A sintaxe no tribunal: apresentando e explorando um corpus jurídico em português anotado sintaticamente segundo o modelo *Universal Dependencies*

Lucelene Lopes, Maria das Graças V. Nunes, Magali S. Duran, Thiago A. S. Pardo

Núcleo Interinstitucional de Linguística Computacional (NILC)
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo
São Carlos – SP – Brasil

lucelene@gmail.com, gracan@icmc.usp.br, magali.duran@gmail.com, taspardo@icmc.usp.br

Abstract*. Legal texts represent a challenge for the field of Natural Language Processing, given the complexity and style characteristic of this type of text. In this paper, we contribute to the development of this area in the context of Brazilian Portuguese through two fronts. First, we present PortJur, a new legal corpus syntactically annotated according to the Universal Dependencies model, resulting, in addition to the relevant linguistic study, in a novel resource for Portuguese. Then, we explore the annotated corpus to create specialized lexical resources, such as lists of content words, verb forms, abbreviations and loanwords, and a gazetteer of named entities.*

Resumo. *Textos jurídicos representam um desafio para o campo do Processamento de Linguagem Natural, dada a complexidade e o estilo característicos desse tipo de texto. Neste artigo, contribuímos com o desenvolvimento dessa área para o português do Brasil em duas frentes. Primeiramente, apresentamos o PortJur, um novo corpus jurídico anotado sintaticamente de acordo com o modelo Universal Dependencies, resultando, para além do estudo linguístico relevante, em um recurso inédito para o português. Em seguida, exploramos o corpus anotado para criar recursos lexicais especializados, como listas de palavras de conteúdo, formas verbais, abreviaturas e estrangeirismos, e um almanaque de entidades nomeadas.*

1. Introdução

A automação de processos no meio jurídico brasileiro tem ganhado destaque nos últimos anos. A disponibilização de documentos digitalizados e a necessidade de acelerar a resposta do sistema jurídico à sociedade têm gerado uma grande demanda pela criação de recursos e sistemas computacionais de apoio. As necessidades incluem desde a simples busca por termos e a classificação de tipos de textos até a extração de informação de jurisprudências, análise das decisões de juízes e simplificação textual, entre outras. Nesse contexto, têm surgido plataformas especializadas que oferecem tecnologia para acesso e manipulação de documentos jurídicos, assim como tem crescido o número de grupos que utilizam corpora para o estudo da linguagem jurídica no Brasil, particularmente nas pesquisas de análise de discurso e nos estudos terminológicos e tradutórios (por exemplo, [Ferrari e Cunha 2022] e [Xavier 2002]).

Extrair informações de textos jurídicos não é tarefa fácil. De um lado, tem-se uma quantidade enorme de longos textos que variam em suas finalidades e características linguísticas – leis, códigos, processos e sentenças, contratos, registros e textos didáticos. De outro lado, cada tipo apresenta especificidades de formato e discurso, vocabulário e termos próprios, que diferem muito da linguagem do dia-a-dia e sua informalidade usual. Se, para o humano, o volume de dados dificulta a extração de informação, para os sistemas computacionais, treinados normalmente para processar linguagem não jurídica, a dificuldade recai sobre o tratamento de sentenças muito longas, estruturadas de maneira não convencional, além do vocabulário próprio e rico em abreviações e expressões raras e, muitas vezes, em latim. É nesse cenário que os corpora de textos jurídicos, bem como léxicos especializados nesse gênero, ganham importância para a área de Processamento de Linguagem Natural (PLN). Neste artigo, avançamos nesse esforço em duas frentes, trazendo contribuições teóricas e práticas.

Em uma primeira frente, coletamos um novo corpus jurídico para o português do Brasil, nomeado PortJur, e o anotamos segundo o modelo gramatical internacional *Universal Dependencies* (UD) [de Marneffe et al. 2021], que tem sido amplamente adotado pela comunidade mundial de PLN. Esse modelo tem revolucionado os estudos sintáticos e permitido o desenvolvimento de sistemas computacionais de análise textual (como *part of speech taggers* e *parsers*) do estado da arte para muitas línguas, incluindo o português. Destaca-se que a anotação de um corpus jurídico traz vários desafios, sendo que o estudo e as decisões de anotação em si já são contribuições relevantes. Pelo que se conhece, não há outros corpora jurídicos anotados com UD para o português. Em outra frente, exploramos o corpus anotado para a construção de recursos lexicais básicos que podem subsidiar outras iniciativas. Em linhas gerais, extraímos listas de palavras de conteúdo mais frequentes, de formas verbais e de abreviaturas e estrangeirismos típicos desses textos. Também compusemos um almanaque, ou seja, um léxico de entidades nomeadas que ocorrem no corpus jurídico em questão.

A seguir, apresentamos uma síntese dos trabalhos relacionados para o português. Na Seção 3, descrevemos o corpus PortJur e sua anotação UD. Na Seção 4, os recursos lexicais são apresentados. Na Seção 5, algumas considerações finais são feitas.

2. Trabalhos relacionados

Há vários corpora jurídicos para o português, sendo alguns compostos exclusivamente por textos desse gênero, como o RulingBR, o LEX-BR-Ius e o Ulysses Tesemõ, enquanto outros, mais diversificados, incluem uma parte dedicada a ele, mostrando a relevância do gênero, como o corpus Carolina. O RulingBR [Feijó e Moreira 2018] é composto por 10.623 sentenças judiciais do Supremo Tribunal Federal, o órgão máximo do Judiciário brasileiro. O LEX-BR-Ius [Ferrari e Marques 2022] é um corpus com cerca de 15 mil textos e 989 mil palavras, de documentos legais federais, e visa aos estudos linguísticos do gênero. Ulysses Tesemõ [Siqueira et al. 2024] é parte do projeto Ulysses, da Câmara dos Deputados, que é um conjunto de iniciativas de IA para aumentar a transparência e prover os legisladores com ferramentas de análises mais complexas. O corpus totaliza 30,7 GiB de texto bruto de dados judiciais, legislativos, acadêmicos, notícias e outros documentos relacionados. Carolina [Sturzeneker et al. 2022] é um corpus de textos em português brasileiro (período de 1970-2021), de diferentes gêneros e domínios, com informações de procedência e tipologia, e de acesso aberto e gratuito. Sua partição de

textos jurídicos compreende cerca de 40 mil textos e 196 milhões de palavras de documentos e notícias do Supremo Tribunal Federal.

Os corpora têm sido utilizados em diversas tarefas de anotação e para criação de sistemas de PLN. Na linha de anotação, [Souza et al. 2024] apresentam corpora que têm sido anotados manual, automática ou semi-automaticamente. Como exemplo, citamos o CDJUR-BR [Brito et al. 2023], um corpus público com 1.216 documentos, e o UlyssesNER-Br [Albuquerque et al. 2022], que é uma partição do Ulysses-Tesemõ, ambos anotados com entidades nomeadas. Na frente de sistemas, em [Castro e Neves, 2024], técnicas de PLN são utilizadas para agrupar acórdãos e identificar semelhanças textuais e possíveis divergências em jurisprudências. [Fama et al. 2024] propõem o uso de representações vetoriais e redes neurais recorrentes para predição de resultados judiciais. [Garcia et al. 2024] treinam um modelo de língua para textos jurídicos e estabelecem um *benchmark* para a área com tarefas de classificação de textos e reconhecimento de entidades nomeadas. [Lins et al. 2024] investigam o problema de sumarização automática abstrativa de textos legais. Vários desses trabalhos citam especificamente as dificuldades provenientes do tamanho maior dos textos (em relação a outros gêneros) e do vocabulário mais especializado.

Apesar dos trabalhos na área, não se tem conhecimento de esforços de anotação sintática de corpora jurídicos para o português segundo a iniciativa UD [de Marneffe et al. 2021]. A UD propõe uma estratégia "universal" para anotação morfológica, morfossintática e sintática de línguas, visando permitir estudos linguísticos tipológicos e abordagens multilíngues e de língua cruzada. Seguindo o paradigma de análise de dependências, a UD fornece 17 etiquetas morfossintáticas/classes gramaticais e 37 relações de dependência. Atualmente, em sua versão 2.16, o projeto conta com mais de 600 participantes e 200 treebanks para mais de 150 línguas. O português está representado por 7 corpora – Porttinari [Pardo et al. 2021; Duran et al. 2023], DANTEStocks [Di Felippo et al. 2024], PetroGold [Souza et al. 2021], Bosque [Rademaker et al. 2017], CINTIL [Branco et al. 2022], GSD e PUD [McDonald et al. 2013] –, sendo que nenhum deles é do gênero jurídico. No total, esses corpora somam mais de 1,5 milhão de *tokens*. O PortJur, aqui descrito, deve integrar o Porttinari.

Como ilustração, a Figura 1 exibe a análise UD para uma sentença do PortJur (os arcos indicam as relações de dependência sintática entre os *tokens* que compõem a sentença; acima dos *tokens*, encontram-se as etiquetas morfossintáticas; e, abaixo dos tokens, listam-se os respectivos lemas e atributos morfológicos). A caracterização completa do corpus PortJur é apresentada na seção a seguir.

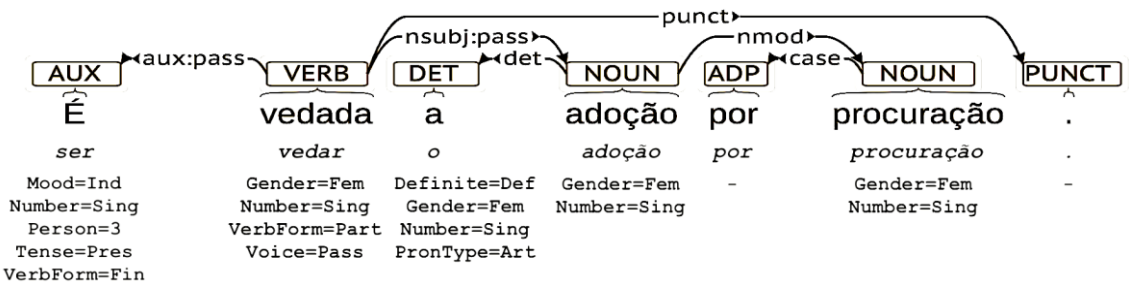

Figura 1. Exemplo de análise UD para a sentença "*É vedada a adoção por procuração.*"

3. Descrição do corpus

Compilou-se um novo corpus jurídico neste trabalho visando-se coletar textos contemporâneos e também privilegiar legislações típicas brasileiras. Foram coletados textos do direito público, de acesso aberto na web, produzidos pelo poder judiciário (principalmente ementas) e pelo legislativo (leis).

Para fins de anotação sintática, é importante que tudo que esteja dentro da lógica sentencial seja agrupado em uma única sentença, ou seja, é preciso desconstruir a estrutura usual do texto jurídico, eliminando quebras de linhas dentro de uma sentença e excluindo os elementos que constituam informação extra-sentencial, como títulos e subtítulos, que normalmente não são sintaticamente analisados. Por esse motivo, o corpus contém metadados que permitem a reconstrução dos textos na íntegra, caso seja de interesse de alguma aplicação. Como exemplo, o Quadro 1 mostra o *antes* e o *depois* do processo de preparação de um trecho do Estatuto da Criança e do Adolescente.

Quadro 1. Antes e Depois do processo de preparação de um texto para anotação

Antes:

Art. 4º É dever da família, da comunidade, da sociedade em geral e do poder público assegurar, com absoluta prioridade, a efetivação dos direitos referentes à vida, à saúde, à alimentação, à educação, ao esporte, ao lazer, à profissionalização, à cultura, à dignidade, ao respeito, à liberdade e à convivência familiar e comunitária.

Parágrafo único. A garantia de prioridade compreende:
a) primazia de receber proteção e socorro em quaisquer circunstâncias;
b) precedência de atendimento nos serviços públicos ou de relevância pública;
c) preferência na formulação e na execução das políticas sociais públicas;
d) destinação privilegiada de recursos públicos nas áreas relacionadas com a proteção à infância e à juventude.

Depois:

É dever da família, da comunidade, da sociedade em geral e do poder público assegurar, com absoluta prioridade, a efetivação dos direitos referentes à vida, à saúde, à alimentação, à educação, ao esporte, ao lazer, à profissionalização, à cultura, à dignidade, ao respeito, à liberdade e à convivência familiar e comunitária.

A garantia de prioridade compreende: a) primazia de receber proteção e socorro em quaisquer circunstâncias; b) precedência de atendimento nos serviços públicos ou de relevância pública; c) preferência na formulação e na execução das políticas sociais públicas; d) destinação privilegiada de recursos públicos nas áreas relacionadas com a proteção à infância e à juventude.

Os textos redigidos por integrantes do poder judiciário constituem 50,72% das sentenças do corpus e 47,47% dos *tokens*. Cada texto de julgamento que integra o corpus possui os seguintes metadados: URL de origem do texto, número do processo no órgão que o disponibilizou, número do processo original, classe processual, relator, órgão julgador, data do julgamento e data da publicação. Como os juízes muitas vezes repetem trechos das sentenças ao julgar um mesmo tema, teve-se o cuidado de variar tanto o tema dos julgamentos quanto os juízes que os proferiram. Integram o corpus julgamentos cujo acesso é livre nos sites do STF, STJ, TSE, TJSP e TJPE.

Os textos jurídicos redigidos por integrantes do poder legislativo (leis) constituem 49,28% das sentenças do corpus e 52,53% dos *tokens*. As leis escolhidas para compor o corpus incluem leis amplamente conhecidas no Brasil e aquelas chamadas “leis extravagantes” (ou seja, leis que versam sobre assuntos até então não contemplados devidamente nos códigos civil e penal), como a Lei Henry Borel, Lei do Marco Civil da

Internet, Lei Maria da Penha, Lei dos Direitos Autorais, Lei da Reforma Agrária, Estatuto da Pessoa Idosa e Estatuto da Criança e do Adolescente.

3.1. Dados numéricos do corpus

O corpus PortJur é composto por 2.005 sentenças, totalizando 101.297 *tokens*, com uma média de 50,52 *tokens* por sentença. Essa média reflete a densidade típica de textos do domínio jurídico, marcados por construções sintáticas longas e complexas, além de um vocabulário especializado. A sentença mais longa do corpus possui 451 *tokens*. O corpus é dividido em 12 subcorpora, cada um representando uma fonte distinta do discurso jurídico. A Tabela 1 apresenta a distribuição de sentenças, *tokens* e média de *tokens* por sentença em cada subcorpus. A distribuição das classes gramaticais revela predominância de substantivos (22,44%), preposições (18,45%) e determinantes (15,48%), o que é coerente com o estilo normativo e descritivo dos textos jurídicos.

Subcorpus	Tipo	Sentenças	*Tokens*	*Tokens*/Sentença
JURECA	Lei	232	10.753	46,35
JUREPI	Lei	190	9.173	48,28
JURJPE	Julgamentos	140	5.850	41,79
JURJSP	Julgamentos	166	4.800	28,92
JURLDA	Lei	221	11.028	49,90
JURLHB	Lei	63	4.240	67,3
JURLRA	Lei	120	7.110	59,25
JURMDI	Lei	64	4.580	71,56
JURMDP	Lei	98	6.331	64,60
JURSTF	Julgamentos	288	13.530	46,98
JURSTJ	Julgamentos	306	15.933	52,07
JURTSE	Julgamentos	117	7.969	68,11

Tabela 1. Distribuição de sentenças e *tokens* por subcorpora do PortJur

3.2. Complexidade textual

Se comparado a outros corpora do português disponíveis na UD – dois jornalísticos (Bosque e Porttinari), um de *tweets*/postagens do X (DANTEStocks) e um acadêmico (Petrogold) –, o PortJur é o mais complexo. Isso pode ser observado na Tabela 2, que mostra algumas métricas de complexidade textual extraídas pela ferramenta NILC-Metrix [Leal et al. 2023]. As três primeiras métricas são índices de leiturabilidade que se aplicam ao escopo de sentenças, sendo que o PortJur foi considerado o mais complexo pelas duas primeiras e o segundo mais complexo pela terceira delas. Além disso, o PortJur tem as sentenças mais longas (linha 4 da Tabela 2) e o maior tamanho médio das palavras de conteúdo (linha 5). O PortJur também possui um percentual de operadores lógicos bem superior aos demais corpora (linha 6), o que significa não apenas que é mais complexo, mas também que, por conter mais conectivos, é o mais logicamente estruturado (o que é esperado de textos que devem evitar dar margem a más interpretações). Por fim, o PortJur apresenta o maior percentual de construções de voz passiva (linha 7). Todas essas características tornam o PortJur um corpus de alta complexidade e, naturalmente, trazem mais desafios para a anotação sintática.

Medidas de complexidade	Bosque	Porttinari	DANTEStocks	Petrogold	PortJur
Índice Flesh de leiturabilidade[1]	45,03988	50,96858	67,46326	28,85404	**12,60864**
Índice GunningFog[2]	7,87051	6,53394	4,1629	9,30437	**13,70189**
Índice DaleChall adaptado[3]	11,16911	10,41144	**13,60309**	12,18864	12,70405
Tokens por sentença[4]	19,33	16,01	10,14	22,87	**33,84**
Média de sílabas por palavra de conteúdo	2,81	2,75	2,32	3,03	**3,22**
% de operadores lógicos	0,035	0,038	0,028	0,031	**0,055**
% de voz passiva	8,09	9,99	2,2	19,66	**29,86**

Tabela 2. Métricas de complexidade do PortJur comparado a outros corpora

3.3. A anotação do corpus

O PortJur foi pré-anotado automaticamente com a versão mais recente do PortParser [Lopes e Pardo 2024], um *parser* do estado da arte para o português (com valores de acurácia acima de 96%). Em sequência, a anotação automática foi revisada e editada por anotadores humanos altamente especializados em UD. Para acompanhamento da qualidade da anotação, foram utilizados a ferramenta de validação do projeto UD e o sistema Verifica-UD [Lopes et al. 2023].

Notou-se um bom desempenho do *parser* em sentenças curtas e médias, porém, quanto maiores as sentenças, mais erros de anotação ocorreram, especialmente quanto à ancoragem correta dos pares *head*-dependente entre *tokens* distantes entre si. Muitas vezes o *parser* acertava a relação de dependência, dado um dependente, porém errava o *head* correspondente, por este estar muito distante (observou-se uma tendência do *parser* de atribuir um *head* elegível que estivesse mais próximo do dependente). Como exemplo, em "*Incumbe ao poder público* ***garantir****, à gestante e à mulher com filho na primeira infância que se encontrem sob custódia em unidade de privação de liberdade,* ***ambiência*** *que atenda às normas sanitárias e assistenciais do Sistema Único de Saúde para o acolhimento do filho, em articulação com o sistema de ensino competente,* ***visando*** *ao desenvolvimento integral da criança.*", o objeto direto "*ambiência*" do verbo "*garantir*", e a oração subordinada adverbial "*visando a…*", cuja oração matriz é "*atenda*", estão separados de seus *heads* por 26 e 28 *tokens*, respectivamente. Há casos em que essa distância é muito maior e esse fato afeta tanto a capacidade de anotação humana quanto a do *parser*.

3.4. Especificidades linguísticas do corpus

O texto jurídico caracteriza-se por uma sintaxe rica e por uma frequência relativamente alta de estruturas não canônicas do português, como VS (Verbo-Sujeito) ao invés de SV

[1] O índice Flesch de leiturabilidade busca uma correlação entre tamanhos médios de palavras e sentenças.
[2] O índice de leiturabilidade Gunning Fog soma a quantidade média de palavras por sentença ao percentual de palavras difíceis no texto e multiplica tudo por 4.
[3] O índice de leiturabilidade de Dale Chall adaptado combina a quantidade de palavras não familiares com a quantidade média de palavras por sentença.
[4] A NILC-Metrix não conta pontuações como *tokens*, o que explica a diferença entre a média de *tokens* do PortJur nessa tabela e a média de *tokens* mencionada nas características do corpus.

(Sujeito-Verbo), mais comuns no dia-a-dia. Há também construções impessoais, com uso de verbos na voz passiva ou com uso do "*se*" como índice de indeterminação do sujeito, recursos que excluem qualquer traço de subjetividade do texto - o que parece ser necessário para um entendimento inequívoco de leis e códigos e para que as sentenças proferidas tenham um caráter impessoal. No que se refere ao léxico, há uso de palavras e expressões pouco frequentes, além de uso de expressões em latim, o que denota tanto erudição quanto reserva de domínio para autores versados no direito. Algumas dessas características acarretam particularidades na anotação UD, detalhadas a seguir.

3.4.1. Itemização

Os textos do PortJur recorrem muito ao recurso da itemização (apresentação de texto em forma de itens). Os itens de uma mesma categoria são coordenados e apenas o primeiro recebe uma função sintática. Há conjuntos de itens que têm função de sujeito, de objeto, de adjunto adverbial, etc. O token itemizador varia muito: ora são empregados numerais ordinais, ora são empregados numerais romanos, ora são empregadas letras. Seguindo orientação da UD, anotaram-se os itemizadores como dependentes da relação *discourse*, uma vez que não possuem função sintática.

3.4.2. Referências a conteúdos legais

Uma característica dos textos jurídicos é a menção pormenorizada ao conteúdo das leis e, curiosamente, os textos das leis normalmente diferem dos textos do judiciário nesse aspecto. Nos textos das leis, a forma de menção faz uso de preposições e parte do mais específico para o mais geral, usando a preposição "*de*", que caracteriza o caso genitivo: "*alínea f do inciso II do art. 61 do Decreto-Lei nº 2.848*". Esses modificadores nominais introduzidos por "*de*" são anotados com a relação *nmod*. Por outro lado, nos textos do judiciário, a forma de menção normalmente não faz uso de preposições e parte do mais geral para o mais específico: "*Decreto 2.848, art. 61, inciso II, alínea f*". Para anotar a relação do todo para a parte não há consenso na UD sobre a relação mais adequada. Por isso, decidiu-se usar a mesma relação de modificador nominal usada para anotar a relação da parte para o todo, ou seja, *nmod*, porém com o acréscimo de uma sub-relação, *nmod:spec*, para marcar esses modificadores nominais especificadores e distingui-los dos modificadores nominais do caso genitivo (da parte para o todo). As Figuras 2 e 3 ilustram a anotação dessas duas formas de referência a conteúdos legais.

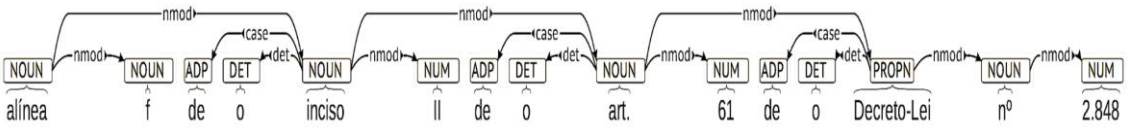

Figura 2. Forma de referenciar conteúdo legal nas próprias leis

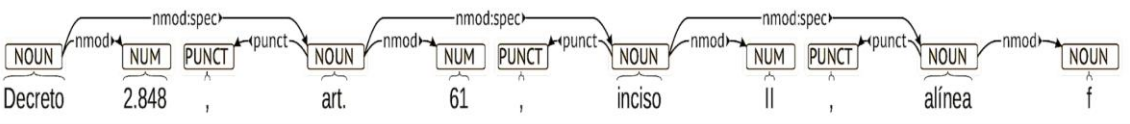

Figura 3. Forma de referenciar conteúdo legal em textos do poder judiciário

3.4.3. Numerais

Os algarismos ordinais e cardinais, à direita de nominais, funcionam como modificadores nominais, têm categoria gramatical NUM e são dependentes da relação *nmod*, por exemplo, "*art. 128*", "*inc. III*", "*Lei nº 6.091/1974*" e "*Recurso Especial 1.750.660/SC*". Algarismos arábicos (cardinais) são quantificadores quando estão à esquerda do nominal e são dependentes da relação *nummod*, por exemplo, "*1 ano de prisão*". Numerais ordinais à esquerda dos nominais têm função adjetiva ("*primeiro*", "*segundo*", etc.), portanto, são classificados como adjetivos (ADJ) e são dependentes da relação *amod* (por exemplo, "*terceira posição*").

3.4.4. Marcadores de condições e exceções

É característica dos textos legais a presença de condições para aplicação de uma lei, bem como de suas exceções. O corpus é muito marcado pelo uso de orações adverbiais reduzidas de particípio com sujeito à direita, com função **condicional** (por exemplo, "*...**observado** o disposto nos §§ 3º e 4º do art. 13*") e **concessiva** ("*A permanência da criança .... não se prolongará por mais de 18 (dezoito meses), **salvo comprovada** necessidade que atenda ao seu superior interesse…"*).

A função de marcar uma exceção pode ser desempenhada por uma palavra única ("*exceto*", "*salvo*", "*menos*", "*excluindo*", etc.) ou por uma locução ("*com exceção de*", "*a não ser*", "*a não ser que*"). No PortJur, há uma preferência pelo uso de "*salvo*" e "*exceto*" como marcadores de exceção. Para simplificar, decidiu-se tratar essas palavras como preposição (ADP), por serem lexicalmente recorrentes e reconhecidas como preposição em vários dicionários e gramáticas, como ilustrado na Figura 4.

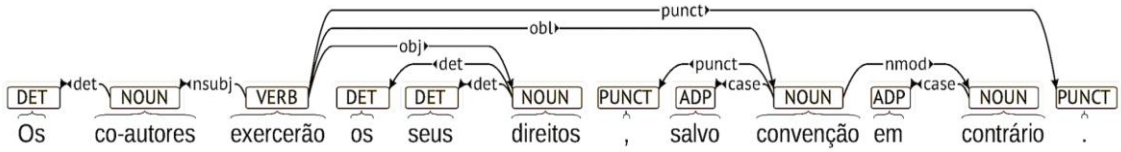

Figura 4. Anotação de "*salvo*" como marcador de exceção

A expressão "*com exceção de*" é tratada como um modificador oblíquo normal (dependente da relação *obl*) e as expressões "*a não ser*", "*a não ser que*" e "*a menos que*" são tratadas como expressões fixas, como ilustrado na Figura 5, na qual "*a não ser que*" é dependente da relação *mark*, usada para conjunções subordinativas.

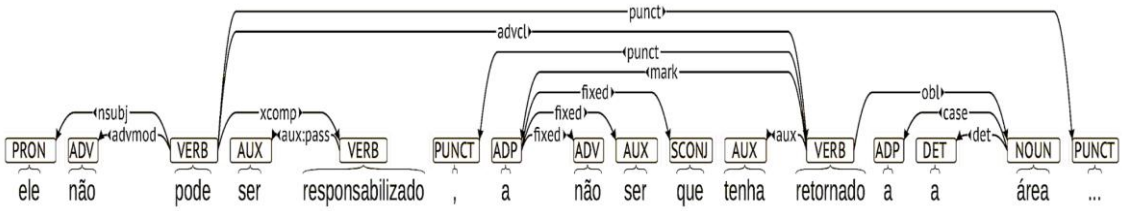

Figura 5. Anotação de "*a não ser que*" como marcador de exceção

3.4.5. Palavras estrangeiras

A ocorrência de muitas expressões estrangeiras, a maioria em latim, cria uma situação peculiar: na UD, palavras estrangeiras recebem a etiqueta X de classe gramatical e não ganham atributos morfológicos, sendo unidas pela relação *flat:foreign* (ou seja, são opacas do ponto de vista de sua sintaxe interna). No entanto, é preciso relacioná-las sintaticamente com o restante da sentença, o que requer que seja considerado seu significado, muitas vezes desconhecido, dificultando a escolha da relação de dependência. Por exemplo, em "*Essas justificativas encerram* ***res inter alios acta*** *em relação ao compromissário adquirente*", a expressão refere-se a um princípio relativo a contratos, e se liga ao predicado "*encerram*" pela relação *obj*. No corpus, as expressões em latim aparecem como dependentes de diferentes relações (*nmod, obl, obj, advmod, nsubj, conj, acl*). Embora frequentes, essas ocorrências tendem a prejudicar o aprendizado computacional de modelos, por sua esparsidade no corpus.

A seguir, os recursos lexicais derivados do corpus anotado são descritos.

4. Recursos lexicais

Além do principal recurso produzido nesse trabalho (o próprio corpus anotado – no difundido formato CoNLL-U), com base no resultado da anotação do corpus PortJur, foram produzidos e organizados diversos recursos lexicais que podem ser utilizados tanto para fins de pesquisa quanto para o desenvolvimento de ferramentas de PLN. Foram criadas e disponibilizadas cinco listas especializadas, descritas a seguir.

- **Lista de Palavras de Conteúdo.** Reúne 4.994 palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) agrupadas por frequência de lemas presentes no PortJur. Pode ser utilizada para subsidiar a análise de utilização de vocabulário técnico e formal, centrado em temas jurídicos fundamentais como "*direito*" (385 ocorrências), "*decisão*" (74), "*processo*" (72) e "*autoridade*" (82), mas também para analisar o vocabulário próprio dos textos do corpus.

- **Lista de Formas Verbais.** Apresenta 3.073 formas verbais utilizadas no corpus, permitindo estudos sobre padrões de uso verbal, aspectos modais e estruturas argumentativas típicas do discurso jurídico. Esta lista permite observar, por exemplo, que as formas verbais no particípio representam a maior parte das ocorrências (2.796 ocorrências de um total de 9.150), ilustrando o caráter objetivo e despersonalizado desse discurso. Também se destaca o uso frequente do infinitivo impessoal (1.783 ocorrências), como em "*considerar*" e "*recorrer*", geralmente associado a construções explicativas ou normativas. Já os verbos no presente do indicativo, como "*prevê*" e "*dispõe*" (1.798 ocorrências), indicam enunciados atemporais ou generalizações, característicos do texto jurídico.

- **Lista de Abreviações.** Contém 22 abreviações comuns neste domínio, como "*art.*" e "*inc.*", respectivamente para "*artigo*" e "*inciso*". Essas formas reduzidas podem ser relevantes para tarefas de normalização, tradução automática e simplificação dos textos.

- **Lista de Expressões Estrangeiras.** Inclui 50 expressões estrangeiras, em sua maioria em latim, como "*habeas corpus*" e "*fumus boni iuris*", que são frequentemente empregadas em decisões e pareceres jurídicos e que marcam

erudição e autoridade. Essa lista pode ser útil para a elaboração de glossários, de terminologias especializadas e para análise semântica.

- **Lista de Nomes Próprios**. Inclui em um almanaque os 627 nomes próprios identificados no corpus, abrangendo pessoas, instituições, localidades e entidades jurídicas. É especialmente útil para tarefas como reconhecimento de entidades nomeadas e desambiguação de referências em textos jurídicos.

A disponibilização desses recursos visa ampliar as possibilidades de investigação linguística no campo jurídico, promover a reprodutibilidade de estudos e fomentar o desenvolvimento de aplicações computacionais que lidem com a complexidade textual do português jurídico.

5. Considerações finais

Este artigo apresenta um recurso inédito para o domínio jurídico em língua portuguesa: o corpus PortJur, anotado sintaticamente segundo a abordagem UD. Esse recurso poderá ser utilizado tanto para treinar sistemas para o gênero jurídico quanto para ser somado a outros corpora para desenvolvimento/treinamento de outras ferramentas e aplicações multigênero. Além disso, somam-se às contribuições desse trabalho os recursos lexicais derivados do corpus anotado e a caracterização sintática necessária para análise textual, resultando em contribuições práticas e teóricas para a área de PLN e de Linguística. Todos os recursos produzidos neste trabalho estão publicamente disponíveis para os interessados no portal web do projeto POeTiSA[5], iniciativa a qual este trabalho se vincula. O PortJur, em especial, deve integrar o *treebank* Porttinari [Pardo et al. 2021].

É interessante observar que, provavelmente devido aos desafios dos textos jurídicos, há relativamente poucos corpora jurídicos conhecidos anotados segundo a UD. Por exemplo, há o corpus CLTT (para o tcheco) [Kríž e Hladká 2018] com 1.121 sentenças jurídicas e alguns corpora parcialmente compostos por sentenças jurídicas, como o GUM (inglês) [Zeldes 2017], o TDT (finlandês) [Haverinen et al. 2014] e o ParTut (inglês, italiano e francês) [Sanguinetti e Bosco 2015]. Pelo que se tem conhecimento, o PortJur é o primeiro corpus do tipo para o português.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Também se agradece ao INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação) (processo CNPq/SECTICS/CAPES/FAPs #408490/2024-1).

[5] https://sites.google.com/icmc.usp.br/poetisa

Referências Bibliográficas

Albuquerque, H. O.; Costa, R.; Silvestre, G.; Souza, E.; Silva, N. F. F.; Vitório, D.; Moriyama, G.; Martins, L.; Soezima, L.; Nunes, A.; Siqueira, F.; Tarrega, J. P.; Beinotti, J. V.; Dias, M.; Silva, M.; Gardini, M.; Silva, V.; Carvalho, A. C. P. L. F.; Oliveira, Adriano L. I. (2022). UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Proceedings of the 15th Computational Processing of the Portuguese Language*, pp. 3-14.

Branco, A.; Silva, J. R.; Gomes, L.; Rodrigues, J. A. (2022). Universal Grammatical Dependencies for Portuguese with CINTIL Data, LX Processing and CLARIN support. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pp. 5617–5626.

Brito, M.; Pinheiro, V.; Furtado, V.; Monteiro Neto, J.; Bomfim, F.; Costa, A.; Silveira, R. (2023). CDJUR-BR - Uma Coleção Dourada do Judiciário Brasileiro com Entidades Nomeadas Refinadas. In *Proceedings of the 14th Symposium in Information and Human Language Technology*, pp. 177-186.

Castro, M.; Neves, A. R. (2024). PLN e Segurança Jurídica: Identificação de divergências jurisprudenciais com Processamento de Linguagem Natural. In *Proceedings of the 15th Symposium in Information and Human Language Technology*, pp. 451-456.

de Marneffe, M. C.; Manning, C. D.; Nivre, J.; Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, v. 47, n. 2, pp. 255-308.

Di Felippo, A.; Nunes, M. G. V.; Barbosa, B. (2024). A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach. In *Proceedings of the 15th Symposium in Information and Human Language Technology*, pp. 192-201.

Duran, M. S.; Lopes, L.; Nunes, M. G. V.; Pardo, T. A. S. (2023). The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In *Proceedings of the 14th Symposium in Information and Human Language Technology*, pp. 115-124.

Fama, I.; Bueno, B.; Alcoforado, A.; Ferraz, T.; Moya, A.; Costa, A. (2024). No Argument Left Behind: Overlapping Chunks for Faster Processing of Arbitrarily Long Legal Texts. In *Proceedings of the 15th Symposium in Information and Human Language Technology*, pp. 129-138.

Feijó, D. V.; Moreira, V. P. (2018). RulingBR: A summarization dataset for legal texts. In *Proceedings of the 13th Computational Processing of the Portuguese Language*, pp. 255-264.

Ferrari, L. A.; Marques, C. G. F. (2022). O LEX-BR-Ius: arquitetura e decisões na compilação de um corpus representativo das leis federais brasileiras. *ANTARES: Letras e Humanidades*, v. 14, n. 34, pp. 40-77.

Ferrari, L. A.; Cunha, E. L. T. P. (2022). Reflexões metodológicas sobre datasets e linguística de corpus: uma análise preliminar de dados legislativos. *Domínios de Lingu@gem,* v. 16, n. 4, pp. 1571-1607.

Garcia, E.; Silva, N.; Siqueira, F.; Gomes, J.; Albuquerque, H. O.; Souza, E.; Lima, E.; Carvalho, A. (2024). RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pp. 374-383.

Haverinen, K.; Nyblom, J.; Viljanen, T.; Laippala, V.; Kohonen, S.; Missilä, A.; Ojala, S.; Salakoski, T.; Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, v. 48, pp. 493-531.

Kríž, V.; Hladká, B. (2018). Czech Legal Text Treebank 2.0. In *Proceedings of the 11th International Conference on Language Resources and Evaluation*, pp. 4501-4505.

Leal S. E.; Duran, M. S.; Scarton, C. S.; Hartmann, N. S.; Aluísio, S. M. (2023). NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, v. 58, pp. 73-110.

Lins, A. A.; Carvalho, C. S.; Bomfim, F. C. J.; Bentes, D. C.; Pinheiro, V. (2024). CLSJUR.BR - A Model for Abstractive Summarization of Legal Documents in Portuguese Language based on Contrastive Learning. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pp. 321-331.

Lopes, L.; Duran, M. S.; Pardo, T. A. S. (2023). Verifica-UD: a Verifier for Universal Dependencies Annotation for Portuguese. In *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pp. 451-460.

Lopes, L.; Pardo, T. A. S. (2024). Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pp. 401-410.

McDonald, R.; Nivre, J.; Quirmbach-Brundage, Y.; Goldberg, Y.; Das, D.; Ganchev, K.; Hall, K.; Petrov, S.; Zhang, H.; Täckström, O.; Bedini, C.; Castelló, N. B.; Lee , J. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pp. 92–97.

Pardo, T. A. S.; Duran, M. S.; Lopes, L.; Di Felippo, A.; Roman, N. T.; Nunes, M. G. V. (2021). Porttinari - a large multi-genre treebank for brazilian portuguese. In *Proceedings of the XIII Symposium in Information and Human Language*, pp. 1-10.

Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017). Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pp. 197-206.

Sanguinetti, M.; Bosco, C. (2015). PartTUT: The Turin University Parallel Treebank. In Basili, R., Bosco, C., Delmonte, R., Moschitti, A., Simi, M. (eds), *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*. Studies in Computational Intelligence, v. 589. Springer, Cham.

Siqueira, F. A.; Vitório, D.; Souza, E.; Santos, J. A. P.; Albuquerque, H. O.; Dias, M. S.; Silva, N. F. F.; Carvalho, A. C. P. L. F.; Oliveira, A. L. I.; Bastos-Filho, C. (2024). Ulysses Tesemõ: a new large corpus for Brazilian legal and governmental domain. *Language Resources and Evaluation*, pp. 1-20.

Souza, E.; Silveira, A.; Cavalcanti, T.; Castro, M. C.; Freitas, C. (2021). PetroGold–Corpus padrão ouro para o domínio do petróleo. In *Proceedings of the 13th Symposium in Information and Human Language Technology*, pp. 29-38.

Souza, E.; Albuquerque, H. O.; Silva, N. F. F.; Cerqueira, M.; Carvalho, A. C. P. L. F.; Oliveira, A. L. I. (2024). PLN no Direito - REN: Reconhecimento de Entidades Nomeadas no Domínio Legal: um Panorama para a Língua Portuguesa. In Caseli, H. M.

and Nunes, M. G. V. (eds), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*. 3a edição BPLN.

Sturzeneker, M. L.; Morales, M. C. R.; Rocha, M. L. S. J.; Finger, M.; Sousa, M. C. P.; Monte, V. M.; Namiuti, C. (2022). Carolina´s Methodology: building a large corpus with provenance and typology information. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing*, v. 3128, p. 53-58.

Xavier, R. C. (2002). *Português no Direito: Linguagem Forense*. Rio de Janeiro: Forense.

Zeldes, A. (2017). The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, v. 51, n. 3, pp. 581-612.