DOI: 10.14393/DL52-v16n4a2022-12

Reflexões metodológicas sobre *datasets* e Linguística de *Corpus*: uma análise preliminar de dados legislativos

Methodological reflections on datasets and Corpus Linguistics: a preliminary analysis of legislative data

Lúcia de Almeida FERRARI[*]
Evandro Landulfo Teixeira Paradela CUNHA[**]

RESUMO: Ferramentas e métodos computacionais são, cada vez mais, importantes aliados para a realização de pesquisas no âmbito das humanidades. Em particular, o uso dessas ferramentas é relevante para a análise linguística diacrônica. Neste estudo, é apresentada uma discussão sobre o uso de *corpora* e *datasets* na linguística, destacando algumas potencialidades e limitações desses recursos. Para ilustrar as possibilidades de uso de um *dataset* para pesquisa linguística, apresenta-se, também, uma análise preliminar da Base de Normas Jurídicas Brasileiras.

PALAVRAS-CHAVE: Processamento de texto. *Dataset* de normas jurídicas. Análise diacrônica. Linguagem e direito.

ABSTRACT: Computational tools and methods are increasingly important for conducting research in the humanities. In particular, these tools are relevant for diachronic linguistic analysis. In this study, we present a discussion about the use of corpora and datasets in linguistics, highlighting some strengths and limitations of these resources. To illustrate the possibilities of using a dataset for linguistic research, a preliminary study employing a dataset of Brazilian legal norms is also presented.

KEYWORDS: Text processing. Legal norms dataset. Diachronic analysis. Language and law.

[*] Doutora em Estudos Linguísticos pela Universidade Federal de Minas Gerais (UFMG). Professora na Faculdade de Letras da UFMG. ORCID: https://orcid.org/0000-0002-9855-0646. ferrari.lu@gmail.com.
[**] Doutor em Linguística pela Universitaet Leiden e em Ciência da Computação pela Universidade Federal de Minas Gerais (UFMG). Professor na Faculdade de Letras da UFMG. ORCID: https://orcid.org/0000-0002-5302-2946. cunhae@ufmg.br.

1 Introdução

A relação entre linguagem e direito é profunda: o direito, enquanto regulamentação do comportamento humano, se manifesta por meio da língua, oral ou escrita; ao mesmo tempo, é por meio da linguagem que se discutem as normas de convivência na sociedade civil[1]. No passado distante, os usos e costumes aceitos e compartilhados por uma sociedade eram transmitidos oralmente. No entanto, já na antiguidade surgiu a demanda pela codificação escrita das noções primordiais dos direitos e deveres dos indivíduos. Prescrições sobre o comportamento e a moral, escritas em tábuas de pedra, são aquelas do decálogo bíblico (Êxodo, 20:1-17; 34:1-5). Historicamente, é à Lei das XII Tábuas da República Romana (*Lex Duodecim Tabularum*, 451 a.C.) que se deve a primeira elaboração articulada de uma legislação escrita à disposição de toda a população, para que não fosse manipulada, e que ainda hoje serve de inspiração, em muitos países, para o Direito Público e o Direito Civil. O sistema legislativo de um Estado é, portanto, uma expressão tangível de fenômenos sociais distintos (direito e língua), mas intimamente ligados entre si (MACIEL, 2001, p. 56).

Uma pesquisa bibliográfica sobre o estudo da linguagem jurídica[2] no Brasil aponta para duas grandes áreas: uma ligada ao mundo do Direito propriamente dito e outra, ainda em expansão, mais próxima aos estudos da linguagem. Além de dicionários de terminologia da área (DINIZ, 1998; SANTOS, 2001; GUIMARÃES, 2013; entre outros), há livros didáticos para o ensino de redação e argumentação em textos jurídicos, direcionados especialmente a estudantes de Direito, nos quais a referência à língua é de tipo instrumental ou sobre seu uso retórico (p. ex., DAMIÃO; HENRIQUES, 2020; AQUINO; DOUGLAS, 2017; PETRI, 2017). Volumes mais

[1] Para uma discussão sobre direito e linguagem, ver Warat (1995) e Ivo (2020).

[2] Empregamos, aqui, o termo "linguagem jurídica" por ser ele o mais utilizado e por possuir um significado mais amplo, apesar de concordarmos com Ramos (2017, *apud* SVOBODOVÁ, 2017) sobre o fato de que a linguagem empregada na área do Direito possa ser dividida em dois tipos: "linguagem legal", utilizada nas normas jurídicas, e "linguagem forense", utilizada no foro, ou seja, em tribunal.

específicos abordam as dificuldades encontradas por possíveis redatores de normas ou atos administrativos sobre o uso do vernáculo, em oposição à terminologia e usos da linguagem do domínio jurídico, não somente do ponto de vista da correção, mas também da compreensão por parte de não especialistas[3]. Reflexões epistemológicas ou semióticas, assim como ponderações linguístico-pragmáticas, são encontradas em obras de juristas que discorrem sobre lógica ou ontologia jurídica (p. ex., WARAT, 1995; BITTAR, 2009; IVO, 2020).

Os trabalhos de cunho linguístico em interface com a área jurídica podem estar, por exemplo, no âmbito da análise do discurso jurídico, que tem tradição no Brasil sobretudo graças ao Grupo de Pesquisa Linguagem e Direito e à Associação de Linguagem & Direito (ALIDI). O projeto TermiSul[4], por sua vez, se ocupa de estudos terminológicos e tradutórios específicos em diversas áreas do conhecimento. O grupo tem publicado, ao longo dos anos, desde dicionários e glossários multilíngues de gestão e direito ambiental (KRIEGER *et al.*, 1998, 2006, 2008) até estudos de léxico especializado da linguagem legal. A partir do início dos anos 2000, o grupo se vale de metodologias de linguística de *corpus* em suas pesquisas. O Projeto CoMET – Corpus Multilíngue para Ensino e Tradução[5] compila e disponibiliza *corpora* técnico-científicos desde 2003. Entre os trabalhos da equipe, se destacam o CorTec – Corpus Técnico-

[3] "Juridiquês" é o termo pejorativo que indica o uso de uma linguagem jurídica especialmente arcaica e repleta de latinismos, expressões consideradas pedantes e desnecessariamente rebuscadas e de raciocínio intricado. A Campanha pela Simplificação da Linguagem Jurídica (https://www.amb.com.br/campanha-pela-simplificacao-da-linguagem-juridica-sera-lancada-as-11-horas/. Acesso em: 20 dez. 2021) tem como objetivo "incentivar os magistrados e operadores do Direito a simplificar a linguagem jurídica, valorizando o uso de um vocabulário mais objetivo, simples e direto para aproximar os cidadãos do Poder Judiciário brasileiro". Os termos técnicos deveriam ser mantidos, mas utilizados de forma adequada e destinados a facilitar a compreensão pelo cidadão comum. As ações vão desde campanhas por meio de concursos para estudantes e magistrados promovidas pela Associação de Magistrados Brasileiros (AMB) a projetos de lei para a elaboração de sentenças em linguagem simples (ver, a esse respeito: https://www12.senado.leg.br/noticias/materias/2012/06/27/guerra-contra-o-2018juridiques2019-pode-levar-a-mudancas-em-projetos-de-lei. Acesso em: 18 dez. 2021).

[4] http://www.ufrgs.br/termisul/. Acesso em: 20 dez. 2021.

[5] https://comet.fflch.usp.br/. Acesso em: 20 dez. 2021.

Científico[6], que contempla uma seção de textos da área do Direito sobre instrumentos contratuais, além de artigos e publicações sobre linguagem jurídica voltados ao público de profissionais, como tradutores, advogados e especialistas (p. ex., CARVALHO, 2006). Além disso, nos últimos anos, pesquisadores brasileiros vêm se inserindo no panorama mundial dos eventos e discussões que reúnem legislação, linguagem e computação: um exemplo é a 18th International Conference on Artificial Intelligence and Law (ICAIL 2021)[7], organizada de forma virtual em São Paulo e que possibilitou grande interação entre estudiosos, analistas, juristas, linguistas e programadores do Brasil e do mundo no debate sobre as possibilidades de análises automáticas e semiautomáticas de textos legislativos e jurídicos.

A proposta do presente artigo é realizar um estudo linguístico exploratório de um conjunto de leis brasileiras, tecendo reflexões de caráter metodológico sobre a utilização desse tipo de dado sob o viés da linguística de *corpus* e sobre a diferença entre a utilização de *corpora* e de *datasets*. O intuito do artigo é discutir as possibilidades investigativas dos conjuntos de leis e realizar uma primeira análise de tipo diacrônica. Os dados utilizados são formados por um *dataset* contendo normas jurídicas federais coletado e disponibilizado por Martim, Lima e Araujo (2018), que foi subdividido em períodos e explorado utilizando *scripts* desenvolvidos em linguagem de programação Python.

O artigo está assim dividido. Na primeira parte, se discute a diferença entre *corpora* e *datasets*. Em seguida, antes de introduzir o *dataset* utilizado no presente trabalho, expomos brevemente como ocorre o processo de elaboração de leis no Brasil. A seção seguinte apresenta as etapas e resultados da análise exploratória. As conclusões discutem as observações efetuadas, inclusive de cunho metodológico, e possibilidades de pesquisas futuras.

[6] https://cortec.fflch.usp.br/. Acesso em: 20 dez. 2021.
[7] https://icail.lawgorithm.com.br/. Acesso em: 21 dez. 2021.

2 *Datasets* e *corpora*

Nesta seção, são apresentados alguns conceitos fundamentais relacionados ao que, em linguística, recebe o nome de *corpus* (plural *corpora*) e sua diferença ao que se pode chamar de *datasets*, ponderando sobre as diferentes metodologias e resultados a que esses diferentes tipos de dados podem levar no contexto dos estudos linguísticos.

Datasets, ou conjuntos de dados, podem ser definidos como coleções de dados agrupados por possuírem características em comum[8]. Frequentemente, esses dados são tabulados, com colunas correspondendo às diferentes variáveis, mas podem ser organizados também em forma de diretórios com arquivos do mesmo tipo e, por vezes, até mesmo contendo interfaces gráficas para a visualização dos dados. Tais dados, geralmente empíricos, podem ser de vários tipos: desde observações espaciais a informações meteorológicas, desde prontuários médicos a informações sobre flutuações monetárias, e são a base para investigações de cientistas e analistas de diversas áreas. No caso de *datasets* para estudos linguísticos e/ou processamento de língua/linguagem natural (PLN), seus conteúdos são igualmente diversificados – como, por exemplo, áudios de fala, arquivos de texto, tabulações com dados linguísticos e/ou extralinguísticos diversos etc. A coleta desses dados pode ser realizada com recursos muito diferentes, como gravações de conversas inteiras ou de trechos curtos (palavras ou frases) e textos dos mais variados gêneros (que podem provir, por exemplo, de páginas da internet ou da digitalização de obras de bibliotecas). Geralmente, grandes *datasets* são a base de aplicações em PLN, as quais têm necessitado de extensos volumes de dados para pesquisas ou para o desenvolvimento de produtos como filtros de spam para e-mails, assistentes virtuais inteligentes, corretores automáticos, tradutores automáticos, entre outros.

Datasets não devem ser confundidos com *databases*, ou bancos de dados, que

[8] Para uma discussão mais ampla sobre como o termo *dataset* é utilizado na literatura científica e técnica, ver Renear, Sacchi e Wickett (2010).

armazenam dados usando uma estrutura mais rígida. O termo remete às fichas catalográficas, por muito tempo utilizadas pelos linguistas, mas também, e especialmente, à importância dada à estrutura em que são arquivados e sucessivamente apresentados os dados (DIMITRIADIS; MUSGRAVE, 2009). Hoje em dia, os próprios catálogos dos sistemas de bibliotecas se comunicam com bancos de dados, na busca por obras e autores, por meio digital. O mesmo ocorre com os bancos de dados linguísticos, que frequentemente utilizam um DBMS (*database management system*). Os dados e os motores de busca ficam armazenados em servidores que podem ser acessados, por meio da internet, pelo usuário final, sem a necessidade de se sobrecarregar o equipamento de cada pesquisador ou organização. Muitas bases de dados não possuem uma interface gráfica amigável para usuários não programadores (ou GUI, *graphical user interface*), mas podem ser acessadas por meio de consultas utilizando linguagens de pesquisa específicas (como a SQL, *Structured Query Language*) ou de aplicativos instalados no computador do usuário que permitem tais buscas. Alguns bancos de dados funcionam com uma interface própria e permitem uma série de pesquisas pela própria página web do serviço[9].

Corpora também são considerados conjuntos de dados, mas, no contexto da linguística de *corpus*, sua arquitetura se relaciona a três conceitos básicos: representatividade, amostragem e balanceamento. Um *corpus* deve ser representativo

[9] Bancos de dados linguísticos para PLN são frequentemente organizados e utilizados por grandes empresas e seu acesso, quando possível, costuma ser pago. Para o português brasileiro, os bancos de dados disponíveis gratuitamente entram na categoria dos *corpora*, pois são compilados por grupos de pesquisa ligados a universidades e seguem uma arquitetura precisa. O portal https://www.linguateca.pt/ (acesso em: 10 dez. 2021) é um site que tem por objetivo catalogar, reunir e disponibilizar (no próprio site ou por meio de links) o maior número possível de recursos disponíveis ao público para o processamento computacional do português. O projeto teve seu financiamento suspenso em 2011 e a própria equipe manteve a página até a última atualização, em 2015. Apesar de não haver nele informações mais recentes, trata-se de um ponto de partida valioso pela riqueza de informações e links para pesquisadores do português em geral e do português brasileiro em específico, indicando e disponibilizando desde *corpora*, ontologias, dicionários, até recursos de buscas e ferramentas, além de informações sobre numerosos projetos e pesquisas.

de uma língua ou variedade linguística: isso envolve conhecimentos sobre a população a ser estudada e decisões sobre o tipo de amostragem estatística que será utilizada, ponto de partida para a coleta dos dados linguísticos. A noção de balanceamento se refere ao tamanho de cada amostra e seu peso relativo com relação ao conjunto dos dados do *corpus*. Outro conceito basilar na arquitetura de um *corpus* é sua finalidade, ou seja, o porquê de sua compilação. Um *corpus* de referência é pensado para ser representativo de dada língua ou variedade em sua totalidade: caso seja escrito, levará em conta as diferentes tipologias textuais e sua distribuição no panorama geral das publicações; caso seja oral, serão as variáveis sociolinguísticas que servirão como ponto de partida para a amostragem da população. *Corpora* de referência são pensados como projetos de muitos anos, que demandam equipes treinadas e coesas para sua compilação, além dos financiamentos necessários para sua implementação. Tais *corpora* permitem pesquisas além daquelas para as quais foram inicialmente planejados e são geralmente disponibilizados à comunidade acadêmica, de forma gratuita ou por meio de acordos entre instituições.

Um *corpus* pode, também, ser compilado para uma pesquisa relativamente restrita e pequena, como em trabalhos de mestrado e doutoramento. *Corpora* para finalidades específicas tendem a ser menores em tamanho, pela demora no trabalho de coleta e preparação dos dados, que muitas vezes recai sobre somente uma pessoa, e nem sempre são adequados ao tipo de pesquisa que se pretende realizar. Muitas vezes, após a conclusão do trabalho, esses *corpora* são abandonados por seus autores, perdendo-se todo o trabalho de compilação. É frequente, portanto, que estudiosos utilizem *corpora* já prontos em suas pesquisas, extraindo amostras ou adaptando-os a suas necessidades.

A utilização de *corpora* subdimensionados ou cuja arquitetura não foi devidamente pensada pode se tornar um problema, pois cada tipo de pesquisa linguística requer uma amostragem não somente da população a ser estudada, mas

também dos textos, além da definição do tamanho de cada um deles – ou seja, seu balanceamento[10]. É de se sublinhar que conseguir atingir a devida representatividade de um *corpus* não é somente uma questão de aumentar seu tamanho ou o tamanho de suas amostras: é importante, ainda, que a arquitetura se preocupe em prever as características necessárias e adequadas à pesquisa que se pretende empreender.

Biber (1988, 1990, 1992, 1993) analisa a distribuição de traços linguísticos[11] frequentes em amostras de registros escritos diferentes. Sua conclusão é de que há uma grande diversificação nos itens linguísticos encontrados à medida que há variação de textos, e estabilização dessa variação quando os textos são muito longos e pouco variados. Componentes linguísticos comuns (como nomes e preposições) apresentam uma distribuição estável entre textos diferentes e podem ser encontrados em amostras relativamente pequenas (com cerca de 2000 palavras). Componentes linguísticos menos frequentes (como verbos conjugados no passado ou no futuro) necessitam de amostras maiores pois sua distribuição varia muito mais dentro dos mesmos textos.

Estudos em *corpora* sobre dispersão lexical (BIBER *et al.*, 2016; EGBERT; BURCH; BIBER, 2020) concluem que subdividir dados de *corpora* em trechos aleatórios de mesmo tamanho para análises pode gerar resultados distorcidos, pois perde-se o contexto específico em que os itens comparecem. Novamente, essas pesquisas apontam para a necessidade de muita cautela na amostragem dos dados. A dispersão lexical resulta ser muito maior pelo efeito do contexto (e dos próprios textos) quando são extraídos de um *corpus* trechos de forma arbitrária do que quando esses fazem parte de agrupamentos linguisticamente relevantes. Em outras palavras, a amostragem deve se preocupar com as tipologias textuais, enquanto a taxa de

[10] Egbert, Larsson e Biber (2020) discutem e exemplificam em detalhes tais problemas. Aqui, nos limitaremos a citar alguns pontos que reputamos importantes sem nos adentrarmos na ampla discussão metodológica dos autores.

[11] Biber (1993) utiliza o termo *features*, traduzido em português como *traços* por Resende e Maverick (2016), para indicar os elementos analisáveis e quantificáveis nas investigações em *corpora*.

dispersão lexical varia muito de texto para texto, sendo necessário levar em conta seus contextos específicos. Análises desse tipo devem se atentar, portanto, a uma cuidadosa seleção dos registros que comporão o *corpus*.

Em *corpora* orais, a representatividade dos textos é obtida geralmente por meio da seleção dos informantes. A tradição sociolinguística (LABOV, 1966, 1972a; BAKER, 2010) fornece domínios específicos para as interações formais: há, portanto, situações como a entrevista estruturada ou semiestruturada, mas também palestras, contextos de sala de aula, interações profissionais, religiosas, políticas, entre outras. Já no caso de interações espontâneas, o domínio é totalmente aberto e a coleta de gravações com ampla variação diafásica tem se mostrado muito produtiva para a obtenção de dados linguísticos extremamente variados e com excelente balanceamento entre sexo e idade, mantendo-se um grau de instrução médio-alto (RASO, 2012; MELLO, 2014). Partir de situações comunicativas heterogêneas e totalmente espontâneas permite: (a) atingir diferenças de natureza diastrática, pois variam muito os informantes; e (b) determinar variações de natureza informacional e ilocucionária, ou seja, atos linguísticos diferentes (SEARLE, 1969), enquanto situações diversas acarretam estruturas organizacionais variadas entre os textos e, como consequência, léxico, morfologia e sintaxe diversificados.

A disponibilização de *corpora* compilados seja por projetos maiores, seja por iniciativas individuais é outro ponto relevante para a pesquisa empírica. Um *corpus*, especialmente de fala ou multimodal, mas mesmo escrito, é um recurso informático que ocupa um espaço virtual grande em termos de armazenamento. Seu compartilhamento, até entre membros da própria equipe, requer que os dados sejam carregados em algum site ou "em nuvem". Não é raro que tais dados sejam retirados do ar após a finalização do projeto ou quando os financiamentos são suspensos, pois manter espaço de armazenamento e páginas na web pode ser relativamente dispendioso.

Em vista de tais problemas, o CQPweb[12] tem se tornado um recurso interessante para vários projetos: trata-se de uma interface gráfica online de elementos de processamento de pesquisa, especificamente do CQP (*Corpus Query Processor*). Sua base é o CWB (*Corpus Workbench*), uma coletânea de ferramentas de código aberto para o processamento e consultas de grandes *corpora* com anotações linguísticas. O Linguateca AC/DC[13] utiliza a interface do CQPweb para a publicação e consultas a seus recursos, que incluem *corpora* do português (inclusive português brasileiro), assim como o CLUL (Centro de Linguística da Universidade de Lisboa)[14], que disponibiliza *corpora* abarcando variedades africanas e asiáticas do português.

A diferença entre se disponibilizar um *dataset* ou *corpus* (com e/ou sem anotações) em uma plataforma para que seja baixado pelo usuário ou em outra que possibilite consultas online se relaciona com a capacidade de processamento e com as competências computacionais do pesquisador[15]. Dados volumosos requerem competências técnicas nem sempre viáveis para todos os linguistas. Como afirma Hardie (2012),

> Alguns pesquisadores em linguística de corpus com maior conhecimento técnico recomendam uma abordagem 'faça você mesmo' para ferramentas de análise de *corpora*, em que, em vez de utilizar um concordanciador pronto, o pesquisador escreve seus próprios programas de computador para processar seus dados. [...] Claramente, essa abordagem 'faça você mesmo' para softwares de análise de *corpora* é a mais poderosa imaginável se a capacidade [de processamento] for equiparada à adaptabilidade – embora os programas 'faça você mesmo' possam ser executados lentamente caso não sejam

[12] Para as especificações da plataforma e uma descrição detalhada de seus recursos, ver Hardie (2012).
[13] Disponível em: https://www.linguateca.pt/ACDC/. Acesso em: 18 dez. 2021.
[14] As buscas nos *corpora* disponibilizados pelo CLUL podem ser realizadas na página: http://gamma.clul.ul.pt/CQPweb/. Acesso em: 5 mar. 2022.
[15] A título de exemplo sobre tais possibilidades, citamos o projeto C-ORAL-BRASIL, que disponibiliza seus *corpora* de fala para download no site http://www.c-oral-brasil.org/ (acesso em: 15 dez. 2021), mas também possui uma base de dados própria, o DB-CoM (*Database for Corpora Multimidia*), que possibilita a consulta online a alguns de seus mini *corpora* etiquetados informacionalmente em http://www.c-oral-brasil.org/db-com (acesso em: 15 dez. 2021).

> acrescentados os 'truques', como indexação, necessários para a alta velocidade em grandes conjuntos de dados. No entanto, essa abordagem é insuficiente em termos de usabilidade para a maioria dos linguistas de *corpus* em potencial, que não podem ou não desejam aprender programação de computadores [...]. Para esses pesquisadores, a usabilidade sempre será mais relevante que a capacidade de processamento" (tradução nossa) (HARDIE, 2012, p. 383)[16].

Ferramentas gratuitas como o AntConc (disponível em https://www.laurenceanthony.net/software/antconc/, acesso em: 10 dez. 2021) conseguem processar *corpora* de tamanhos médios de forma geralmente satisfatória e possibilitam análises cada vez mais sofisticadas. Todavia, quando os dados provêm de *corpora* ou de *datasets* de tamanho muito grande, tais programas podem não suportar o volume de dados. Nesses casos, acaba por se fazer necessária a participação de alguém que saiba produzir seus próprios códigos. Essa tarefa é ainda mais importante quando se pensa que *datasets* não apresentam necessariamente uma arquitetura como, em teoria, os *corpora*, e devem, por sua vez, ser ainda mais cuidadosamente amostrados e selecionados. O trabalho aqui apresentado mostrará, justamente, parte da metodologia e das dificuldades encontradas no processamento de dados previamente coletados e disponibilizados no formato de *dataset*.

Passaremos agora a um breve panorama sobre o processo legislativo brasileiro para, em seguida, seguirmos à apresentação dos dados inicialmente investigados neste artigo e a suas análises.

---

[16] [s]ome technically sophisticated corpus researchers recommend a "do-it-yourself" approach to corpus analysis tools, where rather than exploiting an off-the shelf concordancer, the analyst instead writes their own computer programs to process their data. [...] Clearly this "do-it-yourself" approach to corpus analysis software is the most powerful imaginable if power is equated to maximum scope for adaptability – although "do-it-yourself" programs may run slowly if they do not incorporate the "tricks", such as indexing, needed for high speed on large datasets. However, this approach falls short in terms of usability for the majority of potential corpus analysts, who either cannot or do not wish to learn computer programming [...]. For such researchers, usability will always be more critical than power.

3 O processo legislativo no Brasil e o *dataset* de normas jurídicas brasileiras

O Brasil é uma república federativa presidencialista: a União (governo federal), os 26 estados e o Distrito Federal, além de seus mais de 5000 municípios, formam um estado democrático com divisão entre três poderes. No âmbito federal, o Poder Legislativo é exercido pelo Congresso Nacional, o Executivo pela Presidência da República e o Poder Judiciário pelos Supremo Tribunal Federal, Conselho Nacional de Justiça, Superior Tribunal de Justiça, Tribunal Superior do Trabalho, tribunais regionais federais, tribunais eleitorais, tribunais militares, tribunais de justiça dos estados e do Distrito Federal e territórios.

A Constituição Federal é a lei superior que orienta sobre as competências de cada poder no nível federal, detalhando sobre quais são seus campos de atuação e determinando que um poder fiscalize o outro. O Congresso Nacional, por exemplo, além de legislar, anualmente deve julgar as contas prestadas pela Presidência da República e examinar os relatórios sobre a execução dos planos de governo. A Presidência da República organiza e atua no funcionamento da administração federal, além de se ocupar das relações com os outros países e nomear ministros ou vetar projetos de lei. Ao Poder Judiciário compete zelar pelo cumprimento da Constituição nas diferentes instâncias.

O processo legislativo brasileiro, a partir da Constituição Federal de 1988, prevê que o conjunto de atos para a produção de normas jurídicas (ou leis) passe pelas duas casas do Congresso Nacional: a Câmara dos Deputados e o Senado Federal. O processo bicameral das leis federais estabelece que um projeto de lei, que pode ser proposto por um deputado ou por um senador, após ser aprovado por uma comissão que o avaliará e pela casa proponente, deve passar pela outra casa – a casa revisora –, podendo ser ali modificado. Caso isso ocorra, o projeto retorna à casa proponente, que pode decidir por acatar ou não as mudanças no texto. O quórum necessário para a aprovação do projeto de lei depende do tipo de lei em questão.

Em seguida, o projeto de lei passa para a sanção da Presidência da República: em caso de desacordo, o projeto pode ser vetado em sua integralidade ou em partes. Todavia, caso deputados e senadores não estejam de acordo com o veto, eles podem invalidar, isto é, rejeitar, o veto da Presidência. Em caso de sanção, o projeto é promulgado e, a partir desse ato, se torna lei, dependendo, contudo, de publicação para que tenha validade.

As emendas à Constituição contam com regras próprias, seja em sua proposta, que prevê um quórum mínimo de 1/3 de deputados ou senadores, seja pelos turnos e quóruns necessários à sua aprovação. Há, contudo, algumas cláusulas da Constituição (as cláusulas pétreas) que não podem ser modificadas ou abolidas, sendo elas: o Estado federal; o voto como direito, secreto, universal e periódico; a separação dos poderes; e os direitos e garantias individuais.

O Poder Legislativo, por meio de suas casas, produz e coleta uma quantidade significativa de materiais, pois, além de todos os atos administrativos e de fiscalização, cada projeto de lei é registrado em suas sucessivas modificações até sua eventual sanção.

3.1 O *dataset* de normas jurídicas brasileiras de Martim, Lima e Araujo (2018)

O *Open Government Data* (OGD), em português Dados Governamentais Abertos (DGA), é uma iniciativa internacional que promove a transparência e responsabilidade ética das informações por parte dos governos (cf. https://publicadministration.un.org/en/ogd, acesso em: 10 dez. 2021). Os países que aderem ao DGA se empenham em promover iniciativas e políticas que possibilitem a disponibilização de seus dados a todos os cidadãos, de maneira a tornar as instituições públicas mais transparentes e direcionadas à colaboração democrática de seus habitantes.

Tal abertura se iniciou nos anos 2000 (GRAY, 2014), tendo o Brasil aderido ao

DGA em setembro de 2011, com uma série de ações políticas e técnicas para a publicação na web de dados oficiais, que culminaram no lançamento do Portal Brasileiro de Dados Abertos (https://dados.gov.br, acesso em: 20 dez. 2021). A Controladoria-Geral da União (CGU) se ocupa da gestão e monitoramento da Política de Dados Abertos, através da Infraestrutura Nacional de Dados Abertos (INDA), a qual atua por meio de padrões, tecnologias e orientações para a disseminação e o compartilhamento de dados e informações públicas.

Martim, Lima e Araujo (2018) coletaram e publicaram uma base de normas jurídicas federais acessível online[17]. Trata-se de um conjunto de oito *datasets* que contêm, em princípio, todas as normas legislativas federais desde 4 de outubro de 1946 até 12 de abril de 2017. Segundo os autores, o recorte metodológico selecionou "uma quantidade razoável de normas com maior probabilidade de estarem vigentes. Isso não exclui normas expressamente revogadas, pois o histórico dos textos é importante para a obtenção do texto vigente em uma determinada data" (MARTIM; LIMA; ARAUJO, 2018, p. 137). Os autores do *dataset* optaram por não incluir, na captura de dados, os Decretos, Decretos-Leis, Emendas Constitucionais e outros atos ou proposições normativas.

Os *datasets* que compõem a base, intitulada Base de Normas Jurídicas Brasileiras, estão assim divididos:

(a) *dataset* 1: contém os textos articulados das normas – para cada lei há um arquivo em formato .rtf com o conteúdo legislativo completo;

(b) *dataset* 2: contém as representações LexML dos textos articulados das normas – cada artigo é estruturado com o esquema XML em padrão LexML;

(c) *dataset* 3: nele foram processados em formato .txt os dados do *dataset* 2 e foram subdivididas as sentenças da epígrafe, ementa, preâmbulo, dispositivos e fecho das normas, de maneira a possibilitar pesquisas mais pontuais. Contudo, essa divisão

[17] Disponível em: https://doi.org/10.6084/m9.figshare.c.4029253.v1. Acesso em: 12 nov. 2021.

acabou gerando uma série de sentenças incompletas, fazendo-se necessária a criação do *dataset* 4;

(d) *dataset* 4: é formado por um conjunto de arquivos contendo somente os dispositivos agregadores, ou seja, com acréscimo nas sentenças de incisos e alíneas, apresentados em .txt, como dispositivos distintos;

(e) *dataset* 5: é formado por arquivos em formato .txt de cada dispositivo, com anotação gramatical em formato CoNLL-U;

(f) *dataset* 6: contém, para cada dispositivo, arquivos em formato .json que foram submetidos a análise sintática e identificação de entidades pela API Google Natural Language Processing;

(g) *dataset* 7: são apresentadas, em formato .txt, as sentenças dos dispositivos das normas de maneira completa e, em arquivos separados, suas respectivas ementas;

(h) *dataset* 8: são apresentados os arquivos .json das normas e de suas respectivas ementas, também processados pela API Google Natural Language Processing.

Teixeira *et al.* (2019) utilizaram a Base de Normas Jurídicas Brasileiras para testar um processo de extração de definições utilizando o *dataset* 8. O processo de extração empregou filtros heurísticos, que utilizam um conjunto de regras diversas, para: (a) filtrar sentenças candidatas utilizando expressões em que comparece o verbo *ser* + artigo; (b) isolar as leis que não foram corretamente parseadas pelo esquema LexML; (c) filtrar as sentenças com o verbo *ser* etiquetadas morfossintaticamente; (d) retirar as leis autorizativas, as quais possuem estrutura VSO; (e) extrair do conjunto de dados as entidades nomeadas, ou seja, itens utilizados como designadores específicos como nomes próprios de pessoas, locais ou organizações, datas, valores monetários, entre outros.

O trabalho apresentado na próxima seção utiliza os dados do *dataset* 7 para realizar uma série de análises preliminares e avaliar possíveis caminhos de investigação para pesquisas linguísticas utilizando os dados em questão.

4 Exploração e análise da Base de Normas Jurídicas Brasileiras

Nesta seção, são apresentados resultados preliminares referentes à exploração da Base de Normas Jurídicas Brasileiras disponibilizada por Martim, Lima e Araujo (2018). As análises implementadas aqui utilizaram *scripts* desenvolvidos na linguagem de programação Python[18]. Os principais objetivos foram realizar uma caracterização geral dos dados a partir de uma perspectiva quantitativa e efetuar análises lexicais que levassem em consideração mudanças ao longo do tempo no *dataset* estudado, demonstrando algumas das possibilidades do uso de ferramentas de programação relativamente simples para análise diacrônica. Todas as explorações apresentadas aqui foram realizadas utilizando o *dataset* 7 da Base de Normas Jurídicas Brasileiras – isto é, aquele contendo os textos das normas legislativas completos (em teoria) e, em arquivos separados, suas respectivas ementas, todos em formato .txt, totalizando 25.944 arquivos.

Embora a Base de Normas Jurídicas Brasileiras tenha sido previamente processada pelos seus autores, verificou-se a necessidade de realizar uma etapa adicional de pré-processamento no *dataset* analisado para uma melhor compreensão dos dados. Nessa etapa, notaram-se dois erros que merecem destaque:

(1) no *dataset* original, a Lei 13.407/2016 está duplicada, constando seja no diretório 'LEI-2016-13407' (correto), seja no diretório 'LEI-2016-13047' (com inversão dos dígitos '0' e '4'). Isso acabou gerando conflito com a Lei 13.047/2014 (a 'verdadeira' lei de número 13.047). Para resolver essa questão, optou-se por excluir o diretório 'LEI-2016-13047' e, consequentemente, os dois arquivos que dele faziam parte ('LEI-2016-13047-dispositivos.txt' e 'LEI-2016-13047-ementa.txt');

(2) alguns arquivos estão em branco, inconsistentes ou incompletos. Isso foi constatado

[18] Para fins de reprodutibilidade e para oferecer suporte a pesquisas relacionadas, os *scripts* utilizados durante a realização deste trabalho estão disponíveis em https://github.com/evandrocunha/dados-legislativos.

apenas em arquivos referentes a dispositivos das normas, nunca naqueles referentes às ementas. Por exemplo, o arquivo 'LEI-1979-06765-dispositivos.txt' está em branco, apesar de a ementa correspondente ('LEI-1979-06765-ementa.txt') estar completa. Foram identificados sete arquivos em branco no *dataset*. Detectaram-se, ainda, arquivos contendo dispositivos de normas com um número muito baixo de palavras, o que nos pareceu atípico. Ao analisar os arquivos contendo menos de dez palavras, encontraram-se casos como os dos arquivos 'LEI-1951-01326-dispositivos.txt' (onde se lê apenas "O Quadro de Oficiais Farmacêuticos da Aeronáutica compor-se-á:"), 'LEI-1994-08883-dispositivos.txt' (onde se lê apenas "(VETADO). (VETADO).") e 'LEI-2004-11003-dispositivos.txt' (onde se lê apenas "2.2 - ....."). Nesses e em outros casos, o que consta nos arquivos não corresponde ao conteúdo real das respectivas normas[19]. Ocorrências similares foram identificadas em arquivos contendo entre dez e vinte palavras, mas, nesses casos, havia, também, normas contendo textos efetivamente curtos: é o caso, por exemplo, do arquivo 'LEI-2011-12449-dispositivos.txt', que contém apenas o seguinte texto: "O ator Paulo Autran é declarado Patrono do Teatro Brasileiro. Esta Lei entra em vigor na data de sua publicação.", que corresponde, de fato, ao conteúdo da Lei 12.449/2011[20]. Apesar dessas falhas nos dados, realizou-se a opção metodológica de manter, para a execução das análises aqui apresentadas, todos esses arquivos. Duas razões nos levaram a essa

[19] Para efeito de comparação, os textos integrais das normas mencionadas estão disponíveis em: Lei 1.326/1951: http://www.planalto.gov.br/ccivil_03/leis/1950-1969/l1326.htm; Lei 8.883/1994: http://www.planalto.gov.br/ccivil_03/leis/l8883.htm; Lei 11.003/2004: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2004/lei/l11003.htm. Acesso em: 2 nov. 2021.

[20] Ainda assim, observa-se uma inconsistência na forma como os textos das normas são disponibilizados no *dataset*. Em alguns arquivos, está presente o fechamento do texto; em outros, não. Por exemplo: o arquivo 'LEI-1954-02158-dispositivos.txt' é finalizado com "Rio de Janeiro, em 2 de janeiro de 1954; 133º da Independência e 66º República. GEtúlio VARGAS João Goulart"; já o arquivo contendo os dispositivos da norma que declara o ator Paulo Autran Patrono do Teatro Brasileiro, conforme já indicado, não inclui os dizeres correspondentes (que seriam, nesse caso: "Brasília, 15 de julho de 2011; 190º da Independência e 123º da República. DILMA ROUSSEFF Anna Maria Buarque de Hollanda").

decisão: em primeiro lugar, as análises preliminares realizadas aqui não exigem, necessariamente, acesso ao conteúdo completo das normas; em segundo lugar, a maioria dos mais de vinte mil arquivos parecem estar satisfatórios – portanto, a tendência é que os poucos arquivos problemáticos se percam em meio à grande massa de dados.

A identificação desses erros no *dataset* se deu durante a etapa de caracterização quantitativa dos dados. Ainda que o impacto nos dados analisados não tenha sido grande (apenas dois arquivos foram removidos), isso demonstra a importância dessa etapa, frequentemente negligenciada. Para os estudiosos da linguagem, sua importância é ainda maior quando se sabe que os dados disponibilizados para a pesquisa não foram previamente preparados por uma equipe contendo um/a linguista, como é o caso da Base de Normas Jurídicas Brasileiras utilizada aqui.

Assim, apesar de Martim, Lima e Araujo (2018, p. 143) informarem que "[h]á 25.944 arquivos nesse dataset, sendo metade com arquivos das articulações das normas, metade com arquivos das ementas das normas", o número utilizado neste trabalho, após a remoção da norma duplicada, é 25.942 arquivos (sete deles em branco e muitos deles incompletos) – correspondendo, portanto, a 12.971 normas, cada uma com um arquivo referente à sua ementa e outro aos seus dispositivos.

Para a realização das análises a partir de uma perspectiva diacrônica, o *dataset* foi dividido em três períodos: I (de 04/10/1946 a 31/03/1964), II (de 01/04/1964 a 04/10/1988) e III (de 05/10/1988 a 12/04/2017). Essa periodização leva em consideração importantes mudanças no cenário legislativo brasileiro. O ano de 1946 inaugura o período I por ser a data de promulgação da quinta Constituição brasileira. O próprio *dataset* utilizado aqui tem início com a Lei 1/1946[21], de 04 de outubro daquele ano. O

[21] Conforme informado pelo Centro de Estudos Jurídicos da Secretaria-Geral da Presidência da República, "a partir de 04.10.1946, teve início a numeração de leis ordinárias que vigora até hoje" (informação disponível em: http://www4.planalto.gov.br/centrodeestudos/assuntos/legislacao/reflegis. Acesso em: 20 out. 2021).

ano de 1964, mais precisamente no dia 1 de abril, marca o fim desse período democrático, com a instauração da ditadura militar – ainda que a substituição da Constituição tenha ocorrido apenas em 1967. Finalmente, o período III tem início com a promulgação da Constituição de 1988, a chamada "Constituição Cidadã", no dia 5 de outubro, e vai até o final do período incluído no *dataset*, em 2017. Mesmo que a ditadura militar tenha terminado em 1985, considera-se que apenas com a nova Constituição os direitos sociais tenham sido, de fato, ampliados e universalizados. Adota-se aqui, portanto, a periodização empregada por autores como Santos (1997), cuja análise considera "[d]e um lado, [o período] que se estende de 1946 a 1964, regulado pela Constituição de 1946; de outro, o período atual, marcado pela Constituição de 1988". No caso do nosso estudo, acrescenta-se, ainda, o período intermediário (1964-1988), orientado, inicialmente (1964-1967), por normas emanadas fora do Estado Democrático de Direito; e, posteriormente (1967-1988), pela própria Constituição que institucionalizou e legalizou o regime militar.

O *dataset* analisado contém 6.212.514 palavras, sendo 371.526 nas ementas e 5.840.988 nos dispositivos[22]. Como informado anteriormente, é importante salientar que, em geral, softwares amplamente utilizados em linguística de *corpus* (como o AntConc, o WordSmith Tools, entre outros) podem não ser capazes de processar adequadamente quantidades de dados dessa magnitude, motivo pelo qual a programação de computadores pode ser útil para essas tarefas. Informações gerais sobre o *dataset*, inclusive considerando sua divisão em períodos, são mostradas na Tabela 1. É interessante observar como o valor da média de palavras nos dispositivos difere do valor da mediana para esse mesmo tipo de texto (450,3 *vs*. 120 no *dataset* completo; 246,2 *vs*. 107 no período I; 428,5 *vs*. 130 no período II; 624,4 *vs*. 131 no período III). Isso revela uma grande variabilidade na quantidade de palavras por texto, com

[22] Neste trabalho, o número de palavras contabilizado considera todos os *tokens* exclusivamente alfanuméricos.

certos textos contendo valores atípicos (*outliers*), isto é, apresentando grandes afastamentos dos demais valores presentes no conjunto de dados. O mesmo fenômeno não é observado nas ementas: nesses textos, as médias de palavras se aproximam das medianas, sugerindo uma maior uniformidade na distribuição dos valores ao longo do *dataset*.

Tabela 1 – Caracterização quantitativa do *dataset* 7 da Base de Normas Jurídicas Brasileiras após pré-processamento.

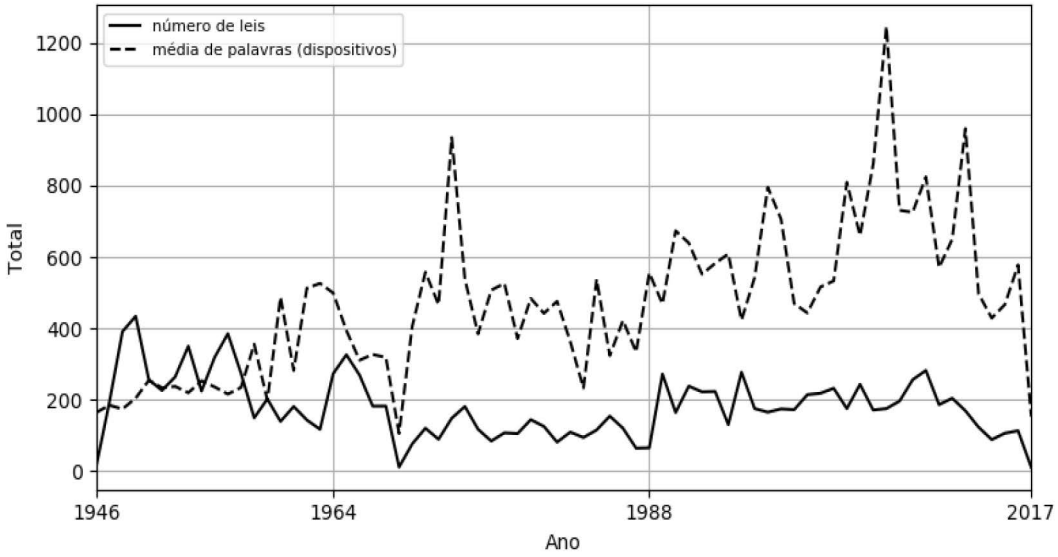
Período	Número de normas	Palavras			
		Tipo dos textos	Total de palavras	Média por texto	Mediana
Dataset completo (1946-2017) [Lei 1/1946 a 13.435/2017]	12.971	Ementas	371.526	28,6	24
		Dispositivos	5.840.988	450,3	120
Período I (1946-1964) [Lei 1/1946 a 4.320/1964]	4.260	Ementas	107.100	25,1	22
		Dispositivos	1.048.748	246,2	107
Período II (1964-1988) [Lei 4.321/1964 a 7.675/1988]	3.301	Ementas	76.838	23,3	21
		Dispositivos	1.414.444	428,5	130
Período III (1988-2017) [Lei 7.676/1988 a 13.435/2017]	5.410	Ementas	187.588	34,7	28
		Dispositivos	3.377.796	624,4	131

Fonte: elaborada pelos autores.

A Figura 1 complementa a Tabela 1 ao mostrar o número de leis e a média de palavras por lei (considerando apenas os dispositivos, não as ementas) em cada ano incluído no *dataset*. Observa-se uma variação natural nos valores; entretanto, é interessante notar alguns padrões. No período II, o número de leis por ano é consistentemente mais baixo que nos períodos I e III. Com relação à média de palavras por lei, convém observar que esse valor mostra uma tendência geral e gradual de crescimento desde 1970, pouco após a emissão do Ato Institucional Número Cinco (AI-

5), em 1968, e da Lei de Segurança Nacional de 1969, que marcaram o início do período que ficou conhecido como "anos de chumbo" na historiografia brasileira (CORDEIRO, 2009). Esse crescimento continua, também, no período III. Os números de leis nos extremos do dataset (anos de 1946 e 2017) são naturalmente mais baixos pois o *dataset* inclui apenas alguns meses desses anos.

Figura 1 – Número de leis e média de palavras por lei (dispositivo) em cada ano incluído no *dataset*.



Fonte: elaborada pelos autores.

O objetivo das análises apresentadas a seguir é comparar alguns aspectos linguísticos observáveis quantitativamente nas normas jurídicas brasileiras nos três períodos considerados. Por se tratar de uma análise preliminar, o *dataset* não passou por nenhum processo de anotação (por exemplo, anotação sintática), nem mesmo por um processo de lematização (convém recordar, no entanto, que versões da Base de Normas Jurídicas Brasileiras lematizadas e anotadas sintaticamente estão disponíveis nos *datasets* 5 e 6, não examinados neste estudo). Os únicos aspectos modificados nos textos foram a padronização ortográfica em caixa baixa e a remoção de pontuação.

4.1 Razão forma/item (type/token ratio)

A razão forma/item – ou, como é mais conhecida, razão *type*/*token* – indica a variabilidade, diversidade ou riqueza lexical do texto. Esse valor é obtido dividindo-se o número de palavras únicas (isto é, sem contar suas repetições – os *types*) pelo total absoluto de palavras (incluindo repetições – os *tokens*). Conforme explica Berber Sardinha (2004, p. 94), quanto maior for seu resultado, mais palavras diferentes (proporcionalmente) o texto conterá: uma razão forma/item próxima ao valor 1 indica um altíssimo grau de variabilidade lexical (pois o número de *types* se aproxima ao de *tokens*), ao passo que uma razão forma/item mais baixa sugere uma maior repetição dos itens lexicais ao longo do texto.

Aqui, comparou-se a razão forma/item entre os três períodos analisados. Os resultados obtidos estão indicados na Tabela 2. Observa-se uma tendência de diminuição na variabilidade lexical ao longo do tempo, passando de 0,0179 (1,79%) no período I a 0,0151 (1,51%) no período II e, posteriormente, a 0,0088 (0,88%) no período III.

Tabela 2 – Razão forma/item (*type/token ratio*) nos três períodos analisados no *dataset*.

Período	Razão forma/item
Período I (1946-1964)	0,0179
Período II (1964-1988)	0,0151
Período III (1988-2017)	0,0088

Fonte: elaborada pelos autores.

Alguns fatores podem explicar, ao menos parcialmente, essa tendência de diminuição. Em primeiro lugar, é preciso considerar o fenômeno conhecido como Lei de Herdan ou Lei de Heaps (HERDAN, 1964; HEAPS, 1978), segundo a qual a probabilidade de uma palavra nova (isto é, que não apareceu ainda) entrar em um

texto/*corpus* é menor à medida que o texto/*corpus* cresce. O motivo para isso é trivial: como o léxico é limitado, sempre que o texto/*corpus* cresce há menos palavras novas disponíveis para ingressar. Já que o *dataset* é menor (em número de palavras) no período I e, progressivamente, cresce nos períodos II e III, essa poderia ser uma explicação para o fenômeno.

Uma outra possível explicação para o fenômeno é que nas últimas décadas foram desenvolvidos manuais de padronização de textos oficiais, inclusive de normas jurídicas. Com isso, a tendência é que haja mais uniformização no léxico presente nesses textos, com uma maior repetição de estruturas fixas e uma maior reincidência de padrões linguísticos.

Por fim, é necessário considerar as discussões recentes sobre a necessidade de simplificação da linguagem jurídica (vide discussão anterior sobre o "juridiquês"). É possível que a diminuição da variabilidade lexical ao longo do tempo esteja relacionada, de fato, a esforços para a simplificação dos textos das normas jurídicas. Uma análise mais aprofundada da diversidade lexical poderá elucidar quais desses fatores possuem, de fato, maior influência na tendência de diminuição da variabilidade lexical observada ao longo do tempo.

4.2 *N*-gramas frequentes

A observação das palavras e sequências de palavras mais frequentes em diferentes trechos do *dataset* pode fornecer sugestões a respeito de elementos relevantes em cada período analisado. Nesta seção, são investigados os *n*-gramas mais frequentes em cada período, com *n* igual a 1 e a 2 – isto é, são analisadas as palavras isoladamente (1-gramas, ou unigramas) e pares de palavras contíguas (2-gramas, ou bigramas).

Para esta análise, foram removidas as *stop words* ("palavras vazias"), ou seja, palavras consideradas irrelevantes para as análises. A lista inicial de palavras vazias

utilizada foi aquela disponibilizada para a língua portuguesa pelo Natural Language Toolkit – NLTK (BIRD; KLEIN; LOPER, 2009) (https://www.nltk.org/, acesso em: 30 nov. 2021), biblioteca para processamento de texto disponível para Python. Essa lista inclui 204 itens, dentre os quais se destacam palavras funcionais/gramaticais (de classe fechada, como artigos, conjunções, preposições, pronomes) e verbos de alta frequência ("estar", "ir", "haver", "ser", "ter").

A primeira análise realizada foi a comparação entre as vinte palavras lexicais (substantivos, verbos, adjetivos e advérbios) mais frequentes em cada período, considerando apenas aquelas compostas exclusivamente por caracteres alfabéticos. Como esperado, em todos os períodos, tais buscas retornaram, sobretudo, termos ligados ao gênero textual legislativo, como "lei", "artigo", "publicação" e "vigor". No período I (1946-1964), o verbo "entrar" utilizado performativamente ("entrará em vigor") é encontrado no futuro, assim como no período II, sendo substituído pelo presente no período III ("entra em vigor"). No período II, destaque para o termo "militar", que também comparece entre os mais frequentes.

Para a análise de bigramas por período, utilizou-se a função *collocations* disponibilizada pelo NLTK, que permite a identificação de bigramas cuja frequência de combinação entre seus elementos é mais alta que o estatisticamente esperado. Os vinte bigramas com associação mais relevante (de acordo com o NLTK) em cada período são apresentados no Quadro 1. Novamente, se destaca a presença de itens próprios do gênero textual legislativo ("sua publicação", "esta lei", "lei entrará", "lei entra", "caput deste", entre tantos outros) – mas, também, de colocações referentes a questões orçamentárias (por exemplo, "mil cruzeiros", "por cento", "crédito suplementar", "orçamento fiscal", "recursos necessários"), instituições e serviços públicos (como "previdência social" no período I, "tribunal regional" no período II, "congresso nacional" no período III), e até mesmo um antropônimo ("juscelino kubitschek", no período I).

Quadro 1 – Bigramas com associação mais relevante de acordo com a função *collocations* do NLTK.

Período	**Bigramas**
Período I (1946-1964)	poder executivo; sua publicação; crédito especial; esta lei; lei entrará; executivo autorizado; pelo ministério; desta lei; juscelino kubitschek; mil cruzeiros; distrito federal; presente lei; para atender; que trata; por cento; outras providências; obras públicas; acôrdo com; nos têrmos; previdência social
Período II (1964-1988)	esta lei; desta lei; sua publicação; poder executivo; distrito federal; que trata; outras providências; bem como; neste artigo; lei entra; lei entrará; por cento; crédito especial; executivo autorizado; presente lei; nesta lei; parágrafo único; tribunal regional; seguinte redação; poderá ser
Período III (1988-2017)	desta lei; que trata; deste artigo; poder executivo; esta lei; sua publicação; lei entra; por cento; nos termos; bem como; caput deste; crédito suplementar; distrito federal; seguridade social; orçamento fiscal; metros até; congresso nacional; neste artigo; recursos necessários; para atender

Fonte: elaborado pelos autores.

Uma forma interessante de visualização geral de *n*-gramas são as nuvens de palavras (*word clouds*), que permitem observar aqueles mais frequentes no conjunto de textos. Para efeito de comparação, a Figura 2 mostra as nuvens de palavras referentes aos três períodos analisados neste estudo. Os resultados encontrados, à primeira vista, não indicam uma mudança tão significativa no padrão de construção dos bigramas frequentes nos três períodos analisados do *dataset*, conforme já havia sido antecipado pelas colocações identificadas no Quadro 1. Uma pesquisa que remova mais elementos irrelevantes para a investigação (por exemplo, que inclua palavras como "lei", "artigo", "publicação", "disposições" e "vigor" na lista de *stop words*) poderá, possivelmente, promover análises mais interessantes da mudança lexical nesse *dataset* ao longo do tempo.

Figura 2 – Nuvens de palavras referentes, respectivamente, aos períodos I, II e III.

Fonte: elaborada pelos autores.

4.3 *Hapax legomena*

Nesta análise preliminar, também foram considerados os *hapax legomena*, isto é, ocorrências que só aparecem uma vez no *dataset*. Em todos os períodos foram encontrados erros de digitação (ou de digitalização), mas foi possível também perceber ocorrências que podem ser úteis como indicadores de mudanças diacrônicas a serem verificadas.

No período I (1946-1964), encontram-se erros de digitação (ou de digitalização) como: "atiântica" por "atlântica", "ciqüenta" por "cinqüenta", "dafender" por "defender", "federai" por "federal", "fiacalização" por "fiscalização", "jalgamento" por "julgamento", entre outros, mas também palavras com grafias diferentes ("datilógrafo" e "dactilógrafo"). Na grande maioria dos casos, os *hapax legomena* encontrados são nomes próprios de pessoa ("thyco", "teodoro", "lucélia", "izidoro") ou topônimos ("uaupés", "urucará", "turiaçú"). Foram observados vários termos estrangeiros, seja antropônimos ("truman", referente a Sally Truman; "kirkons", de Kirkons Nodhjalp), seja nomes de empresas e instituições ("thomas", da firma Thomas de la Rue & Company Limited; "telefonaktisebolaget", da firma Telefonaktisebolaget - l.m. Ericsson de Stokolmo), seja utilizados como terminologia de campos específicos ("truks", em "caixas de graxas para truks de carros e vagões"; "royalty", também grafado como "rorgarties").

No Quadro 1, observou-se que entre as colocações mais relevantes desse período está o nome do presidente Juscelino Kubitschek. Curiosamente, a grafia de seu nome está presente de formas muito variadas, tendo sido encontrada como "juscelimo" e "juscellino", e o sobrenome com nove variações de ocorrência única: "kubigtschek", "kubitscbek", "kubitschet", "kubitsckek", "kubitshek", "kubltsckek", "kubsitschek", "kubstischek", "kubstschek".

Para finalizar, no período I foram encontrados três interessantes *hapax legomena* não presentes nos períodos II e III: "chefatura" (com significado de "chefia", no caso em

questão da polícia), "locupletamento" (no período II há uma ocorrência de "locupletaram", do verbo "locupletar", com o significado de "enriquecer, não necessariamente de forma lícita") e "leprólogos" (peritos em hanseníase, antigamente/popularmente conhecida como "lepra").

A análise dos *hapax legomena* do período II (1964-1988) inclui erros de digitação (ou de digitalização) como "acôdo" por "acordo", "administraç/ão" por "administração", "ainciso" por "inciso", "aiterações" por "alterações", "anônino" por "anônimo", "biihetes" por "bilhetes", "canpos" por "campos", "exêcito" por "exército", "imcompatibilidade" por "incompatibilidade", além de grafias diferentes ("balisamento" e "balizamento", "autárquias" uma única vez frente a "autarquias"). Nesse período também há muitos nomes próprios de pessoa ("adolfo", "agripino", "doris") e topônimos ("afuá", "alpercata", "botuverá", "cajamar", "erechim", "iepê", "pojuca"). Comparecem, aparentemente em número menor, termos estrangeiros: nomes próprios ("charles", de Charles de Gaulle), de empresas e instituições (como da empresa Deutsche Ibero-Amerika Stiftung, ou da Standard Elektrik Arktiengeselischaft), ou utilizados como terminologia de campos específicos ("broadcastinge", em "estação de broadcastinge de televisão"). Assinalamos, ainda, uma presença considerável de siglas, muitas delas referentes a órgãos, instituições ou programas estatais, como: "capes" (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), "cbdf" (Corpo de Bombeiros do Distrito Federal), "cdfa" (Comissão Desportiva das Forças Armadas), "cbee" (Companhia Brasileira de Energia Elétrica), "celf" (Centrais Elétricas Fluminenses), "codeplan" (Companhia de Desenvolvimento do Planalto Central), "dag" (Departamento de Administração Geral), "dnff" (Departamento Nacional de Estradas de Ferro), "iee" (Instituto de Educação do Excepcional), "pebe" (Programa Especial de Bôlsas de Estudo), "pgpm" (Política de Garantia de Preços Mínimos), entre muitíssimas outras. O termo "nosológico" comparece com uma única ocorrência e não se encontra nos outros dois períodos.

A varredura dos *hapax legomena* do período III (1988-2017) mais uma vez evidenciou erros de digitação (ou de digitalização), sempre numerosos (entre os vários, citamos: "vidaou" por "vida ou", "uperávir" por "superávit", "títulares" por "titulares", "trêsvaras" por "três varas", "trêsreais" por "três reais", "trubutária" por "tributária", "tratameto" por "tratamento", "prsidente" por "presidente", "cenros" por "centros"). Enquanto os topônimos de ocorrência única também se mantiveram numerosos (como "xexéu", "urupá", "uruarama", "uarini", "sooretama", "mozarlândia", "iraci", "cujubim", "aracoiaba"), os nomes próprios de pessoa foram relativamente poucos ("thaumaturgo", "teotônio", "sidney", "januário", "ezequiel", "euclydes"), assim como os nomes de instituições em língua estrangeira (como Morgan Guaranty Trust Company of New York). Foram novamente encontrados termos ingleses utilizados em campos específicos, como "sweepstakes", "pellets", "loft", "lockout". Mais uma vez houve uma presença considerável de siglas e acrônimos, como: "zees" (zoneamentos ecológico-econômicos), "utn" (usinas termonucleares), "tsb" (técnico em saúde bucal), "transpetro" (Petrobras Transporte S.A.), "telpe" (Telecomunicações de Pernambuco S.A.), "ted" (transferência eletrônica de dados), "srbf" (Secretaria da Receita Federal do Brasil), "snsm" (Sistema Nacional de Sementes e Mudas), "serse" (Secretaria Especial da Região Sudeste), "renavam" (Registro Nacional de Veículos Automotores), "qofarm" (Quadro de Oficiais Farmacêuticos), "cgee" (Centro de Gestão e Estudos Estratégicos), "boh" (Boletim de Ocupação Hoteleira), entre muitos outros. Destaque para o número crescente de siglas referentes a instituições de ensino e de pesquisa, tais quais: "usp" (Universidade de São Paulo), "unirio" (Universidade do Rio de Janeiro), "unifesp" (Universidade Federal de São Paulo), "ufla" (Universidade Federal de Lavras), "ufg" (Universidade Federal de Goiás), "inpa" (Instituto Nacional de Pesquisas da Amazônia), "fua" (Fundação Universidade do Amazonas), entre tantas outras.

5 Considerações finais

Neste artigo, discutiu-se a relação entre linguagem e direito, tecendo um pequeno panorama sobre os estudos de interface entre as duas áreas no Brasil. Em seguida, após uma reflexão sobre a utilização de *corpora* e *datasets* em pesquisas linguísticas, é apresentada uma análise lexical preliminar da Base de Normas Jurídicas Brasileiras, que contém o texto de quase treze mil leis promulgadas no Brasil entre 1946 e 2017. Apesar de possuir inconsistências e imperfeições, o *dataset* disponibilizado por Martim, Lima e Araujo (2018) se mostra uma fonte de dados interessante para estudos que possuam como objetivo investigar aspectos (sejam eles linguísticos e/ou sociais) da legislação brasileira, em particular sob uma perspectiva diacrônica, como a que é apresentada aqui.

Labov (1972b, p. 100) afirma que "A grande arte do linguista histórico é tirar o melhor proveito desses dados ruins – 'ruins' no sentido de que podem ser fragmentários, corrompidos ou muitas vezes removidos das produções reais dos falantes nativos" (tradução nossa)[23]. Em um certo sentido, o trabalho do linguista que lida com grandes *datasets* não é muito diferente. Como pode ser observado neste trabalho, dados disponibilizados ao pesquisador, sobretudo quando foram obtidos sem a devida curadoria linguística, podem se revelar falhos, incompletos e com os mais variados tipos de erros. Ainda assim, a viabilização desse tipo de dado se mostra importante por permitir que sejam realizadas análises em grande escala e a um custo relativamente baixo – seja de tempo, seja propriamente financeiro.

Conforme se discutiu anteriormente, *corpora* compilados para estudos linguísticos preveem que seja pensada uma arquitetura específica para as finalidades da pesquisa que se pretende empreender, por meio de uma amostragem adequada dos dados e seu devido balanceamento, a fim de se obter a representatividade de uma

[23] [t]he great art of the historical linguist is to make the best of this bad data – 'bad' in the sense that it may be fragmentary, corrupted, or many times removed from the actual productions of native speakers.

determinada variedade linguística. Tal processo implica que haja conhecimento, por parte da equipe de compilação, das características específicas de uma linguagem setorial (neste caso, das leis), assim como do contexto de sua produção. A compilação de um *corpus* de referência da linguagem jurídica do português brasileiro, recentemente empreendida por uma das autoras, tenta cobrir justamente esses pontos. O LEX-BR-Ius (FERRARI; MARQUES, em preparação) será a seção legislativa de tal *corpus* e incluirá, também, textos completos de normas legislativas federais brasileiras[24]. Além disso, (a) as normas que farão parte do corpus deverão estar em vigor no momento da coleta; e (b) o uso efetivo das leis está sendo verificado através de uma análise prévia com base no critério de relevância, o que será fator decisivo na escolha por aquelas que irão entrar ou não no corpus. O balanceamento entre textos que possuem, por sua própria natureza, dimensões muito variáveis se dará pela criação de subseções relativas às diferentes tipologias legislativas. As subseções terão números de palavras similares, mantendo a integridade dos textos, mas garantindo, assim, sua comparabilidade interna em termos quantitativos. A extração dos dados será seguida por diferentes tipos de tratamento computacional, o que permitirá que o *corpus* esteja disponível em sua versão bruta (*raw text*), completo com anotação textual (*markup* textual) e anotação POS (classe gramatical). O texto bruto passará por uma limpeza prévia que eliminará expressões formulaicas típicas do gênero, assim como os campos de assinaturas finais de cada norma. Tais informações, contudo, serão mantidas nos metadados, que incluirão data de extração, pessoas responsáveis por sua promulgação, número de palavras, artigos em que é subdividida cada norma, ementa, assunto e alterações. De tal forma, a varredura dos metadados possibilitará ao pesquisador uma definição prévia do tipo de lei que será interessante para sua

[24] A decisão por não extrair excertos das normas de tamanho igual (o que facilitaria o balanceamento) respeita a integridade do texto, por entender que a língua necessita de seu contexto próprio de uso para que as regularidades e padrões não fiquem enviesados (BIBER, 1998; BIBER; CONRAD, 2009; BERBER SARDINHA, 2010; BIBER *et al*., 2016; EGBERT; LARSSON; BIBER, 2020).

pesquisa (por legislatura, por período, por assunto etc.).

Embora não tenha sido compilado utilizando critérios tão rigorosos quanto esses, o *dataset* disponibilizado por Martim, Lima e Araujo (2018) e aqui apresentado nos permitiu vislumbrar padrões interessantes que podem constituir o ponto de partida para pesquisas futuras. Salientamos a necessidade de cautela nas generalizações dos resultados, mas acreditamos que, com os devidos cuidados, esta pesquisa possa abrir uma série de oportunidades para análises futuras utilizando esses mesmos dados. Do ponto de vista lexical, pretende-se refinar, em um trabalho posterior, a análise da variação da frequência de determinadas palavras e expressões ao longo do tempo – observando, inclusive, a conceitualização em torno desses itens lexicais, conforme apresentado, por exemplo, por Cunha *et al*. (2018). Para isso, poderá ser importante adicionar procedimentos como a lematização do texto, além da definição de periodizações mais finas (por exemplo, pode-se trabalhar com períodos de dez ou cinco anos, ou ainda menos, em vez de apenas três períodos, como foi realizado aqui). Uma possibilidade promissora de investigação é a utilização do método apresentado por Cunha e Wichmann (2021) para a identificação das datas de fixação e de obsolescência de palavras e expressões no conjunto de dados analisado: assim, será possível identificar, por exemplo, em que momento determinados *n*-gramas passaram a ser mencionados nas leis e quando outros deixaram de aparecer nas normas brasileiras. Por fim, uma melhor compreensão dos fenômenos será beneficiada por uma pesquisa mais aprofundada acerca dos períodos históricos aqui apresentados e seus desdobramentos nas leis; e, ao mesmo tempo, por um aprofundamento na relevância de determinadas leis em cada período, sua validade e sua efetiva utilização na jurisprudência.

Referências

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

BAKER, P. **Sociolinguistics and Corpus Linguistics**. Edinburgh: Edinburgh University Press, 2010.

BERBER SARDINHA, T. **Lingüística de Corpus**. Barueri: Manole, 2004.

BERBER SARDINHA, T. A abordagem metodológica da análise multidimensional. **Gragoatá**, v. 15, n. 29, p. 107-125, 2010. DOI https://doi.org/10.22409/gragoata.v15i29.33077

BIBER, D. **Variation across speech and writing**. Cambridge: Cambridge University Press, 1988. DOI https://doi.org/10.1017/CBO9780511621024

BIBER, D. Methodological issues regarding corpus-based analyses of linguistic variation. **Literary and Linguistic Computing**, v. 5, n. 4, p. 257-269, 1990. DOI https://doi.org/10.1093/llc/5.4.257

BIBER, D. On the complexity of discourse complexity: A multidimensional analysis. **Discourse Processes**, v. 15, n. 2, p. 133-163, 1992. DOI https://doi.org/10.1080/01638539209544806

BIBER, D. Representativeness in Corpus Design. Literary and Linguistic Computing, v. 8, n. 4, p. 243-257, 1993. DOI https://doi.org/10.1093/llc/8.4.243

BIBER, D; CONRAD, S. **Register, genre, and style**. Cambridge: CUP, 2009. DOI https://doi.org/10.1017/CBO9780511814358

BIBER, D.; REPPEN, R.; SCHNUR, E.; GHANEM, R. On the (non)utility of Juilland's *D* to measure lexical dispersion in large corpora. **International Journal of Corpus Linguistics**, v. 21, n. 4, p. 439-464, 2016. DOI https://doi.org/10.1075/ijcl.21.4.01bib

BIRD, S.; KLEIN, E.; LOPER, E. **Natural language processing with Python**: analyzing text with the Natural Language Toolkit. Sebastopol: O'Reilly, 2009.

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

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

CORDEIRO, J. M. Anos de chumbo ou anos de ouro? A memória social sobre o governo Médici. **Estudos Históricos**, v. 22, n. 43, p. 85-104, 2009. DOI https://doi.org/10.1590/S0103-21862009000100005

CUNHA, E.; MAGNO, G.; CAETANO, J.; TEIXEIRA, D.; ALMEIDA, V. Fake news as we feel it: perception and conceptualization of the term "fake news" in the media. *In*: STAAB, S.; KOLTSOVA, O.; IGNATOV, D. I. (ed.). **Social informatics** [Lecture Notes in Computer Science, n. 11185]. Cham: Springer, 2018. p. 151-166. DOI https://doi.org/10.1007/978-3-030-01129-1_10

CUNHA, E. L. T. P.; WICHMANN, S. An algorithm to identify periods of establishment and obsolescence of linguistic items in a diachronic corpus. **Corpora**, Edinburgh, v. 16, n. 2, p. 205-236, 2021. DOI https://doi.org/10.3366/cor.2021.0218

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

DIMITRIADIS, A.; MUSGRAVE, S. Designing linguistic databases: a primer for linguists. *In*: EVERAERT, M.; MUSGRAVE, S.; DIMITRIADIS, A. (ed.). **The use of databases in cross-linguistic studies**. Berlin/New York: Mouton de Gruyter, 2009. p. 13-75.

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

EGBERT, J.; BURCH, B.; BIBER, D. Lexical dispersion and corpus design. **International Journal of Corpus Linguistics**, v. 25, n. 1, p. 89-115, 2020. DOI https://doi.org/10.1075/ijcl.18010.egb

EGBERT, J.; LARSSON, T.; BIBER, D. **Doing Linguistics with a Corpus**. Methodological Considerations for the Everyday User. Cambridge: Cambridge University Press, 2020. DOI https://doi.org/10.1017/9781108888790

FERRARI, L. A; MARQUES, C. G. F. **O corpus LEX-BR-Ius, seção legislativa das leis federais brasileiras**: arquitetura e primeiras análises. Em preparação.

GRAY, J. Towards a Genealogy of Open Data. *In*: **GENERAL CONFERENCE OF THE EUROPEAN CONSORTIUM FOR POLITICAL RESEARCH**. Glasgow, 2014. DOI https://dx.doi.org/10.2139/ssrn.2605828

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

HARDIE, A. CQPweb — combining power, flexibility and usability in a corpus analysis tool. **International Journal of Corpus Linguistics**, v. 17, n. 3, p. 380-409, 2012. DOI https://doi.org/10.1075/ijcl.17.3.04har

HEAPS, H. S. **Information retrieval**: computational and theoretical aspects. New York: Academic Press, 1978.

HERDAN, G. **Quantitative linguistics**. London: Butterworth, 1964.

IVO, G. O direito e a inevitabilidade do cerco da linguagem. *In*: CARVALHO, P. de B.; CARVALHO, A. T. de (org.). **Constructivismo lógico-semântico**. 2. ed. revista v. 1. São Paulo: Noeses, 2020. p. 65-91.

KRIEGER, M. G.; MACIEL, A. M. B.; BEVILACQUA, C. R.; ROCHA, J. C.; FINATTO, M. J. B. **Dicionário de Direito Ambiental**. Porto Alegre: Editora da Universidade, 1998.

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.

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.

LABOV, W. **The social stratification of English in New York City**. Washington: Center for Applied Linguistics, 1966.

LABOV, W. **Sociolinguistic patterns**. Philadelphia: University of Pennsylvania Press, 1972a.

LABOV, W. Some principles of linguistic methodology. **Language in Society**, v. 1, n. 1, p. 97-120, 1972b. Disponível em: https://www.jstor.org/stable/4166672. Acesso em: 12 nov. 2021. DOI https://doi.org/10.1017/S0047404500006576

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.

MARTIM, H.; LIMA, J. A. O.; ARAUJO, L. C. Base de Normas Jurídicas Brasileiras: uma iniciativa de Open Government Data. **Perspectivas em Ciência da Informação**, v. 23, n. 4, p. 133, 2018. DOI https://doi.org/10.1590/1981-5344/3567

MELLO, H. Methodological issues for spontaneous speech corpora compilation: the case of C-ORAL-BRASIL. *In*: RASO, T.; MELLO, H. (org.). **Spoken Corpora and Linguistic Studies**. Amsterdam: John Benjamins, 2014. v. 1, p. 27-68. DOI https://doi.org/10.1075/scl.61.01mel

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

RAMOS, J. J. S. C. **Ocorrência e interpretação dos verbos modais 'dever' e 'poder' em contexto jurídico**: contributos para uma análise juslinguística. 207 f. Tese (Doutorado) – Filozofická Fackultá, Univerzita Karlova, Praha, Rep. Tcheca, 2017 *apud* SVOBODOVÁ (2017).

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

RENEAR, A. H.; SACCHI, S.; WICKETT, K. M. Definitions of *dataset* in the scientific and technical literature. **Proceedings of the American Society for Information Science and Technology**, v. 47, n. 1, 2010. DOI https://doi.org/10.1002/meet.14504701240

RESENDE, S. V.; MAVERICK, R. Planejamento, compilação e organização de corpora. *In*: **Anais do EBRALC 2015 & ELC 2015** [Blucher Social Science Proceedings, n. 3, v. 2]. São Paulo: Blucher, 2016. p. 27-35. DOI https://doi.org/10.5151/sosci-viiieblc-xiii-elc-06_artigo_03

SANTOS, F. Patronagem e Poder de Agenda na Política Brasileira. **Dados**: Revista de Ciências Sociais, v. 40, n. 3, 1997. DOI https://doi.org/10.1590/S0011-52581997000300007

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

SEARLE, J. R. **Speech Acts**. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press, 1969. DOI https://doi.org/10.1017/CBO9781139173438

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

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.

WARAT, L. A. **O direito e sua linguagem**. Porto Alegre: Sergio Antonio Fabris Editor, 1995.

Artigo recebido em: 30.12.2021 Artigo aprovado em: 30.05.2022