Construções sintáticas do português que desafiam a tarefa de *parsing*: uma análise qualitativa

Magali S. Duran[1], Maria das Graças V. Nunes[1,2], Thiago A. S. Pardo[1,2]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
[2]Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (USP)
magali.duran@uol.com.br,gracan@icmc.usp.br,taspardo@icmc.usp.br

Abstract. *When used to train a parser, an annotated corpus reveals its strengths and weaknesses. Based on a qualitative analysis of the performance of a parser trained on an annotated corpus in the Universal Dependencies scheme, this paper points out some errors motivated by the non-canonical order of constituents in Portuguese: postposed subjects and determiners and anteposed adjectives. By using illustrations of syntactic trees before and after manual correction of these errors, the article aims to highlight the importance of having a reasonable number of sentences with these non-canonical structures in order to increase the probability that the parser learns to analyze them correctly.*

Resumo. *Ao ser usado para treinar um parser, um córpus anotado mostra suas qualidades e suas deficiências. Baseado em uma análise qualitativa do desempenho de um parser treinado em córpus anotado no esquema Universal Dependencies, este artigo discute alguns erros motivados pela ordem não canônica dos constituintes em Português: sujeitos e determinantes pospostos e adjetivos antepostos. Usando ilustrações de árvores sintáticas antes e depois da correção manual desses erros, o artigo tem por objetivo destacar a importância de haver uma quantidade razoável de sentenças com essas estruturas não canônicas a fim de aumentar a probabilidade de que o parser aprenda a analisá-las corretamente.*

1. Introdução

Em tempos em que o Aprendizado de Máquina (AM) é a abordagem dominante nos sistemas de Inteligência Artificial (IA) e de Processamento de Línguas Naturais (PLN), torna-se importante identificar os fenômenos que apresentam dificuldade para os algoritmos de aprendizagem. Em tarefas como *parsing* (ou seja, análise sintática automática), que alcançam precisão expressiva com essa tecnologia, é de se esperar que construções que dependam de algum conhecimento semântico ofereçam maior desafio para os algoritmos, por exemplo, os conhecidos *PP-attachments* e as coordenações em contextos com opções variadas. Entretanto, esses não são os únicos casos. Posições menos frequentes de elementos na sintaxe da língua, como sujeitos pospostos, determinantes pospostos e adjetivos antepostos, também são responsáveis por erros recorrentes.

Este trabalho tem por objetivo discutir problemas de natureza exclusivamente sintática observados durante a análise qualitativa do *parser* UDPipe 2[1] [Straka 2018]

[1] https://ufal.mff.cuni.cz/udpipe/2

treinado sobre o córpus Porttinari-base (que compõe o *treebank* Porttinari [Pardo et al. 2021]), um córpus de 8.418 sentenças (168.080 tokens), extraídas do córpus jornalístico Folha-Kaggle[2], que foram inicialmente anotadas com relações de dependências segundo o modelo *Universal Dependencies* (UD) [de Marneffe et al. 2021] [Nivre at al. 2020] e posteriormente revisadas utilizando-se como parâmetro dois manuais de anotação: o Manual de Anotação de *PoS Tags* [Duran 2021] e o Manual de Anotação de Relações de Dependência [Duran 2022].

A análise qualitativa que revelou os casos aqui relatados teve o objetivo de identificar e relatar erros recorrentes para propor melhorias no córpus de treinamento que pudessem incrementar o aprendizado automático. Mesmo pequenas melhorias são relevantes em larga escala, pois o *parser* resultante deverá ser usado para anotar automaticamente o córpus Folha-Kaggle inteiro (com 3.964.292 sentenças e 84.795,823 tokens), bem como outros córpus, de outros gêneros, visando aumentar os recursos baseados em sintaxe e desenvolver ferramentas e aplicativos para a língua portuguesa do Brasil com vistas a alcançar o estado da arte mundial nessa área. Ao leitor interessado, sugere-se a leitura do relatório contendo a análise qualitativa completa [Duran, Nunes & Pardo, 2023].

Na Seção 2, introduzimos brevemente a UD; na Seção 3, apresentamos a metodologia utilizada; na Seção 4, analisamos os resultados; na Seção 5, concluímos com algumas recomendações e trabalhos futuros.

2. *Universal Dependencies* (UD)

É importante introduzir o esquema de anotação UD, pois é o esquema que utilizaremos para ilustrar graficamente nossas anotações. Em sua versão atual, a UD possui dezessete etiquetas morfossintáticas[3] ou *Part-of-Speech* (PoS*) tags*. Também possui 37 etiquetas de relações de dependência[4] – *deprel* (de ***dep****endency* ***rel****ation*). Uma *deprel* é uma relação que liga dois a dois os elementos (*tokens*) de uma sentença tal que:

- um deles é chamado de *head* (cabeça, governante ou núcleo da relação) e o outro é chamado de **dependente**;
- um *token* pode ser *head* de mais de uma relação;
- um *token* pode ser dependente de uma relação e *head* de outra;
- um *token* **não** pode ser dependente de mais de uma relação;
- o nome da relação está sempre associado à função que o dependente desempenha em relação ao *head*;
- graficamente, uma seta parte sempre do *head* em direção ao dependente da relação;
- um *head* é sempre uma palavra de conteúdo (verbo, substantivo, adjetivo, pronome, numeral e advérbio) – exceções são símbolos que podem ser expressos por palavras, como R$ (reais), % (por cento) e § (parágrafo);

[2] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol
[3] https://universaldependencies.org/u/pos/index.html
[4] https://universaldependencies.org/u/dep/index.html

- palavras funcionais (determinantes, preposições, conjunções) e sinais de pontuação, por sua vez, deverão ser apenas dependentes e nunca *head* de relações;
- algumas relações são permitidas apenas em um sentido, enquanto outras relações são admitidas nos dois sentidos;
- quando o dependente tiver forma oracional, o elemento apontado pela seta será o núcleo do predicado da oração dependente;
- toda sentença tem uma raiz (normalmente o predicado da oração principal), marcada como dependente da *deprel* **root** – a *deprel* **root** é a única que não tem *head*, apenas dependente (é a partir do **root** que é construída a árvore sintática, da qual cada nova *deprel* constitui um galho).

A atribuição de relações de dependência deve observar o princípio da projetividade, ou seja, os arcos das relações *não devem* se cruzar. As diretrizes da UD estão disponíveis em sua *homepage*[5], assim como os mais de 200 córpus já anotados. Essas diretrizes já foram instanciadas para a língua portuguesa [Duran 2021, 2022] e há alguns córpus de português brasileiro revisados manualmente já disponíveis no site da UD, como o Bosque-UD [Rademaker et al. 2017] e o PetroGold [Souza et al. 2021].

A Figura 1 ilustra uma sentença anotada com relações de dependência UD.

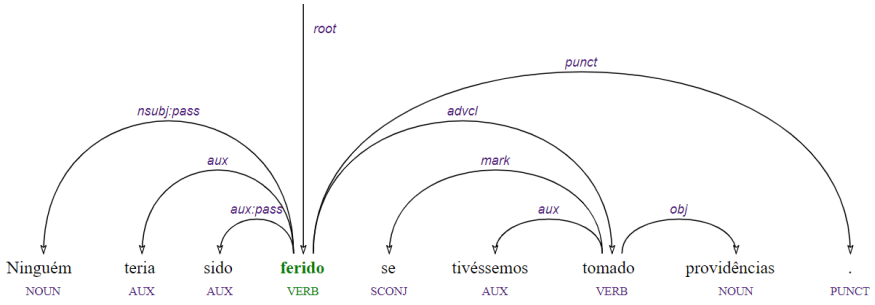

Figura 1 - Exemplo de árvore de dependências anotada com etiquetas da UD

3. Metodologia

Os dados exibidos neste trabalho são parte dos resultados de uma avaliação intrínseca na qual analisou-se o quanto o resultado do *parser* está em conformidade com as diretrizes definidas nos manuais de anotação. O conjunto avaliado consiste de uma amostra aleatória de 600 sentenças, com 12.076 tokens e tamanho médio de 20 tokens (entre 5 e 52 tokens), as quais fazem parte do Porttinari-check, um subcórpus do *treebank* Porttinari, cujos tamanhos de sentenças, *PoS tags* e *deprel* possuem distribuição similar aos do córpus Porttinari-base, sendo, por esse motivo, uma boa amostra desse córpus manualmente revisado. O Porttinari-check foi anotado automaticamente pelo *parser* UD-Pipe treinado sobre o córpus Porttinari-base. O tamanho da amostra (equivalente a 7,13% do córpus Portinari-base) não prejudica a análise qualitativa; pelo contrário,

[5] https://universaldependencies.org/

evidencia que solucionar os problemas recorrentes nesta amostra proporcionará melhorias no treinamento do córpus completo.

Primeiramente, procedeu-se à revisão manual da anotação dessas sentenças, usando a interface do Arborator-NILC[6] [Miranda e Pardo 2022]. Em seguida, foi feita a análise dos erros, buscando agrupá-los de forma lógica para construir uma tipologia de erros. Computou-se um erro para cada alteração feita pelo anotador humano, seja uma mudança só de nome da *deprel*, seja uma mudança só de *head* da *deprel*, ou seja uma mudança de nome e de *head* da *deprel*.

Na avaliação, foram computados 722 erros em 298 sentenças. Das 600 sentenças da amostra analisada, 302 (50%) estavam totalmente corretas. Com base na análise realizada, os erros foram divididos em três categorias: 1) erros provenientes de problemas de pré-processamento (tokenização, lematização, segmentação de sentenças e anotação morfossintatica); 2) erros de escolha de *head* de *deprel* (relacionados à semântica[7]); e 3) erros de natureza puramente sintática, ou seja, aqueles que não são motivados por erros de pré-processamento ou que não são dependentes de informações semânticas.

Na categoria de erros puramente sintáticos, foram identificados e agrupados: erro de identificação de sujeito; erro de **amod** (adjetivo) anteposto; erro de **det** (determinante) posposto; erro de reconhecimento de **fixed** (expressões fixas); erro de reconhecimento de **flat:name** (nomes próprios compostos) e erro de **root**. Os erros de **fixed**, **flat:name** e **root** são exclusivos do esquema de anotação UD. Por esse motivo, decidimos tirá-los do foco das reflexões apresentadas neste artigo. Na próxima seção analisamos os erros que são foco deste artigo, ou seja, aqueles que dizem respeito a ordens de elementos não canônicas na língua portuguesa.

4. Análise dos erros

Discutiremos, a seguir, os erros de sujeito e determinante pospostos e adjetivo anteposto, a fim de levantar subsídios que possam ser úteis àqueles que se dedicam a criar ferramentas automáticas para analisar a língua portuguesa.

4.1 Erro na identificação de nsubj, nsubj:pass e csubj

De modo geral, observa-se que sentenças com seus constituintes na ordem canônica do português - SVO (sujeito, verbo, objeto) - são anotadas corretamente, ao passo que sentenças com elipses de constituintes e/ou inversão da ordem canônica apresentam mais erros.

O problema na identificação do sujeito ocorreu 38 vezes na amostra analisada e compreende três *deprels*: **nsubj** (sujeito da voz ativa), **nsubj:pass** (sujeito da voz

[6] https://arborator.icmc.usp.br/

[7] Só a interpretação semântica pode determinar qual é o *head* de um *token* e, dependendo da *POS tag* do *head*, qual é a relação de dependência. Casos ambíguos incluem **acl** e **advcl** (oração adjetiva e oração adverbial), **advmod** (advérbios simples), bem como **nmod** e **obl** (modificador nominal e oblíquo, respectivamente).

passiva) e **csubj** (sujeito oracional). Para fins de interpretação da relevância desse número, é válido ressaltar que, nas 600 sentenças, houve 599 *deprels* de sujeito atribuídas, onze das quais foram atribuídas incorretamente (não eram sujeitos), ou seja, 588 sujeitos foram atribuídos corretamente. Além disso, houve 27 casos de sujeitos não identificados pelo *parser*. Tanto a identificação incorreta quanto a não identificação do sujeito estão relacionadas, principalmente, à ocorrência do sujeito posposto ao verbo, ou seja, em construções do tipo VS (Verbo Sujeito) ou OVS (Objeto Verbo Sujeito), que, embora sejam ordens não canônicas no português, apresentam frequência significativa no córpus.

Um dos casos comuns de sujeito posposto na voz ativa ocorre com verbos intransitivos, como “estrear”, “faltar”, “sobrar”, “restar”, “ocorrer”, “bastar” e “existir”, e transitivos indiretos, como “caber” e “constar”. Verbos que admitem o papel semântico de tema na posição de sujeito também costumam apresentar sujeitos pospostos, principalmente na negativa, como “não interessa X” e “não importa X”, em que X tem papel de sujeito.

A Figura 2 traz o exemplo de um verbo intransitivo, “sobrar”, e a Figura 3 traz o exemplo de um verbo transitivo indireto, “constar”, em locução verbal com o modal “dever”, que é o *head* no **nsubj**, por ser o verbo que concorda com o sujeito.

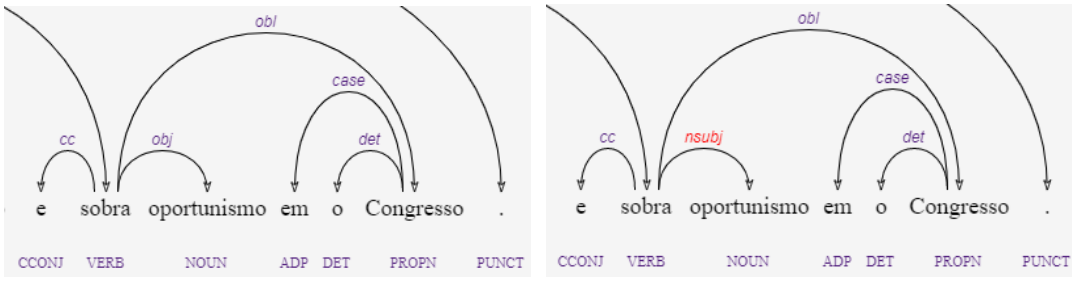

Figura 2 - Sentença com verbo intransitivo anotada pelo *parser* (esq) e corrigida (dir)

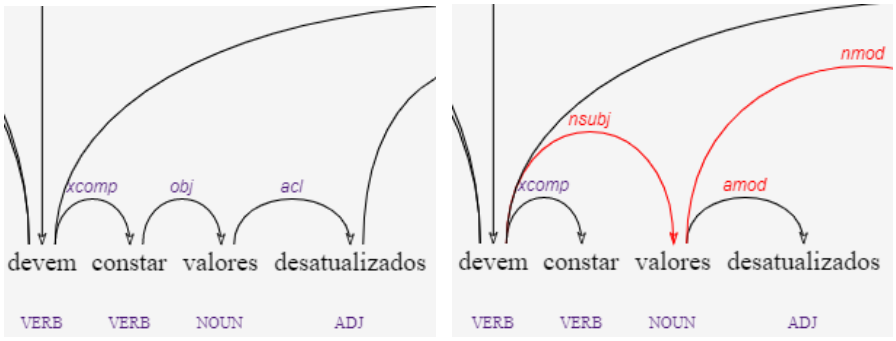

Figura 3 - Sentença com verbo trans. indireto anotada pelo *parser* (esq) e corrigida (dir)

Outro caso muito frequente de sujeito posposto é com verbos na voz passiva analítica, principalmente quando a oração é reduzida e o verbo auxiliar de passiva está elíptico, como em “assim que [for] recebida a denúncia” (Figura 4)[8].

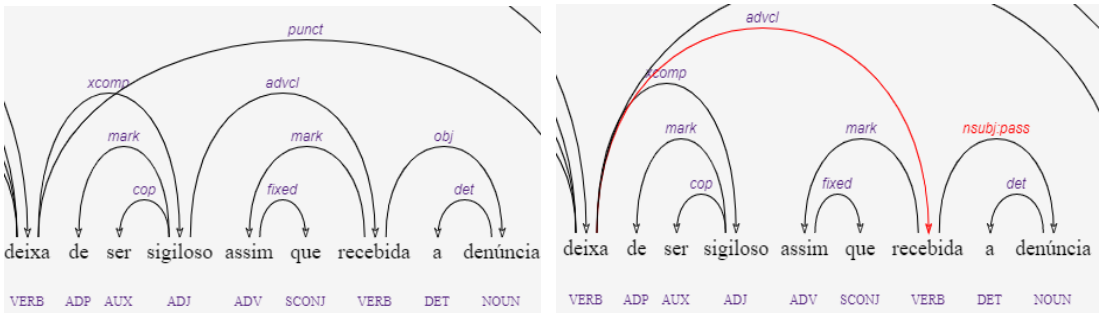

Figura 4 - Sentença na voz passiva analítica anotada pelo *parser* (esq) e corrigida (dir)

Também é frequente o sujeito posposto com verbos na voz passiva sintética, como mostrado na Figura 5.

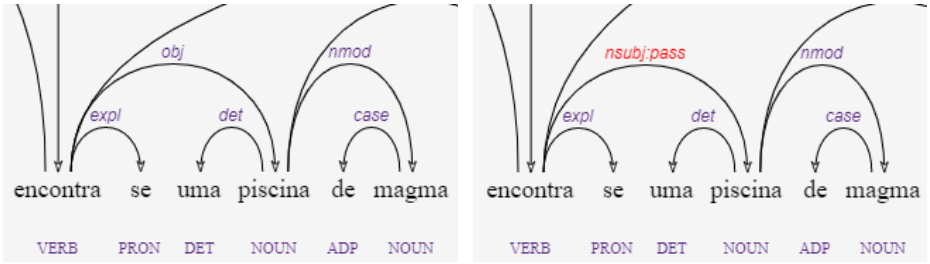

Figura 5 -Sentença na voz passiva sintética anotada pelo *parser* (esq) e corrigida (dir)

Além desses casos de predicados verbais, há também casos de predicados nominais com sujeito posposto. Nesses casos, o verbo de cópula inicia a oração, sendo seguido pelo predicativo e pelo sujeito, na maioria das vezes um sujeito oracional (**csubj**). No geral, o *parser* aprendeu muito bem a identificar esses sujeitos, por isso são raros erros como o ilustrado na Figura 6 e corrigido na Figura 7.

[8] Curiosamente, todos os sujeitos pospostos recorrentes têm papel semântico de tema, papel que também é comumente atribuído ao objeto dos verbos transitivos diretos.

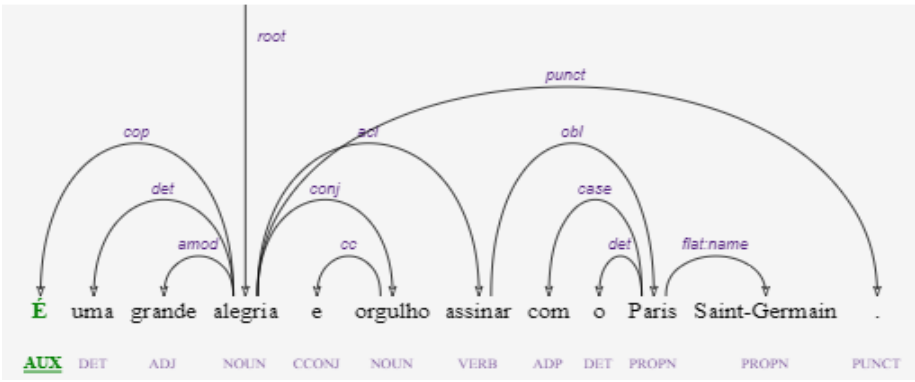

Figura 6 - Sentença com sujeito posposto anotado pelo *parser*

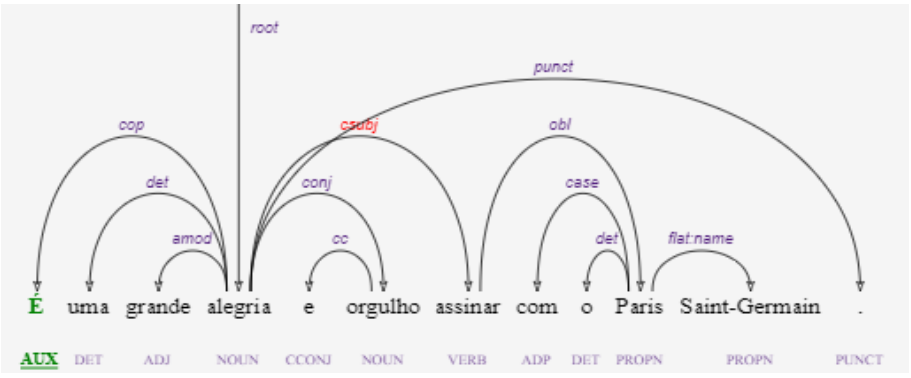

Figura 7 - Sentença com sujeito posposto corrigido manualmente

No português, o *parser* tem que lidar com a possibilidade de o sujeito estar elíptico, ou seja, casos em que a *deprel* **nsubj** não será usada. Quando há um candidato a sujeito à esquerda do verbo, o *parser* quase sempre acerta a atribuição. Entretanto, quando não há um candidato à esquerda e há um candidato à direita, o *parser* se confunde, pois sintagmas nominais à direita podem ser objeto ou sujeito. A Figura 8[9] ilustra uma confusão desse tipo em que o *parser* atribuiu **nsubj** a um token que é **obj**.

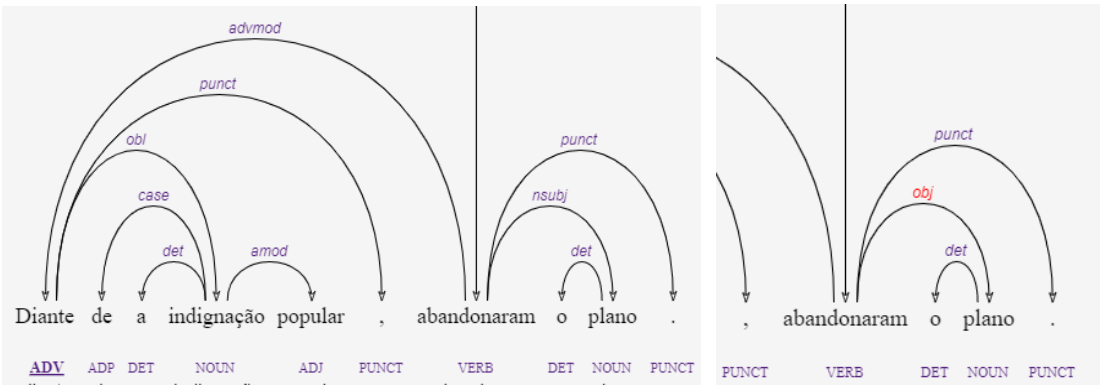

Figura 8 - Sentença com sujeito elíptico, anotada pelo *parser* (esq) e corrigida (dir)

[9] Não se anotou "diante de" como expressão fixa com valor de preposição, mas sim como advérbio predicativo, porque é possível inserir palavras entre as partes (ex: "diante não apenas de x, mas também de y").

4.2 Erro na anotação de modificador amod anteposto

Em português, os adjetivos que modificam substantivos (*deprel* **amod**) podem ocorrer tanto à esquerda quanto à direita (antepostos ou pospostos). Contudo, a ocorrência posposta é extremamente mais frequente. Casos de dois adjetivos, um anteposto e um posposto ao substantivo modificado, costumam confundir o *parser*, mesmo com a anotação correta de *PoS tags*, como pode ser visto na Figura 9, que apresenta um substantivo com terminação típica de adjetivo: “contencioso”.

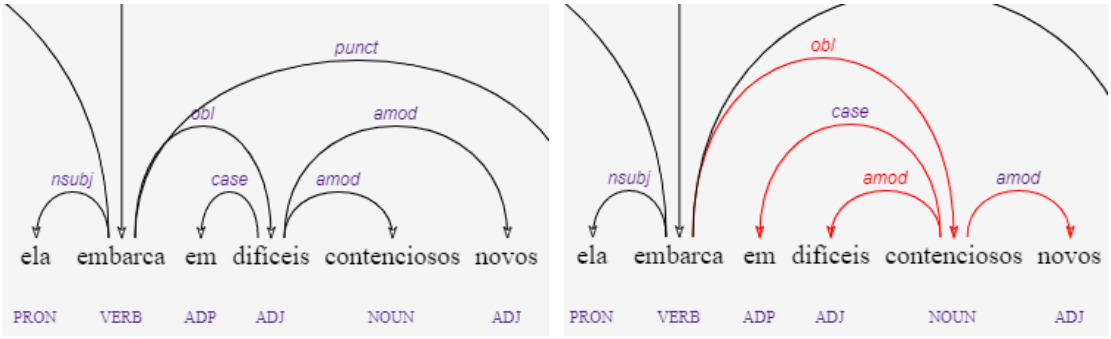

Figura 9 - Sentença com adjetivo anteposto anotada pelo *parser* (esq) e corrigida (dir)

O problema ocorreu pouco na amostra avaliada: seis vezes, incluindo o caso já mencionado de “difíceis contenciosos novos”. São exemplos de **amod** anteposto não identificado pelo *parser* na amostra: “**demasiados** entraves”, “**exímio** pianista”, “**enorme** e **exagerada** reação negativa”, “**impressionante** simulação de violência” e “**legítimo** interesse”. A possibilidade de um ADJ ocorrer anteposto é uma característica lexical e, por isso, enriquecer o córpus com exemplos de ADJ que admitem anteposição pode melhorar o aprendizado automático.

4.3 Erro na anotação de modificador det posposto

Esse tipo de erro é raro (ocorreu uma única vez na amostra), pois em português os determinantes (**det**) ocorrem majoritariamente antes do substantivo. Entretanto, seria interessante ter mais dados de **det** posposto (*deprel* **det** com sentido da esquerda para a direita) para que o *parser* aprenda a anotá-lo, como o caso ilustrado na Figura 10, que apresenta o pronome intensificador "mesmo"[10]. Embora pouco frequente no córpus, é muito normal um **det** posposto, como em “Espelho *meu*”, “proposta *esta* que…”, “ele *próprio*”, “eu *mesmo*” e “eles *todos*”.

[10] Os pronomes que modificam substantivos são anotados como determinantes na UD. A palavra "mesmo", em outros contextos, poderia ser pronome substantivo ou advérbio, situações em que receberia outra anotação.

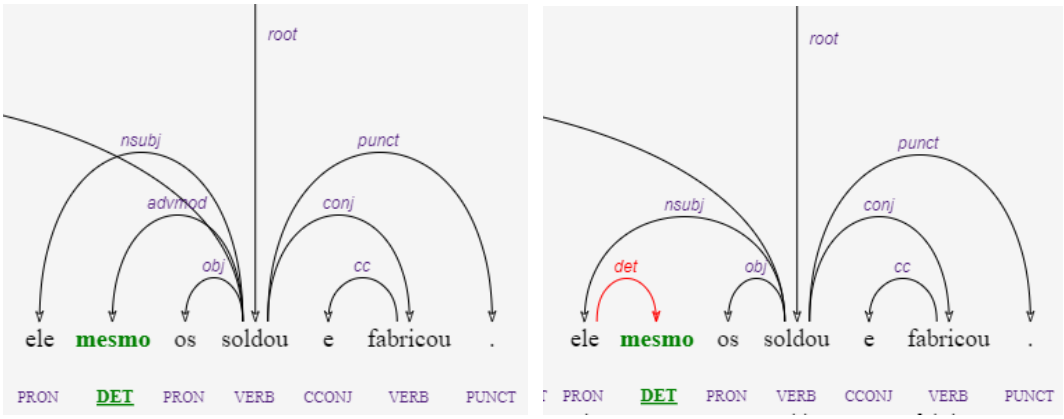

Figura 10. Sentença com det. posposto anotada pelo *parser* (esq) e corrigida (dir)

5. Conclusões

Sob o ponto de vista qualitativo, o *parser* avaliado apresentou excelente desempenho, o que deverá ser ratificado por resultados quantitativos assim que sua versão final for divulgada. Acredita-se que os problemas de natureza exclusivamente sintática possam ser minimizados se o córpus de treinamento vier a ser aumentado usando técnicas recentes de aumento de dados (*data augmentation)* para diminuir a esparsidade de alguns fenômenos [Shorten & Khoshgoftaar, 2019], como sujeitos pospostos (relações **nsubj** da esquerda para a direita), adjetivos antepostos (relações **amod** da direita para a esquerda) e determinantes pospostos (relações **det** da esquerda para a direita). Outro caminho interessante pode ser a proposta de sistemas simbólicos de pós-edição de árvores sintáticas produzidas automaticamente, visando a sua correção.

O estudo apresentado neste artigo deve subsidiar novas iniciativas sintáticas para o processamento computacional do português, mas não apenas isso. Acredita-se que possa também ser a base para outros estudos linguísticos que visem explorar questões de frequência dos fenômenos discutidos ou mesmo explorar tais fenômenos no contexto de ensino da língua portuguesa.

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.

Referências bibliográficas

1. Duran, M.S. (2021) “Manual de Anotação de *PoS tags*: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD)”. Relatório Técnico do ICMC 434. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, Setembro, 55p.
2. Duran, M.S. (2022) “Manual de Anotação de Relações de Dependência –Versão Revisada e Estendida”. Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Outubro, 166p.
3. Duran, M.S.; Nunes, M.G.V.; Pardo, T.A.S. (2023). Avaliação qualitativa do analisador sintático UDPipe 2 treinado sobre o córpus jornalístico Porttinari-base. Relatório Técnico do ICMC 442. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Abril, 58p.
4. de Marneffe, M.; Manning, C.; Nivre, J.; Zeman, D. (2021) “Universal Dependencies”, In: Computational Linguistics 47 (2). MIT PRESS, p. 255-308.
5. Miranda, L.G.M.; Pardo, T.A.S. (2022) “An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction”, In: Proceedings of the PROPOR Demonstrations Workshop, p.1-3.
6. Nivre, J.; de Marneffe, M.; Ginter, F.; Hajič, J.; Manning, C.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020) “Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection”. In: Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC 2020), p. 4034-4043.
7. Pardo, T.A.S.; Duran, M.S.; Lopes, L.; Di Felippo, A.; Roman, N.T.; Nunes, M.G.V. (2021) “Porttinari - A large multi-genre treebank for Brazilian Portuguese”. In: Proceedings of the XIV Symposium in Information and Human Language (STIL 2021), p. 1-10.
8. Rademaker, A.; Chalub, F.; Real, Livy; Freitas, C.; Bick, E.; Paiva, V. (2017) “Universal Dependencies for Portuguese”. In: Proceedings of the Fourth International Conference on Dependency Linguistics. Linköping University Electronic Press, p. 197-206.
9. Shorten, C., & Khoshgoftaar, T. M. (2019). “A survey on Image Data Augmentation for Deep Learning”. In: Journal of Big Data, 6(1), 60.
10. Straka, M. (2018) “UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task”. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium: Association for Computational Linguistics, p. 197-207.