Do pequeno Asteróide B-612 ao universo: a anotação de O Pequeno Príncipe segundo o modelo *Universal Dependencies*

Magali Sanches Duran, Lucelene Lopes, Maria das Graças Volpe Nunes,

Thiago Alexandre Salgueiro Pardo

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

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

Abstract. *In this paper, we report the annotation process of the iconic book The Little Prince according to the Universal Dependencies international model, as part of an effort to expand the currently available annotated data for Brazilian Portuguese. In particular, we present the annotation protocol and the linguistic decisions involved, as well as a quantitative analysis of the corpus.*

Resumo. *Relatamos, neste artigo, o processo de anotação do icônico livro O Pequeno Príncipe segundo o modelo internacional Universal Dependencies, como parte de um esforço de expansão dos dados anotados atualmente disponíveis para o português do Brasil. Em especial, apresentamos o protocolo de anotação e as decisões linguísticas envolvidas, além de uma análise quantitativa do córpus.*

1. Introdução

A iniciativa internacional *Universal Dependencies* (UD) (de Marneffe et al., 2021), ao agregar mais de 600 contribuidores da Linguística e da Computação de diversas línguas e nacionalidades, vem avançando os estudos da análise de dependências e revolucionando o tratamento sintático na área de Processamento de Línguas Naturais (PLN). Em sua versão atual[1], a UD já conta com a participação de 193 línguas e 353 córpus anotados no mesmo formato[2], o que cria um cenário rico para estudos contrastivos[3] entre línguas e famílias linguísticas, bem como para o desenvolvimento de recursos de PLN multilíngues, como os parsers UDPipe (Straka, 2018) e Stanza (Qi et al., 2020). Graças a esforços recentes, a língua portuguesa passou a ocupar uma posição de destaque, sendo a 5ª língua em termos de número de palavras[4] (segundo dados na listagem online do projeto), contabilizando mais de 1,5 milhão de palavras, sendo que o

[1] https://universaldependencies.org/ (acesso em 23/05/2026)

[2] O formato de anotação da UD é conhecido pela designação CoNLL-U, derivado dos formatos usados nas *shared tasks* da *Conference on Natural Language Learning.*

[3] A comunidade UD organiza um grande evento bianual, o SyntaxFest, dividido em vários sub-eventos que discutem sintaxe empírica, anotação linguística, análise estatística de línguas e PLN. O SyntaxFest está atualmente em sua quinta edição: https://syntaxfest.github.io/syntaxfest27/

[4] Apesar de ser mais usual encontrar esse tipo de contagem em tokens, a página do projeto utiliza o termo *words*, que é o que optamos por mencionar nesta passagem do artigo, para sermos fiéis aos dados oficiais do projeto que estamos citando. A diferenciação entre token e palavra na UD não é uma questão trivial, dado considerar diversos alfabetos: https://universaldependencies.org/u/overview/tokenization.html

ranque é dominado pelo tcheco, com 4,1 milhões de palavras. Em termos de número de córpus, o português também fica em 5º lugar com 7 córpus anotados, sendo que o 1º lugar é ocupado pelo inglês, que conta com 13 córpus anotados.

Para além de sua relevância para os estudos linguísticos, os dados anotados em português têm sido utilizados para diversos fins de PLN, como treinamento de anotadores morfossintáticos (*taggers*) e analisadores sintáticos (*parsers*) (por exemplo, Lopes & Pardo, 2024), embasamento de estudos de níveis linguísticos mais avançados, como relações sintáticas aprimoradas (no inglês, *enhanced relations*) (Duran et al., 2025) e papéis semânticos (Freitas & Pardo, 2025), e desenvolvimento de pesquisas aplicadas variadas, como extração de informação aberta (Oliveira et al., 2023) e caracterização linguística de notícias falsas produzidas por humanos e por máquinas (Andrade et al., 2026).

Visando contribuir para as iniciativas existentes e avançar os estudos para o português do Brasil, relatamos neste artigo o processo de anotação de um córpus construído a partir do livro "O Pequeno Príncipe", de Antoine de Saint-Exupéry. A motivação para essa escolha é o fato de a obra ter sido previamente utilizada por Anchiêta & Pardo (2018) para receber anotação semântica segundo a abordagem *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013) a qual, combinada à anotação UD, pode alavancar muitos estudos na interface entre a sintaxe e a semântica.

As contribuições da presente pesquisa residem não apenas na disponibilização de mais um córpus anotado de acordo com o modelo UD, mas também no fato de que é um córpus de gênero diferente daqueles já anotados em português. Embora a obra seja uma tradução, ela é tomada aqui como legítima representante do gênero literário. Por isso, discutimos seus padrões morfossintáticos e sintáticos como típicos de uma obra literária e acreditamos que decisões de anotação relatadas poderão servir de exemplo para outras anotações do mesmo gênero. É possível, contudo, que futuramente, quando houver outros córpus literários anotados, diferenças entre obras originais e obras traduzidas venham a ser percebidas.

A seguir, apresentamos brevemente um panorama da língua portuguesa no âmbito da UD, assim como as principais características deste modelo. Na Seção 3, relatamos a anotação de córpus que conduzimos, apresentando os dados do córpus, o processo de anotação, as decisões linguísticas tomadas e uma análise quantitativa dos dados anotados. Na Seção 4, tecemos algumas considerações finais.

2. O modelo *Universal Dependencies* e a língua portuguesa

A iniciativa UD (de Marneffe et al., 2021) consiste em um arcabouço para anotação consensual das línguas em nível morfológico, morfossintático e sintático. Além de centenas de atributos morfológicos, são previstas 17 etiquetas morfossintáticas (classes gramaticais; no inglês, *universal part of speech tags* - UPOS tags) e 37 relações de dependência sintática (*dependency relations* - DEPRELs)[5].

[5] As 37 DEPRELs constituem um conjunto fixo de relações de dependência, porém a UD permite que as línguas acrescentem sub-relações para especificá-las. As sub-relações não precisam ser universais, por isso não serão discutidas neste artigo.

Como exemplo de análise UD, reproduzimos na Figura 1 um exemplo de Pagano et al. (2026). As etiquetas morfossintáticas encontram-se abaixo dos tokens, enquanto as relações de dependência estão nos arcos direcionados que conectam os tokens (os atributos morfológicos das palavras não são exibidos).

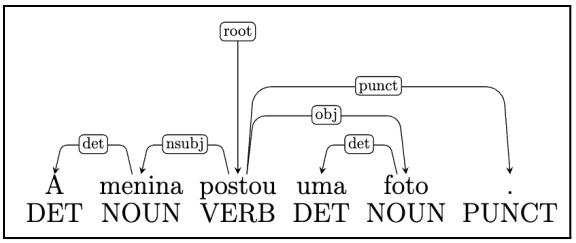

Figura 1. Exemplo de sentença com análise UD (fonte: Pagano et al., 2026)

Pelo que se tem conhecimento, o trabalho pioneiro para o português remonta à pesquisa de Rademaker et al. (2017), que anota um córpus de textos jornalísticos, o Bosque. Desde então, outras iniciativas surgiram, como o PetroGold (Souza et al., 2021), composto por textos acadêmicos do domínio de óleo e gás, e o treebank multigênero Porttinari (Pardo et al., 2021), composto pelos subcórpus Porttinari-base (Duran et al., 2023), com sentenças de notícias jornalísticas da Folha de São Paulo, DANTEStocks, com postagens do Twitter (antes de ser renomeado para X) do domínio financeiro (Di Felippo & Roman, 2026), e PortJur (Lopes et al., 2025), com leis e textos jurídicos (ementas) brasileiros, entre outros. Vários desses esforços foram baseados em manuais e diretrizes de anotação desenvolvidos para o português, como o relatado por Duran et al. (2022)[6].

Até o momento, nenhum córpus de gênero literário em português brasileiro foi disponibilizado no site da UD, lacuna que começará a ser preenchida pelo córpus O Pequeno Príncipe. A seguir, descrevemos o trabalho relativo à anotação desse córpus, que passa a integrar o conjunto de dados anotados com UD para a língua portuguesa, enriquecendo as pesquisas na área.

3. Materiais e Métodos

O córpus O Pequeno Príncipe foi constituído a partir da obra traduzida do francês para o português por Dom Marcos Barbosa, em 1952, e contém 1.527 sentenças e 12.703 tokens. Esse córpus foi criado originalmente por Anchiêta & Pardo (2018) para fazer anotação semântica segundo o modelo AMR (Banerescu et al., 2013) em português.

Aproveitamos, do córpus de Anchieta & Pardo (2018), apenas as sentenças. O alinhamento entre a anotação semântica produzida por Anchiêta & Pardo e a anotação UD, descrita neste artigo, deverá ser feito em trabalhos futuros, pois o formato dessas anotações diferem, o que impede que uma anotação seja simplesmente acrescentada à outra. A anotação AMR não considera todos os tokens, apenas os tokens que expressam conceitos (às vezes um único conceito é expresso por uma combinação de tokens). Já a

[6] A UD evita alterar as diretrizes de anotação a fim de não exigir muita manutenção nos córpus já anotados. O presente trabalho utilizou a versão 2 das Guidelines da UD, a mesma utilizada pelos córpus que constituem a v.2.18 da UD, lançada em 16/05/2026.

anotação UD considera todos os tokens, inclusive os que não são palavras, como é o caso das pontuações.

Normalmente a anotação UD parte de uma anotação automática usando o melhor *parser* disponível no momento, anotação essa que é revisada e corrigida por anotadores humanos, seguindo diretrizes compartilhadas previamente. A anotação automática, por sua vez, pode incluir ou não as fases de segmentação em sentenças e tokenização. Aproveitamos a segmentação feita no córpus original, porém, para a aplicação das diretrizes da UD, a tokenização foi alterada para desmembrar as contrações, por exemplo, "ao" torna-se "a" + "o".

O córpus foi pré-anotado usando o Portparser (Lopes & Pardo, 2024), um *parser* disponível gratuitamente que reporta uma acurácia de 99% na anotação morfossintática e 96% na anotação sintática[7] quando testado no mesmo gênero em que foi treinado, ou seja, no gênero jornalístico. Em seguida, um anotador experiente revisou as sentenças anotadas no ambiente Arborator-Grew (Gibbon et al., 2020), consultando manuais de anotação (Duran et al., 2022). Sentenças que envolviam ambiguidades ou para as quais não havia uma regra clara de anotação foram submetidas à análise de um linguista experiente em UD para decisão sobre a melhor forma de anotar e, se cabível, definição de diretrizes para casos típicos do gênero literário.

Submetemos o córpus revisado ao sistema Verifica-UD (Lopes et al., 2023), uma ferramenta baseada em regras que verifica inconsistências entre a anotação com revisão humana e as regras de anotação em língua portuguesa, produzindo um relatório que orienta a análise e as correções por um linguista. As inconsistências são principalmente detalhes de anotação, como, por exemplo, um verbo anotado com voz passiva (Voice=Pass) cujo sujeito anotado não é o de voz passiva (nsubj:pass). Na sequência, submetemos o córpus ao validador da UD[8], que verifica a aderência da anotação às exigências que a UD faz para todas as línguas antes de incorporar um córpus à sua coleção disponível na web. Por exemplo, todas as expressões fixas têm que ter o Atributo ExtPos anotado, explicitando qual a função morfossintática da expressão como um todo. Esse passo também produziu um relatório de inconsistências que foram resolvidas por um linguista.

3.1. Destaques do processo de anotação

A revisão da anotação automática exige muito conhecimento prévio sobre as diretrizes UD e um olhar atento, pois a tela do Arborator-Grew exibe várias informações e cabe ao anotador/revisor conferir as atribuições de etiquetas e corrigir o que for necessário. A Figura 2 ilustra as etiquetas atribuídas aos tokens de uma sentença (morfossintáticas e sintáticas), além do lema de cada token e atributos dentro das classes gramaticais. O token "*meu*", por exemplo, recebeu a etiqueta morfossintática **DET** (de determinante), a etiqueta sintática **det** (função de determinante), o lema "*meu*" e atributos de tipo de pronome (Prs=pessoal), gênero (Masc=masculino), número (Sing=singular) e pessoa (1=primeira pessoa). Todas essas informações foram geradas automaticamente pelo *parser* e precisaram ser conferidas pelo anotador humano.

[7] https://huggingface.co/spaces/NILC-ICMC-USP/Portparser.v2
[8] https://github.com/UniversalDependencies/tools

Uma grande diferença apresentada pelo gênero literário em comparação a outros gêneros já anotados com UD é a presença de muitas frases interrogativas. O ponto de interrogação ocorre 161 vezes no córpus. Isso representou um desafio quando a pergunta envolvia um pronome interrogativo ou um advérbio, pois o *parser* treinado em outros gêneros nem sempre acertava a atribuição das etiquetas de relações de dependência. A Figura 3 mostra a anotação automática e a versão corrigida de uma sentença do córpus.

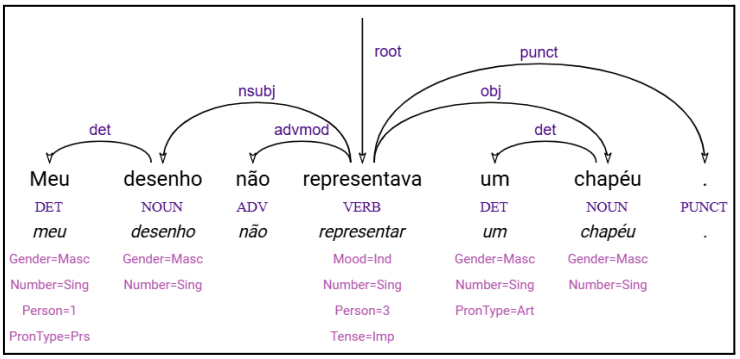

Figura 2. Exemplo de anotação no Arborator-Grew

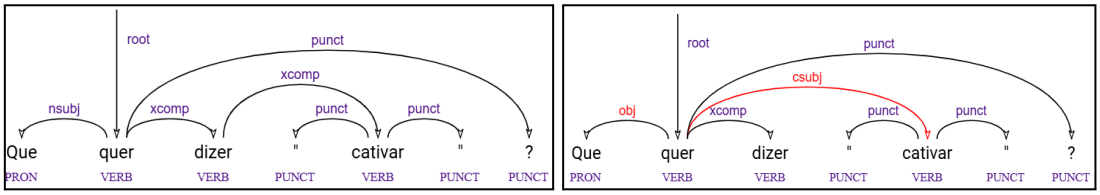

Figura 3. Sentença antes e após a correção

Para realizar essa correção, o anotador precisa perceber que o papel do "*que*" é de objeto direto da locução verbal "*querer dizer*", e "*cativar*" é um sujeito oracional da mesma locução: "*Cativar quer dizer isso*", sendo que "*isso*", no exemplo da interrogativa, é representado pelo pronome "*que*"[9].

Seguindo as diretrizes para português publicadas por Duran et al. (2022), quando se tratava de construção de cópula na interrogativa, havendo dois candidatos a sujeito, anotamos a incógnita como predicativo *(root)*, e o outro constituinte, como sujeito (*nsubj*), como na Figura 4. Na UD não há nenhuma diretriz específica para construções de cópula interrogativas, apenas a diretriz arbitrária para que, numa construção de cópula, o primeiro termo seja anotado como sujeito e o segundo termo como predicativo. Contudo, em construções como "*Onde você está?/Onde está você?*", "*Quando você esteve aqui?*", "*De onde você é?/De onde é você?*", "*Como você está? Como está você?*", é claro que o primeiro termo não poderia ser sujeito, pois não é um nominal. As diretrizes para o português, portanto, estendem essa análise das interrogativas (predicado-sujeito-cópula e predicado-cópula-sujeito) aos pronomes "*que*", "*quem*", "*qual*" e "*quais*": "*Que é isso?*", "*Quem é você?*", "*Qual é o irmão dele?*", "*Quais são as alternativas?*", criando uma recorrência de padrão que beneficia o

[9] "Cativar" foi anotado como VERB e não como NOUN porque, para ser um verbo substantivado, teria que admitir o uso de um determinante, como o artigo "o": *"Que quer dizer [o] "cativar"?".

aprendizado automático sem comprometer a análise linguística[10]. Essa análise, aliás, parece ser compatível com estudos linguísticos funcionalistas, que consideram a incógnita focalizada o elemento mais importante da oração interrogativa e, portanto, o predicativo (v. Fontes, 2012).

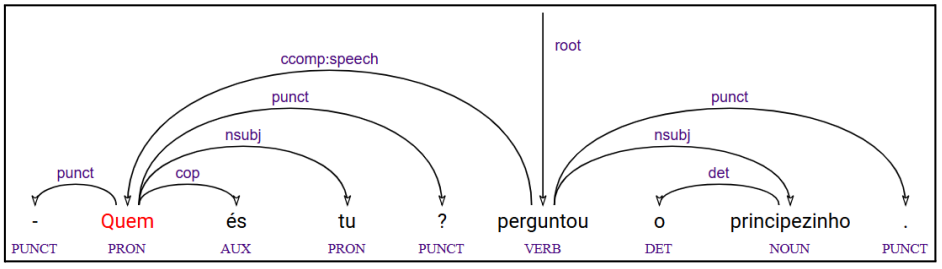

Figura 4. Anotação de construção de cópula interrogativa

O córpus tem o padrão da língua culta vigente nos anos 50 do século XX, com uso de muitos pronomes acusativos e dativos, bem como da segunda pessoa do singular e do plural (*tu* e *vós*), dois fenômenos pouco frequentes no português brasileiro contemporâneo. No córpus, houve 838 ocorrências de verbos na primeira pessoa, 330 na segunda pessoa e 2.210 na terceira pessoa. Em córpus de outros gêneros, como jurídico e jornalístico, é rara a ocorrência de verbos na primeira e na segunda pessoas. A segunda pessoa, em particular, é rara em textos do português brasileiro contemporâneo, uma vez que as formas "*você*"/"*vocês*", mais usuais, constituem uma forma de se referir à segunda pessoa do discurso mas utilizam a flexão de terceira pessoa.

Outras marcas diacrônicas no córpus são encontradas no léxico. A colocação "*fazer medo*" é usada para expressar o que sincronicamente é expresso por "*dar medo*": "*Respondera-me : "Por que é que um chapéu faria medo" ?*".

Encontramos também o uso do clítico junto ao verbo em substituição a um genitivo, o que caracteriza um uso pouco frequente atualmente: "*As únicas montanhas que conhecera eram os três vulcões que* ***lhe*** *davam pelo joelho.*" (Em português brasileiro moderno: "*... que davam/batiam no joelho* ***dele***."). Esse clítico, por ser um pronome objetivo indireto, foi anotado como **iobj** (objeto indireto), apesar de a estrutura argumental do verbo não prever um objeto indireto.

Exceto pelas formas "*lhe*" e "*lhes*", os pronomes acusativos e dativos têm formas iguais e foi preciso atentar para a regência verbal para classificar o pronome como **obj** (case=acc) ou **iobj** (case=dat). Por exemplo, em "*Será uma peça que te prego…*", o pronome "*te*" teve que ser corrigido de **obj** para **iobj** de "*prego*".

A Figura 5 ilustra um problema de tokenização. Se fosse um pronome demonstrativo, "*deste*" deveria ser desmembrado para "*de*" + "*este*". Contudo, nesse córpus ocorreu a forma "*deste*" como a segunda pessoa do singular do passado do verbo "*dar*". Nesse caso, o desmembramento teve que ser desfeito manualmente por não se tratar de uma contração (Figura 5, antes e depois).

[10] Essa análise, contudo, só funciona se o outro termo da cópula for um nominal. Se for um adjetivo ou um advérbio, o pronome é anotado como sujeito. Ex: "*Quem* é mais bonito?", "*Qual* deles é perto daqui?"

Na sentença “*E de novo, sem compreender* ***porque*** *[sic], eu sentia um estranho pesar*”, tokenizamos a forma “*porque*” desmembrando-a em dois tokens: “*por*” + “*que*” uma vez que “*porque*” aí não é uma conjunção, mas sim “*por que*”, um sintagma preposicionado.

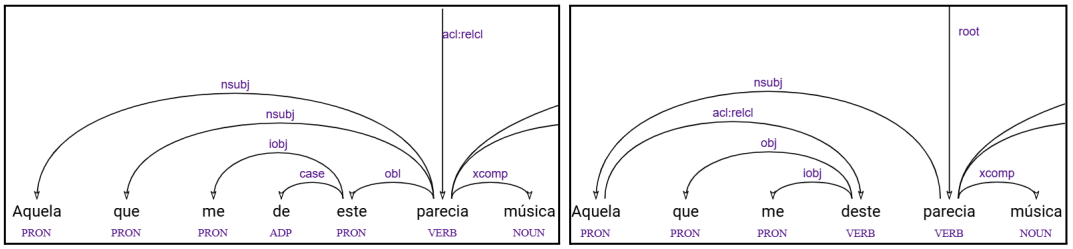

Figura 5. Restauração da forma “deste”, do verbo “dar”

O grande diferencial desse córpus fica por conta da combinação, em uma mesma sentença, de trechos de discurso direto e indireto, o que inexiste nos gêneros já anotados com UD no português. Essa característica acarreta muitas ocorrências de sinais de pontuação raros em outros gêneros, como o ponto de exclamação, o ponto de interrogação e o travessão. Ademais, o uso de inicial maiúscula no interior da sentença (quando há mudança do tipo de discurso), desafia o processamento automático, uma vez que costuma estar correlacionado a início de sentença. O mesmo ocorre com os pontos de interrogação e de exclamação no interior da sentença, pois costumam estar correlacionados a final de sentença.

3.3. Análise quantitativa do córpus anotado

O córpus anotado é composto por 1.527 sentenças. Após os ajustes da tokenização exigidos pela UD, o córpus passou a totalizar 16.802 tokens, representando acréscimo de 4.099 unidades às originais 12.703. O córpus apresenta média de 11 tokens por sentença, considerando palavras, numerais e sinais de pontuação. A Figura 6 apresenta a distribuição do tamanho das sentenças.

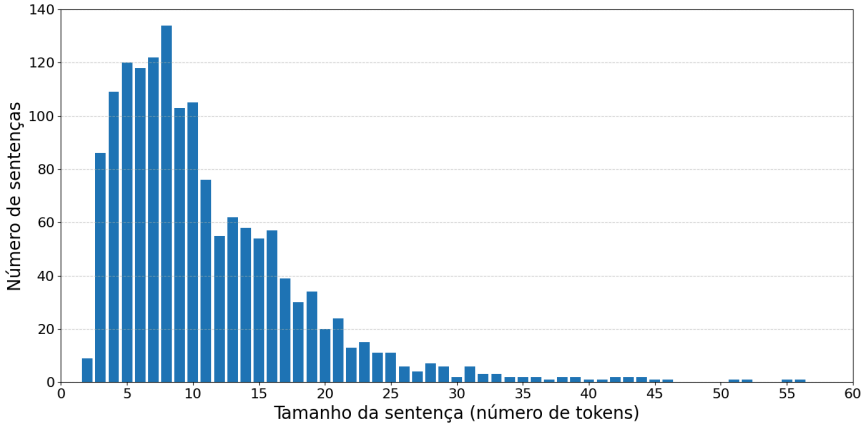

Figura 6. Histograma do número de sentenças por tamanho

Comparado a córpus de outros gêneros, o tamanho médio das sentenças desse córpus pode ser considerado pequeno. No gênero jornalístico, essa média é de 24 tokens (Duran et al., 2022), e no gênero jurídico, de 50 tokens (Lopes et al., 2024).

A UPOS mais frequente do córpus foi PUNCT (3.325 ocorrências), seguida de NOUN (2.585) e VERB (2.215). O fato de o número de verbos (VERB) quase se equiparar ao número de substantivos (NOUN) denota baixa complexidade textual e trabalhos futuros poderão comprovar isso usando métricas (Leal et al., 2023).

As Figuras 7 e 8 apresentam distribuições percentuais comparativas de O Pequeno Príncipe com os três integrantes atuais do córpus multigênero Porttinari: Porttinari-base (textos jornalísticos), PortJur (textos jurídicos) e DANTEStocks (postagens no Twitter do domínio financeiro).

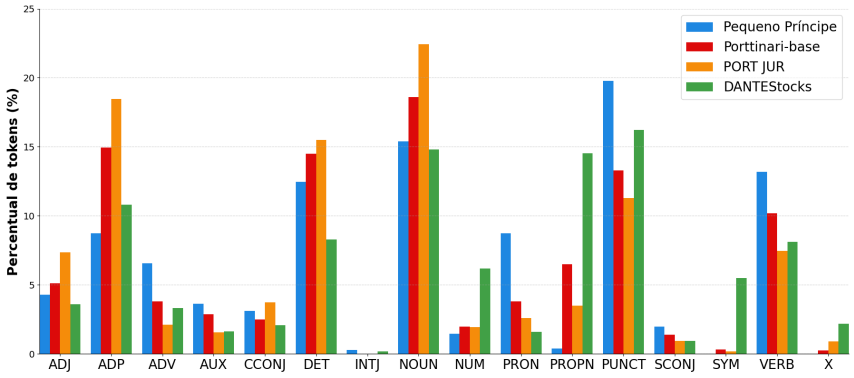

Figura 7. Distribuição das etiquetas UPOS em diferentes córpus

Em relação aos demais gêneros[11], observa-se em O Pequeno Príncipe uma maior frequência de verbos (VERB) e uma menor frequência de substantivos (NOUN) e preposições (ADP). De fato, as nominalizações são uma marca de discursos mais complexos, como o jurídico que, como mostra a Figura 8, é o que apresenta maior percentual de NOUN e ADP. Na Figura 9, a consequência do baixo índice de nominalizações pode ser observada pela frequência da etiqueta de **nmod** (modificador nominal) que, em O Pequeno Príncipe, é a menor de todos os córpus. Observa-se, ainda, frequência relativamente elevada de PRON no córpus, bem como da relação **iobj** (objeto indireto pronominal), o que pode ser explicado pelo uso recorrente de pronomes clíticos dativos, característica da língua padrão vigente à época da tradução. Esse uso, conforme destacam Morais & Salles (2024), está praticamente extinto no português contemporâneo para as formas "*lhe*" e "*lhes*", mas ainda resiste para as formas de primeira e segunda pessoas.

A maior frequência de pronomes de O Pequeno Príncipe em relação aos demais córpus pode também estar relacionada ao fato de se tratar de uma obra traduzida do francês. Como o francês não é uma língua que admite a elipse do sujeito[12], todo sujeito

[11] As barras do gráfico são percentuais de cada etiqueta em relação ao total de tokens de cada córpus. Isso significa que as diferenças de tamanho dos córpus pouco interferem na comparação.
[12] Característica de línguas chamadas *pro-drop*.

pronominal acaba sendo traduzido. Em trabalhos futuros, seria interessante comparar o percentual de pronomes nominativos de O Pequeno Príncipe com o mesmo percentual em uma obra original do português da mesma época. Segundo Othero & Lazzari (2024), desde que o português brasileiro passou a simplificar as flexões verbais, utilizando a conjugação da terceira pessoa para a segunda pessoa (*você/vocês*), a elipse do sujeito vem diminuindo. Contudo, como a conjugação da segunda pessoa é usada no córpus em questão, poderia ensejar a elipse, como ocorre no português europeu.

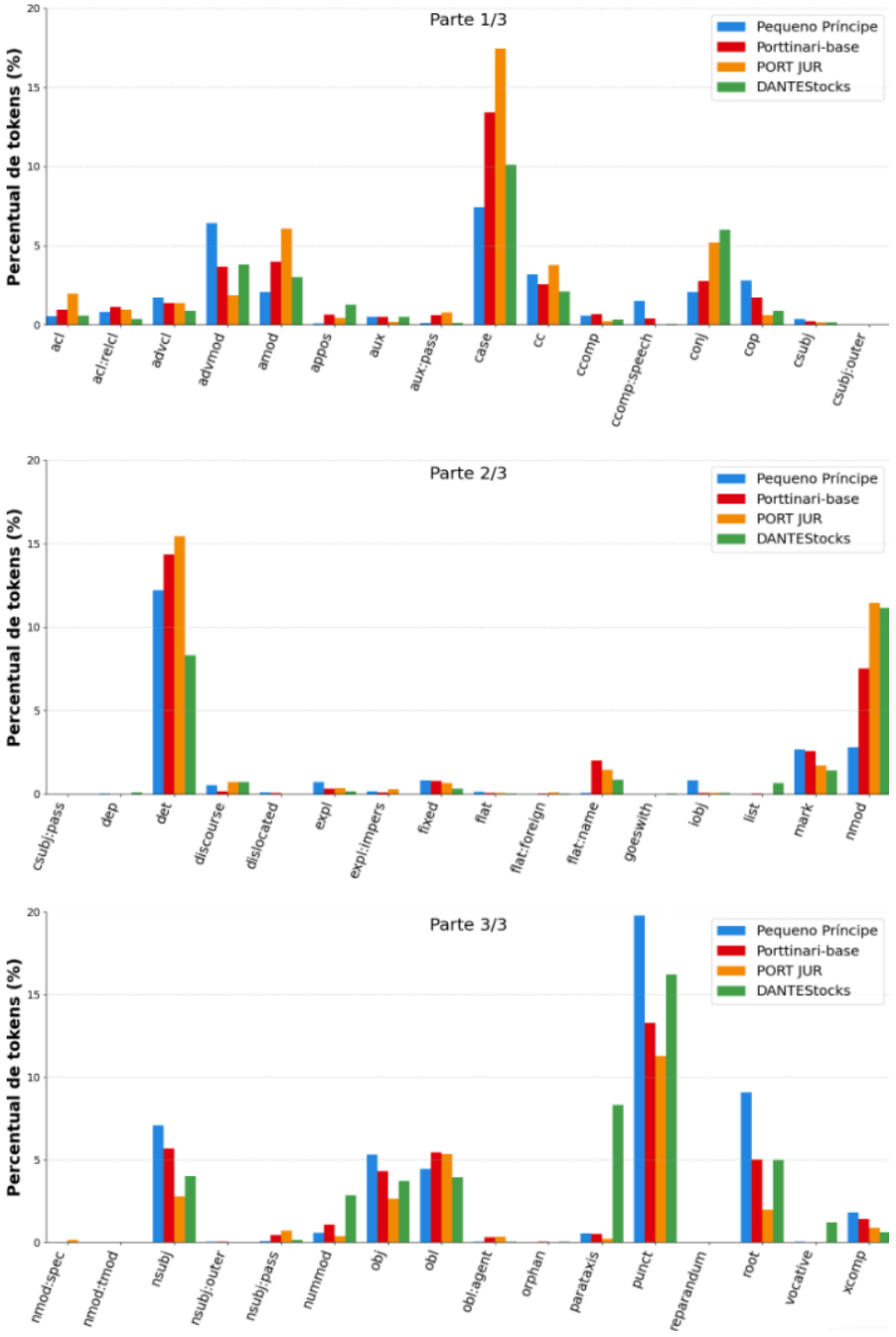

Figura 8. Distribuição das etiquetas DEPREL em diferentes córpus

Reunindo todos os dados, observamos que há correlação entre o fato de O Pequeno Príncipe ter a menor média de tamanho de sentença, e o fato de apresentar menor percentual de substantivos e modificadores nominais, menor percentual de preposições e maior percentual de verbos. Essas características conferem ao córpus maior simplicidade textual se comparado aos demais córpus de português na UD.

4. Considerações finais

A anotação sintática do córpus O Pequeno Príncipe é uma nova conquista dentro do objetivo de diversificar os gêneros disponíveis na coleção de córpus de português anotados com UD. São muitas as contribuições do gênero literário para a diversidade de textos. Uma delas é trazer muitas orações interrogativas, raras nos gêneros já anotados. Outra, é o uso expressivo da conjugação verbal de primeira e de segunda pessoas, no singular e no plural, característico do discurso direto. Além disso, duas marcas de diacronia são observadas no Pequeno Príncipe: o uso de "*tu*" e "*vós*" e o uso recorrente de pronomes clíticos, hoje pouco frequentes até no registro culto do português brasileiro.

Além de contribuir para aumentar o acervo do português em UD, o córpus O Pequeno Príncipe, por já ter sido anotado semanticamente por Anchiêta & Pardo (2018), abre a oportunidade para a exploração automática da correlação entre anotação sintática e anotação semântica. Para o leitor interessado, o córpus anotado e outras informações relevantes podem ser encontrados no portal web do projeto POeTiSA, em https://sites.google.com/icmc.usp.br/poetisa/resources-and-tools.

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 apoio do INCT TILD-IAR (Inteligência Artificial Responsável para Linguística Computacional, Tratamento e Disseminação de Informação - https://tildiar.dcc.ufmg.br/) (processo CNPq/SECTICS/CAPES/FAPs #408490/2024-1).

Referências

Anchiêta, R. and Pardo, T. (2018). Towards AMR-BR: A SemBank for Brazilian Portuguese language. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) (pp. 449–460). European Language Resources Association (ELRA).

Andrade, P. L. C. de, Silva, R. M. and Pardo, T. A. S. (2026). Caracterização lexical e sintática de notícias falsas em português produzidas por humanos e por máquinas. In Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 2 (pp. 148–158). Salvador, Brazil.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M. and Schneider, N. (2013). Abstract meaning representation for sembanking. In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse (pp. 178–186). Sofia, Bulgaria.

Di Felippo, A. and Roman, N. T. (2025). DANTEStocks: A multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. Revista Brasileira de Linguística Aplicada, 25(1), 1–32.

Duran, M. S., Nunes, M. G. V., Lopes, L. and Pardo, T. A. S. (2022). Manual de anotação como recurso de processamento de linguagem natural: O modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@gem, 16(4).

Duran, M., Lopes, L., Nunes, M. G. and Pardo, T. (2023). The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion. In Proceedings of the 14th Brazilian Symposium in Information and Human Language Technology (pp. 124–133). Belo Horizonte, Brazil.

Duran, M. S., de Souza, E. A., Nunes, M. G. V., Pagano, A. S. and Pardo, T. A. S. (2025). Extending the enhanced Universal Dependencies: Addressing subjects in pro-drop languages. In Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025) (pp. 143–152). Ljubljana, Slovenia.

Fontes, M. G. (2012). A clivagem do constituinte interrogativo em sentenças interrogativas do português brasileiro: uma abordagem diacrônica. Signum: Estudos da Linguagem, 15(3), 149–170.

Freitas, C. and Pardo, T. A. S. (2025). PropBanks e representações semânticas: O que temos, o que queremos e o que podemos. Linguamática, 17(2), 3–31.

Guibon, G., Courtin, M., Gerdes, K. and Guillaume, B. (2020). When collaborative treebank curation meets graph grammars. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 5291–5300). Marseille, France.

Leal, S. E.; Duran, M. S.; Scarton, C. E.; Hartmann, N.; Aluísio, S. M. (2023) NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. Lang Resources & Evaluation

Lopes, L., Duran, M. S. and 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). Belo Horizonte, Brazil.

Lopes, L. and Pardo, T. (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 - Vol. 1 (pp. 401–410). Santiago de Compostela, Galicia/Spain.

Lopes, L., Nunes, M. G. V., Duran, M. S. and Pardo, T. A. S. (2025). A sintaxe no tribunal: Apresentando e explorando um corpus jurídico em português anotado sintaticamente segundo o modelo Universal Dependencies. In Proceedings of the XVI Symposium in Information and Human Language Technology (STIL) (pp. 220–232). Fortaleza/CE, Brazil.

de Marneffe, M.-C., Manning, C. D., Nivre, J. and Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255–308.

Morais, M. A. T.; Salles, H. M. M. (2024). Resistência do dativo de primeira pessoa na batalha (quase) perdida dos clíticos pronominais do português brasileiro. Revista de Estudos da Linguagem, [S. l.], v. 30, n. 4, p. 1621–1657, 2024.

Oliveira, L., Claro, D. B. and Souza, M. (2023). DptOIE: A Portuguese open information extraction based on dependency analysis. Artificial Intelligence Review, 56, 7015–7046.

Othero, G. A.; Lazzari, M. (2024) Sujeitos e objetos nulos em português brasileiro: correlações e mudança. Revista de Estudos da Linguagem, [S. l.], v. 30, n. 4, p. 1831–1854, 2024.

Pagano, A. S., Rassi, A. and Pagano, A. C. S. (2026). A ordem e a função das palavras em uma sentença: Sintaxe. In H. M. Caseli and M. G. V. Nunes (Eds.), Processamento de linguagem natural: Conceitos, técnicas e aplicações em português (Vol. 1, 4th ed., chap. 6). BPLN.

Pardo, T., Duran, M., Lopes, L., Di Felippo, A., Roman, N. and Nunes, M. G. (2021). Porttinari: A large multi-genre treebank for Brazilian Portuguese. In Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology (pp. 1–10). Porto Alegre, Brazil.

Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E. and de Paiva, V. (2017). Universal Dependencies for Portuguese. In Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017) (pp. 197–206). Pisa, Italy: Linköping University Electronic Press.

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

Straka, M. (2018). UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, pp. 197-207, Association for Computational Linguistics, Stroudsburg, PA, USA.