Antares
LETRAS E HUMANIDADES ISSN: 1984-4921
Programa de Pós-Graduação em Letras e Cultura
UCS UNIVERSIDADE DE CAXIAS DO SUL

DOI: http://dx.doi.org/10.18226/19844921.v14.n34.02

O LEX-BR-Ius: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras

The LEX-BR-Ius: architecture and decisions on the compilation of a Brazilian federal statutory laws representative corpus

Lucia de Almeida Ferrari*
Carolina Godoi de Faria Marques**

Resumo: O artigo apresenta o LEX-BR-Ius, um corpus representativo das leis federais brasileiras, em fase de compilação. São introduzidas as plataformas de busca das normas legislativas brasileiras e a necessidade da compilação de um corpus para a pesquisa linguística do gênero. Explicamos sobre as diversas espécies normativas presentes na legislação brasileira para descrever as escolhas metodológicas na compilação do corpus. Discutem-se em seguida as etapas de compilação e as decisões em aberto. Os critérios de escolha das normas legais que farão parte do corpus são (a) estar em vigor no momento da coleta; (b) serem textos inteiros, sem recortes, para não interferir em sua textualização e representatividade interna (SINCLAIR, 2004); (c) serem selecionadas com base em sua frequência de uso.
O corpus possui marcação textual em Modest XML (HARDIE, 2014), que permite refinar as buscas e cabeçalho em XML com informações precisas sobre cada norma e será disponibilizado também em sua versão com texto limpo e com lematização e anotação morfossintática. São apresentadas algumas possibilidades investigativas a partir dos dados coletados.

Palavras-chave: *LEX-BR-Ius*, Corpus. Legislação Federal Brasileira. Arquitetura. Metodologia.

Abstract: The article presents the LEX-BR-Ius, a Brazilian federal statutory laws representative corpus, in compilation. We introduce search engines for Brazilian statutory laws and the necessity to compile a specific corpus for linguistic research purposes. We explain the different types of laws in Brazilian legislation to describe the methodological choices in compiling the corpus. Following up, the compilation steps and some open issues are approached. The corpus sample criteria include that laws: (a) must be in force during data collection; (b) must be whole texts, without samplings, so as not to interfere with the internal textualization and representativeness (SINCLAIR, 2004); (C) must be sampled based on their frequency of usage. The corpus presents a

* Universidade Federal de Minas Gerais (UFMG).
** Universidade Federal de Minas Gerais (UFMG).

textual markup in Modest XML (HARDIE, 2014) that will allow refining analysis and XML headers with precise information about each law, and it will be also available in its raw text version and in a lemmatized and PoS annotated version. We also present some possible and ongoing research on the collected data.

Keywords: *LEX-BR-Ius*. Corpus. Brazilian Federal Statutory Laws. A*rchitecture*. Methodology.

Introdução

O mundo do Direito, e a interseção entre Direito e Linguagem, incluem uma pluralidade de realidades textuais (escritas e orais) que apresentam características distintas e englobam registros e domínios muito diferentes (COULTHARD; JOHNSON, 2007; 2010; CARAPINHA, 2018). Há a linguagem utilizada para a criação das leis, nas legislações em si, nas interações judiciárias (escritas e orais), em artigos acadêmicos etc. Tem-se ainda leis que discutem sobre linguagem e direitos e/ou crimes linguísticos, assim como a análise de textos legais e suas traduções. Esse conjunto de linguagens diversas é definido como linguagem jurídica (COULTHARD; JOHNSON, 2007) e abarca a "análise do texto legal escrito e do discurso oral em sala de audiências" (CARAPINHA, 2018, p. 92), assim como a ação do linguista em pareceres para as investigações e tribunais e na perícia de fala ou de escrita.

Entre tal vasto mundo de contextos, nossa contribuição para a descrição do gênero jurídico pretende concentrar-se na diamesia escrita e, especificamente, nas normas legais, conhecidas comummente como leis.

A caracterização desse mundo requer um duplo trabalho por parte do linguista: uma perspectiva próxima, na detecção de possíveis traços específicos do gênero, o que requer a leitura atenta das normas, e um olhar mais distante, do conjunto dos textos em seus atributos quantitativos gerais, para depois, eventualmente, retornar

aos dados em busca dos padrões indicados pela análise quantitativa. Trata-se, nos dois tipos de estudos, da avaliação de dados empíricos e autênticos. O primeiro tipo de investigação geralmente se debruça sobre um, ou uns poucos textos. O segundo requer um grande volume de dados que retrate com fidelidade o conjunto. A linguística de corpus pode ser uma ferramenta útil em tal tipo de exame, enquanto permite seja a visualização do todo, seja a análise do específico, possibilitando ao mesmo tempo a verificação de características previamente identificadas, seja a descoberta de novos padrões.

Neste artigo apresentamos a arquitetura, algumas questões metodológicas e as possibilidades investigativas do *LEX-BR-Ius*, um corpus representativo da legislação federal brasileira, em fase de coleta. O corpus será composto por sete seções com amostragem de textos inteiros das diferentes divisões convencionadas no mundo jurídico. Cada seção disponibilizará, para cada norma, quatro arquivos: (a) o texto limpo (*raw text*) em ".txt"; (b) os metadados da norma em ".xml"; (c) o texto com marcação textual ".xml" seguindo a proposta de Hardie (2014), com etiquetas criadas especificamente para o projeto; (d) o texto lematizado e com anotação *Part of Speech*. Uma vez compilado, o corpus será disponibilizado gratuitamente online à comunidade acadêmica, acompanhado de sua descrição.

O artigo está assim dividido: em uma primeira parte discutimos sobre a disponibilização de dados legislativos aos operadores do direito e ao público em geral. Passamos então à apresentação da linguística de corpus e seu interesse pela linguagem jurídica, de maneira a introduzir o capítulo seguinte, que justifica a necessidade de criação de um corpus de normas legislativas brasileiras para a pesquisa linguística. O quinto capítulo explica o funcionamento do processo legislativo federal brasileiro e suas subdivisões internas, para a seguir, no capítulo sexto, apresentar a proposta do LEX-BR-Ius. Os dois capítulos subsequentes

detalham o processo de compilação do corpus e algumas questões metodológicas que ainda estão em aberto. Para finalizar, discute-se brevemente sobre as possibilidades exploratórias do LEX-BR-Ius.

Dados legislativos e seu acesso

Os textos legais são alterados e editados constantemente. As versões impressas desses textos, oferecidas por editoras, além de serem pagas, não acompanham as mudanças na legislação, sendo atualizadas apenas uma vez por ano no início desse período. Para remediar esse problema algumas editoras oferecem, com a aquisição da versão impressa, o acesso a uma versão online alterada com maior frequência. Entretanto, seu preço, a falta de atualizações e o enorme volume de dados produzido e processado no âmbito legislativo, fazem com que as plataformas online de busca de legislação sejam preferidas pelo público. Utilizadas majoritariamente pelos operadores do direito, tais como advogados, juízes, promotores e estudantes, mas também pelo cidadão comum em caso de dúvidas ou curiosidades sobre os textos legais, essas plataformas armazenam os textos legais editados nos âmbitos federal, estadual e/ou municipal. De fácil acesso, seu conteúdo é constantemente atualizado, aberto e disponibilizado gratuitamente, podendo ser acessado através de mecanismos de busca.

Há tanto domínios governamentais das mais diversas esferas e poderes quanto sites de provedores privados. No âmbito federal, destaca-se o Portal da Legislação do Planalto Brasileiro[3], utilizado para compilar nosso corpus. Nele estão disponíveis gratuitamente e são atualizados diariamente, entre outros, os textos legais brasileiros federais (com exceção das resoluções e dos decretos legislativos) e estaduais, além de textos federais de caráter administrativo. Temos

[3] Disponível em: http://www4.planalto.gov.br/legislacao/. Acesso em: 05 jan. 2022.

ainda os portais da Câmara dos Deputados[4] e do Senado Federal[5] que disponibilizam tanto a legislação federal quanto a legislação interna de cada casa. Já o portal *normas.leg*[6], idealizado pelo Congresso Nacional, disponibiliza as leis federais em uma linha do tempo, permitindo não só a busca legislativa sincrônica, mas também diacrônica.

Nos âmbitos estadual e municipal é costumeiro disponibilizar no site oficial do executivo a Constituição Estadual (no caso dos estados) ou Lei orgânica (no caso do DF e dos municípios) e, muitas vezes, também a Constituição Federal de 1988. Além disso, nos sites do poder legislativo de cada esfera é possível encontrar a relação completa das normas por eles editadas assim como os projetos de lei. Como exemplo, citamos o site da Assembleia Legislativa de Minas Gerais[7] que disponibiliza, entre outros, a Constituição Estadual, por escrito e áudio, e as Leis Estaduais, assim como a Constituição Federal.

No âmbito privado, por sua vez, o portal mais utilizado é o *JusBrasil*[8], um site aberto agregador de conteúdo relacionado ao direito brasileiro. Nele é possível buscar não só as legislações, mas também jurisprudências, artigos, notícias e, em geral, todos os textos de acesso livre disponíveis na internet que envolvam o mundo jurídico brasileiro.

Dada sua facilidade de acesso, tais portais seriam, à primeira vista, um rico campo de pesquisa para a coleta e análise linguística. Contudo, e justamente pela grande mutabilidade dos dados legislativos, pois os arquivos são modificados diariamente, sua extração não é tão simples. Garantir sua validade ecológica, ou seja, que os dados sejam atualizados e representativos do conjunto de normas, requer

4 Disponível em: https://www.camara.leg.br/legislacao. Acesso em: 05 jan. 2022.
5 Disponível em: https://www12.senado.leg.br/hpsenado. Acesso em: 05 jan. 2022.
6 Disponível em: https://normas.leg.br/busca. Acesso em: 05 jan. 2022.
7 Disponível em: https://www.almg.gov.br/consulte/legislacao/index.html?aba=js_tabLegislacaoMineira. Acesso em: 05 jan. 2022.
8 Disponível em: https://www.jusbrasil.com.br/home. Acesso em: 05 jan. 2022.

um cuidadoso trabalho de curadoria com especificações precisas de suas características. Isso porque cada língua e cada tradição jurídica apresenta suas particularidades na edição e aplicação de suas normas que se diferenciam por espécie e assunto regulado, tendo extensões muito distintas.

Para a pesquisa linguística faz-se necessário uma diversidade e grande quantidade de textos de forma a obter dados linguísticos significativos, sendo o uso de corpora o mais indicado. Corpora são conjuntos de textos autênticos, coletados de acordo com critérios específicos de arquitetura para que sejam representativos de uma língua ou de uma determinada variedade linguística, armazenados e submetidos a tratamento computacional, possibilitando buscas automáticas ou semiautomáticas. Diferentemente dos corpora, os bancos de dados e *datasets* têm como finalidade armazenar informações, coletando dados sobre determinado assunto, no caso das plataformas de busca em questão, as legislações, que são agrupados segundo características comuns, não havendo uma preocupação com os três pilares da compilação de corpora: representatividade, amostragem e balanceamento. (SINCLAIR, 1991; MCENERY e WILSON, 1996; TOGNINI-BONELLI, 2001; BAKER e HARDIE, 2006)

Para ilustrar essa questão, destacamos o trabalho dos pesquisadores de ciências da computação do Senado Federal, Martim, Lima e Araújo (2018). Eles coletaram e publicaram a Base de Normas Jurídicas Federais, composta por um conjunto de oito *datasets* que contêm todas as normas legislativas federais desde 4 de outubro de 1946 até 12 de abril de 2017[9]. Em estudo recente, Ferrari e Cunha (2022) fizeram uma varredura preliminar do *dataset* e mostraram a diferença entre as possibilidades investigativas de um corpus e

[9] Disponível em https://doi.org/10.6084/m9.figshare.c.4029253.v1. Acesso em: 05 jan. 2022.

aquelas de bancos de dados. Em específico, o estudo apontou para a dificuldade de identificação de características mais finas nos dados, como as modificações pelas quais as leis passaram ou a seleção de trechos por partes específicas de cada texto.

As análises preliminares de Ferrari e Cunha (2022) nos fizeram perceber a necessidade de compilarmos um corpus próprio para as investigações que pretendemos implementar, mas que também possa servir de base para outros pesquisadores. Como mostraremos ao longo do artigo, a intenção é que o corpus seja representativo da linguagem jurídica da esfera legislativa federal brasileira e permita análises nos níveis lexical e morfossintático assim como textual e do discurso, podendo-se refinar a extração de dados em sincronia e diacronia e por níveis internos ao texto, graças à etiquetagem textual que estamos implementando. Nossa intenção é que, uma vez finalizado, o corpus seja disponibilizado para pesquisas à comunidade acadêmica.

A linguística de corpus e os corpora jurídicos

Por linguística de corpus entende-se: (a) um ramo da linguística, que concebe a língua explicitamente como formada por dados empíricos coletados seguindo uma arquitetura rigorosa, e que leva em conta os princípios de amostragem, representatividade, balanceamento e disponibilização computacional, (b) uma metodologia que se vale dos preceitos acima, mas especialmente de todo um conjunto de ferramentas estatísticas e computacionais de análise. Ela permite a coleta e análise de um grande volume de amostras representativas da linguagem em uso *(“real life language”)* a serem usadas nos mais diversos estudos linguísticos empíricos via corpora (SINCLAIR, 1991; ATKINS, CLEAR e OSLER, 1992; MCENERY e WILSON, 1996; TOGNINI-BONELLI, 2001; MEYER, 2002; TAGNIN, 2004; BERBER

SARDINHA, 2004; MCENERY e HARDIE, 2012; STEFANOVICH, 2020).

Os corpora podem ser de diferentes tipos e diamesias. Há corpora escritos, orais e multimodais, podendo ser sincrônicos (quando retratam um momento histórico específico), diacrônicos (quando dois ou mais momentos são retratados para comparações), contemporâneos (retrata-se o momento presente) ou ainda, históricos (retrata-se algum momento passado). Temos também os corpora especializados que representam uma variedade linguística ou linguagem especializada e os de referência que representam uma língua em geral. Eles podem ser monolíngues (constituídos por uma única língua) ou multilíngues (constituídos por mais de uma língua), paralelos (o mesmo texto é traduzido de uma língua para outra(s) e geralmente são apresentados de forma alinhada, possibilitando comparações de traduções diferentes) ou comparáveis – a mesma tipologia textual ou o mesmo gênero é apresentado em dois ou mais corpora de línguas diferentes para análises interlinguísticas (BERBER SARDINHA, 2004; MCENERY e HARDIE, 2012).

Geralmente a representatividade do corpus é alcançada selecionando os textos por amostragem: escolhe-se uma amostra específica da população, nesse caso, da língua a ser representada, buscando o balanceamento, ou seja, a distribuição dos textos de maneira semelhante à realidade representada, através de estudos prévios sobre a língua ou variedade que se quer estudar (BIBER, 1988; BERBER SARDINHA, 2004; MCENERY; HARDIE, 2012).

Em termos de linguagem jurídica, existem corpora especializados desde os anos 1990. Pontrandolfo (2012) e Giampieri (2018) fornecem um levantamento geral dos projetos e corpora legais disponíveis, com especial atenção à língua inglesa, que apresenta o maior número de materiais e pesquisas, e ao espanhol e italiano, seus objetos de

pesquisa específicos. Uma lista de corpora legais com algum nível de descrição e links está disponível também no site https://legal-linguistics.net/data-collections/.

Dentre os corpora de língua inglesa, Pontrandolfo (2012), destaca o *Cambridge Corpus of Legal English*, com cerca de 20 milhões de palavras: trata-se de um sub-corpus do *Cambridge English Corpus* (*CEC),* previamente conhecido como *Cambridge International Corpus* (XIAO, 2008), uma iniciativa da Cambridge University Press, não disponível para a comunidade. Além desse, são citados o *House of Lords Judgments Corpus* (HOLJ) da Universidade de Edimburgo, cujo objetivo é a formulação de resumos em automático, compilado com uma seleção de julgamentos da *House of Lords*, e o *Proceedings of the Old Bailey* (a Corte Criminal Central de Londres), um corpus diacrônico de linguagem judicial de julgamentos criminais, que contém cerca de 127 milhões de palavras. Assinalamos ainda o *American Law Corpus* (ALC), um corpus de mais de 5 milhões de palavras de sete gêneros legais americanos compilados por Goźdź-Roszkowski (2011), da Universidade de **Łódz**, na Polônia.

A Espanha também tem investido fortemente na compilação de corpora legais. Pontradolfo (2012) cita o projeto de pesquisa JUD-GENTT, seção legal do projeto GENTT (*Textual Genres for Translation*), que está compilando um corpus comparável multilíngue (inglês, espanhol, alemão e francês) de diferentes gêneros textuais (direito, medicina e outras linguagens técnicas). No Instituto de Linguística Aplicada da Universidade Pompeu Fabra de Barcelona (IULA) o projeto CORPUS (*Multilingual Specialised Textual Corpus*, conhecido também como *Technical Corpus*) compilou e analisa um corpus paralelo de domínios diferentes: direito, economia, ambiente, medicina e ciências informáticas. Na Universidade de Vigo foi desenvolvido o CLUVI (*Linguistic Corpus of the University of Vigo*), um corpus paralelo

de registros especializados (ficção, computação, jornalismo, direito e administração entre outros) que soma mais de 27 milhões de palavras: dentro do CLUVI, os subcorpora LEGA e LEGE-BI são especializados em linguagem jurídica.

Na Holanda citamos o *Corpus Juridisch Nederlands*[10] que disponibiliza 5856 textos legais datados desde 1814 até 1989, coletados a partir do *N-Lex*, uma base de dados da legislação holandês. Os dados são acessíveis, mas não encontramos uma documentação precisa quanto às escolhas metodológicas da coleta. Já na Estônia destaca-se o *Reference corpus of Estonian: Legislation*[11] que apresenta 391 arquivos (1.8 milhões de tokens) com as leis da Estônia e 5431 arquivos (9.6 milhões de tokens) com traduções de leis da União Europeia em estoniano. Uma breve descrição do corpus informa que a extração foi automática e há especificações sobre as principais etiquetas e como acessar os arquivos.

Na Itália, o projeto pioneiro na compilação de corpora legais foi o *Bononia Legal Corpus* (BoLC) um corpus bilíngue (italiano e inglês) representativo da linguagem jurídica das duas realidades linguísticas e legais (*civil law* e *common law*), iniciado em 1997 na Universidade de Bolonha. O subcorpus inglês perfaz cerca de 21 milhões de palavras, enquanto o subcorpus italiano chega a mais de 33 milhões de palavras. Todos os documentos foram coletados entre 1968 e 1995. Outro corpus compilado na Universidade de Bolonha foi o CODIS (*Corpus Dinamico dell'Italiano Scritto*), que possui um subcorpus de linguagem jurídica com cerca 10 milhões de palavras. (ROSSINI; TAMBURINI; DE SANTIS, 2002)

[10] Disponível em: http://hdl.handle.net/10032/tm-a2-u2. Acesso em: 20 de jun. 2022.

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

Junto à Universidade de Bergamo, o projeto *Corpus of Academic English* (CADIS) coletou dados de quatro grandes áreas (linguística, economia, direito e medicina), divididos por sua vez entre resumos, livros, revisões, editoriais e artigos. Uma grande parte do corpus é de língua inglesa e uma parte menor em italiano, para comparações interlinguísticas. Sempre na Itália, o corpus *Jus Jurium* (BARBERA, 2005; ONESTI, 2011) nos inspirou em várias decisões metodológicas, contudo **não** conseguimos encontrar uma descrição muito precisa do projeto ou se houve interrupção da coleta ou das análises. As informações mais completas se encontram na plataforma: http://www.bmanuel.org/projects/ju-HOME.html, onde é possível realizar algumas buscas **básicas.**

Ferrari e Cunha (2022) traçam um breve panorama das pesquisas em curso no Brasil sobre a linguagem jurídica. Encontram-se duas vertentes: uma está inserida no âmbito da análise do discurso jurídico, com o Grupo de Pesquisa Linguagem e Direito e a Associação de Linguagem e Direito (ALIDI). Outros grupos se inserem nos estudos terminológicos ou tradutórios, como o projeto TermiSul e o Projeto CoMET (Corpus Multilíngue para Ensino e Tradução), em seu subcorpus de linguagem legal comercial (CorTec) com cerca de um milhão de palavras.

Por que um corpus jurídico do português brasileiro?

Uma busca entre trabalhos acadêmicos e publicações científicas revelou que a utilização de corpora para o estudo da linguagem jurídica do português brasileiro está em fase de crescimento, sobretudo na pesquisa lexicográfica e na área de traduções, em específico de e para o inglês (FERRARI; CUNHA, 2022). Entretanto, segundo investigações prévias, não encontramos um corpus representativo de textos legais brasileiros. O que há são plataformas para busca de legislação que,

conforme exposto, não são adequadas para a realização de pesquisa linguística. A compilação de um corpus é um trabalho demorado e, quando feito por um único pesquisador, acaba frequentemente por ser abandonado com consequente perda desses dados. A possibilidade de contar com uma equipe treinada e levar adiante um projeto que se preocupa com a representatividade e curadoria dos dados, poderá auxiliar os estudiosos da linguística, mas, acreditamos, também aqueles da área do direito ou do ensino/aprendizagem da linguagem jurídica. Compilar e disponibilizar um corpus representativo da linguagem jurídica do português brasileiro pode auxiliar nos avanços dos estudos na área e na preparação de matérias para diversas finalidades.

O *LEX-BR-Ius* propõe-se como um corpus representativo da legislação federal brasileira. Ele está inserido no projeto *BR-Ius* que visa compilar e analisar uma série de corpora representativos da linguagem jurídica brasileira. O projeto inclui três etapas: a primeira está coletando e analisando dados legislativos federais (*LEX-BR-Ius*); a segunda se ocupará de dados da jurisprudência (*JUR-BR-Ius*) e a terceira de dados da doutrina (*DOC-BR-Ius*). A implementação de cada etapa está prevista para um prazo de dois a três anos a partir de seu início e inclui a publicação dos resultados de análises parciais além de sua descrição, e da disponibilização do corpus em uma plataforma de livre acesso para que outros pesquisadores possam utilizá-lo.

Os critérios de coleta dos textos para o *LEX-BR-Ius* seguem os seguintes preceitos e estão processando:

(a)textos legais federais, por serem aplicáveis em todo o território nacional;

(b)textos inteiros, pois apenas uma parte de um texto não é necessariamente representativa do todo (SINCLAIR, 2004), ainda mais em se tratando de textos legais, em que vários assuntos podem ser expostos na mesma norma e divididos em vária seções;

(c)textos vigentes na data da coleta. Como as normas legais são modificadas frequentemente, um texto vigente hoje, pode não o ser amanhã. É importante sublinhar que os textos vigentes atualmente podem ter sido criados tanto na vigência da Constituição presente quanto em períodos anteriores a ela e terem passado por modificações ao longo dos anos. A título de exemplo, citamos a Consolidação das Leis do Trabalho (BRASIL, 1943), promulgada sob Getúlio Vargas em 1943, ou seja, anterior à Constituição de 1988, mas que continua válida e cuja última alteração é de maio de 2022.

Essas particularidades nos levaram à decisão de anotar nos metadados as especificidades da coleta, incluindo a data da extração do texto e das suas modificações, possibilitando ao pesquisador refinar por data suas pesquisas, assim como a criação de uma versão da norma com marcação textual em ".xml" das diferentes seções e modificações pelas quais passou, possibilitando análises finas por seção, além de análises diacrônicas precisas.

Para garantir sua representatividade, a seleção dos textos está sendo feita com base em sua frequência de uso. Uma análise prévia está verificando, para cada seção, as citações de todas as legislações federais vigentes nas plataformas *Jusbrasil* e *Google Brasil*, para estabelecer quais são as mais utilizadas: a primeira seleção ordenou citações no *Jusbrasil* (o portal não refina citações acima de 10.000 ocorrências), para em seguida verificá-las no *Google Brasil* e estabelecer quais normas entrarão no corpus.

A seguir explicaremos algumas noções fundamentais sobre o processo legislativo brasileiro e como isso determinou as decisões metodológicas do *LEX-BR-Ius*.

O processo legislativo brasileiro

Os textos legais, também denominados normas jurídicas, legislações ou, popularmente, leis são textos editados pelo poder legislativo com a finalidade de regular a vida em sociedade em seus mais diversos aspectos, estabelecendo regras, direitos, deveres, sanções, entre outros, ao qual todos estão sujeitos. A sua criação se dá a partir de discussões teóricas no âmbito legislativo e seu conteúdo deve abranger o maior número de cenários e questões imaginárias possíveis, devendo ser geral e abstrata. O resultado dessas discussões é formalizado por escrito na forma das espécies normativas previstas pela Constituição que são então promulgados e publicados adquirindo status legal e passando a produzir efeitos (GONÇALVES, 2018; SILVA, 2020).

Para tanto faz-se necessário observar o processo legislativo, uma série de atos legislativos realizados seguindo as regras procedimentais previstas no Título IV, Capítulo I, Seção VIII da Constituição da República Federativa do Brasil de 1988 (BRASIL, 1988) pelos atores por ela legitimados (senadores, vereadores, presidente etc.). Trata-se de um processo complexo com particularidades próprias de cada espécie normativa a ser editada, cuja inobservância implica vício (formal ou material) e consequente inconstitucionalidade da norma. Vale ressaltar que, de forma a garantir a lisura, impedir vícios e inconstitucionalidades e fiscalizar a legitimidade do processo legislativo, a Constituição estabelece ainda um rigoroso processo de controle de constitucionalidade que implica, entre outros, um controle prévio/preventivo realizado seja pelo legislativo que pelo executivo nas fases de elaboração da norma e um controle repressivo/ posterior após a edição da norma (SILVA, 2020; LENZA, 2020).

De forma geral, o processo legislativo envolve três grandes fases:

a) Fase de iniciativa: dá início ao processo com a proposição do projeto de lei por aqueles a quem a Constituição legitima (art. 14, 60 a 62, CR/88) (BRASIL, 1988), seguindo as hipóteses e procedimentos por ela estabelecidos (iniciativa geral, concorrente, privativa, popular, conjunta, parlamentar ou extraparlamentar) (LENZA, 2020);
b) Fase constitutiva: análise, discussão, revisão e votação, em cada casa (Câmara dos Deputados e Senado), não necessariamente nessa ordem, do projeto de lei no âmbito legislativo pelos parlamentares, primeiramente pelas comissões e depois, a depender da espécie, pelo Plenário (deliberação parlamentar). Em caso de aprovação do projeto ele segue para apreciação do chefe do executivo que decide 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) (LENZA, 2020);
c) Fase Complementar: última fase do processo, composta pela promulgação e a publicação. A primeira é o ato, geralmente realizado pelo chefe do executivo, que transforma o projeto de lei em lei ao atestar sua validade, trazendo-a à existência. Após a promulgação há a publicação do texto legal no Diário Oficial para dar conhecimento a todos da existência e conteúdo da nova norma e estabelecer quando ela deverá passar a ser cumprida (LENZA, 2020).

Esclarece-se ainda que o Brasil é uma Federação, o que significa que o poder é distribuído entre os entes federados (União, Estados Membros, Distrito Federal e Municípios) que são dotados de autonomia recíproca o que permite, entre outros, que cada ente elabore suas próprias legislações. Isso implica que os textos legais podem ser federais, estaduais ou municipais, sendo a Constituição hierarquicamente superior às demais normas, aquelas federais hierarquicamente superiores às estaduais e às municipais e as estaduais superiores às municipais. Independente da esfera (federal, estadual ou municipal) é a nossa Carta Magna que determina as espécies e matérias a serem regulamentadas por cada uma em seus artigos 59 e seguintes. É importante esclarecer que, apesar de não estar elencada no texto constitucional, a Constituição também é uma espécie normativa (SILVA, 2020; LENZA, 2020).

Segundo eles podem ser editadas sete espécies normativas, necessariamente por meio do processo legislativo:

a) Emendas à constituição: normas que alteram o texto constitucional. É importante ressaltar que certas partes da Constituição (cláusulas pétreas, artigos que regulam a forma federativa do estado, os direitos e garantias individuais, etc.) não podem ser alteradas, conforme estabelecido pelo próprio texto constitucional. (LENZA, 2020);
b) Leis complementares: normas cuja criação e finalidade está prevista taxativamente, ou seja, de forma explícita, na Constituição (LENZA, 2020);
c) Leis ordinárias: normas de competência residual, regulam tudo o que não for objeto de lei complementar, decreto legislativo ou resolução (LENZA, 2020);
d) Leis delegadas: normas editadas pelo Presidente da República mediante de solicitação ao Congresso Nacional para 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);
e) Medidas provisórias: normas editadas pelo Presidente da República por iniciativa própria, sob a justificativa de relevância e urgência. 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. (LENZA, 2020);
f) Decretos legislativos: normas que regulam matérias de competência exclusiva do Congresso Nacional (art. 49 e 62, § 3.º, CR/88) podendo 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 que não participa do processo (LENZA, 2020);
g) Resoluções: normas que regulamentam as matérias de competência privativa da Câmara dos Deputados e do Senado Federal (art. 51 e 52, CR/88 e regimentos internos). 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).

Vale ressaltar que as plataformas de busca de legislação utilizam uma nomenclatura para os textos legais um pouco diferente daquela utilizada na Constituição. Por exemplo, no Portal da Legislação do Planalto as categorias disponíveis 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.

Faz-se importante esclarecer que o termo "código", como em "Código penal" não se encontra na Constituição, não sendo uma espécie normativa. A palavra "código" seria o assunto da norma, sendo um termo adotado pelos operadores do direito para se referir a normas que regulam de forma ampla determinada área do direito, por exemplo o Código Civil (BRASIL, 2002), que regula as normas de Direito Civil. Essas normas que ficaram conhecidas como códigos pertencem a diferentes espécies normativas; o Código Civil (BRASIL, 2002), por exemplo, é uma lei ordinária, já o Código Penal (BRASIL, 1940) é um Decreto Lei. O mesmo se aplica aos chamados "estatutos": sua criação não está prevista na Constituição de 1988. Trata-se de um termo utilizado para se referir a normas que regulam grupos específicos, por exemplo o Estatuto do idoso (BRASIL, 2003), que regula questões próprias da parcela da população classificada como idosa, ou seja aquela com 60 anos ou mais.

Já os Decretos-Leis são uma espécie normativa prevista na Constituição anterior à nossa que extinguiu essa espécie normativa. Eles não se confundem com os decretos legislativos previstos pela Constituição (BRASIL, 1988). Se trata de institutos diversos. Os Decretos Lei são textos com força de lei que eram editados pelo Presidente da República antes de 1988, logo anteriores à Constituição

atual. Entretanto, alguns desses decretos ainda produzem efeitos, entre eles os de maior destaque são: o Decreto-Lei No 2.848, de 7 de dezembro de 1940, que institui o Código Penal (BRASIL, 1940), o Decreto-Lei Nº 3.689, de 3 de outubro de 1941, que institui o Código de Processo Penal (BRASIL, 1941) e o Decreto-Lei Nº 5.452, de 1º de maio de 1943, que institui a Constituição das Leis Trabalhistas (BRASIL, 1943).

Quanto à classificação "Decretos", essa se refere às normas editadas pelo presidente da república que têm natureza administrativa. Eles regulamentam as leis e dispõem sobre a organização da administração pública. Da mesma forma, os "Decretos não Numerados" também são administrativos, mas diferentemente dos demais, eles não possuem caráter normativo. Logo não podem ser considerados textos legais. São textos editados pelo presidente que dispõe sobre questões da administração pública, tais como abertura de crédito e criação de grupos de trabalho. De tal forma, decidiu-se não os incluir em nossa coleta.

A proposta do *LEX-BR-Ius* corpus

A explicação acima quer mostrar a complexidade não somente do processo legislativo em si, mas também das diferenças situacionais e estruturais entre as espécies normativas. Para que seja representativo da legislação brasileira, decidiu-se que o corpus deveria incluir somente normas federais. A segunda decisão metodológica recaiu sobre como dividir internamente o corpus e como balanceá-lo. Após uma série de avaliações a respeito da sua possível arquitetura, optou-se por seguir a terminologia difundida no mundo jurídico. O *LEX-BR-Ius* terá 7 seções: (a) Constituição Federal, (b) Códigos, (c) Estatutos, (e) Emendas à Constituição, (f) Medidas provisórias, (g) Leis ordinárias e (h) Leis complementares.

Como dissemos, as legislações são editadas em uma frequência quase diária, por isso a relação de textos legais em vigor e sua quantidade é extremamente volátil. Para fins ilustrativos, a seguir apresentamos o número de textos legais federais em janeiro de 2022, quando foi dado início à compilação do corpus.

Tabela 1 – Textos legais federais por categoria em 28/01/2022

Categoria	**Nº textos**
Constituição	1
Emendas à Constituição	114
Códigos	17
Leis complementares	192
Leis ordinárias	13491
Medidas provisórias	1713
Estatutos	18
TOTAL	15546

Fonte: autoras (2022)

Como pode ser observado há uma discrepância entre o número de textos por espécie normativa, desde os Códigos que giram em torno de 17 textos até leis ordinárias que na época da formulação da tabela era de 13.491 textos. Além dessa diferença em termos de quantidade de textos por categoria, também a extensão dos textos é altamente variável e se torna ainda mais óbvia quando comparamos textos de espécies distintas. Por exemplo: a Emenda Constitucional 66, de 13/7/2010 (BRASIL, 2010) tem 43 palavras, já o Estatuto da Criança e do Adolescente (BRASIL, 1990) conta com 35.241 palavras contra 167.415 palavras da Consolidações das Leis do Trabalho (BRASIL, 1943).

Após os primeiros sete meses da primeira etapa, foram compiladas quatro das sete seções previstas, para um total de 989.090 palavras:

a) Constituição, com um texto de 97.082 palavras;

b) Códigos: foram escolhidos, segundo os critérios citados acima, 13 textos para um total de 716.411palavras;
c) Emendas à Constituição: foram compilados 47 dos 114 textos para um total de 50.241 palavras;
d) Estatutos: foram selecionados 10 dos 18 textos para um total de 125.356 palavras.

Como especificado, a seleção dos textos por seção foi efetuada pela frequência. Uma outra escolha metodológica que pesou, especialmente no que se refere à seção "Códigos" foi o fato que excluímos deliberadamente códigos que possam ser pensados para parcelas específicas da população, como é o caso do Código Penal Militar (BRASIL, 1969), apesar de sua alta frequência de uso.

Pensamos que, para uma melhor utilização, deveríamos disponibilizar o corpus em diferentes versões dos mesmos textos, de maneira a facilitar análises futuras. Além da lematização e do processo de etiquetagem em *Part of Speech* (PoS), quisemos implementar uma marcação textual que levasse em conta as características dos textos legais. Para tal, foi feito um estudo prévio das diferentes seções em que um texto legislativo é subdividido e preparado um manual de etiquetagem específico (MARQUES, em preparação). Para o projeto reputamos que uma marcação em ".xml" completa (*full XML*) utilizando ou não o sistema TEI seria demasiadamente detalhada, por isso optamos por seguir a proposta de Hardie (2014), um sistema de regras que permite adicionar ou isolar informações no texto através de etiquetas personalizáveis que são delimitadas por parênteses angulares (< e >) utilizando a extensão ".xml".

Seguiu-se uma nova discussão metodológica sobre qual língua utilizar para tais etiquetas. O inglês, em um primeiro momento, pareceu-nos uma boa opção por permitir uma abrangência maior do corpus, permitindo análises interlinguísticas. Contudo, a terminologia

de cada país é muito específica e no caso do português jurídico e do inglês jurídico temos uma diferença de sistemas jurídicos que torna o estabelecimento de equivalentes tradutórios uma tarefa complexa. Isso porque o inglês jurídico tem como contexto o *common law*, o sistema jurídico anglo-saxão, enquanto no Brasil impera o *civil law* (LENZA, 2020). Como consequência, muitos institutos brasileiros não existem ou são aplicados diferentemente no *common law* e vice-versa, sendo muitas vezes necessário recorrer a adaptações e explicações para que as particularidades dos sistemas não interfiram na transmissão da mensagem.

A título de exemplo trazemos o uso da abreviação "Art." nas legislações brasileiras. O artigo é a unidade básica do texto, sendo utilizada a abreviação "Art." para introduzir e enumerar o conteúdo de determinado assunto trazido no texto legal, na mesma linha do texto e geralmente ocupando poucas linhas. Em inglês existe a tradução literal "*Article*" que também é usado nas legislações nessa língua, mas esta não corresponde ao significado utilizado no Brasil. *Article* é sempre escrito na integra, nunca abreviado, em letra maiúscula e centralizada, estando separada do texto, e por fim, seu papel na estrutura é de subdividir partes da lei, incluindo vários assuntos diversos em seu interior. O termo mais próximo do "Art." tal qual usado nas nossas leis em inglês jurídico seria *"section",* cuja tradução literal seria seção. Entretanto, nas leis Brasileiras temos uma subdivisão intitulada "seção", cuja função se assemelharia ao *Article* usado no inglês jurídico, mas que não tem correspondente exato. (CASTRO, 2013)

Para uma melhor visualização dessa diferença, se observarmos o texto da Constituição Brasileira (BRASIL, 1988), este tem 368 artigos e é subdividido em preambulo, título (9), capítulo (33), seção (50) e subseção (5). Já o texto da Constituição Americana (EUA, 1788) tem 21 *Sections*, sendo que 3 de seus *Articles* não tem *Sections* e a maior

parte deles ocupa ao menos uma página e é subdividido em *Preamble, Articles (7) e Amendments* (27).

Diante dessas questões o uso de termos em inglês se mostrou inviável já que poderia abrir espaço para ambiguidades e interpretações incorretas dos termos utilizados. Decidiu-se, portanto, criar etiquetas em português, seguindo a nomenclatura própria da área jurídica brasileira. Optou-se por criar etiquetas para identificar os textos, armazenar seus metadados em forma de cabeçalho, separar as seções dos textos normativos e seus artigos. Para evitar problemas de codificação e processamento todas as etiquetas são em letras minúsculas e não possuem acentos ou sinais gráficos.

O processo de compilação do corpus

Para a compilação das várias seções do corpus, cada norma é acessada no Portal da Legislação do Planalto[12] e copiada por inteiro. As modificações pelas quais passam as normas são evidenciadas, no texto online, com um tachado. Para manter tal informação e possibilitar a inserção adequada das etiquetas correspondentes, o texto copiado é inicialmente salvo em formato ".docx". Foram testadas algumas expressões regulares para conseguir importar o tachado diretamente no texto em ".txt", mas, até o momento, não encontramos nenhuma forma de semi-automatizar o processo que se mostre suficientemente confiável. O texto é então nomeado com um código que possa facilmente identificar o tipo de norma, seu número e a data de publicação. Por exemplo, o Código de Defesa do Consumidor (BRASIL, 1990) apresenta como sigla C8.078_11.09.1990, ou seja: C de Códigos, 8.078 que é o número da norma, e 11.09.1990 que se refere à data em que foi publicada.

[12] Disponível em: http://www4.planalto.gov.br/legislacao. Acesso em: 05 jan. 2022.

A partir do texto em ".docx" é preenchido, no o *software Notepad++* (HO,2020), o cabeçalho em XML que contém os metadados: (a) o código de identificação da norma, (b) o nome da norma, (c) sua ementa, (d) o tipo da norma, (e) o assunto, (f) a área a que pertence, (g) o chefe do executivo sob o qual foi promulgada, (h) as datas de promulgação, publicação e início da vigência, (i) as alterações pelas quais passou, (j) o número de artigos que contém, (k) o número de palavras, (l) o portal do qual a norma foi extraída, (m) a data da extração, (n) a qual seção pertence, (o) a equipe de compilação responsável e os revisores. A imagem abaixo ilustra tal processo, com o cabeçalho do Código Florestal (BRASIL, 2012).

Figura 1- 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: autoras (2022)

Em seguida, sempre a partir do texto em ".docx", que contém a norma tal como na data da coleta, são criadas duas versões da mesma com codificação UTF-8 no *software Notepad++* (HO,2020): uma de texto limpo (*raw text*) em ".txt", e outra com anotação de marcação textual (*textual markup*) segundo nossa adaptação de Hardie (2014) em ".xml".

O texto em sua versão limpa *(raw text*) passa por um processo manual de limpeza com auxílio de algumas expressões regulares no software em questão. São retirados os elementos extratextuais, tabelas, anexos, espaços e linhas em branco. Entre os elementos retirados, citamos: o Brasão das Armas Nacionais da República Federativa do Brasil, além de vários dizeres (ex.: "Câmara dos Deputados", "Centro de Documentação e Informação", "Presidência da República", "Secretaria-Geral"), os hiperlinks, o nome da lei, as frases de abertura e de promulgação e publicação, as divisões textuais ("capítulo", "seção", "título", etc.), as modificações (ex.: "VETADO"; "Caput" do artigo com redação dada pela Lei) e o nome do chefe de estado e outros políticos envolvidos na criação da norma. As informações relevantes para a pesquisa são preservadas, pois constam nos metadados e na anotação de *markup* textual. Essa versão é pensada para facilitar sua utilização em qualquer ferramenta de análise (como o *AntConc* (ANTHONY,2019)) por parte dos pesquisadores que queiram extrair estatísticas gerais de cada texto ou do corpus inteiro sem ter que se preocupar com um processo prévio de limpeza, mas também para o processamento das operações de lematização e etiquetagem em *PoS* que estão começando a serem implementadas com a parte do corpus já pronta.

A versão com marcação textual passa por um processo manual de anotação, seguindo o guia que orienta o projeto. Nela apagam-se as informações extratextuais, tabelas, anexos, espaços e linhas

em branco e procede-se à comparação com o texto em ".docx" para a marcação das etiquetas textuais por nós elaboradas. Na figura abaixo ilustramos essa versão com a Emenda Constitucional nº 39 de 19/12/2002.

Figura 2– Emenda Constitucional nº 39 com marcação textual

<texto id = "EC39_19.12.2002">
<norma>
EMENDA CONSTITUCIONAL Nº 39, DE 19 DE DEZEMBRO DE 2002
</norma>
<ementa>
Acrescenta o art. 149-A à Constituição Federal (Instituindo contribuição para custeio do serviço de iluminação pública nos Municípios e no Distrito Federal).
</ementa>
<abertura>
As Mesas da Câmara dos Deputados e do Senado Federal, nos termos do § 3º do art. 60 da Constituição Federal, promulgam a seguinte Emenda ao texto Constitucional:
</abertura>
<artigo>
Art. 1º A Constituição Federal passa a vigorar acrescida do seguinte art. 149-A:
"Art. 149-A Os Municípios e o Distrito Federal poderão instituir contribuição, na forma das respectivas leis, para o custeio do serviço de iluminação pública, observado o disposto no art. 150, I e III.
Parágrafo único. É facultada a cobrança da contribuição a que se refere o caput, na fatura de consumo de energia elétrica."
</artigo>
<artigo>
Art. 2º Esta Emenda Constitucional entra em vigor na data de sua publicação.
</artigo>
<promulgacao>
Brasília, em 19 de dezembro de 2002
</promulgacao>
<assinatura>

Fonte: autoras (2022)

A correta inserção das etiquetas não é de domínio imediato, é necessário um treinamento prévio, seja em relação ao funcionamento do esquema XML, que não permite abertura e fechamento de etiquetas com ou sem atributos sem os devidos cuidados, seja pelas particularidades do texto normativo em si, que requer um conhecimento da terminologia jurídica e sua utilização. Nossa equipe

atual é composta por uma mestranda formada em Direito e por alunos de Iniciação Científica da Faculdade de Letras. Antes do início da fase de coleta, uma série de discussões norteou a preparação do guia. Houve encontros para a apresentação do projeto, para o treinamento sobre a utilização das etiquetas e constantemente é ativo um grupo de discussão para resolução de dúvidas quanto ao trabalho. Como há rotatividade no grupo, esses treinamentos deverão ser constantes para que haja uma formação adequada de cada nova leva de jovens pesquisadores. A fim de garantir a confiabilidade de nosso trabalho, estamos implementando um processo de revisão de todas as etapas efetuado pelos membros mais experientes do grupo. Ao mesmo tempo estamos pensando em possíveis testes estatísticos a serem aplicados ao corpus inteiro ou a parte dele para verificar sua confiabilidade e significância.

As próximas etapas da compilação do corpus serão (a) a etiquetagem em *PoS* das partes do corpus que ficarem prontas – inicialmente prevemos utilizar o etiquetador Palavras (BICK, 2000): isso permitirá o início da etapa de análises mais finas que pretendemos implementar; (b) uma Análise Multidimensional (AMD) que possibilite identificar as características linguísticas e situacionais dos textos legais federais brasileiros compilados até o momento e seus padrões de co-ocorrência de elementos léxico-gramaticais (dimensões) segundo a proposta de Biber (1988). Tal estudo nos ajudará a definir de maneira mais clara como balancear; (c) a coleta e processamento dos dados das próximas seções do corpus.

Questões metodológicas em aberto

Ao longo desse primeiro período de coleta e processamento do corpus, se apresentaram várias questões que merecem uma

discussão à parte, por terem determinado alguns passos operacionais sucessivos ou mudanças de rumo no projeto.

Em primeiro lugar citamos a criação do guia de etiquetas: as discussões prévias nortearam a criação das etiquetas que, como dissemos, são na língua portuguesa e seguiram a terminologia adotada nas legislações brasileiras. À medida que as normas jurídicas começaram a ser processadas e marcadas, percebemos, contudo, um fenômeno não esperado. As normas, que em teoria deveriam seguir uma certa estrutura hierárquica e padrão de organização e divisão interna, existindo inclusive manuais de redação tanto da Câmara dos Deputados (BRASIL, 2004), quanto do Senado (BRASIL, 2006) e do Planalto (BRASIL, 2018), que entre outros, abordam questões gramaticais e estruturais específicas para a redação de textos legislativos, não o são. Foram encontrados erros ortográficos, de digitação, trechos repetidos e também falta de padronização na formatação e no uso de alguns termos dentro de um mesmo texto legal, assim como o uso de estruturas, informações extratextuais, como observações e expedientes e termos exclusivos de algumas legislações que não se enquadravam nas nossas etiquetas. Isso não somente dificultou o processamento dos dados, mas também nos compeliu à criação de novas etiquetas dos casos não previstos. Um exemplo que levou à criação de uma nova etiqueta foi um trecho do Código Comercial (BRASIL, 1850):

> 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. (BRASIL, 1850)

Para resolver essa questão criou-se a etiqueta <outros>, que está sendo usada para delimitar informações que não fazem parte

do texto da lei em si, mas que foram reputadas importantes de serem preservadas via marcação textual.

Uma outra questão refere-se ao balanceamento do corpus. Inicialmente tínhamos pensado em dividi-lo nas seções que apresentamos acima, manter os textos completos, coletar aqueles com maior frequência de uso e balancear as diversas seções mantendo um número de palavras similar entre elas. Após a compilação das primeiras quatro seções percebemos que as dimensões de cada norma não nos permitem realizar o balanceamento dessa forma. Nas investigações iniciais estamos também buscando entender o quanto as diferentes seções realmente se diferenciam entre si em termos de características linguísticas e situacionais, se seria mais oportuno separar grupos mais abrangentes ou, até mesmo, eliminar tal divisão interna. Até o momento esse fator é o que mais pesa em termos de decisões metodológicas: contamos que o trabalho de Marques (em preparação) possa nos dar respostas sobre a melhor forma de proceder.

A terceira questão em aberto é relativa ao tipo de lematização e anotação PoS ao qual submeteremos o corpus. Os dados compilados até o momento estão sendo processados com o *parser* PALAVRAS (BICK, 2000) para a sucessiva realização da AMD (BIBER, 1988). Um corpus lematizado e anotado morfossintaticamente é fundamental para possibilitar pesquisas linguísticas. O PALAVRAS é utilizado por importantes corpora do português, como o projeto AC/DC (SANTOS; BICK 2000), o Linguateca (SANTOS et al., 2004), o C-ORAL-BRASIL (BICK, 2012), o Corpus Brasileiro de Variação de Registro (BERBER SARDINHA *et al*., 2014) entre outros e apontado como uma das alternativas com mais alto grau de confiabilidade. Todavia, em se tratando de um gênero específico e de uma linguagem altamente especializada, nossa dúvida recai sobre a efetiva precisão do *parser* para tal variedade ou se outros etiquetadores, devidamente treinados,

seriam capazes de uma acurácia maior. É de nosso interesse levar adiante uma análise quantitativa e qualitativa de nossos dados testando outros etiquetadores como, por exemplo o *TreeTagger* (SCHMID, 1994 e 1995), rotinas em Python (PYTHON SOFTWARE FOUNDATION, 2020) com a biblioteca Spacy) e comparar os resultados, visando assim contribuir para uma avaliação das ferramentas disponíveis para o português brasileiro.

Possibilidades investigativas

Como afirma Carapinha (2018, p. 93), a "variação interna da linguagem do Direito nunca despertou, com efeito, muita curiosidade, e só recentemente a variabilidade e a complexidade dessas linguagens, bem como a sua articulação, começaram a ser alvo de interesse". De fato, os estudos sobre as normas jurídicas, inclusive aqueles em corpora, têm se dedicado majoritariamente às análises lexicais. Essa com certeza é uma das possibilidades investigativas de nosso corpus e que o próprio grupo está levando adiante em uma de suas frentes. Seus dados poderão também ser comparados àqueles de variedades do português e a outras línguas, seja para efeitos de análise intralinguística seja para servir como base para traduções ou compilações de glossários e dicionários (GOTTI, 2003, 2011, 2016).

Como dissemos acima, avaliar nossos dados através da AMD é algo que já está sendo efetuado e que possibilitará um entendimento melhor da variabilidade da língua utilizada nas normas legais, não somente em suas particularidades lexicais, mas também na complexidade sintática que a caracteriza, assim como na sua precisão semântica e nos diversos processos morfológicos que comparecem nesse tipo de texto[13].

[13] Carapinha (2018) aponta, em sua análise de Códigos do português europeu, por exemplo, para um elevado número de nominalizações e processos de prefixação entre as características morfológicas. Outro aspecto identificado pela autora

Do ponto de vista sintático nos interessa averiguar o fenômeno da impessoalidade, identificável pelas construções passivas e pelo uso de pronomes específicos. Estudos ainda incipientes apontam para a forte presença de tais traços no corpus, mas estamos aguardando a etiquetagem morfossintática para poder corroborar tais achados.

Para além das investigações de cunho propriamente linguístico, vemos também aplicações práticas para o corpus: (a) como fonte de dados confiáveis e limpos para o processamento da linguagem natural (PLN) na criação de ontologias específicas do domínio legislativo ou na extração de terminologias; (b) como material a ser utilizado na produção de materiais didáticos para o ensino da linguagem jurídica ou para a sua simplificação, em projetos de interação multidisciplinares.

Conclusão

Nesse artigo apresentamos o LEX-BR-Ius corpus, um corpus representativo da linguagem legal utilizada nas leis federais brasileiras. São expostas as plataformas de acesso às normas legais assim como suas divisões em espécies normativas diferentes, que determinaram as subdivisões e a nomenclatura adotada no corpus. É apresentado o estado da arte sobre os corpora jurídicos disponíveis e é motivada a escolha por um corpus de legislações brasileiras. São introduzidas as escolhas metodológicas do corpus e as características da parte até agora compilada, além de serem detalhadas as etapas de compilação e algumas dificuldades enfrentadas em sua organização. Por fim, são ilustradas possíveis utilizações do corpus e trabalhos que foram iniciados junto ao grupo de compilação.

é o comparecimento de listas de verbos, nomes e adjetivos, com significado próximo, na mesma frase quando há enunciados definitórios, fenômeno esse observado também por nós.

Referências

ANTHONY, Lawrence. *AntConc* (Version 3.5.8) [Computer Software]. Tokyo: Waseda University, 2019. Disponível em: Acesso em: 5 mar. 2020.

ATKINS, Sue; CLEAR, Jeremy; OSTLER, Nicholas. Corpus Design Criteria. *Literary e Linguistic Computing*, v. 7, n. 1, p. 1-16, 1992.

BAKER, P.; HARDIE, Andrew; MCENERY, T. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press, 2006.

BARBERA, Manuel*. Jus Jurium, corpus giuridico italiano*. 2005. Disponível em: http://www.bmanuel.org/Projects/ju-HOME.html. Acesso em: 20 jan. 2022.

BARBERA, Manuel; ONESTI, Cristina. Scheda Progetto di ricerca n. 9. Corpus Jus Jurium. *In*: DIADORI, Pierangela. *Progetto JURA: la formazione dei docenti di lingua e traduzione in ambito giuridico italo*. Perugia: Guerra Edizion, 2009. p. 349-351.

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

BERBER SARDINHA T.; KAUFFMANN C.; ACUNZO C. *A multi-dimensional analysis of register variation in Brazilian Portuguese*. Corpora.n. 9, p. 239-271, 2014.

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

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

BICK, E. *The parsing system palavras*. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 505f. Tese (Doutorado em Linguística) – Department of Linguistics, University of Arhus, Arhus, Dinamarca, 2000.

BICK, E. A anotação gramatical do C-ORAL-BRASIL. *In*: RASO T.; MELLO H. (Orgs.). C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal. Belo Horizonte, Editora UFMG, 2012. p. 223-254.

BIEL, Łucja. *Corpus-Based Studies of Legal Language for Translation Purposes*: Methodological and Practical Potential. Online proceedings of the XVII European LSP Symposium, 2009. Disponível em: https://www.researchgate.net/publication/216576418_CorpusBased_Studies_of_Legal_Language_for_Translation_Purposes. Acesso em: 6 set. 2020.

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 set. 2020.

BRASIL. *Decreto-lei nº 1.001, de 21 de outubro de 1969* (Código Penal Militar). Disponível em: http://www.planalto.gov.br/ccivil_03/Decreto-Lei/Del1001.htm. Acesso em: 10 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 set. 2020.

BRASIL. *Decreto-lei nº 3.689, de 3 de outubro de 1941* (Código de Processo Penal). Disponível em: http://www.planalto.gov.br/ccivil_03/Decreto-Lei/Del3689.htm. Acesso em: 10 set. 2020.

BRASIL. *Decreto-lei nº 5.452, de 1º de maio de 1943* (Consolidação das Leis do Trabalho). Disponível em: http://www.planalto.gov.br/ccivil_03/Decreto-Lei/Del5452.htm. Acesso em: 10 set. 2020.

BRASIL. *Lei nº 8.069, de 13 de julho de 1990* (Estatuto da Criança e do Adolescente). Disponível em: http://www.planalto.gov.br/ccivil_03/leis/l8069.htm. Acesso em: 10 set. 2020.

BRASIL. *Lei nº 8.078, de 11 de setembro de 1990* (Código de Defesa do Consumidor). Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/L8078.htm. Acesso em: 10 set. 2020.

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. Acesso em: 10 set. 2020.

BRASIL. *Lei nº 10.741, de 1º de outubro de 2003* (Estatuto do Idoso). Disponível em: http://www.planalto.gov.br/ccivil_03/Leis/2003/L10.741.htm. Acesso em: 10 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 set. 2020.

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

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

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

CARAPINHA, C. A linguagem jurídica. Contributos para uma caracterização dos Códigos Legais. *Redis: Revista de Estudos do Discurso*, n. 7, p. 91-119, 2018.

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

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

COULTHARD, Malcolm; JOHNSON, Alison. *An Introduction to Forensic Linguistics: Language in Evidence*. New York: Routledge, 2007

COULTHARD, Malcolm; JOHNSON, Alison. *The Routledge handbook of forensic*

linguistics. Madison Ave: New York, 2010.

DAVIES, Mark; FERREIRA, Michael. *Corpus do Português*: Historical Genres. Disponível em: http://www.corpusdoportugues.org/hist-gen/. Acesso em: 23 set. 2021.

DUTCH LANGUAGE INSTITUTE*. Corpus Juridisch Nederlands* (Version 1.0) (September 2021) [Online service]. Disponível em: http://hdl.handle.net/10032/tm-a2-u2. Acesso em: 23 set. 2021.

ESTADOS UNIDOS DA AMÉRICA. *Constitution of the United States*. Disponível em: https://www.senate.gov/civics/constitution_item/constitution.htm#amendments. Acesso em: 20 set. 2022.

FANTINUOLI C.; ZANETTIN F. Creating and using multilingual corpora in translation studies. In: FANTINUOLI C. e ZANETTIN F. (Org.). *New directions in corpus-based translation studies.* Berlim: Language Science Press, 2015. p. 1-10.

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.

FRANCIS, W. N.; KUCERA H. *Computational Analysis of Present-Day American English*. Providence, RI: Brown University Press, 1967.

GIAMPIERI, Patrizia. Online Parallel and Comparable Corpora for Legal Translations. *Altre modernità / Otras modernidades / Autres modernités / Other Modernities*, Milão, n. 20, p. 237-252, 2018. Disponível em: https://www.researchgate.net/publication/329365608_Online_Parallel_and_Comparable_Corpora_for_Legal_Translations. Acesso em: 6 set. 2020.

GÓMEZ Guinovart, X.; SACAU Fontenla E. Parallel Corpora for the Galician Language: Building and Processing of the CLUVI (Linguistic Corpus of the University of Vigo). *In*: LINO, Maria Teresa; XAVIER, Maria Francisca; FERREIRA, Fátima; COSTA, Rute; SILVA, Raquel (Org.). *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisboa: European Language Resources Association (ELRA), 2004. p. 1179-1182.

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

GOTTI, Maurizio. *Specialized Discourse*: Linguistic Features and Changing Conventions. Bern: Peter Lang, 2003.

GOTTI, Maurizio. Globalisation in the legal field: Adopting and adapting international comercial arbitration rules. *In*: PÉREZ-LLANTADA, Carmen; WATSON, Maida (Org.). *Specialised*

Languages in the Global Village: A Multi-Perspective Approach. Newcastle upon Tyne: Cambridge Scholars, 2011. p. 79-101.

GOTTI, Maurizio. The translation of legal texts: Interlinguistic and intralinguistic perspectives. *ESP Today,* v. 4, n. 1, p. 5-21, 2016.

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

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

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

GOŹDŹ-ROSZKOWSKI, Stanisław. Corpus Linguistics in Legal Discourse. I*nternational Journal for the Semiotics of Law – Revue internationale de Sémiotique juridique*, 34, p. 1515-1540, 2021.

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

GOŹDŹ-ROSZKOWSKI, Stanisław; PONTRANDOLFO, Gianluca. Legal Phraseology Today: Corpus-based Applications Across Legal Languages and Genres*. International Journal of Specialized Comunication*, v. XXXVII, p. 130-138, 2015. Disponível em: https://www.academia.edu/18714805/Legal_Phraseology_Today_Corpus_based_Applications_Across_Legal_Languages_and_Genres. Acesso em: 6 set. 2020.

HARDIE, A. Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal*, v. 38, n. 1, p. 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 mar. 2020.

LENZA, Pedro. *Direito Constitucional esquematizado*. São Paulo: Saraiva, 2020.

LORZ R. A. Creating Law with Language – Crossing Borders and Connecting Disciplines from the Perspective of Legislative Practice. *In*: VOGEL F. (Ed). *Legal Linguistics Beyond Borders: Language and Law in a World of Media, Globalisation and Social Conflicts Relaunching the International Language and Law Association (ILLA).* Berlim: Duncker e Humblot GmbH, 2019. p. 5-8.

MARQUES, C.G.F. Uma análise multidimensional do LEX-BR-IUS, um corpus representativo da legislação federal brasileira. Dissertação (Mestrado em Estudos Linguísticos) – Faculdade de Letras da Universidade Federal de Minas Gerais, em preparação.

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.

MCENERY, T.; WILSON A. *Corpus Linguistics:* An Introduction. Edinburgh: Edinburgh University Press, 1996.

MCENERY T., XIAO R.; TONO Y. *Corpus-based Language Studies:* An Advanced Resource Book. London/New York: Routledge, 2006.

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

MELLO, H. Os corpora orais e o 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. 31-54.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, London, v. 18, n. 5, p. 544-551, 2011.

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

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO. *Corpus Brasileiro.* Disponível em: http://corpusbrasileiro.pucsp.br/cb/Inicial.html. Acesso em: 23 ago. 2021.

PONTRANDOLFO, Gianluca. Legal Corpora: an overview**.** *Rivista Internazionale di Tecnica della Traduzione*, Trieste, v. 14, p. 121-136, 2012. Disponível em: https://www.openstarts.units.it/bitstream/10077/9783/1/12Pontrandolfo.pdf**.** Acesso em: 6 set. 2020.

PYTHON SOFTWARE FOUNDATION. *Python Language Site: Documentation*, 2020. Página de documentação. Disponível em: <https://www.python.org/doc/>. Acesso em: 06 nov. de 2020.

RICHARD, Isabelle. Is legal lexis a characteristic of legal language? *Lexis* [Online], 11, p.1-14, 2018. Disponível em: https://www.researchgate.net/publication/324949333_Is_legal_lexis_a_characteristic_of_legal_language**.** Acesso em: 6 set. 2020.

ROSSINI, R. Favretti; TAMBURINI F.; DE SANTIS, C.. CORIS/CODIS: A corpus of written Italian: a defined and dynamic model. *In*: WILSON, A.; RAYSON, P.; MCENERY, T. (Org.). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World.* Munich: Lincom-Europa, 2002.

ROSSINI, R. Favretti; TAMBURINI F.; MARTELLI, E. Words from Bononia Legal Corpus. *Text Corpora and Multilingual Lexicography,* Amsterdam: John Benjamins, v. 8, p. 11-30, 2007.

SANTOS D.; BICK E. “Providing Internet access to Portuguese corpora: the AC/DC project”. *In*: GAVRILIDOU, Maria; CARAYANNIS, George; MARKANTONATOU, Stella; PIPERIDIS, Stelios; STAINHAUER, Gregory (Org.). *Proceedings of the Second International Conference on Language Resources and Evaluation* (LREC 2000) (Atenas, Grécia, 31 de Maio a 2 de Junho de 2000), p. 205-210.

SANTOS D.; SIMÕES A.; FRANKENBERG-GARCIA, A.; PINTO A.; BARREIRO A.; MAIA B.; MOTA C.; OLIVEIRA D.; BICK E.; RANCHHOD E.; DIAS A. J. J.; CABRAL L.; COSTA L.; SARMENTO L.; CHAVES M.; CARDOSO N.; ROCHA P.; AIRES R.; SILVA R.; VILELA R.; AFONSO S.. “Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa”. *In*: LUNA, Guillermo de Ita; CHÁVEZ, Olac Fuentes;

GALINDO, Mauricio Osorio (Org.). *Proceedings of the international workshop "Taller de Herramientas y Recursos Linguísticos para el Espanõl y el Portugués", IX Iberoamerican Conference on Artificial Intelligence* (IBERAMIA 2004) (Puebla, México, Novembro de 2004), pp. 147-154.

SCHMID, H. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994, p. 44-49.

SCHMID, H. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland, 1995.

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

SINCLAIR, John (Org.). *Corpus, Concordance, Collocation*. Oxford, UK: Oxford University Press, 1991.

SINCLAIR, John*. Trust the Text: Language, Corpus and Discourse*. London: Routledge, 2004.

TAGNIN, S. E. O., e TEIXEIRA, E. D. Lingüística de Corpus e Tradução Técnica – Relato da montagem de um corpus multivarietal de culinária. *Tradterm*, 10, p. 313-358, 2004.

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

TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam/ Philadelphia: John Benjamins, 2001.

UNIVERSITÀ DI BOLOGNA, *Bononia Legal Corpus*, Disponível em: http://corpora.dslo.unibo.it/BOLCCorpQuery.html. Acesso em: 10 set. 2020.

VIANA V. e TAGNIN S.E.O. (eds). *Corpora no ensino de línguas estrangeiras*. São Paulo: HUB editorial, 2011.

Recebido em: 30/09/2022
Aprovado em: 21/10/2022