First person
76 Verbs: Future present tense
77 Verbs: Future preterite tense
78 Verbs: Future subjunctive mood
79 Verbs: Gerund form, all
80 Verbs: Imperative mood
81 Verbs: Imperfect
82 Verbs: Infinitive
83 Verbs: Mental
84 Verbs: Past indicative tense
85 Verbs: Past participle
86 Verbs: Past subjunctive mood
87 Verbs: Present subjunctive mood
88 Verbs: Private
89 Verbs: Progressive preceded by infinitive
90 Verbs: Public
91 Verbs: Second person
92 Vocabulary distribution: Average word length

93 Vocabulary distribution: Type-token ratio

Fonte: Berber Sardinha, Kauffmann, Acunzo (2014, p. 67 – 69)

ANEXO C – Estatística descritiva do CBVR

#	*Feature*	Min.	Max.	*Mean*	*Std. Dev.*
1	*Adjectives: Affiliative*	0	17.54	1.46	2.51
2	*Adjectives: All*	9.89	157.4	60.5	22.19
3	*Adjectives: Attributive position*	1.75	107	37.66	18.66
4	*Adjectives: Attributive, post-modifying*	0	97.28	27.18	15.86
5	*Adjectives: Attributive, pre-modifying*	0	31.73	9.37	5.34
6	*Adjectives: Augmentative*	0	8.99	0.58	0.95
7	*Adjectives: Color*	0	24.66	0.57	1.72
8	*Adjectives: Evaluative*	0	27.4	5.57	4.27
9	*Adjectives: Except evaluative*	0	70.53	25.38	12.46
10	*Adjectives: Predicative position*	0	35.34	8.82	5.27
11	*Adjectives: Relational*	0	43.66	10.98	6.23
12	*Adjectives: Size*	0	24.9	3.5	2.97
13	*Adjectives: Time*	0	16.98	2.57	2.5
14	*Adjectives: Topical*	0	38.8	6.31	6.19
15	*Adverbs: Não*	0	95.87	7.97	6.73
16	*Adverbs: All*	2.58	210.5	59.63	31.88
17	*Adverbs: Amplifier*	0	34.92	7.04	5.05
18	*Adverbs: Attitudinal*	0	14.93	1.59	1.71
19	*Adverbs: Comparative*	0	45.39	9.38	5.15
20	*Adverbs: Compound*	0	31.75	7.8	4.48
21	*Adverbs: Downtoner*	0	10.66	1.28	1.51
22	*Adverbs: Emphatic*	0	27.15	5.06	3.95
23	*Adverbs: Exc. time, manner and place*	0	125.9	25.37	14.46
24	*Adverbs: Factive*	0	109	13.93	10.77
25	*Adverbs: Hedge*	0	12.11	0.88	1.3
26	*Adverbs: Intensity*	0	31.1	6.59	4.91
27	*Adverbs: Likelihood*	0	15.57	1.51	1.67
28	*Adverbs: Long (10 letters or more)*	0	62.75	9.82	10.67
29	*Adverbs: Manner*	0	17.58	2.06	2.64
30	*Adverbs: Negative, except não*	0	6.59	0.42	0.84
31	*Adverbs: Non-factual*	0	14.96	1.75	1.8
32	*Adverbs: Place*	0	42.33	4.08	5.01
33	*Adverbs: Time*	0	32.97	5.58	4.49
34	*Articles: Definite*	27.34	192.9	113.6	28.08
35	*Articles: Indefinite*	0	48.16	14.02	7.84
36	*Conjunctions: Coordinating (ou)*	0	26.69	3.39	4.22

37	*Conjunctions: Coordinating (additive)*	1.78	73.3	31.5	10.05
38	*Conjunctions: Coord. (adversative)*	0	17.01	2.34	2.32
39	*Conjunctions: Coordinating (all)*	7.46	83.14	34.05	10.57
40	*Conjunctions: Coordinating (clausal)*	0	40.86	11.37	6.04
41	*Conjunctions: Coord. (conclusive)*	0	14.26	1.64	2.05
42	*Conjunctions: Coordinating (phrasal)*	0	41.57	6.3	5.36
43	*CU clause cntrld. by verb*	0	17.58	1.55	1.6
44 I	*Infinitive clause cntrld. by adjective*	0	13.61	0.58	1.06
45 I	*Inf. clause cntrld. by affective adjective*	0	4.04	0.01	0.16
46 I	*Inf. clause cntrld. by attitude adjective*	0	8.33	0.22	0.62
47 I	*Inf. clause cntrld. by attitudinal noun*	0	7.8	0.2	0.55
48 I	*Inf. clause cntrld. by causative verb*	0	5.53	0.2	0.53
49 I	*Inf. clause cntrld. by causative verb*	0	2.27	0.04	0.19
50 I	*Inf. clause cntrld. by certainty adjective*	0	1.73	0.01	0.08
51 I	*Inf. clause cntrld. by cognitive verb*	0	6.94	0.06	0.34
52 I	*Inf. clause cntrld. by desire verb*	0	3.96	0.07	0.3
53 I	*Inf. cl. cntrld. by ease or difficulty adj.*	0	8.77	0.23	0.62
54 I	*Inf. clause cntrld. by evaluation adj.*	0	4.54	0.15	0.43
55 I	*Inf. clause cntrld. by factive adjective*	0	8.33	0.06	0.47
56 I	*Inf. clause cntrld. by factual noun*	0	3.28	0.12	0.34
57 I	*Inf. cl. cntrld. by likelihood adjective*	0	4.87	0.1	0.35
58 I	*Inf. clause cntrld. by likelihood noun*	0	4.53	0.12	0.41
59 I	*Inf. clause cntrld. by non-factual noun*	0	1.87	0.04	0.17
60 I	*Infinitive clause cntrld. by preposition*	0	31.97	9.1	4.66
61 I	*. Infinitive clause cntrld. by probability verb*	0	1.7	0.01	0.08
62 I	*Infinitive clause cntrld. by speech verb*	0	1.77	0.02	0.1
63 I	*Inf. cl. cntrld. by willingness adjective*	*.4*	4.16	0.15	0.42
64	*Modals: All*	0	30.54	5.75	4.51
65	*Modals: Conseguir*	0	6.54	0.31	0.71
66	*Modals: Dever*	0	16.38	1.21	1.94
67	*Modals: Haver que/de*	0	2.36	0.05	0.23
68	*Modals: Obligation*	0	17.11	2.26	2.37
69	*Modals: Parecer*	0	3.79	0.1	0.37
70	*Modals: Poder*	0	21.85	3.07	2.89
71	*Modals: Precisar*	0	7.14	0.34	0.79
72	*Modals: Ter que/de*	0	7.09	0.66	1.08
73	*Nouns: Abstract*	1.78	142.9	52.1	27.21
74	*Nouns: All, except nominalizations*	56.64	354.8	159.7	37.52
75	*Nouns: Animate*	0	54.15	15.57	10.45
76	*Nouns: Cognition*	0	47.09	6.85	5.76
77	*Nouns: Compound*	0	66.98	16.08	10.46

78	*Nouns: Concrete*	1.42	84.54	20.95	12.25
79	*Nouns: In subject position*	0	62.39	25.82	10.36
80	*Nouns: Institution*	0	30.76	3.08	3.63
81	*Nouns: Nominalization*	0	156.1	46.1	25.27
82	*Nouns: Nominaliz. in subject position*	0	35.05	6.95	5.16
83	*Nouns: Place*	0	61.96	7.97	6.82
84	*Nouns: Proper*	0	735.7	55.9	60.37
85	*Nouns: Quantity*	0	69.92	14.04	8.89
86	*Nouns: Technical*	0	50.19	3.84	4.89
87	*Other: Por passive*	0	6.27	0.4	0.8
88	*Other: Se passive*	0	6.14	0.43	0.8
89	*Other: Agentless passive*	0	14.32	1.95	2.38
90	*Other: Passive postnominal modifier*	0	25.81	4.2	3.84
91	*Other: Contractions*	0	36.84	1.33	4.3
92	*Other: Discourse marker*	0	56.13	5.41	5.97
93	*Other: Focus marker*	0	11.7	0.76	1.26
94	*Other: Subject omission*	0	87.75	13.57	11.07
95	*Prepositions: All*	49.15	316.5	147.2	33.79
96	*Pronouns: Demonstrative*	0	51.44	11.51	6.94
97	*Pronouns: 1st p. plural, in subject pos.*	0	18.84	0.56	1.52
98	*Pronouns: 1st p. sing., in subject pos.*	0	47.86	3.09	5.94
99	*Pronouns: 1st person, object position*	0	99.79	2.11	5.29
100	*Pronouns: Nominal in subject position*	0	7.56	0.64	1.06
101	*Pronouns: Possessive*	0	56.39	10.66	8.98
102	*Pronouns: Quantifier*	0	63.35	18.84	9.98
103	*Pronouns: Rare in object position*	0	13.72	0.88	1.66
104	*Pronouns: Relative qual or cujo*	0	7.91	0.6	1.03
105	*Pronouns: Relative que*	0	32.32	10.21	5.5
106	*Pronouns: Relative prec. by preposition*	0	9.52	1.05	1.3
107	*Pronouns: 2nd p. plural, subject pos.*	0	6.96	0.24	0.67
108	*Pronouns: 2nd p sing., subject pos.*	0	30.4	2.32	4
109	*Pronouns: 2nd p., object position*	0	86.4	5.06	8.81
110	*Pronouns: 3rd p. plural, subject pos.*	0	9.89	0.56	1.1
111	*Pronouns: 3rd p. sing., subject pos.*	0	43.2	2.69	4.41
112	*Pronouns: 3rd p., object position*	0	41.35	7.37	5.37
113	*Que clause cntrld. by adjective*	0	5.38	0.45	0.76
114	*Que clause cntrld. by adjective*	0	7.03	1.24	1.28
115	*Que clause cntrld. by adj. of certainty*	0	3.49	0.02	0.17
116	*Que cl. cntrld. by adj. of ease or diffc.*	0	2.32	0.03	0.16
117	*Que cl. cntrld. by adj. of evaluation*	0	3.14	0.08	0.28
118	*Que clause cntrld. by adverb*	0	21.48	0.53	1.25

119	*Que clause cntrld. by attitude adjective*	0	2.21	0.02	0.12
120	*Que clause cntrld. by attitude noun*	0	2.21	0.03	0.16
121	*Que clause cntrld. by cognitive verb*	0	6.24	0.2	0.51
122	*Que clause cntrld. by desire verb*	0	3.97	0.04	0.2
123	*Que clause cntrld. by factive adjective*	0	4.3	0.05	0.26
124	*Que clause cntrld. by factual noun*	0	3.55	0.08	0.29
125	*Que cl. cntrld. by likelihood adjective*	0	2.32	0.03	0.19
126	*Que clause cntrld. by likelihood noun*	0	5.1	0.13	0.4
127	*Que clause cntrld. by non-factual noun*	0	3.6	0.02	0.17
128	*Que clause cntrld. by noun*	0	18.48	5.85	3.63
129	*Que clause cntrld. by preposition*	0	8.08	1.05	1.28
130	*Que clause cntrld. by probability verb*	0	1.02	0	0.04
131	*Que clause cntrld. by speech verb*	0	8.33	0.23	0.67
132	*Que clause cntrld. by verb*	0	24.34	4.95	4.36
133	*Que cl. cntrld. by indicative verb*	0	24.34	3.76	3.87
134	*Que clause cntrld. by subjunctive*	0	1.68	0.01	0.09
135	*Que or inf. clause cntrld. by adjective*	0	16.67	0.48	1.08
136	*Que or infinitive clause cntrld. by noun*	0	8.5	0.73	1.02
137	*Que or infinitive clause cntrld. by verb*	0	8.93	0.87	1.23
138	*Questions: QU questions*	0	18.1	1.36	2.49
139	*Questions: Tag questions*	0	40.09	0.98	4.17
140	*Questions: Yes or No question*	0	56.07	3.7	7.63
141	*Reduced progressive clause*	0	21.43	3.26	2.73
142	*Sub. cl., exc. causal /conc./cond.*	0	11.88	2.13	1.77
143	*Subordinating (all)*	0	54.7	13.99	8.94
144	*Subordinating (causal)*	0	39.06	9.92	6.99
145	*Subordinating (concessive)*	0	4.25	0.31	0.59
146	*Subordinating (conditional)*	0	13.99	1.63	1.98
147	*Subordinating (conformative)*	0	1.97	0.04	0.18
148	*Subordinating (final)*	0	1.7	0.01	0.08
149	*Subordinating (proportional)*	0	1.7	0.01	0.09
150	*Subordinating (temporal)*	0	11.88	2.07	1.75
151	*Verbs: Ir future*	0	33.84	2.73	3.96
152	*Verbs: Ser or estar*	0	90.48	19.51	12.66
153	*Verbs: Action*	5.89	111.3	36.99	14.52
154	*Verbs: All*	25.18	364.4	156.2	46.02
155	*Verbs: Aspectual*	0	13.16	0.46	0.94
156	*Verbs: Auxiliary*	0	24.44	5.12	3.59
157	*Verbs: Communication*	0	37.11	8.13	6.41
158	*Verbs: Existence or relation*	0	111.2	43.69	19.36
159	*Verbs: Facilitation*	0	23.42	3.81	2.98

160	*Verbs: First person*	0	109.6	16.22	18.77
161	*Verbs: Future present tense*	0	62.5	3.92	6.18
162	*Verbs: Future preterit tense*	0	25.59	1.81	2.61
163	*Verbs: Future subjunctive mood*	0	28.22	2.05	3
164	*Verbs: Gerund form, all*	0	75.25	8.27	5.45
165	*Verbs: Imperative mood*	0	62.79	3.48	6.98
166	*Verbs: Imperfect*	0	87.64	8.32	11.82
167	*Verbs: Indicative mood*	0	238.3	84.73	40.78
168	*Verbs: Infinitive*	0	21.07	2.55	2.3
169	*Verbs: Mental*	0	88.17	16.87	11.15
170	*Verbs: Occurrence*	0	18.83	3.32	2.48
171	*Verbs: Past indicative tense*	0	89.02	19.56	17.32
172	*Verbs: Past participle*	0	62.89	18.63	10.68
173	*Verbs: Past subjunctive mood*	0	21.32	1.02	1.67
174	*Verbs: Perfect aspect*	0	100.5	38.53	17.83
175	*Verbs: Personal infinitive*	0	19.16	2.13	2.15
176	*Verbs: Pluperfect tense*	0	11.92	0.34	1.01
177	*Verbs: Present indicative tense*	0	214.8	53.62	32.47
178	*Verbs: Present subjunctive mood*	0	71.3	6.31	7.53
179	*Verbs: Private*	0	52.88	12.41	8.11
180	*Verbs: Progressive*	0	6.05	0.38	0.72
181	*Verbs: Progressive preceded by inf.*	0	22.56	0.16	0.84
182	*Verbs: Public*	0	37.7	9.76	6.46
183	*Verbs: Second person*	0	35.32	1.36	3.91
184	*Verbs: Split auxiliary*	0	3.06	0.08	0.26
185	*Verbs: Suasive*	0	27.14	2.66	2.52
186	*Verbs: Third person*	7.36	203.1	83.85	29.32
187	*Vocabulary distr.: avg. clause length*	0.08	100	8.73	8.65
188	*Vocabulary distr.: Average word length*	3.62	6.57	4.78	0.43
189	*Vocabulary distr.: Type-token ratio*	0.17	0.75	0.53	0.07
190	*Vocabulary distribution: Word count*	401	107911.00	5879	13257

Fonte: Berber Sardinha, Kauffmann, Acunzo (2014, p. 69 – 74)

ANEXO D – Dimensões do CBVR

Dimensão 1: *Oral versus Literate Discourse*

Feature [label]	***Loading***
Pronouns: Second person, in object position [prn2obl]	0.91
Verbs: First person [vb1]	0.826
Verbs: Mental [vbment]	827
Verbs: Ir future [vbfutir]	0.821
Pronouns: Second person singular, in subject position [prn2sngsubj]	0.763
Verbs: Private [vbpriv]	0.702
Verbs: Action [vbact]	0.689
Adverbs: Não (no) [advnao]	0.674
Pronouns: First person singular, in subject position [prn1sngsubj]	0.633
Adverbs: Time [advtime]	0.629
Pronouns: First person, object position [prn1obl]	0.622
Pronouns: Quantifier [prnqtf]	0.616
Adjectives: Evaluative [adjeval]	0.612
Pronouns: Possessive [prnposs]	0.6
Adverbs: Intensity [advints]	0.552
QU questions [qsqu]	0.539
Adverbs: Amplifier [advampl]	0.522
Adverbs: Emphatic [advemph]	0.476
Que clause controlled by verb in indicative mood [vbqueindic]	0.467
(Adverbs: Place [advpl]	.460)
Adjectives: Predicative position [adjpred]	0.456
(Yes or no question [qsyn]	.421)
Verbs: Infinitive [vbinf]	0.407
Adverbs: Manner [advmanner]	0.373
Verbs: Communication [vbcomm]	0.356
(Discourse marker [discmrkr]	.350)
Verbs: Gerund form, all [vbgerall]	0.345
(Subject omission [subjdrop]	.343)
Modals: Precisar (need to) [mdprecisar]	0.338
Que clause controlled by adverb [clqueeadv]	0.337
Verbs: Progressive preceded by infinitive [vbproginf]	0.33
(Verbs: Future subjunctive mood [vbsubfut]	.323)
Adverbs: Negative, except não [advneg]	0.32
(Subordinating (conditional) clause [cjcond]	.313)
Pronouns: Nominal in subject position [prnnomsubj]	0.308
Reduced progressive clause [vbprogphr]	-0.309

Pronouns: Relative qual or cujo [prnqualcujo]	-0.315
(Adjectives: Affiliative [adjaffi]	-.336)
Agentless passives [clpassless]	-0.377
Adjectives: Relational [adjrela]	-0.422
Past participle [vbpastprt]	-0.497
Nominalization in subject position [nominlzsubj]	-0.505
Adjectives: Topical [adjtopi]	-0.511
Nouns: Abstract [nabst]	-0.521
Average word length [wl]	-0.529
Adjectives: Attributive position [adjattr]	-0.589
Nouns: Compound [ncomp]	-0.651
Articles: Definite [artdef]	-0.739
Prepositions: All [prpall]	-0.776

Dimensão 2: *Argumentation*

Feature [label]	***Loading***
Que clause controlled by noun [nounque]	0.593
Pronouns: Relative que [prnque]	0.529
Adverbs: Comparative [advcomp]	0.473
Nouns: Cognition [ncogn]	0.451
Que or infinitive clause controlled by noun (stance) [nqueinfcl]	0.447
Infinitive clause controlled by adjective [clinfadj]	0.447
Que clause controlled by preposition [clqueeprp]	0.426
Pronouns: Demonstrative [prndem]	0.406
Infinitive clause controlled by preposition [clinfprp]	0.395
Que clause controlled by adjective (stance) [adjque]	0.378
(Adjectives: Predicative position [adjpred]	.353)
(Pronouns: Quantifier [prnqtf]	.352)
(Pronouns: Third person, object position [prn3obl]	.343)
(Modals: Poder [mdpoder]	.336)
Infinitive clause controlled by ease or difficulty adjective [clinfadjease]	0.334
Adverbs: Hedge [advhedg]	0.331
Articles: Indefinite [artindef]	0.325
Verbs: Future preterit tense [vbfutpret]	0.311
Conjunctions: Coordinating (adversative) [cjadv]	0.31

Dimensão 3: *Involved versus Informational Production*

Feature [label]	***Loading***
Tag questions [qsttag]	0.795
Contractions [contrac]	0.714
Discourse marker [discmrkr]	0.671
Questions: Yes or No question [qsyn]	0.547
Pronouns: Third person singular, in subject position [prn3sngsubj]	0.498
Pronouns: Third person plural, in subject position [prn3plusubj]	0.481
Conjunctions: Coordinating (conclusive) [cjcncl]	0.475
Adverbs: Place [advpl]	0.462
Modals: Ter que/ter de (have to, ought to) [mdter]	0.315
(Pronouns: Demonstrative [prndem]	.362)
(Que clause controlled by verb in indicative mood [vbqueindic]	.316)
(Pronouns: First person singular, in subject position [prn1sngsubj]	.314)
Type-token ratio [ttr]	-0.346
(Adjectives: Attributive position [adjattr]	-.357)
(Pronouns: Possessive [prnposs]	-.431)

Dimensão 4: *Directive discourse*

Feature [label]	***Loading***
Verbs: Present subjunctive mood [vbsubpres]	0.821
Verbs: Imperative mood [vbimp]	0.774
Nouns: Concrete [nconc]	0.565
Subject omission [subjdrop]	0.545
Verbs: Facilitation [vbfacil]	0.485
Conjunctions: Coordinating (clausal) [cjcoorcls]	0.465
Adverbs: Manner [advmanner]	.345)

Dimensão 5: *Future versus Past Orientation*

Feature [label]	***Loading***
Verbs: Future subjunctive mood [vbsubfut]	0.616
Conjunctions: Coordinating (ou) [cjou]	0.611
Verbs: Future present tense [vbfutpres]	0.513
Modals: Dever [mddever]	0.474
Modals: Poder [mdpoder]	0.426
Subordinating (conditional) clause [cjcond]	0.39
Adverbs: Likelihood [advlikl]	0.389
Conjunctions: Coordinating (phrasal) [cjcoorphr]	0.322

(Adjectives: Relational [adjrela]	.303)
Nouns: Place [nplac]	-0.301
Verbs: Past subjunctive mood [vbsubpast]	-0.308
(Articles: Indefinite [artindef]	-.320)
Adjectives: Affiliative [adjaffi]	-0.355
Verbs: Imperfect [vbimpf]	-0.375
Verbs: Past indicative tense [vbpast]	-0.554

Dimensão 6: *Reported Discourse*

Feature [label]	***Loading***
Pronouns: Rare in object position [objprnrare]	0.628
Verbs: Second person [vb2]	0.466
Pronouns: Possessive [prnposs]	0.424
Subordinating (final) clause [cjfinal]	0.413
(Que clause controlled by preposition [clqueeprp]	.380)
Pronouns: Third person, object position [prn3obl]	0.371
(Pronouns: Relative que [prnque]	.340)
Verbs: Public [vbpubl]	0.327
Modals: Haver que/haver de (have to, ought to) [mdhaver]	0.311
(Adjectives: Evaluative [adjeval]	-.318)
(Yes or no question [qsyn]	-.330)
(Adverbs: Amplifier [advampl]	-.340)
(Adverbs: Intensity [advints]	-.341)

Fonte: Berber Sardinha, Kauffmann, Acunzo (2014, p. 44 – 59)