DOI: 10.14393/DL52-v16n4a2022-13

Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo *Universal Dependencies* em língua portuguesa

An annotation manual as a Natural Language Processing resource: the Universal Dependencies model in Portuguese

Magali Sanches DURAN[*]
Maria das Graças Volpe NUNES[**]
Lucelene LOPES[***]
Thiago Alexandre Salgueiro PARDO[****]

RESUMO: Com o avanço da área de Processamento de Linguagem Natural (PLN), *corpora* são recursos que têm tido um lugar de destaque. Mais do que subsidiar estudos linguísticos, eles constituem as bases para o treinamento de modelos de Aprendizagem de Máquina e para o desenvolvimento de aplicações computacionais de ponta. Particularmente, há grande necessidade de *corpora* anotados, porém sua geração requer outro recurso essencial, o manual de anotação, que instancia o modelo de anotação de interesse para a língua em questão e delineia as decisões de anotação que devem ser adotadas. Neste artigo,

ABSTRACT: With the advances of the Natural Language Processing area, corpora are resources that have had a prominent place. More than subsidizing linguistic studies, they constitute the basis for training Machine Learning models and developing cutting-edge computational applications. In particular, there is a great need for annotated corpora, but their production requires another essential resource, the annotation manual, which instantiates the annotation model of interest for the language in question and outlines the annotation decisions that should be adopted. In this paper, we explore issues related to the

[*] Doutora em Estudos Linguísticos pela UNESP de São José do Rio Preto e pesquisadora de pós-doutorado no NILC. ORCID: https://orcid.org/0000-0002-3843-4600. magali.duran@uol.com.br

[**] Professora Doutora do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, no campus de São Carlos. ORCID: https://orcid.org/0000-0002-2776-6140. gracan@icmc.usp.br

[***] Doutora em Ciência da Computação pela PUC Rio Grande do Sul e pesquisadora de pós-doutorado no NILC. ORCID: https://orcid.org/0000-0003-0314-140X. lucelene@gmail.com

[****] Professor Doutor do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, no campus de São Carlos. ORCID: https://orcid.org/0000-0003-2111-1319. taspardo@icmc.usp.br

exploramos questões relacionadas ao desenvolvimento de manuais para a anotação de *corpus* em português brasileiro segundo o modelo internacional *Universal Dependencies*, amplamente adotado na área. Partimos da discussão da evolução do PLN e o uso de *corpora*, passamos pelas questões, recursos e ferramentas fundamentais relacionados à representação sintática, discutimos o modelo *Universal Dependencies* e apresentamos as principais decisões tomadas na instanciação de suas diretrizes no português brasileiro. Por questões práticas e de didática, dividimos o manual em duas partes: o Manual de Anotação de *PoS tags* (anotação morfossintática) e o Manual de Anotação Relações de Dependência. Ambos foram resultado do processo relatado neste artigo e estão disponíveis para livre acesso no site do projeto POeTiSA na Web.

PALAVRAS-CHAVE: *Corpora* anotados. Manual de anotação. *Universal Dependencies*. Árvores de dependência. Português brasileiro.

development of manuals for the annotation of Brazilian Portuguese corpora according to the Universal Dependencies model, widely adopted in the field. We discuss the evolution of NLP and the use of corpora, the fundamental issues, resources and tools related to syntactic representation, the Universal Dependencies model, and the main decisions made in the instantiation of UD guidelines in Brazilian Portuguese. For practical and didactic reasons, we divided the manual into two parts: the PoS Tag Annotation Manual (morphosyntactic annotation) and the Dependency Relations Annotation Manual. Both resulted from the process reported in this paper and are available for free access on the POeTiSA project's Web site.

KEYWORDS: Annotated corpora. Annotation manual. Universal Dependencies. Dependency trees. Brazilian Portuguese.

1 Introdução

Neste artigo, apresentamos o manual de anotação como recurso de Processamento de Linguagem Natural (PLN), o que não é uma ideia muito difundida nem nos círculos de linguistas nem nos círculos de cientistas da computação. Para fundamentar tal afirmação, dividimos esta introdução em três subseções, iniciando com uma retrospectiva do papel do linguista ao longo da evolução do PLN (1.1), passando pela ascensão dos *corpora* anotados como recurso valioso para os novos métodos de PLN (1.2) e terminando por discorrer sobre a importância dos manuais de anotação como parte indissociável de um esquema de anotação (1.3).

1.1 Processamento de Linguagem Natural: a evolução rumo aos *corpora*

O PLN desenvolveu-se muito nas últimas décadas e o papel do linguista nos projetos de PLN também se modificou. Até a década de 1990, acreditava-se que a representação do conhecimento linguístico na forma de léxicos, ontologias, gramáticas, regras e outros formalismos de representação de conhecimento seria capaz de fornecer todo o conhecimento necessário para o PLN.

A partir dos anos 1990, ocorreram alguns avanços tecnológicos e mudanças de comportamento social que revolucionaram toda essa área. O advento da internet e de sua interface fez aumentar exponencialmente a quantidade de documentos em línguas naturais na Web e, consequentemente, aumentou a demanda por tarefas de processamento: de busca, de classificação, de auxílio à produção, de análise e correção, de consulta e de sumarização, entre muitas outras. Os avanços em matéria de *hardware* aumentaram de forma gigantesca a capacidade de processamento e de armazenagem de dados, permitindo o emprego de métodos sofisticados, antes proibitivos por consumirem muito tempo e espaço de memória. E, por fim, mais recentemente, o surgimento das redes sociais aumentou o conteúdo gerado por usuários, exigindo que o PLN passasse a lidar com textos que fogem em muito do registro formal da língua.

Essas mudanças de cenário, de uma forma ou de outra, cada uma com sua influência, trouxeram o PLN para a era da Aprendizagem de Máquina (AM) (MITCHELL, 1997). Com a AM, é possível que a máquina aprenda tarefas baseadas em conhecimento humano sem que seja necessário que o humano represente e estruture formal e explicitamente esse conhecimento. Isso é vantajoso principalmente no caso de problemas cuja solução requer conhecimento cuja formalização é difícil ou altamente custosa. O importante para a AM é ter um conjunto suficientemente grande e representativo de exemplos sobre o qual os algoritmos possam aprender. No caso de PLN, esses exemplos constituem os *corpora* de texto ou de fala transcrita, que,

alinhando-se à tradição empirista de pesquisa, representam um recorte de interesse da língua (MANNING; SCHÜTZE, 1999; SARDINHA, 2000).

A introdução da AM no PLN foi revolucionária e inaugurou um ritmo acelerado de novas conquistas. Um dos mais modernos modelos de AM, as Redes Neurais Profundas (*Deep Learning*) (GOODFELLOW; BENGIO; COURVILLE, 2016) alavancaram o desempenho dos tradutores automáticos - a primeira grande tarefa de PLN - e têm sido usadas, com excelentes resultados, nos mais variados cenários: *chatbots*, reconhecimento e geração de fala, geração de documentos, análise de sentimentos, detecção de discurso de ódio e *fake news*, entre muitos outros.

Nesse atual modo de fazer PLN, o papel do linguista passa a incluir o de subsidiar o projeto do *corpus*, o qual fornecerá o conhecimento aos algoritmos de aprendizagem. Há um compromisso entre o *corpus* fornecido, o algoritmo de AM utilizado e os resultados obtidos. Assim, quanto mais "informados" forem os exemplos fornecidos, melhores serão os resultados alcançados. Diversos atributos podem ser anotados em textos, tanto associados a análises linguísticas (como os atributos morfológicos, sintáticos e semânticos), quanto associados a outros tipos de análise, como análise de sentimentos e de opiniões.

Se, pelo lado da computação, o PLN evolui por meio da melhoria dos métodos de aprendizagem automática, do lado da linguística o PLN evolui por meio da produção de *corpora* anotados de forma cada vez mais lógica, completa e consistente, guiando e solidificando modelos e teorias linguísticas.

1.2 *Corpus* no coração do PLN

A anotação de *corpus* é uma ciência (IDE, 2017). Ela começa pela modelagem do fenômeno a ser anotado, que culmina na definição de um esquema de anotação, com um conjunto de etiquetas a serem atribuídas aos segmentos de texto. A modelagem de

fenômenos para anotação é um trabalho árduo e, por isso, é comum haver reaproveitamento de esquemas de anotação bem sucedidos na comunidade de PLN.

As línguas que possuem recursos mais avançados de PLN criam modelos de anotação e esses costumam ser replicados e adaptados por outras línguas que têm menos recursos humanos dedicados à criação de modelos. Foi assim que proliferaram, por exemplo, clones e adaptações da Framenet (BAKER; FILLMORE; LOWE, 1998) e do Propbank (PALMER; GILDEA; KINGSBURY, 2005), que modelaram, de formas diferentes, a teoria de papéis semânticos de Fillmore (1968), e da Verbnet (KIPPER-SCHULER, 2005), que modelou a teoria de classes verbais de Levin (1993).

Embora em um dado momento da história o reaproveitamento de esquemas tinha só o objetivo de poupar esforço de modelagem, não se demorou a perceber o grande potencial que a utilização de um mesmo esquema de anotação poderia fornecer para o PLN, no sentido de facilitar a comparação entre as línguas e permitir o desenvolvimento de ferramentas e aplicações de PLN multilíngues.

Em 2006, um desafio de PLN para criar um *parser* de dependências multilíngue (BUCHHOLZ; MARSI, 2006) exigiu que os organizadores adotassem um esquema de anotação de *corpus* razoavelmente flexível para anotar as treze línguas envolvidas da competição. Devido a sua simplicidade, o esquema escolhido foi o apresentado em Nivre *et al*. (2006). Os resultados no treinamento de *parsers* multilíngues a partir desses *corpora* foram promissores e, em 2015, um esquema de anotação "universal" de dependências sintáticas já estava sendo utilizado por várias línguas. Esse esquema, chamado de *Universal Dependencies* (UD) (NIVRE *et al*., 2015, 2020; MARNEFFE *et al*., 2021) pretende ser independente de língua e tem ganhado adeptos em diversas partes do mundo. Mais de cem línguas, das mais diferentes famílias, já possuem *corpora* anotados no esquema UD.

A disponibilidade de *corpora* anotados em diversas línguas seguindo um mesmo esquema de anotação abriu oportunidade para o desenvolvimento de vários

parsers[1] cujos métodos são independentes de língua. Como exemplos, podemos citar o UDPipe (STRAKA; STRAKOVA, 2017), o UDify (KONDRATYUK; STRAKA, 2019), o Stanza (QI et al., 2020) e o *parser* do spaCy (HONNIBAL; JOHNSON, 2015). Esses *parsers*, por terem uma capacidade muito boa de aprendizado[2], mesmo a partir de poucos dados, muitas vezes são usados para pré-anotar novos *corpora* no formato UD, que são revisados por humanos. Assim, num ciclo virtuoso, com mais dados para retreinamento, os *parsers* podem ser continuamente aprimorados.

O reaproveitamento de um esquema de anotação não se dá, no entanto, sem esforço. Pustejovsky, Bunt e Zaene (2017) distinguem o que chamam de *markables*[3] dos *non-markables* em um esquema de anotação. Os *markables* estão diretamente associados aos conjuntos de etiquetas (elementos explícitos na anotação) e os *non-markables* são as partes de um esquema de anotação expressas por meio de diretrizes que devem ser seguidas a fim de se alcançar consistência na anotação (elementos implícitos na anotação). Os *markables* não podem ser alterados, pois garantem a comparabilidade física entre as línguas que o utilizam. Sem os *non-markables*, contudo, corre-se o risco de atribuir etiquetas iguais a fenômenos diferentes e isso também compromete a comparabilidade entre as línguas.

Os *non-markables* são expressos por meio de diretrizes de anotação e, como as diretrizes fazem uso de muitos exemplos, elas normalmente são dependentes de língua. Assim, seja para reaproveitar um esquema de anotação feito para uma outra língua, seja para utilizar um esquema "universal", é essencial instanciar as instruções de anotação na língua em que se vai aplicar o esquema.

---

[1] Anotadores automáticos de funções sintáticas.

[2] O desempenho de um *parser* é, em grande parte, função da qualidade e quantidade de dados usados para seu treinamento. Por isso, um mesmo *parser* multilíngue pode apresentar desempenhos diferentes em línguas diferentes, sendo relativamente melhor em línguas que disponibilizam *corpora* maiores e mais consistentemente anotados.

[3] Pustejovsky et al. (2017) dizem que os *markables* são a parte do modelo que consome etiquetas (*consuming tags*) e os *non-markables* são a parte do modelo que não consome etiquetas (*non-consuming tags*).

1.3 Recursos essenciais aos *corpora*: diretrizes e manuais

As diretrizes em uma nova língua têm por objetivo tornar claro para os anotadores como o conjunto de etiquetas deve ser utilizado, com rica exemplificação que contemple desde casos comuns e frequentes até casos mais raros e difíceis de anotar.

As diretrizes da UD[4], por não serem específicas de língua, deixam grandes lacunas a serem preenchidas pelos linguistas que as instanciam numa língua. Um fórum mantido pela UD e disponível no *github*[5] mostra o quanto essa instanciação é desafiadora, pois centenas de tópicos são discutidos e diversas opiniões concorrem para solucionar as dúvidas.

Para o português, já foram disponibilizados no site da UD três *corpora*[6]: o PUD, o GSD e o Bosque-UD, este último descrito em Rademaker *et al*. (2017). É desejável que novos *corpora*, maiores e de novos domínios, sejam anotados seguindo o mesmo modelo, pois isso vai contribuir para a melhoria do desempenho dos *parsers* de português. Contudo, a falta de um manual de anotação dificulta as iniciativas de produção de novos *corpora* de UD em português, exigindo que cada novo projeto tenha que enfrentar o desafio de instanciar as diretrizes da UD para a língua portuguesa. Além disso, se cada projeto adota decisões diferentes para lidar com desafios iguais na língua, os *corpora* anotados deixam de ser comparáveis e isso pode impedir que sejam "somados" no esforço de treinar *parsers* mais robustos.

Foi essa dificuldade que enfrentamos ao iniciarmos um projeto de anotação de *corpus* seguindo o modelo UD: para guiar a anotação, precisávamos de um manual que contivesse explicações e exemplos de cada uma das etiquetas dos conjuntos de

[4] https://universaldependencies.org/guidelines.html
[5] https://github.com/universaldependencies/docs/issues
[6] https://universaldependencies.org/#download

etiquetas da UD, bem como discussões sobre como resolver ambiguidades que frequentemente geram dúvidas.

O relatório produzido após a anotação do *corpus* Bosque-UD (SOUZA *et al*., 2020) foi de grande auxílio para iniciarmos nosso trabalho, pois reporta diversos dos problemas enfrentados durante a anotação do esquema UD em língua portuguesa. Contudo, tal relatório não parece ter a pretensão de servir como um manual de anotação UD, pois não é organizado de forma a contemplar de forma sistemática cada uma das etiquetas dos conjuntos de etiquetas da UD.

Diante dessa lacuna, empreendemos a construção de um manual de anotação de *corpus* em UD, no qual estão instanciadas diretrizes da UD, englobando decisões de projeto e estudos acerca de palavras altamente ambíguas que, no decorrer do processo de anotação, se mostraram foco de dúvidas e discordâncias entre anotadores. Neste artigo, exploramos detalhadamente as questões relacionadas ao desenvolvimento desse manual.

O artigo está organizado em seis seções além desta introdução. Na Seção 2, fazemos uma breve revisão sobre a anotação sintática de *corpus* de língua portuguesa e sua relação com o desenvolvimento de *parsers* do português. Dedicamos a Seção 3 à discussão sobre os motivos que levaram a comunidade de PLN a consagrar a teoria de Tesnière (1959, 2015) como vantajosa para a anotação sintática. Na Seção 4, apresentamos o esquema de anotação da UD. Na Seção 5, apresentamos os conjuntos de etiquetas morfossintáticas (5.1) e de relações sintáticas (5.2) da UD, relacionando-as às classes morfossintáticas e funções sintáticas das gramáticas normativas do português. Na Seção 6, discutimos quais das diretrizes da UD são facilmente adotáveis no português, quais são problemáticas e exigiram tomada de decisão no projeto com base em fundamentos linguísticos e computacionais, e quais exigiram tomada de decisões arbitrárias, apenas para manter consistência na anotação, mas que poderão

ser alteradas no futuro caso uma solução melhor se apresente. Na Seção 7, tecemos considerações finais e discutimos trabalhos futuros.

2 *Parsers* e *treebanks* de português

Durante bastante tempo foi possível construir sistemas de PLN que prescindiam da análise sintática. Quase todo conhecimento era provido pelos itens lexicais e suas posições relativas (mesmo que, indiretamente, isso espelhasse a sintaxe). Para muitas aplicações mais simples e limitadas, como os primeiros corretores ortográficos, isso foi suficiente. Para outras aplicações mais avançadas, especialmente as que envolvem qualquer nível de interpretação do conteúdo, como, por exemplo, sistemas de pergunta e resposta, a análise sintática mostrou-se fundamental. À medida que se avançava nas tarefas de PLN, a análise sintática foi se tornando um ponto crítico, pois de sua qualidade dependia toda uma série de processamentos subsequentes.

Nos anos de 1990, os *parsers* estatísticos marcaram o início de uma nova era: a da geração de *parsers* por meio de aprendizagem automática. Para treinar esses *parsers*, eram necessários *corpora* anotados sintaticamente, de preferência revisados por humanos, pois isso garantiria que o aprendizado não se daria sobre desvios de anotação. A fim de facilitar o aprendizado automático, praticamente toda anotação adotava o formato de árvores, seja de constituintes, seja de dependências, e é por isso que os *corpora* anotados sintaticamente são chamados *treebanks*.

A estratégia era a seguinte: os *corpora* eram anotados com os *parsers* existentes, baseados em regras; em seguida, eram revisados por humanos e disponibilizados como exemplo para o aprendizado automático.

O primeiro *parser* para português largamente conhecido e utilizado em PLN tanto no Brasil quanto em Portugal foi o Palavras (BICK, 2000), e sua primeira versão comercial é de meados dos anos 1990. Ele é constituído por centenas de regras escritas

pelo autor e que seguem o paradigma de *Constraint Grammar* (KARLSSON, 1990). Inicialmente, o Palavras só gerava saída de árvores de dependências, mas a partir dos anos 2000 passou também a gerar uma saída de árvores de constituintes. E foi essa saída de constituintes a escolhida para pré-anotar o *corpus* Floresta Sintá(c)tica (AFONSO *et al*., 2002), primeira iniciativa bem-sucedida de gerar um *treebank* de língua portuguesa.

Na ocasião de seu lançamento, apenas 10% do *corpus* Floresta Sintá(c)tica havia sido revisado, porém todos os desvios encontrados na porção revisada serviram para melhorar as regras do *parser* Palavras. Como a parte não revisada do *corpus* Floresta foi reanotada com a versão aperfeiçoada do Palavras, pode-se dizer que a revisão beneficiou todo o *treebank*. Uma vez que o Floresta Sintá(c)tica possui partes totalmente revisadas, partes parcialmente revisadas e partes não revisadas, essas partes passaram a ter nomes distintos[7].

O Bosque é a parte do *corpus* Floresta Sintá(c)tica que foi totalmente revisada. Ele possui 9.368 sentenças, extraídas de *corpora* jornalísticos, das quais cerca de metade são de português brasileiro (CetenFolha) e metade do português europeu (Cetempúblico) e até hoje é o único *treebank* revisado que contém português brasileiro. O Selva, que teve suas árvores parcialmente revisadas, possui cerca de 30.000 sentenças e contém amostras de língua falada, bem como de textos literários e científicos, separadamente. O Floresta Virgem possui cerca de 96 mil sentenças não revisadas. E por fim, o Amazônia, anexado posteriormente, possui cerca de 275 mil sentenças não revisadas extraídas de *blogs* e, portanto, é uma porção do *treebank* que destoa das demais por ser constituída de conteúdos produzidos por usuários da Web e apresentar um registro de língua coloquial (FREITAS; ROCHA; BICK, 2008).

Em 2016, o Palavras passou a gerar também uma saída na forma de dependências em formato UD (BICK, 2016). Esse conversor de formato foi usado sobre

[7] https://www.linguateca.pt/Floresta/*corpus*.html

o *corpus* Bosque e o resultado foi a base para a revisão manual que culminou na produção do *corpus* Bosque-UD (RADEMAKER *et al*., 2017). O *corpus* Bosque-UD, disponibilizado no site da UD[8], passou a ser utilizado para treinar *parsers* de dependência que utilizam técnicas independentes de língua[9], que já vinham apresentando alta capacidade de aprendizado em outras línguas (como os já citados UDPipe, UDify, SpaCy e Stanza). A vantagem desses *parsers* é que são livres (enquanto o Palavras não é) e, por isso, podem ser retreinados por outras equipes, com novos *corpora* e em diferentes gêneros.

Outros *parsers* e *treebanks* surgiram ao longo dos últimos anos, principalmente por iniciativa da comunidade linguística portuguesa, como o LX-Parser (SILVA *et al*., 2010) e o treebank Cintil, mas o Palavras continuou sendo muito utilizado. Seu papel no processamento do português é inegável e, graças a ele, pôde-se fazer a transição entre um *parser* baseado em regras para os *parsers* baseados em AM. Da mesma forma, é inegável o papel do *corpus* Bosque, primeiro *treebank* do português cujas árvores sintáticas foram totalmente revisadas por humanos e, por isso, ideal para ser usado para aprendizado automático.

Embora no site da UD hoje existam três *corpora* já disponibilizados em formato UD, apenas o Bosque é constituído de sentenças originalmente produzidas em português. O *corpus* PUD tem apenas 1.000 sentenças e é uma versão em português do *corpus* paralelo produzido para a competição CoNLL 2017 de *parsers* multilíngues[10]. O *corpus* GSD tem 12.078 sentenças da variante brasileira do português, resultado da conversão para UD da anotação do *corpus* do Google, um *corpus* paralelo de traduções para várias línguas.

---

[8] https://lindat.mff.cuni.cz/services/udpipe/

[9] Esses *parsers* são dependentes de modelo: eles só aprendem a partir de *corpus* anotado no modelo UD, e desde que uma língua tenha um *corpus* anotado no modelo UD, eles podem ser treinados para se tornarem um parser para essa língua.

[10] *CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies* (http://universaldependencies.org/conll17/).

Quando comparamos o tamanho e a variedade de gêneros dos *treebanks* que temos com os dos *treebanks* disponibilizados nas línguas mais bem providas de recursos de PLN, vemos que há muito a ser feito no português e, especialmente, no português do Brasil.

3 O PLN e a estrutura sintática de dependências

As árvores sintáticas de dependência concorreram alternativamente com as árvores de constituintes durante muito tempo no PLN. No entanto, quando passou-se a utilizar árvores sintáticas para comparar línguas automaticamente, percebeu-se que as árvores de dependência eram mais adequadas. Isso se deve principalmente ao fato de que a estrutura de dependências permite anotar relações sintáticas entre as palavras independentemente da posição em que elas ocorrem na oração, ao passo que a estrutura de constituintes é mais rígida quanto à ordem de realização de seus constituintes.

Essa característica da estrutura de dependências é fundamental quando o objetivo é relacionar línguas posicionais (línguas como o Português, em que a ordem dos constituintes é relevante, pois o papel sintático está em grande parte atrelado à posição das palavras na oração) a línguas desinenciais (línguas como o Latim, em que a ordem dos constituintes da oração é livre, pois o papel sintático das palavras é marcado por desinências).

Além disso, em termos de técnicas de AM, a estrutura de dependências permite um aprendizado mais lexicalizado (por ser apoiado em tokens e não em sintagmas), o que tem se mostrado vantajoso em várias aplicações de PLN.

Por esses motivos, as gramáticas de dependência (inspiradas na proposta de Tesnière (1959)), passaram a ser revalorizadas como um modelo promissor para o PLN.

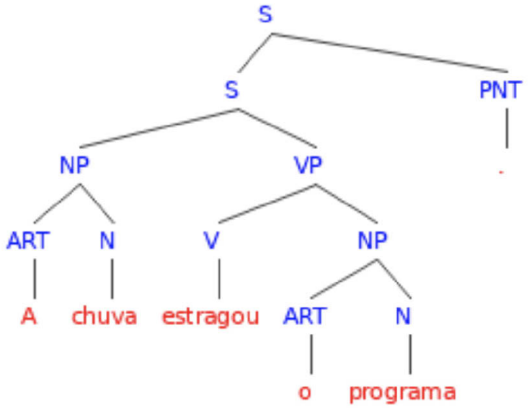
Tesnière (1959) rejeitava a divisão dualista da estrutura sintática em sujeito e predicado. Para ele, o verbo é a raiz de toda estrutura sintática e o sujeito e o objeto são seus subordinados. Ele percebeu que uma sentença não era composta apenas das palavras que a integravam (elementos explícitos), mas também de relações implícitas entre essas palavras, relações que tinham uma hierarquia (um governante e um subordinado) e uma direção (que marca qual elemento está governando a relação). Para explicitar essas relações de dependência, Tesnière (1959) introduziu a ideia de representar a sintaxe por meio de uma árvore, que ele chamava de *stemma*, na qual estão representados todos os elementos de uma sentença e todas as dependências que ligam esses elementos em uma única estrutura.

As árvores de dependência são o cerne do esquema de anotação da UD, cujo objetivo é, em última instância, subsidiar a anotação de *corpora* que serão insumo para desenvolver aplicações de PLN multilíngues. A iniciativa UD motivou, assim, a disseminação tardia do trabalho de Tesnière em língua inglesa (TESNIÈRE[11], 2015).

Para fins de ilustração, apresentamos um exemplo de representação da sintaxe por meio de árvore de constituintes (Figura 1) e um exemplo de representação por árvores de dependências (Figura 2). Contrastando as duas figuras, é possível perceber que a segunda é mais simples, pois contém menos nós no total, relacionando mais diretamente os tokens.

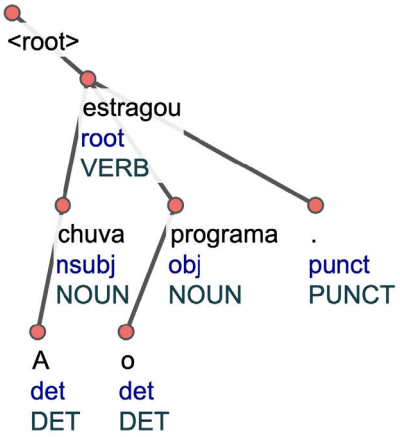
[11] A tradução de 2015 para o inglês é comentada à luz do emprego atual da teoria.

Figura 1 — Análise de "A chuva estragou o programa" sob forma de árvore de constituintes.



Fonte: árvore sintática produzida pelo parser LX-Parser[12].

Figura 2 — Análise de "A chuva estragou o programa" sob forma de árvore de dependências.



Fonte: árvore sintática em formato UD produzida no software UDPipe LINDAT/CLARIN[13].

4 O modelo UD

O modelo de anotação UD é lexicalista, o que significa que todas as etiquetas são associadas diretamente aos *tokens*. Em uma relação de dependência, apenas um *token* é o *head* (governante) e apenas um *token* é o dependente. Todos os *tokens*, sem exceção, recebem uma etiqueta morfossintática e participam de pelo menos uma relação de dependência. A UD define que apenas as palavras de conteúdo (substantivos, pronomes substantivos, verbos, adjetivos, advérbios e numerais)

[12] https://portulanclarin.net/workbench/lx-parser/
[13] http://lindat.mff.cuni.cz/services/udpipe/run.php

podem ser *head* de relações (e podem ser dependentes de relações também). Já as palavras funcionais (como auxiliares, preposições, determinantes e conjunções), símbolos e sinais de pontuação só podem ser dependentes de relações.

O *corpus* a ser anotado deve ser previamente processado para: 1) fazer a separação do texto em sentenças (pois cada sentença é anotada individualmente) e 2) fazer a tokenização, que inclui separar os sinais de pontuação colados às palavras (exceto: pontos das abreviações, pontos separadores das milhares, vírgulas das casas decimais dos numerais e hifens das palavras compostas), separar os verbos dos pronomes enclíticos e mesoclíticos, eliminando os hifens que os juntavam e separar os *tokens* que constituem contrações na língua (ex: deste = de este, comigo = com mim, naquela = em aquela). As palavras compostas (unidas por hífen) são aceitas normalmente como *tokens* da língua, mas *tokens* que integram multipalavras, como "panela de pressão" e "energia solar", são anotados como *tokens* independentes. A UD não trabalha com o conceito de multipalavras no nível morfossintático e, no nível sintático, tem três relações de dependência para anotar apenas as multipalavras que não apresentam relações sintáticas entre os *tokens* que as integram.

A anotação no esquema UD segue o formato definido na conferência Computational Natural Language Learning (CoNLL) e por isso é chamado CoNLL-U (U para "universal"). Várias ferramentas de anotação de interface gráfica e amigável estão disponíveis para anotar nesse formato[14], não representando, portanto, uma dificuldade para os projetos. É importante, contudo, entender as partes que constituem esse formato. Apesar de o objetivo ser a anotação de relações de dependência sintática, o CoNLL-U permite que se registre muito mais informações além dessas relações. Isso porque, quanto mais rica a anotação, mais atributos podem contribuir para a distinção entre as etiquetas no aprendizado automático.

[14] https://universaldependencies.org/tools.html

O formato CoNLL-U é constituído por uma tabela de 10 colunas com informações associadas aos tokens nas linhas. As colunas mais importantes para o anotador são a quarta, que se chama UPOS (preenchida com uma das 17 etiquetas morfossintáticas da UD), e a oitava, que se chama DepRel (preenchida com uma das 37 etiquetas de relações de dependência sintática da UD). As demais colunas contêm: o identificador de posição de cada token, a forma, o lema, os atributos morfológicos (flexões, por exemplo), o *head* da relação de dependência e três colunas de preenchimento opcional. Esse formato, ao mesmo tempo que restringe o conteúdo da maioria das colunas, caracterizando a anotação segundo a abordagem UD, prevê algum espaço para que também sejam anotadas características específicas de cada língua (a quinta coluna, XPOS, é reservada para etiquetas morfossintáticas específicas da língua e a décima coluna, MISC, é de uso livre). A Tabela 1 mostra um exemplo de sentença no formato CoNLL-U.

Tabela 1 - Exemplo de sentença anotada no formato CoNLL-U.

# text = **O estúdio também divulgou novo poster de o filme .**									
id.	*forma*	*lema*	*UPOS*	*XPOS*	*Atributos morfológicos*	*head*	*DelRel*	*Deps*	*MISC*
1	**O**	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det	_	_
2	**estúdio**	estúdio	NOUN	_	Gender=Masc|Number=Sing	4	nsubj	_	_
3	**também**	também	ADV	_	_	4	advmod	_	_
4	**divulgou**	divulgar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
5	**novo**	novo	ADJ	_	Gender=Masc|Number=Sing	6	amod	_	_
6	**poster**	poster	NOUN	_	Foreign=Yes|Gender=Masc|Number=Sing	4	obj	_	_
7	**de**	de	ADP	_	_	9	case	_	_
8	**o**	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	9	det	_	_
9	**filme**	filme	NOUN	_	Gender=Masc|Number=Sing	6	nmod	_	_
10	**.**	.	PUNCT	_	_	4	punct	_	SpacesAfter=\s\s\n

5 A relação entre o esquema UD e a Gramática Normativa

Olhando apenas para o conjunto de 17 etiquetas morfossintáticas (*part-of-speech tags* ou simplesmente *PoS tags*) e de 37 relações sintáticas da UD, poderíamos nos guiar pelos conhecimentos adquiridos com o estudo das Gramáticas Normativas e arriscar-nos a fazer a atribuição de grande parte delas às palavras do texto, pois os nomes das etiquetas muitas vezes coincidem com termos utilizados na Nomenclatura Gramatical Brasileira (HENRIQUES, 2009). No entanto, as diretrizes da UD nos mostram que a

interpretação dessas etiquetas nem sempre é simples de ser feita a partir apenas de conhecimentos prévios de gramáticas. Para exemplificar como os conceitos das etiquetas da UD diferem dos conceitos largamente difundidos pelas gramáticas normativas, abordaremos as etiquetas morfossintáticas e as etiquetas de relações de dependência em duas seções diferentes (da mesma forma como abordamos em dois manuais separados as diretrizes para atribuição desses dois conjuntos de etiquetas).

Embora apresentadas em seções diferentes, é importante destacar que na UD há alta correlação entre as etiquetas morfossintáticas e as etiquetas de relações sintáticas. Na Seção 5.2 mostraremos como essa correlação se traduz num conjunto de restrições que as relações sintáticas impõem sobre as etiquetas morfossintáticas.

Ressaltamos que este artigo exemplifica apenas algumas das questões contempladas nas diretrizes que integram nossos manuais de anotação. Os manuais foram elaborados a partir de experiências iniciais de anotação e, como são geridos pelas mesmas pessoas que anotam o *corpus*, sofrem manutenções periódicas com vistas a refletir novos fatos observados no *corpus* e que ainda não tenham sido contemplados nas diretrizes[15].

5.1 Etiquetas morfossintáticas da UD

As etiquetas morfossintáticas padrões da UD são:

- Para classes abertas: ADJ (adjetivos), ADV (advérbios), INTJ (interjeições), NOUN (substantivos), PROPN (nomes próprios) e VERB (verbos);
- Para classes fechadas: ADP (preposições), AUX (verbos auxiliares), CCONJ (conjunções coordenativas), DET (determinantes), NUM (numerais cardinais), PART (partículas), PRON (pronomes) e SCONJ (conjunções subordinativas);

[15] A versão do manual divulgada neste artigo baseia-se nas diretrizes da UD vigentes em abril de 2022. Novas diretrizes da UD que venham a ser divulgadas serão in*corpora*das em futuras versões do manual.

- Outros tokens que não se enquadram nas classes abertas e fechadas: PUNCT (pontuações), SYM (símbolos) e X (demais casos, incluindo onomatopeias e palavras em língua estrangeira).

As principais diferenças entre as etiquetas morfossintáticas da UD e as classes morfossintáticas da gramática normativa do português brasileiro são:

- DET: a UD trabalha com o conceito de determinante e sob essa etiqueta reúne os artigos e os pronomes não nominais (demonstrativos e possessivos);
- NUM: apenas os numerais cardinais devem ser anotados como NUM, enquanto os numerais ordinais devem ser anotados como ADJ;
- PRON: apenas os pronomes nominais são anotados com essa etiqueta, e os demais pronomes, desde que estejam modificando um nominal, devem ser anotados como DET;
- NOUN: apenas os substantivos comuns são anotados sob essa etiqueta, pois os nomes próprios são anotados como PROPN;
- VERB: apenas os verbos considerados plenos e passíveis de serem classificados como predicados verbais devem ser anotados com essa etiqueta.
- AUX: verbos auxiliares e verbos de cópula "altamente gramaticalizados", ou seja, sem carga semântica significativa, devem ser anotados com essa etiqueta;
- PROPN: etiqueta destinada a anotar nomes próprios, desde que não coincidam com palavras comuns da língua.

Em nosso manual de anotação de *PoS tags* (etiquetas morfossintáticas) cada etiqueta é tratada em uma seção, com definição, exemplos, léxico das palavras (no caso

de classes fechadas) e diretrizes de desambiguação em relação a outras etiquetas que costumam ser confundidas com ela em determinados contextos. A ambiguidade entre etiquetas morfossintáticas é, em grande parte, dependente de língua, e por isso não é contemplada nas diretrizes da UD. Por exemplo, a palavra "menos", antes de um substantivo, com o sentido de "menor quantidade de", é um ADJ, pois modifica um NOUN. Quando, porém, a palavra "menos" modifica um VERB, um ADJ ou um ADV, com o sentido de "com menor intensidade" ou "em menor grau" ela é anotada como ADV. Há contextos, porém, em que "menos" ocorre entre um VERB e um NOUN e o anotador desavisado pode se confundir na atribuição. Exemplo:

O plano favoreceu **menos** classes D e E do que classe média. (aqui "menos" é um ADV que modifica o verbo "favorecer": "favoreceu em menor grau as classes D e E")

O plano atual favoreceu **menos** pessoas que o plano anterior. (aqui "menos" é um ADJ e modifica o substantivo "pessoas": "uma menor quantidade de pessoas").

5.2 Relações de dependência da UD

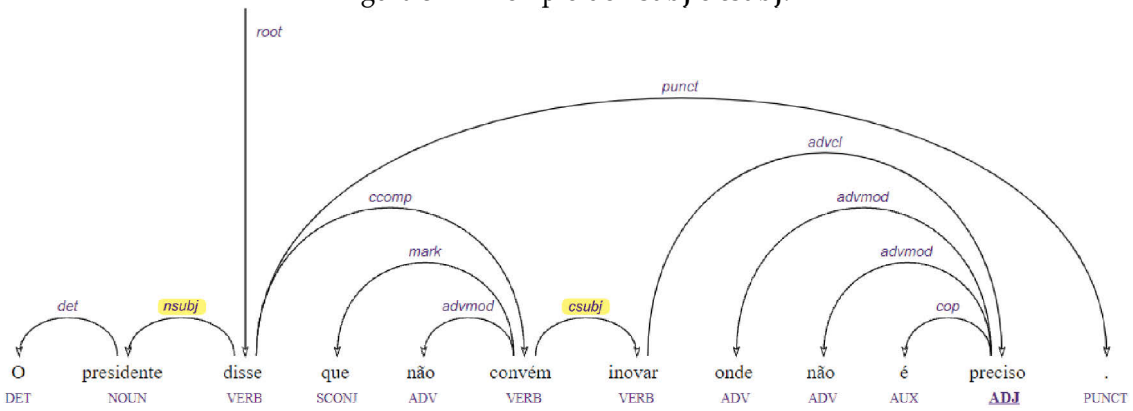
A UD possui um conjunto de 37 etiquetas para marcar as relações de dependência[16] e o descasamento entre essas relações e os componentes da análise sintática descritos nas gramáticas normativas é ainda maior do que o observado no

[16] Um quadro completo das relações de dependência do UD está disponível em https://universaldependencies.org/u/dep/index.html

conjunto de etiquetas morfossintáticas. A seguir serão discutidas as principais diferenças entre o esquema de anotação sintática da UD e as funções sintáticas que constam das gramáticas normativas.

A UD tem relações separadas para as funções sintáticas preenchidas por orações (tanto predicados verbais quanto nominais) e as mesmas funções sintáticas preenchidas por palavras núcleos de sintagmas. Assim, por exemplo, um sujeito pode ser **nsubj**, se for um nominal, ou **csubj**, se for constituído por uma oração, conforme pode ser observado na Figura 3.

Figura 3 — Exemplo de **nsubj** e **csubj**.

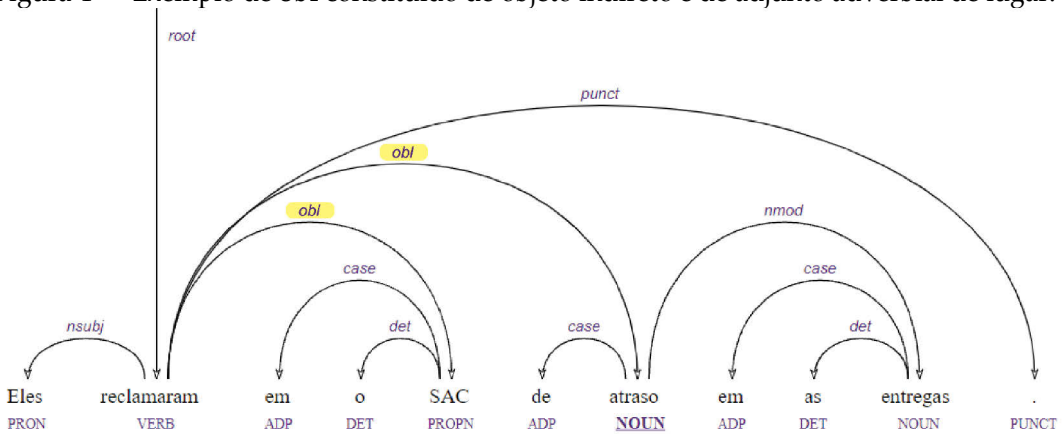


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew[17].

Na UD, os adjuntos adverbiais cujo núcleo é um NOUN (em sua maioria introduzidos por preposição) não são discriminados dos objetos indiretos introduzidos por preposição, pois ambos são considerados modificadores verbais do tipo nominal e anotados com a relação **obl** (ilustrados na Figura 4).

[17] https://arborator.icmc.usp.br/#/

Figura 4 — Exemplo de **obl** constituído de objeto indireto e de adjunto adverbial de lugar.

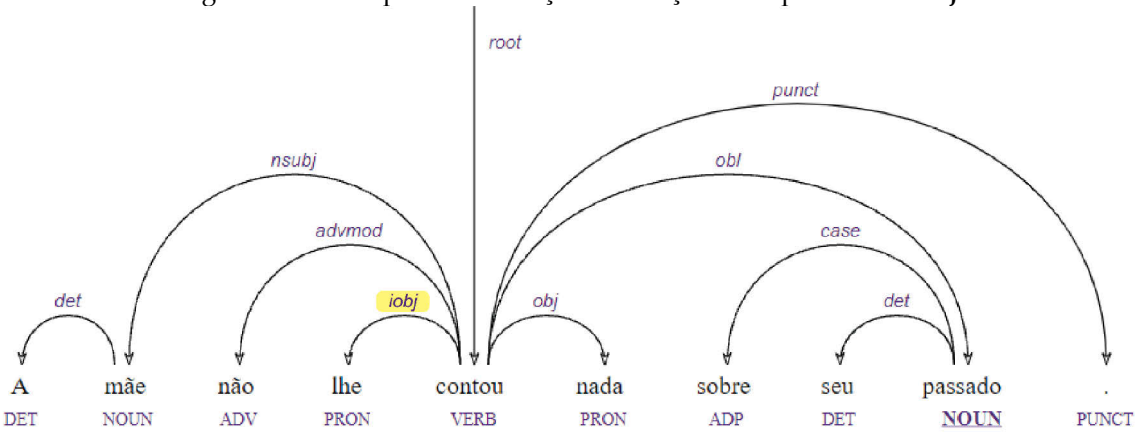


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Talvez uma das relações da UD mais fáceis de serem mal interpretadas seja **iobj** (objeto indireto). Essa relação só deve ser atribuída a objetos indiretos não preposicionados, o que é um contrassenso se pensarmos no que significa "indireto" no nome dessa função sintática. Esse é um resquício das origens das relações na anotação em língua inglesa, pois o inglês tem casos de dois objetos não preposicionados[18] e ambos podem ser alçados à posição de sujeito na diátese de voz passiva. No português, só devemos utilizar a relação **iobj** para anotar objetos indiretos na forma de pronomes oblíquos (*me, te, se, lhe, nos, vos, lhes*), como mostrado na Figura 5.

[18] Dois objetos não preposicionados podem ocorrer nos verbos dativos do inglês. Em "John sent Mary a message", ambos os objetos podem ser alçados a sujeito da passiva: "A message was sent to Mary" e "Mary was sent a message".

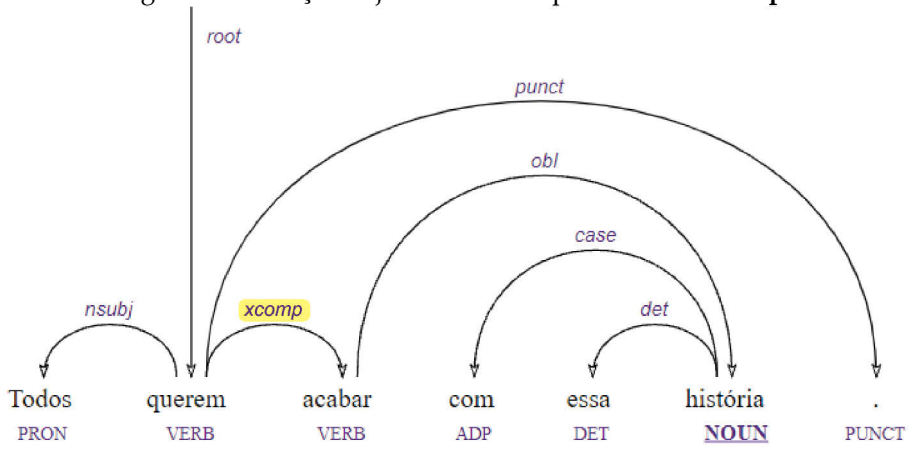
Figura 5 — Exemplo de atribuição da relação de dependência **iobj**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Na UD, objetos diretos são anotados com a relação **obj**, mas, se forem realizados por meio de oração, recebem dois tipos de anotação diferentes. As orações objetivas diretas que têm sujeito NULL[19] governado pelo sujeito ou pelo objeto da oração matriz, são anotadas como dependentes da relação **xcomp** (Figura 6), e as que têm sujeito realizável (embora possa estar elíptico), são anotadas com a relação **ccomp** (Figura 7).

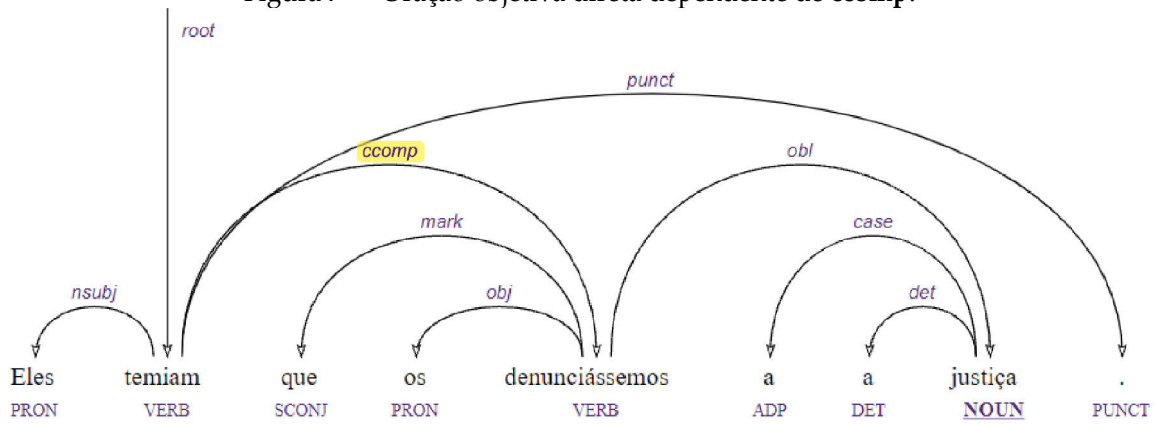
Figura 6 — Oração objetiva direta dependente de **xcomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

[19] Para um aprofundamento sobre o conceito de sujeito NULL e xcomp, recomendamos a leitura de Bresnan (1982).

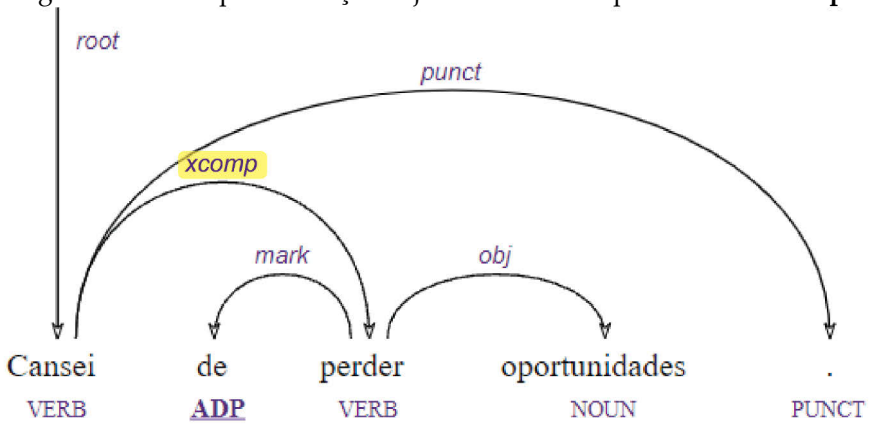
Figura 7 — Oração objetiva direta dependente de **ccomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

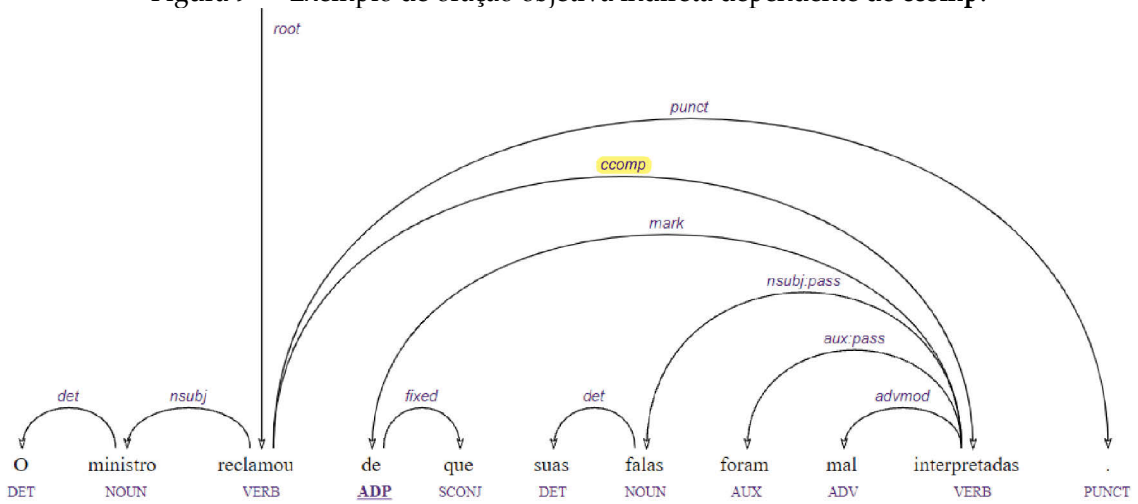
O mesmo ocorre com os objetos indiretos realizados sob forma de oração (subordinada substantiva objetiva indireta): se a oração subordinada tiver sujeito NULL governado pelo sujeito ou pelo objeto da oração matriz, será dependente de **xcomp** (Figura 8) e, se tiver sujeito realizável, será dependente de **ccomp** (Figura 9).

Figura 8 — Exemplo de oração objetiva indireta dependente de **xcomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

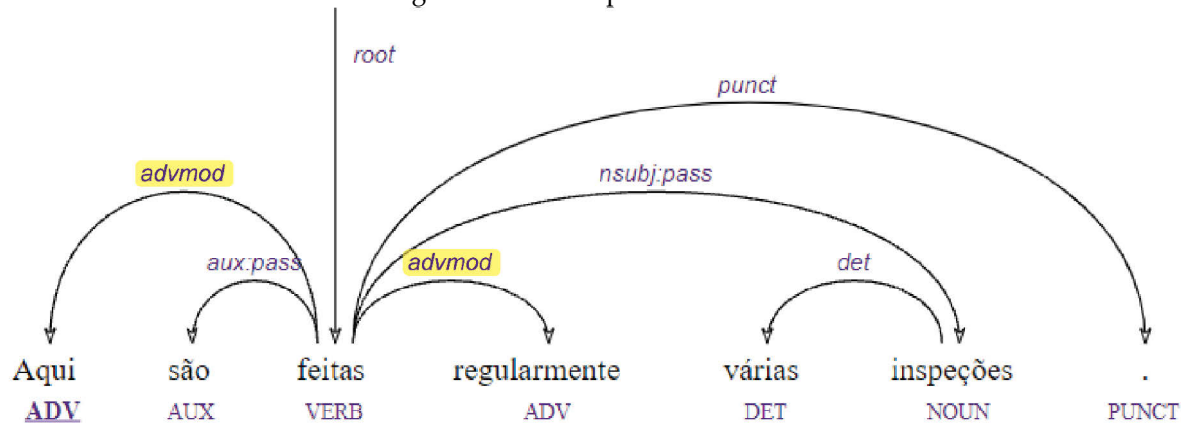
Figura 9 — Exemplo de oração objetiva indireta dependente de **ccomp**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

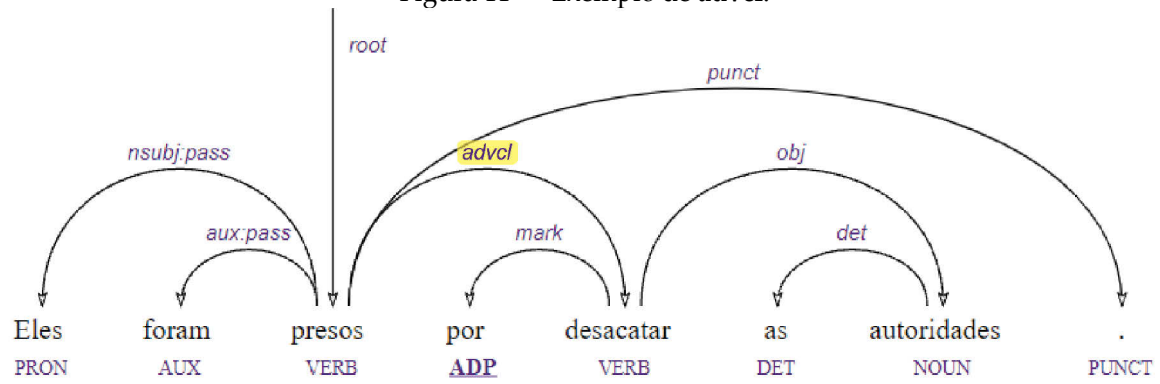
Os modificadores constituídos por advérbios (ADV) são anotados com a relação **advmod** (Figura 10). Porém, em se tratando de uma oração adverbial, a relação será **advcl** (Figura 11).

Figura 10 — Exemplo de **advmod**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

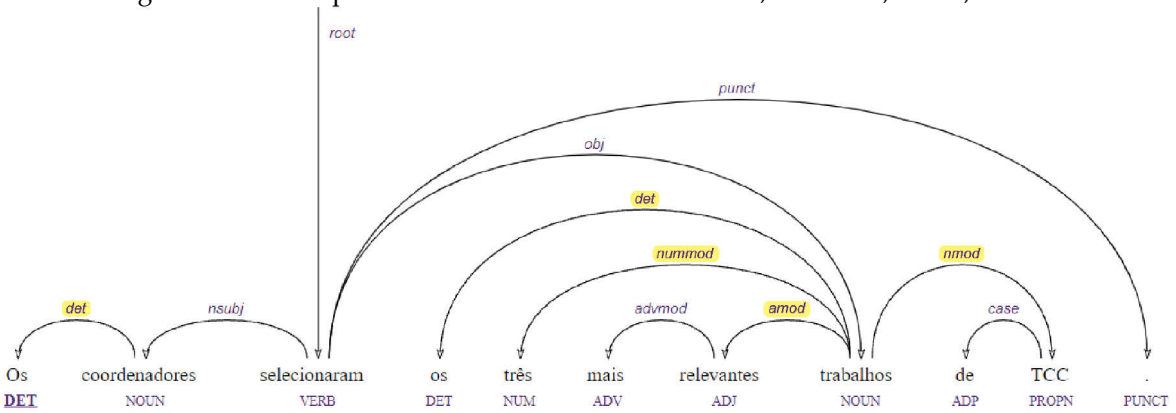
Figura 11 — Exemplo de **advcl**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

A UD também tem uma granularidade maior para os modificadores nominais do que as gramáticas normativas. Os modificadores nominais podem ser: **amod** (para ADJ), **nmod** (para NOUN e PROPN) e **nummod** (para NUM). As preposições (ADP) que introduzem um **nmod** são ligadas pela relação **case** ao nominal *head* da relação. Os determinantes (DET) são ligados aos nominais que determinam por meio da relação **det**, como pode ser observado na Figura 12.

Figura 12 — Exemplos de modificadores nominais: det, nummod, amod, nmod.

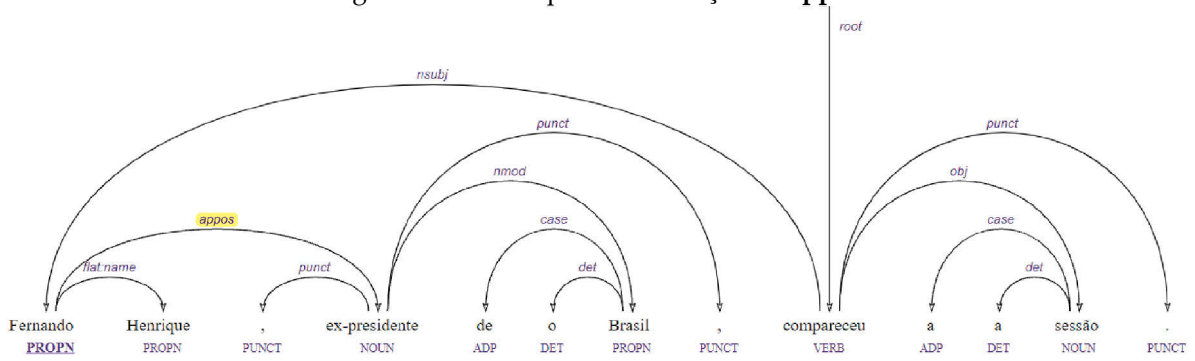


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew

Se, contudo, um modificador nominal estiver sob forma de oração, a relação de dependência será sempre **acl**, pois a UD não distingue orações completivas nominais de orações adjetivas.

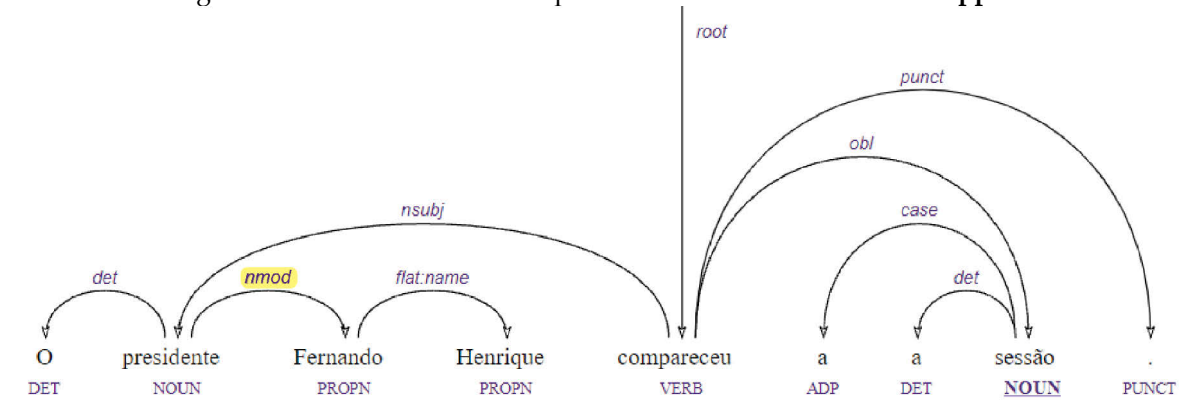
Outra diferença que merece ser destacada é o conceito de aposto na UD. Segundo as instruções da UD, só deve ser anotada como **appos** a relação entre dois elementos intercambiáveis, que sejam independentes e que possam ser trocados de ordem sem prejuízo para a gramaticalidade da sentença, como na Figura 13. Casos como o mostrado na Figura 14, contudo, tradicionalmente tratados como aposto, são anotados como um atributo na UD, **nmod**.

Figura 13 — Exemplo de atribuição de **appos**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Figura 14 — Atributo de **nmod** que costuma ser confundido com **appos**.

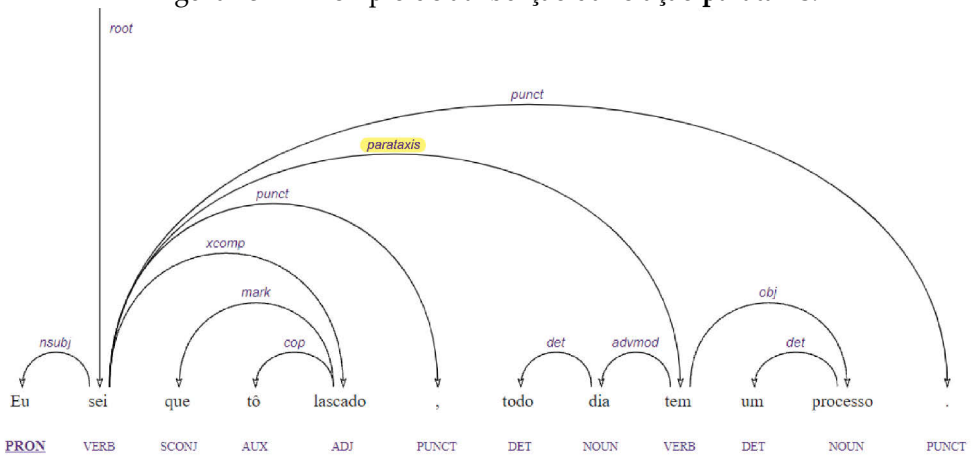


Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Além das relações de dependência associadas a funções sintáticas da gramática normativa, a UD prevê uma série de relações para ligar os elementos da sentença que não têm relação sintática com outros elementos, como as relações **parataxis**, **discourse**, **reparandum**, **dislocated** e **intj**.

A relação de **parataxis** (ilustrada na Figura 15), por exemplo, serve para ligar duas orações sem coesão lógica explícita dentro da mesma sentença.

Figura 15 — Exemplo de atribuição da relação **parataxis**.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

As diretrizes da UD apresentam também três relações de dependência para anotar multipalavras que não possuem relação sintática entre si (**compound**, **flat** e **fixed**). O uso dessas etiquetas deve ser parcimonioso, pois elas simplesmente atestam que as palavras ligadas por elas não possuem nenhuma relação sintática entre si, o que as torna opacas para que o *parser* aprenda sintaxe com elas. Por exemplo:

João Paulo - PROPN PROPN ligados pela relação **flat**
além de - ADV ADP ligados pela relação **fixed**
ir embora - VERB ADV ligados pela relação **compound**.

Não se pode dizer, portanto, que as diretrizes da UD determinem a anotação de todas as multipalavras, pois aquelas que possuem relações sintáticas entre os tokens que as compõem são anotadas com relações comuns. Por exemplo:

efeito estufa é um NOUN NOUN ligado pela relação **nmod**, cujo *head* é *efeito* e o dependente (modificador) é *estufa*
protetor solar é um NOUN ADJ ligado pela relação **amod**, cujo *head* é *protetor* e o dependente (modificador) é *solar*
casa de câmbio é um NOUN ADP NOUN ligado pelas relações **nmod** (entre *casa* e *câmbio*) e **case** (entre *câmbio* e *de*)

Algo que deve ser observado nos exemplos apresentados nesta seção é o fato de que algumas relações sintáticas só admitem dependentes que tenham uma determinada etiqueta morfossintática: **det** só admite DET, **case** só admite ADP, **advmod** só admite ADV, **nmod** só admite NOUN ou PROPN, **amod** só admite ADJ, **nummod** só admite NUM, **aux** só admite AUX, **cop** só admite AUX, **punct** só admite PUNCT, **intj** só admite INTJ e **mark** só admite SCONJ e ADP. Isso ilustra como a UD restringiu automaticamente a anotação, impedindo a ocorrência de grande parte dos erros de atribuição de etiquetas.

6 Os desafios da instanciação das diretrizes UD e da construção do manual de anotação

A tarefa de produzir os dois manuais de anotação (de etiquetas morfossintáticas e de etiquetas de relações dependência) iniciou-se com a leitura minuciosa das diretrizes da UD, complementada pela leitura de outras fontes de consulta sobre a atribuição das etiquetas do esquema UD, como artigos e fórum de discussão disponível no *github* da UD. Realizamos o trabalho em duas etapas , uma vez que a anotação das etiquetas morfossintáticas seria realizada em uma fase anterior à fase de anotação de relações de dependência (por isso organizamos as diretrizes em dois manuais).

Antes de iniciarmos a anotação de cada fase, já tínhamos uma versão preliminar dos respectivos manuais, os quais foram utilizados no treinamento dos anotadores. Essas versões preliminares dos manuais foram elaboradas por uma linguista e revisadas por uma pesquisadora de PLN (não linguista) que participa do processo de anotação. A equipe de anotadores foi ampliada na sequência, e as pessoas responsáveis pela versão preliminar do manual continuaram na equipe, participando da anotação.

Durante cada fase do processo de anotação, as necessidades de melhoria das definições e de maior detalhamento das instruções tornaram-se evidentes. Utilizamos as divergências entre os anotadores como subsídio para aperfeiçoar os manuais, assim como utilizamos exemplos do *corpus* para ilustrar questões que se mostraram desafiadoras para a anotação.

Outra fonte de consulta foram os *corpora* já anotados em português e em outras línguas com o modelo UD. Para isso, utilizamos o buscador Grew-Match[20], que permite buscas por palavras, por lemas, por etiquetas morfossintáticas e por relações de dependência. Essa prática foi importante tanto para descobrirmos soluções adotadas em outros *corpora* que poderiam ser aplicáveis no nosso *corpus* quanto para percebermos que 1) diferentes *corpora* de uma mesma língua às vezes adotam formas diferentes de atribuir uma mesma etiqueta ou relação; 2) *corpora* de diferentes línguas adotam, eventualmente, etiquetas e relações diferentes para anotar fenômenos semelhantes e 3) muitas vezes um padrão não é reconhecido pelos anotadores e ele acaba sendo anotado de diferentes maneiras ao longo do mesmo *corpus*, gerando inconsistência.

Durante a confecção dos manuais, nos deparamos com três tipos de situações: 1) aquelas em que a orientação da UD era clara e facilmente aplicável no português; 2) aquelas em que havia lacunas na orientação da UD, mas que, seguindo princípios da linguística, conseguimos preencher com decisões claras; e 3) aquelas em que as

[20] http://match.grew.fr/

diretrizes da UD não eram claras ou apresentavam lacunas e a fundamentação linguística não era suficiente para garantir uma decisão que pudesse ser considerada "correta", situações em que tivemos que tomar decisões arbitrárias a fim de garantir consistência na anotação.

Por exemplo, os *corpora* de UD de português anotam quantidades diferentes de verbos com AUX: o PUD anota 13, o GSD anota 59 e o Bosque anota 6 verbos[21]. Nas outras línguas, há grande variação na atribuição dessa etiqueta também. O inglês, por exemplo, anota modais como *would, should, can, could, may, might* e *must*, enquanto o francês anota também o causativo *faire* como AUX.

A questão dos verbos auxiliares e de cópula foi uma das que exigiram que tomássemos uma decisão arbitrária (e provisória até que solução melhor se apresente), já que as diretrizes da UD deixam a cargo de cada língua fazer essa decisão. A única recomendação da UD é a de que só devem ser anotados como AUX os verbos auxiliares altamente gramaticalizados e os verbos de cópula esvaziados de semântica. Atualmente, estamos anotando como AUX apenas os auxiliares de tempo e de voz passiva e os verbos de cópula *ser* e *estar*.

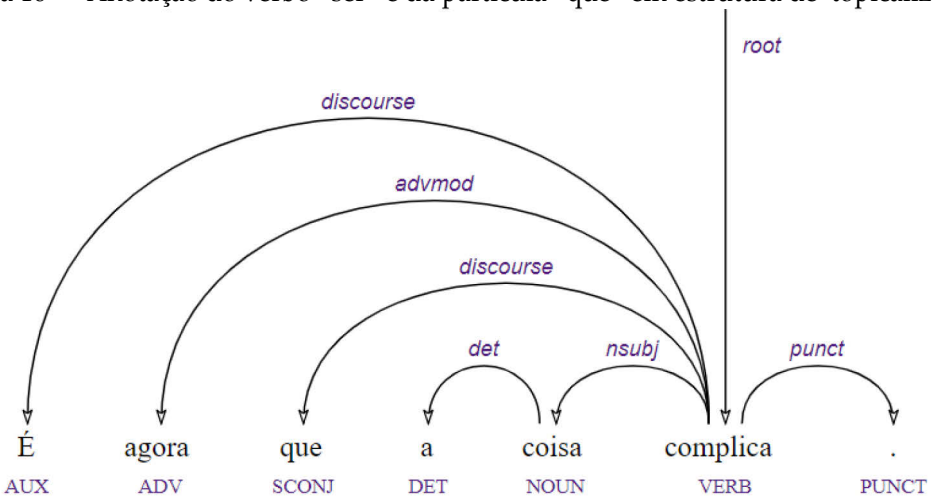
Já uma questão que exigiu uma decisão baseada em critérios linguísticos foi a anotação das preposições que introduzem orações reduzidas de infinitivo como ADP e não como SCONJ. A UD só diz que SCONJ é usada para *complementizers* (no português, *que* e *se*) e palavras que introduzem orações adverbiais (como *quando*, nas temporais, e *como*, nas causais e conformativas). Segundo Azeredo (2013, item 291) as conjunções subordinativas são utilizadas para introduzir orações desenvolvidas e não orações reduzidas, o que nos permitiu inferir, portanto, que as preposições que introduzem orações subordinadas reduzidas de infinitivo devem continuar sendo anotadas como ADP, embora a respectiva relação sintática seja **mark**. Essa anotação pode beneficiar futuramente tarefas de anotação de papéis semânticos, pois as

[21] Para uma comparação detalhada: https://universaldependencies.org/treebanks/pt-comparison.html

preposições que introduzem nominais ou orações no infinitivo são importantes pistas para discriminar os papéis semânticos na estrutura argumental de um verbo.

A anotação do verbo *ser* e da conjunção *que* com função de topicalização foi um caso em que reproduzimos a decisão arbitrária tomada no *corpus* Bosque-UD: anotamos o verbo como AUX, a conjunção como SCONJ e a relação de ambos com o predicado como **discourse**.

Figura 16 — Anotação do verbo "ser" e da partícula "que" em estrutura de topicalização.



Fonte: árvore sintática em formato UD produzida no software Arborator-Grew.

Um caso em que não seguimos cegamente as diretrizes da UD foi o da anotação dos nomes próprios. A UD prescreve que os nomes próprios que constituem substantivos e adjetivos comuns da língua sejam anotados com as etiquetas próprias de suas categorias de origem e sua relação sintática seja anotada com as relações sintáticas usadas para palavras comuns. Assim, *Maria das Dores* seria composto por PROPN, ADP, DET e NOUN, e *Companhia de Água e Esgoto* por NOUN, ADP, NOUN, CCONJ e NOUN. No entanto, fomos conservadores e mantivemos como PROPN todas as palavras grafadas em letra maiúscula e as ligamos com a relação **flat:name**, pois não queríamos limitar tarefas de PLN que dependem do resultado do *parser*, como o reconhecimento de entidades nomeadas.

7 Considerações finais e trabalhos futuros

Conforme exposto, o trabalho de construir o manual de anotação do modelo UD em *corpus* do português exigiu bastante esforço e constantes aperfeiçoamentos. Por razões práticas e didáticas, separamos as diretrizes em dois manuais, um para anotação de etiquetas morfossintáticas (Manual de Anotação de *PoS tags*) e outro para relações de dependência (Manual de Anotação de Relações de Dependência). Esses dois manuais, que se completam, estão disponíveis no site do projeto POeTiSA[22], e poderão abreviar o esforço necessário para novos empreendimentos de anotação sintática de *corpus* seguindo o modelo UD. No mesmo site serão publicadas novas versões desses manuais, sempre que manutenções no conteúdo se façam necessárias.

Como nosso projeto prevê a anotação de outros *corpora*, em diferentes domínios, é de se esperar que novas diretrizes venham a ser somadas ao material existente, na forma, por exemplo, de apêndices dedicados a domínios específicos.

Agradecimentos

Os autores agradecem o apoio do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI-http://c4ai.inova.usp.br/), financiado pela IBM e pela FAPESP (processo#2019/07665-4).

Referências Bibliográficas

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: a treebank for Portuguese. *In*: RODRÍGUEZ, M. G.; ARAUJO, C. P. S. (ed.), **Proceedings of the Third International Conference on Language Resources and Evaluatio**n (LREC 2002). European Language Resources Association, 2002. p. 1698-1703.

AZEREDO, J. C. S. **Fundamentos de Gramática do Português**. Rio de Janeiro: Jorge Zahar Editores, 2013. E-book.

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

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley Framenet Project. *In*: **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics**, Vol. 1. Quebec: Association for Computational Linguistics, 1998. p. 86-90. DOI https://doi.org/10.3115/980845.980860

BICK, E. **The Parsing System PALAVRAS**: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Aarhus University Press, 2000.

BICK, E. Constraint grammar-based conversion of dependency treebanks. *In*: **Proceedings of the 13th International Conference on Natural Language Processing** (ICON). Varanasi: NLP Association of India, 2016, p. 109–114.

BRESNAN, J. Control and Complementation. **Linguistic Inquiry**, Vol. 13, n. 3, p. 343-434, 1982.

BUCHHOLZ, S.; MARSI, E. CoNLL-X Shared Task on Multilingual Dependency Parsing. **Proceedings of the Tenth Conference on Computational Natural Language Learning**. New York: Association for Computational Linguistics, 2006. p. 149–164. DOI https://doi.org/10.3115/1596276.1596305

FILLMORE, C. J. The Case for Case. *In*: BACH, E.; HARMS R. T. (ed.) **Universals in Linguistic Theory**. London: Holt, Rinehart and Winston. p. 1-88, 1968.

FREITAS, C.; ROCHA, P.; BICK, E. "Floresta Sintá(c)tica: Bigger, Thicker and Easier". *In*: TEIXEIRA, A.; LIMA, V. L. St. de; OLIVEIRA, L. C. de; QUARESMA, P. (ed.), **Proceedings of Computational Processing of the Portuguese Language, 8th International Conference,** (PROPOR 2008), vol. 5190. Springer Verlag, 2008. p. 216-219. DOI https://doi.org/10.1007/978-3-540-85980-2_23

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.. **Deep Learning**. Cambridge MA: MIT Press, 2016.

HENRIQUES, C. C. **Nomenclatura Gramatical Brasileira**: cinquenta anos depois. São Paulo: Parábola, 2009.

HONNIBAL, M.; JOHNSON, M. An Improved Non-monotonic Transition System for Dependency Parsing. *In*: **Proceedings of the Conference on Empirical Methods in Natural Language Processing** (EMNLP). Lisboa: Association for Computational Linguistics, 2015. p. 1373-1378. DOI https://doi.org/10.18653/v1/D15-1162

IDE, N. Introduction: The Handbook of Linguistic Annotation. In: Ide, Nancy; Pustejovsky, James (ed). **Handbook of Linguistic Annotation**. Springer, 2017. DOI https://doi.org/10.1007/978-94-024-0881-2

KARLSSON, F. Constraint Grammar as a Framework for Parsing Unrestricted Text. *In*: KARLGREN, H. (ed.), **Proceedings of the 13th International Conference of Computational Linguistics**, Vol. 3. ACM Digital Library, 1990. p. 168-173. DOI https://doi.org/10.3115/991146.991176

KIPPER-SCHULER, K. **VerbNet:** A Broad-Coverage, Comprehensive Verb Lexicon. Tese de Doutorado (Ciência da Computação). University of Pennsylvania, 2005.

KONDRATYUK, D.; STRAKA, M. 75 Languages, 1 Model: Parsing Universal Dependencies Universally. *In*: INUI, K.; JIANG, J.; NG, V.; WAN, X. (ed.) **Proceedings of the 2019 Conference on Empirical Metho1982)ds in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing** (EMNLP), Hong Kong, China. Association for Computational Linguistics, 2019. p. 2779–2795. DOI https://doi.org/10.18653/v1/D19-1279

LEVIN, B. **English Verb Classes and Alternations:** A Preliminary Investigation. University of Chicago Press, 1993.

MANNING, C.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA: MIT Press, 1999.

MARNEFFE, M.-C. de; MANNING, C. D.; NIVRE, J.; ZEMAN, D.. Universal Dependencies. **Computational Linguistics** 47 (2), p. 255–308, 2021.

MITCHELL, T. Machine Learning. New York: McGraw Hill, 1997.

NIVRE, J. Towards a Universal Grammar for Natural Language Processing. *In*: **Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics** (CICLing), 2015. p. 3-16. DOI https://doi.org/10.1007/978-3-319-18111-0_1

NIVRE, J.; HALL, J.; NILSSON, J.; ERYIĞIT, G.; MARINOV, S. Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines. *In*: **Proceedings of the Tenth Conference on Computational Natural Language Learning** (CoNLL-X). New York: Association for Computational Linguistics, 2006. p. 221–225. DOI https://doi.org/10.3115/1596276.1596318

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; HAJIČ, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. *In*: **Proceedings of the 12nd International Conference on Language Resources and Evaluation** (LREC 2020). Marseille: European Language Resources Association, 2020. p. 4034-4043.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The Proposition Bank: An Annotated *Corpus* of Semantic Roles. **Computational Linguistics**, 31:1., p. 71-105, March, 2005. DOI https://doi.org/10.1162/0891201053630264

PUSTEJOVSKY, J.; BUNT, H.; ZAENE, A. Designing Annotation Schemes: From Theory to Model. *In*: IDE, N.; PUSTEJOVSKY, J. (ed.). **Handbook of Linguistic Annotation**. Springer, 2017. DOI https://doi.org/10.1007/978-94-024-0881-2_2

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. *In*: JURAFSKY, D.; CHAI, J.; SCHLUTER, N.; TETRAULT, J. R. (ed.). **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.** Association for Computational Linguistics, 2020. p. 101-108. DOI https://doi.org/10.18653/v1/2020.acl-demos.14

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal Dependencies for Portuguese. *In*: **Proceedings of the Fourth International Conference on Dependency Linguistics.** Linköping University Electronic Press, 2017. p. 197-206.

SARDINHA, T. B. Linguística de *corpus*: Histórico e Problemática. **Documentação e Estudos em Linguística Teórica e Aplicada** (DELTA), 16:2, 2000. p. 323-367. DOI https://doi.org/10.1590/S0102-44502000000200005

SILVA, J.; BRANCO, A.; CASTRO, S.; REIS, R.. Out-of-the-Box Robust Parsing of Portuguese. *In*: **Proceedings of the 9th International Conference on the Computational Processing of Portuguese**. Springer, 2010. p. 75–85. DOI https://doi.org/10.1007/978-3-642-12320-7_10

SOUZA, E. de; CAVALCANTI, T.; SILVEIRA, A.; EVELYN, W.; FREITAS, C. **Diretivas e documentação de anotação UD em português (e para língua portuguesa)**. Rio de Janeiro: PUC-RIO, 2020. Disponível em: http://comcorhd.letras.puc-rio.br/recursos/.

STRAKA, M.; STRAKOVA, J. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *In*: **Proceedings of the CoNLL 2017 Shared Task**: Multilingual

Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, 2017. p. 88-99. DOI https://doi.org/10.18653/v1/K17-3009

TESNIÈRE, L. **Éléments de Syntaxe Structurale**. Paris: Librarie C. Klincksieck, 1959.

TESNIÈRE, L. **Elements of Structural Syntax.** Tradução de OSBORNE, T.; KAHANE, S. Amsterdam: John Benjamins, 2015. DOI https://doi.org/10.1075/z.185

Artigo recebido em: 19.10.2021 Artigo aprovado em: 25.04.2022