UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

ISSN 0103-2569

AVALIAÇÃO QUALITATIVA DO ANALISADOR SINTÁTICO UDPIPE 2 TREINADO SOBRE O CÓRPUS JORNALÍSTICO PORTTINARI-BASE

MAGALI SANCHES DURAN
MARIA DAS GRAÇAS VOLPE NUNES
THIAGO ALEXANDRE SALGUEIRO PARDO

Nº 442

RELATÓRIOS TÉCNICOS

São Carlos – SP
Abr./2023

POeTiSA

POrtuguese processing – Towards Syntactic Analysis and parsing

Avaliação qualitativa do analisador sintático UDPipe 2 treinado sobre o córpus jornalístico Porttinari-base

Magali Sanches Duran
Maria das Graças Volpe Nunes
Thiago Alexandre Salgueiro Pardo

Abril de 2023

Índice

1. Introdução **2**
2. Metodologia **4**
3. Análise dos Resultados **7**
3.1 Erros provenientes de problemas de pré-processamento 10
3.1.1 Erros de Sentenciação 10
3.1.2 Textos sem sintaxe ou em forma não oracional 11
3.1.3 Erros de lematização e PoS tagging 13
3.2 - Erros de escolha de head 16
3.2.1 Erro de head de nmod 18
3.2.2 Erro de head de obl 19
3.2.3 Erro de head de acl 22
3.2.4 Erro de head de advcl 24
3.2.5 Erro de head de conj 25
3.2.6 Erro de head de modificadores de elementos coordenados 28
3.3 Erros de natureza sintática 30
3.3.1 Erro na atribuição do root da árvore sintática 31
3.3.2 Erro na identificação de nsubj, nsubj:pass e csubj 33
3.3.3 Erro na anotação de modificador amod anteposto 41
3.3.5 Erro na identificação de expressões fixed 44
3.3.6 Erro na identificação de flat:name 46
4. Conclusões **49**
Agradecimentos **55**
Referências Bibliográficas **56**

1. Introdução

O presente relatório integra o conjunto de trabalhos desenvolvidos no âmbito do projeto POeTiSA (*POrtuguese processing - Towards Syntactic Analysis and parsing*), que faz parte da iniciativa de Processamento de Línguas Naturais (NLP2 - *Natural Language Processing for Portuguese*) do Centro de Inteligência Artificial (C4AI - *Center for Artificial Intelligence*) da Universidade de São Paulo, financiado pela IBM e pela FAPESP (projeto nr. 2019/07665-4). Este projeto também é 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 Residência em TIC 13, DOU 01245.010222/2022-44.

O POeTiSA é um projeto de longo prazo que visa aumentar os recursos baseados em sintaxe e desenvolver ferramentas e aplicações relacionadas à língua portuguesa do Brasil, visando alcançar resultados de ponta nessa área, incluindo a produção de um córpus multigênero grande e abrangente, anotado segundo o modelo *Universal Dependencies* (UD) (Marneffe et al., 2021; Nivre at al., 2020).

O primeiro córpus anotado com relações de dependências UD no Projeto POeTiSA é o Porttinari-base (um dos integrantes do grande córpus Porttinari, descrito por Pardo et al., 2021), um córpus de 8.418 sentenças (168.075 tokens) extraídas do córpus Folha-Kaggle[1]. Esse córpus, anotado preliminarmente com o UDPipe treinado no córpus Bosque (Rademaker et al., 2017), foi revisado manualmente utilizando-se como parâmetro dois manuais de anotação: o Manual de Anotação de PoS tags (Duran, 2021[2]) e

[1] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol
[2] https://drive.google.com/file/d/1BddPswn-_Ioo-A5GsldA1cO1kqbcCahb/view?usp=sharing

o Manual de Anotação de Relações de Dependência (Duran, 2022[3]). A anotação está em conformidade com as mais recentes diretrizes da UD, divulgadas em maio de 2022. O conjunto de etiquetas da UD está descrito em pormenores nesses manuais técnicos de anotação e, por isso, não será descrito aqui.

Este relatório técnico tem por objetivo descrever especificamente a análise qualitativa do analisador sintático automático (parser) UDPipe 2[4] (Straka, 2018) treinado sobre o córpus Porttinari-Base. A avaliação quantitativa do parser, usando métricas consagradas em Processamento de Línguas Naturais (PLN), está sendo feita paralelamente e será divulgada oportunamente.

A avaliação qualitativa aqui relatada foi realizada a partir de um conjunto de 1.684 sentenças. Essas sentenças constituem um *testbed* (conjunto de dados para avaliação), equivalendo a 20% do tamanho do córpus de treinamento, e apresentam a mesma distribuição de tamanhos de sentenças desse córpus.

O objetivo desta análise qualitativa é identificar os erros recorrentes e relatá-los, de modo que futuros usuários tenham mais elementos para prever o comportamento do parser em suas aplicações. Como objetivo secundário, a análise visa propor melhorias no córpus de treinamento que possam incrementar o aprendizado automático e evitar, em futuras novas versões do parser, muitos dos erros observados . Isso é muito relevante para o projeto POeTiSA, pois o parser será usado para anotar automaticamente outros córpus, de outros gêneros, que serão revisados por humanos, bem como o córpus Folha-Kaggle inteiro (com 3.611.476 sentenças e 102.996.263 tokens) que, obviamente, não será revisado por humanos.

[3] https://drive.google.com/file/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view?usp=sharing
[4] https://ufal.mff.cuni.cz/udpipe/2

2. Metodologia

A avaliação empreendida é intrínseca, ou seja, analisa-se o quanto o resultado do parser está em conformidade com as diretrizes definidas nos manuais de anotação.

Foram montados, aleatoriamente, 30 pacotes de 20 sentenças extraídas do *testbed* de 1.684 sentenças, ou seja, uma amostra de 600 sentenças. Essa amostra de 600 sentenças contém 12.076 tokens, o que resulta em um tamanho médio de sentença de 20 tokens (o tamanho máximo é de 52 tokens e o mínimo é de 5 tokens), como pode ser observado no gráfico da Figura 1.

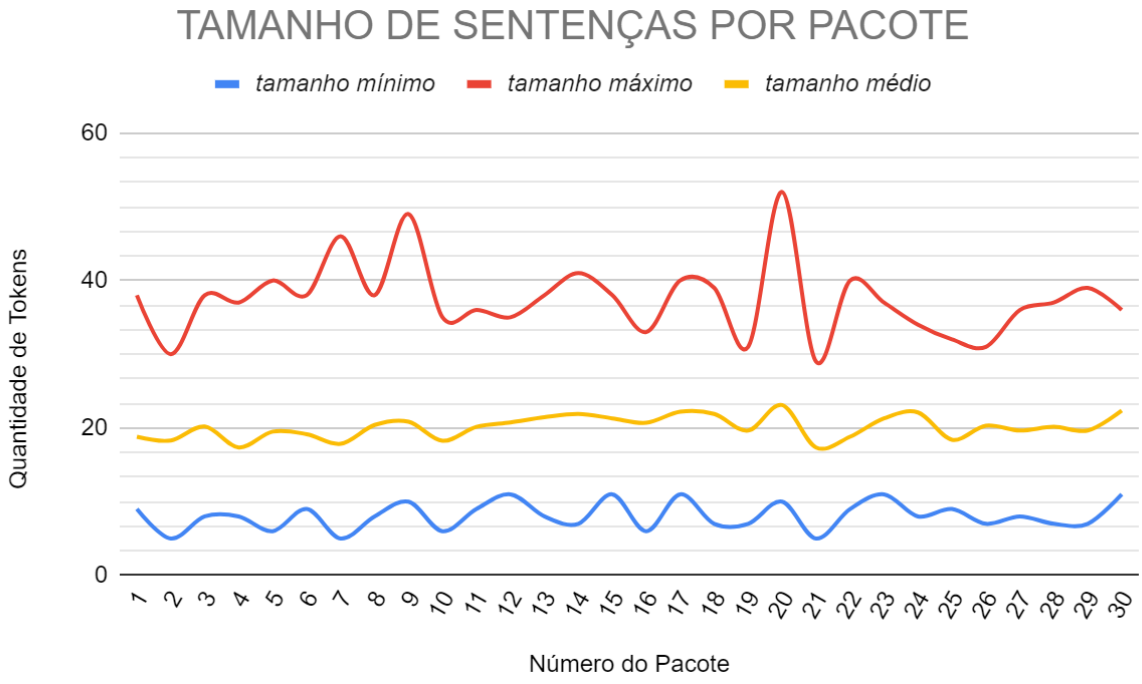


Figura 1. Tamanho mínimo, máximo e médio de sentença por pacote

Esses pacotes foram revisados manualmente, usando a interface do Arborator-NILC[5] (Miranda e Pardo, 2022) e, em seguida, procedeu-se à análise dos erros, buscando agrupá-los de forma lógica para construir uma tipologia de erros.

É interessante destacar a motivação para a construção de pacotes aleatórios distintos de avaliação (em vez de um único grande pacote de avaliação). A proposta dos pacotes é tentar mensurar com que sistematicidade os erros acontecem. Assim, ao usar o parser, usuários podem esperar encontrar erros que ocorrem em todos ou em quase todos os pacotes. Erros que ocorrem em alguns ou nenhum pacote devem ser menos frequentes também para os usuários do parser.

Primeiramente, foram contados todos os erros. Computou-se um erro para cada alteração de arco sem alteração de deprel (forma reduzida utilizada normalmente para se referir ao termo *dependency relation*), ou alteração de deprel sem alteração de arco (ou seja, a aresta entre as palavras conectadas pela deprel), ou alteração de arco e deprel simultaneamente. Sob o ponto de vista qualitativo, contudo, julgou-se mais relevante descrever erros genuínos, desconsiderando os erros acarretados por eles.

Por exemplo, diante do não reconhecimento de uma expressão **fixed**, vários erros podem ser acarretados, dependendo de quantos tokens a expressão contém, porém, em essência, trata-se de um único erro de reconhecimento de expressão **fixed**. Por esse motivo, uma nova contagem foi realizada, computando apenas os tipos de erros principais, com suas respectivas categorias, e não os erros acarretados.

A ideia de não ter uma tipologia de erros *a priori* foi muito importante para não influenciar a análise com expectativas construídas em experiências passadas.

---

5 https://www.google.com/url?q=https%3A%2F%2Farborator.icmc.usp.br%2F&sa=D&sntz=1&usg=AOvVaw06E4ct_ovM2CgvxSeSnqfP

Foi importante também separar a fase de identificação e correção dos erros da fase de análise dos erros, pois, ao anotar, cada erro parece ser único, ao passo que, ao analisar todos de uma vez, as semelhanças "saltaram aos olhos" e generalizações tornaram-se possíveis.

Os 30 pacotes corrigidos, contendo 600 sentenças, serão disponibilizados juntamente com as 1.684 sentenças que constituem o *testbed* para parsers treinados no Porttinari-base. Para referência, esse *testbed* é chamado de Porttinari-check. Juntos, o Porttinari-base, o Porttinari-check e o restante do córpus Folha-Kaggle anotado automaticamente (chamado de Porttinari-automatic) devem compor a porção jornalística do grande treebank Porttinari (Pardo et al., 2021).

3. Análise dos Resultados

Usando os critérios descritos na seção anterior, foram computados 722 erros, em 298 sentenças. Das 600 sentenças da amostra analisada, 302 (50%) estavam totalmente corretas. Nas 298 sentenças que apresentaram erros, observou-se que a moda estatística é de 1 erro somente (137 sentenças, ou seja, 22,8% do total de sentenças avaliadas).

A Tabela 1 traz a quantidade de erros em cada uma das 20 sentenças dos 30 pacotes analisados (em destaque as sentenças com 3 ou mais erros).

Tabela 1. Quantidade de erros por sentença nos 30 pacotes

Sentença	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Quant. de erros por pacote	Quant. de sentenças sem erros
pacote 01	1	0	0	4	0	0	1	0	1	1	1	0	1	1	1	0	0	5	1	9	27	8
pacote 02	0	0	1	0	8	0	2	0	0	0	0	0	1	0	0	3	0	0	3	0	18	14
pacote 03	0	0	0	0	0	1	0	2	2	0	1	1	2	0	1	0	0	1	2	0	13	11
pacote 04	1	0	0	3	0	1	1	0	2	0	0	2	0	0	0	1	1	4	0	1	17	10
pacote 05	2	1	0	3	2	2	8	2	2	0	0	2	0	0	3	0	0	0	0	1	28	9
pacote 06	3	0	0	0	0	1	0	0	2	0	1	0	0	4	0	0	0	0	8	1	20	13
pacote 07	0	0	2	0	0	0	0	0	0	0	1	0	1	1	0	0	2	0	3	2	12	13
pacote 08	1	0	0	1	0	1	1	1	1	0	0	1	0	0	3	1	0	0	0	0	11	11
pacote 09	2	0	1	0	0	0	1	2	0	1	0	0	0	0	0	1	1	1	3	4	17	10
pacote 10	0	6	2	1	0	0	1	0	1	5	0	0	4	0	0	0	1	1	0	0	22	11
pacote 11	0	0	3	1	0	0	2	0	0	1	0	5	1	0	0	0	1	1	0	0	15	12
pacote 12	0	0	1	1	1	0	0	1	0	0	0	4	0	0	0	2	0	0	0	2	12	13
pacote 13	1	2	5	0	2	3	3	1	0	0	0	0	2	1	1	0	0	1	3	2	27	7
pacote 14	1	0	1	7	4	2	0	0	0	1	0	0	2	0	2	6	2	2	3	3	36	7
pacote 15	1	1	0	4	0	0	0	1	1	2	0	0	3	0	0	2	2	1	1	2	21	8
pacote 16	7	0	4	5	1	0	5	0	0	0	0	1	5	1	0	5	0	1	1	0	36	9
pacote 17	0	1	0	0	0	0	0	0	0	0	3	0	0	0	3	0	0	1	0	0	8	16
pacote 18	0	5	4	1	6	1	0	1	0	0	0	3	1	0	1	1	1	2	1	2	30	6
pacote 19	7	0	0	1	1	1	0	1	0	8	2	6	1	1	0	2	1	0	1	0	33	7
pacote 20	0	3	0	1	0	0	0	1	0	1	0	0	0	1	0	1	1	3	5	2	19	10
pacote 21	0	3	5	0	2	0	3	0	1	0	0	0	4	2	2	0	0	0	0	0	22	12
pacote 22	4	0	0	3	2	0	0	1	0	0	4	5	2	2	2	1	0	0	0	2	28	9

pacote 23	1	2	2	0	8	0	1	1	1	3	0	0	3	8	0	0	7	0	1	0	38	8
pacote 24	9	0	12	0	0	0	0	0	3	0	0	0	1	2	0	9	1	6	1	0	44	11
pacote 25	2	1	0	3	0	3	2	0	0	4	1	0	2	2	3	0	2	0	0	1	26	8
pacote 26	1	0	0	0	7	8	0	0	7	0	0	0	0	0	0	1	3	4	1	1	33	11
pacote 27	2	2	0	0	2	1	1	0	0	3	0	0	1	1	0	0	0	1	12	0	26	10
pacote 28	0	1	2	1	0	0	2	1	0	0	0	3	0	0	1	2	0	0	2	1	16	10
pacote 29	0	1	0	8	2	3	3	1	0	5	0	2	0	2	1	0	0	0	1	0	29	9
pacote 30	0	1	0	1	0	4	3	1	3	3	9	0	0	0	11	1	3	0	0	0	40	9
Totais																					724	302

Obs. Média de erros por sentença na amostra = 1,2 e desvio-padrão = 1,9

Na Tabela 1, as linhas discriminam os pacotes de 1 a 30 e as colunas discriminam as sentenças, numeradas sequencialmente de 1 a 20. Na última coluna é apresentado o total de erros e a quantidade de sentenças sem erro de cada pacote. Esses dados foram levantados manualmente durante a avaliação.

Nas sentenças que apresentaram 3 ou mais erros (destacadas em azul), observou-se uma diferença de tipos de erro: havia os erros dos tipos observados nas sentenças com 1 ou 2 erros, porém havia também casos de erros acarretados por outros erros. Por exemplo, erros de **root** (ou seja, aqueles em que a raiz da árvore sintática não é corretamente identificada), erros de identificação de expressões **fixed** ou erros de pré-processamento acabavam acarretando outros erros, pelo fato de a correção exigir várias mudanças de head de relações (o termo "head", amplamente adotado na área de pesquisa, é utilizado para indicar o elemento cabeça/dominante envolvido em uma relação, ou seja, aquele que "governa" o outro termo).

As Figuras 2a e 2b ilustram um caso assim: embora haja 6 erros computados pelos critérios estabelecidos, os erros primários que os causaram são apenas 2: erro de reconhecimento da expressão **fixed** "de vez em quando", que funciona como **advmod**, e erro de atribuição do **root** ao verbo da sentença, "faz". É importante observar que, para referência, os exemplos mostrados contêm a identificação da sentença no córpus (por

exemplo, na Figura 2a, a sentença mostrada é aquela identificada por FOLHA_DOC005019_SENT030).

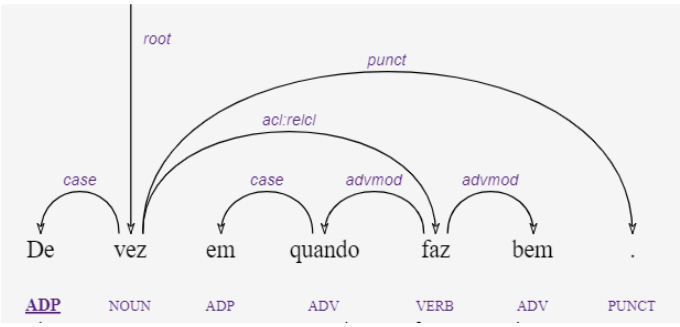


Figura 2a. FOLHA_DOC005019_SENT030 - anotada pelo parser

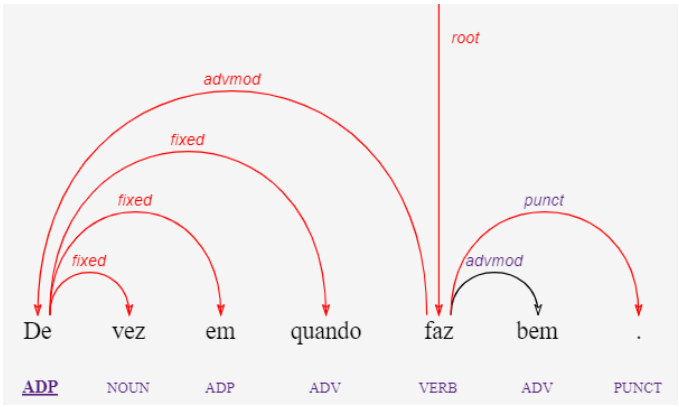


Figura 2b. FOLHA_DOC005019_SENT030 - corrigida manualmente

Sendo assim, na análise subsequente, por tipos de erros, concentrou-se em erros primários, desprezando-se os erros acarretados.

Com base na análise realizada, os erros foram divididos em três categorias, sendo apenas uma delas constituída de erros puramente sintáticos:

- Erros provenientes de problemas de pré-processamento (tokenização, lematização, sentenciação e PoS *tagging*);
- Erros de escolha de head de deprel (relacionados à semântica);
- Erros de natureza genuinamente sintática.

A seguir, essas três categorias são detalhadas e os erros que cada uma apresenta são exemplificados com sentenças extraídas do material avaliado.

3.1 Erros provenientes de problemas de pré-processamento

Houve erros de *parsing* que provavelmente foram decorrentes de erros de pré-processamento, ou seja, falhas nas fases de sentenciação, tokenização, lematização ou PoS *tagging*. Embora esses tipos de erro não sejam objeto desta análise, é relevante mencioná-los a fim de ressaltar a importância do pré-processamento para o desempenho do parser.

3.1.1 Erros de Sentenciação

O principal erro de sentenciação que ocorre no córpus é a presença de título ou subtítulo agregado ao início da sentença, característica comum em córpus jornalísticos como o Folha-Kaggle, fonte das sentenças do Porttinari-base. Infelizmente, o córpus Folha-Kaggle não possui marcação de títulos e subtítulos, o que poderia facilitar a extração exclusiva de sentenças. Como os títulos e subtítulos não são separados da sentença por pontos, seria necessário que o sentenciador reconhecesse o início de cada sentença pelo uso de maiúsculas, o que é uma heurística limitada, dada a possível ocorrência de nomes próprios nessa posição.
Dado que títulos e subtítulos frequentemente não são oracionais e contêm pelo menos um NP (*Noun Phrase* ou, em português, sintagma nominal), o parser interpreta esse NP como sujeito, simplesmente pelo fato de estar à esquerda do verbo. As Figuras 3a e 3b mostram, respectivamente, os erros provocados pela presença do título na sentença e sua correção (em vermelho).

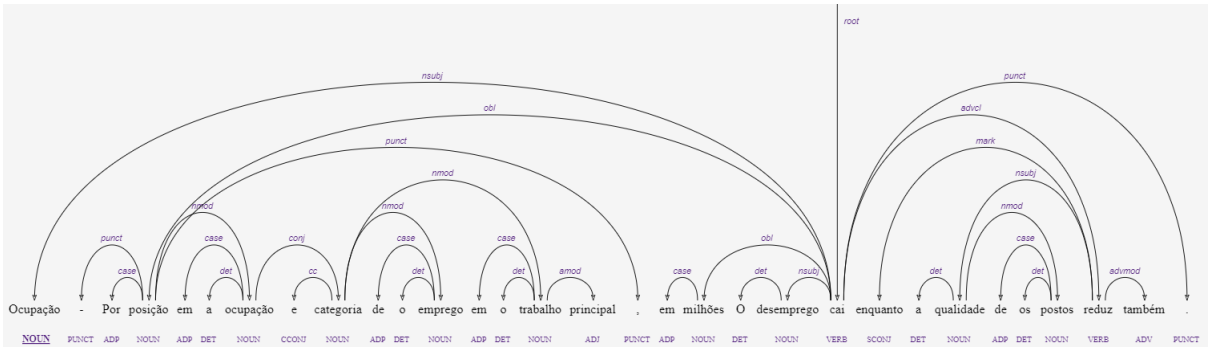


Figura 3a. FOLHA_DOC005028_SENT010 - anotada pelo parser

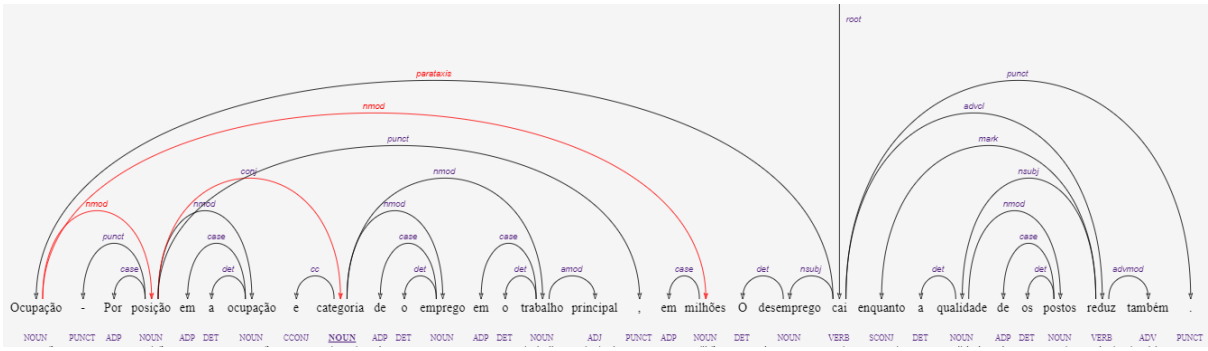


Figura 3b. FOLHA_DOC005028_SENT010 - corrigida manualmente

A UD orienta a anotação de segmentos não relacionados sintaticamente à sentença com a deprel **parataxis**, mas nem sempre o parser consegue fazer essa distinção, principalmente se a posição de sujeito à esquerda do verbo estiver desocupada e o segmento não relacionado é um candidato sintaticamente apto a ocupar a posição de sujeito.

3.1.2 Textos sem sintaxe ou em forma não oracional

Há, entre as sentenças, textos que não apresentam estrutura em forma de orações, o que torna difícil estabelecer relações sintáticas entre suas partes. Um desses casos (um anúncio) é ilustrado nas Figuras 4a e 4b, que apresentam, respectivamente, a anotação do parser e a anotação corrigida manualmente.

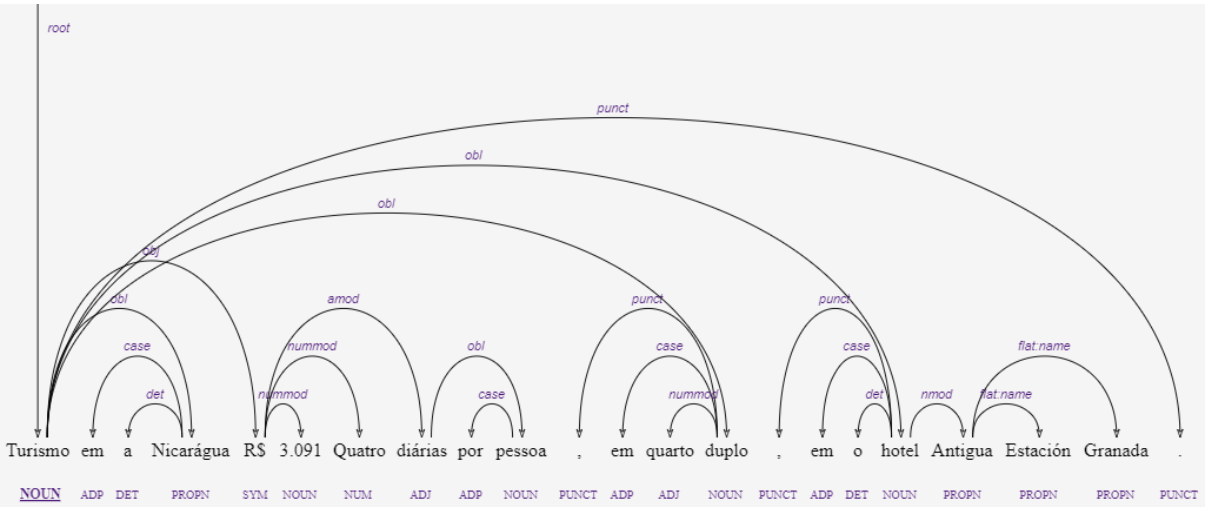


Figura 4a. FOLHA_DOC005054_SENT018 - anotada pelo parser

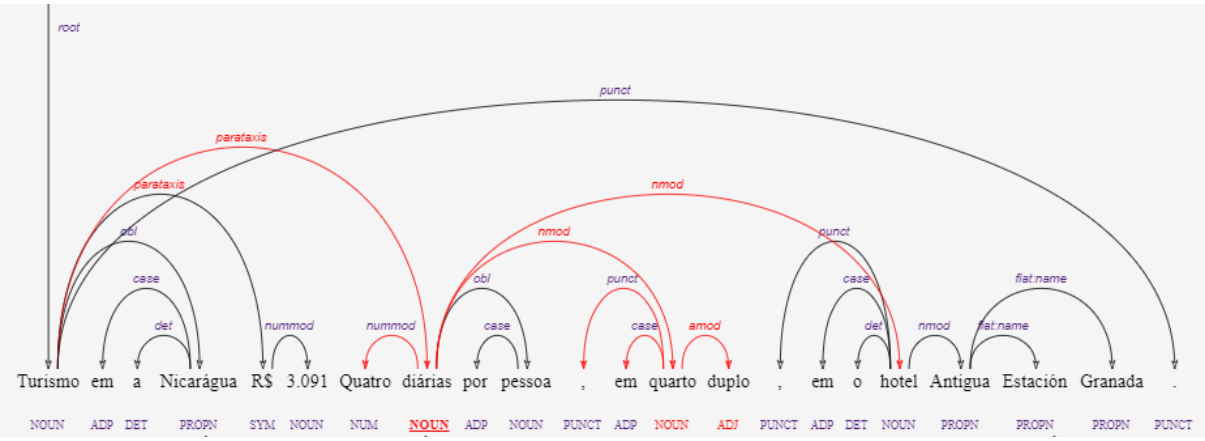


Figura 4b. FOLHA_DOC005054_SENT018 - corrigida manualmente

Conteúdos não oracionais podem ser adequadamente analisados pelo parser, desde que haja sintaxe entre suas partes (como grandes sintagmas nominais, por exemplo). Contudo, se o conteúdo for desprovido de sintaxe, o parser apresentará problemas de desempenho, como ilustrado anteriormente.

3.1.3 Erros de lematização e PoS *tagging*

Se o lema e/ou a PoS tag de uma palavra da sentença estiverem incorretos, o parser também poderá incorrer em erro. É o caso da sentença ilustrada na Figura 5a, na qual a palavra “termos” foi lematizada como forma do verbo “ter” e recebeu a PoS tag VERB, quando o correto seria o lema “termo” e a PoS tag NOUN, como mostrado em vermelho na Figura 5b.

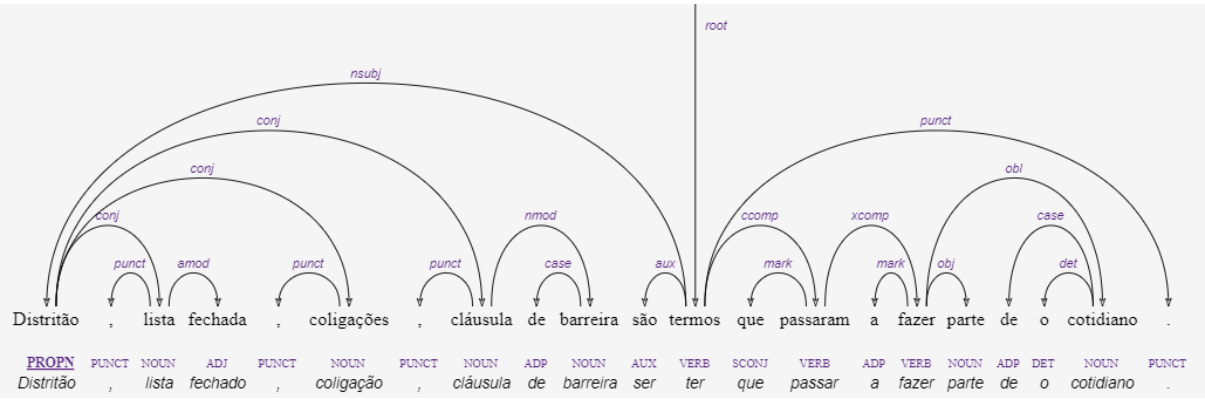


Figura 5a. FOLHA_DOC005026_SENT005 - anotada pelo parser

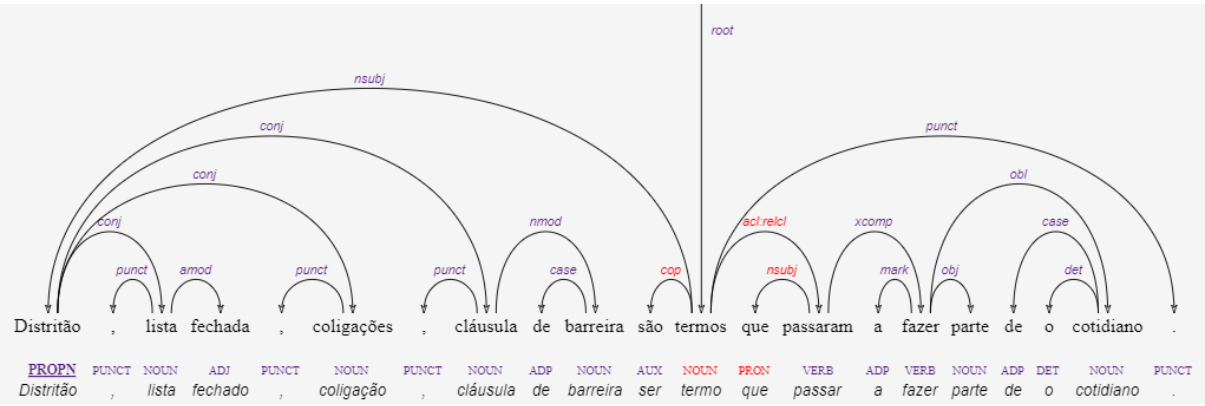


Figura 5b. FOLHA_DOC005026_SENT005 - corrigida manualmente

Outro erro de lematização e PoS tag é ilustrado na Figura 6a, na qual a forma “tê” foi lematizada como “tê” e recebeu a PoS NOUN, quando o correto seria o lema “ter” e a PoS AUX; a forma “banida”, por sua vez,

recebeu o lema “banido” e a PoS ADJ, quando o correto seria o lema “banir” e a PoS VERB, conforme mostrado em vermelho na Figura 6b.

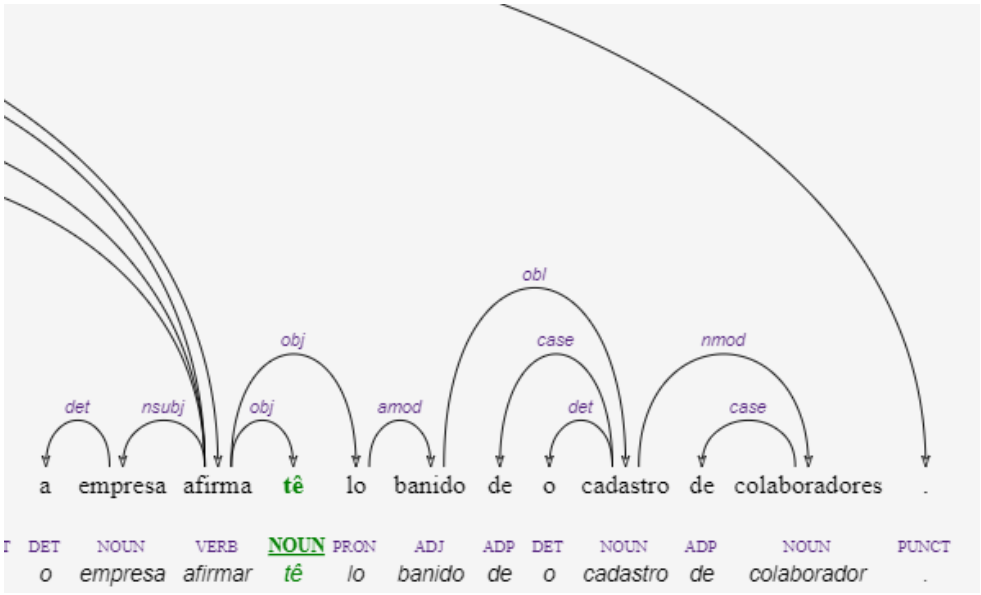


Figura 6a. FOLHA_DOC005029_SENT019 - anotada pelo parser

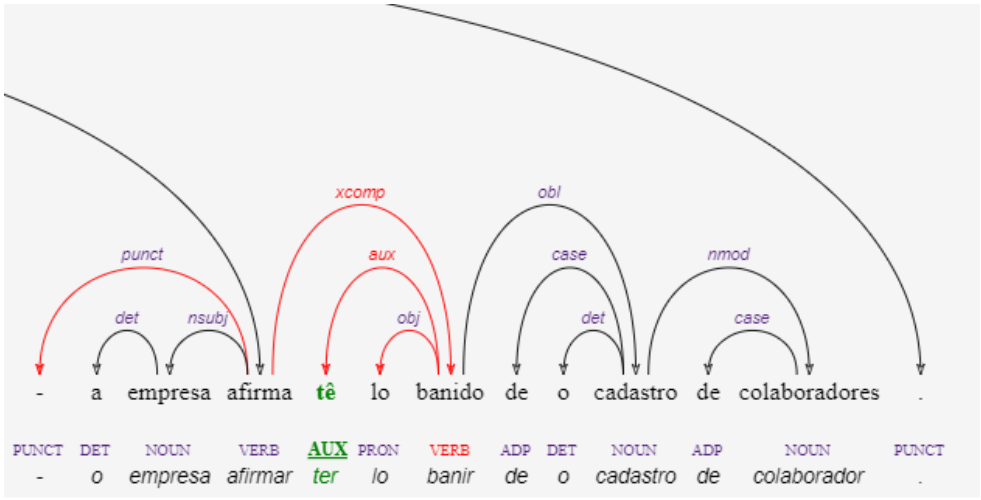


Figura 6b. FOLHA_DOC005029_SENT019 - corrigida manualmente

Por fim, vale ressaltar que há erros anteriores a qualquer processamento, ou seja, erros de escrita, que dificultam o *parsing*. Por exemplo, a sentença cuja árvore é ilustrada nas Figuras 7a e 7b tem dois erros: um de concordância e um de regência verbal. O aposto de “barras bravas” está no singular – “integrante” (deveria ser “integrantes”) – e o verbo “conter” foi

utilizado com a regência e o sentido do verbo “impedir” (impedir alguém *de* fazer alguma coisa, o que não é uma possibilidade). Conclui-se que, sempre que o insumo contém erros, o resultado do *parsing* pode ser afetado. Isso é importante de ser ressaltado, pois em alguns tipos de texto (como por exemplo UGC - *User Generated Content*), a probabilidade de haver erros de escrita é maior.

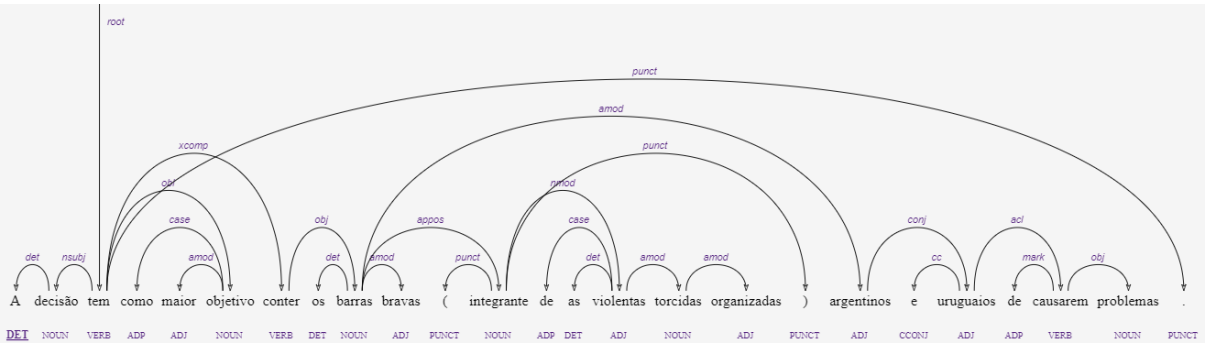


Figura 7a. FOLHA_DOC005006_SENT005 - anotada pelo parser

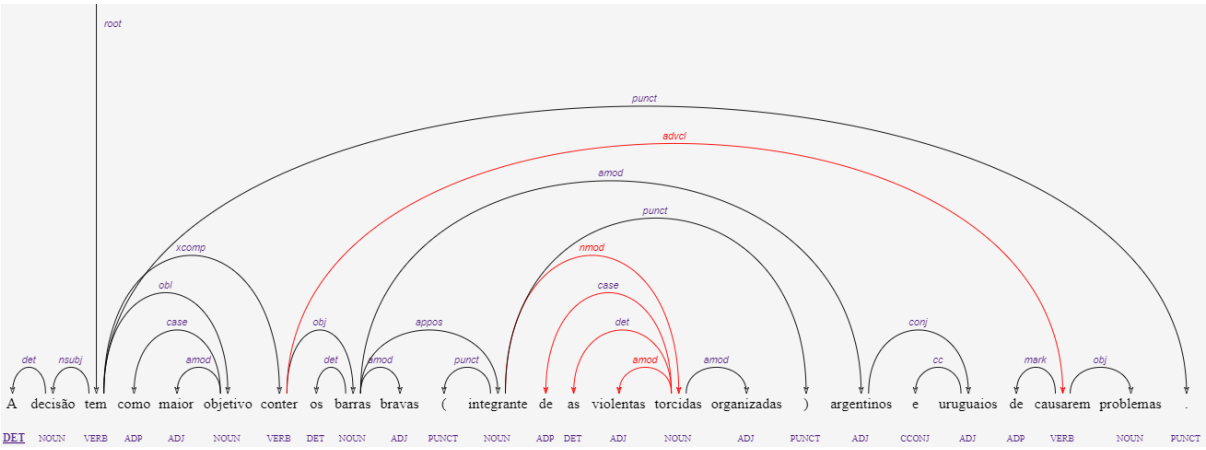


Figura 7b. FOLHA_DOC005006_SENT005 - corrigida manualmente

A simples falta de uma preposição pode confundir o parser, como pode ser observado nas Figuras 8a e 8b. Um humano pode deduzir que “1, 2 e 3 setembro” é o mesmo que “1, 2 e 3 *de* setembro”, mas a máquina não.

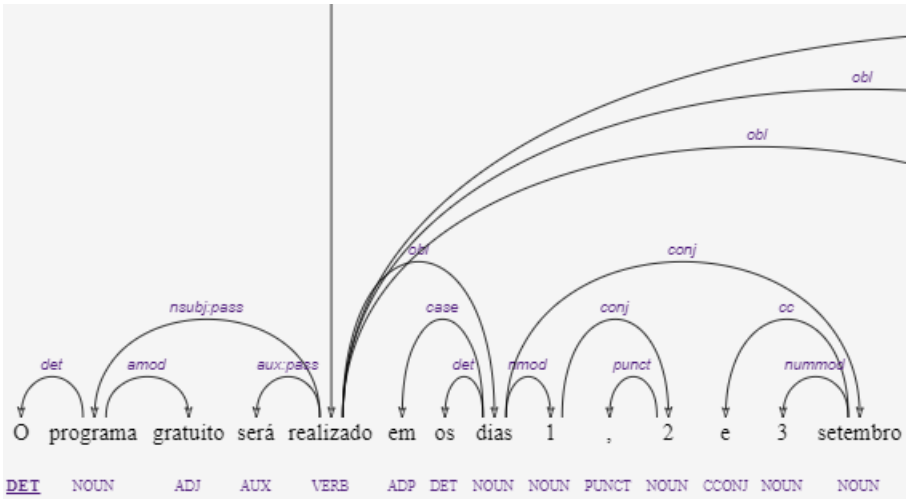


Figura 8a. FOLHA_DOC005012_SENT003 - anotada pelo parser

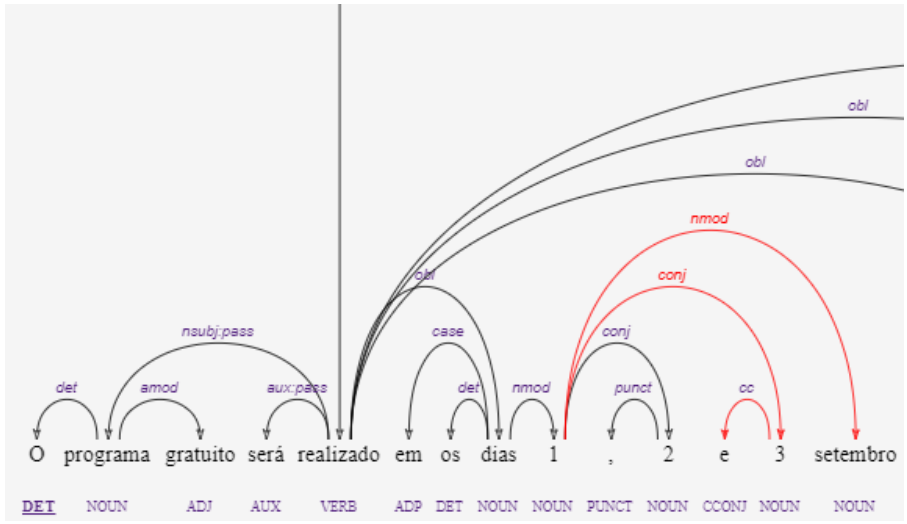


Figura 8b. FOLHA_DOC005012_SENT003 - corrigida manualmente

3.2 - Erros de escolha de head

Uma espécie bem comum desse erro é o conhecido erro de "*PP attachment*", pois, na maioria das vezes, o dependente da deprel é um sintagma preposicionado. Porém, o problema ocorre sempre que há dois ou mais candidatos aptos sintaticamente a assumir o head de um segmento

modificador, seja ele um sintagma preposicionado, um advérbio, um adjetivo ou uma oração modificadora.

A escolha do head tem uma restrição sintática: modificador (dependente da deprel) e modificado (head da deprel) devem concordar sempre que ambos tiverem marcas de número e gênero. Se o modificador e o modificado não tiverem marcas de número e gênero, apenas conhecimento de semântica lexical e conhecimento de mundo são capazes de auxiliar um anotador humano a decidir qual é o head, muitas vezes persistindo a ambiguidade.

Uma das maiores dificuldades na atribuição do head das relações decorre do fato de que modificado e modificador não ocorrem necessariamente em posições contíguas. Informações de frequência lexical podem auxiliar o aprendizado da estrutura argumental de verbos, substantivos, adjetivos e advérbios, porém pouco ajudam quando se trata de um modificador circunstancial, pois modificadores desse tipo podem modificar qualquer predicado.

O parser pode "adquirir" esses conhecimentos a partir da frequência lexical no córpus ou a partir de informações de semântica lexical, mas o problema extrapola a questão sintática e um parser não deveria ser penalizado por atribuir erroneamente o head nesses casos, pois sintaticamente há mais de uma forma possível de estabelecer a relação, mesmo que nem todas sejam semanticamente corretas. Aliás, uma hipótese a ser testada é a de que a anotação de papéis semânticos talvez possa retroalimentar as decisões de um parser com relação ao head das deprel.

A seguir são comentados e ilustrados vários tipos de erro de atribuição de head de relações de dependência.

3.2.1 Erro de head de **nmod**

A deprel **nmod** é altamente frequente em córpus de português. Só nas 600 sentenças avaliadas, foram atribuídos 453 **nmod**, 69 dos quais apresentaram erro. Erros em seu head podem ser corrigidos apenas mudando o NOUN modificado, como ilustrado nas Figuras 9a e 9b, o que ocorreu 30 vezes na amostra. No entanto, quando o head correto é um VERB, também a deprel deve ser corrigida, passando de **nmod** para **obl**, como ilustrado nas Figuras 10a e 10b, o que ocorreu 39 vezes.

No primeiro caso, trata-se de um atributo do substantivo "área", que só conhecimento de mundo ajuda a identificar. No segundo, trata-se de um adjunto de modo: "pelo vestibular tradicional", mais simples de ser aprendido automaticamente, pois o verbo "selecionar" frequentemente é acompanhado de um modificador de modo ou instrumento iniciado pela preposição "por".

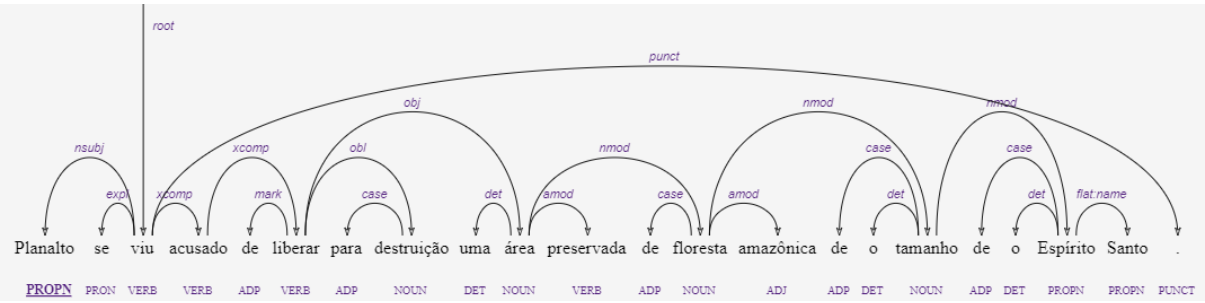


Figura 9a. FOLHA_DOC005053_SENT005 - anotada pelo parser

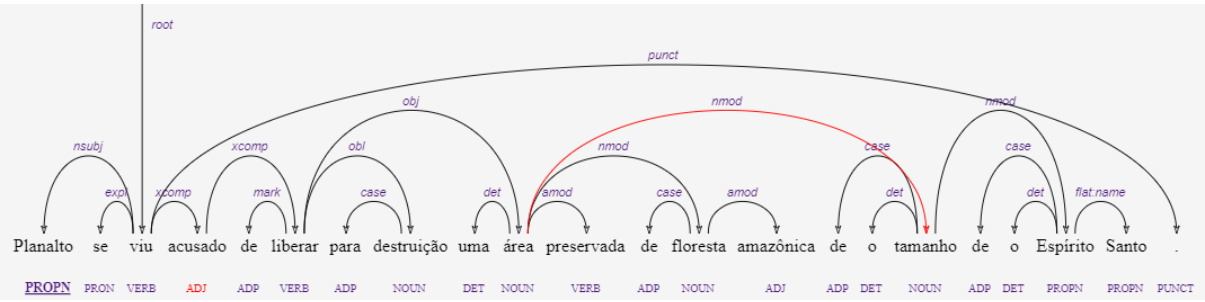


Figura 9b. FOLHA_DOC005053_SENT005 - corrigida manualmente

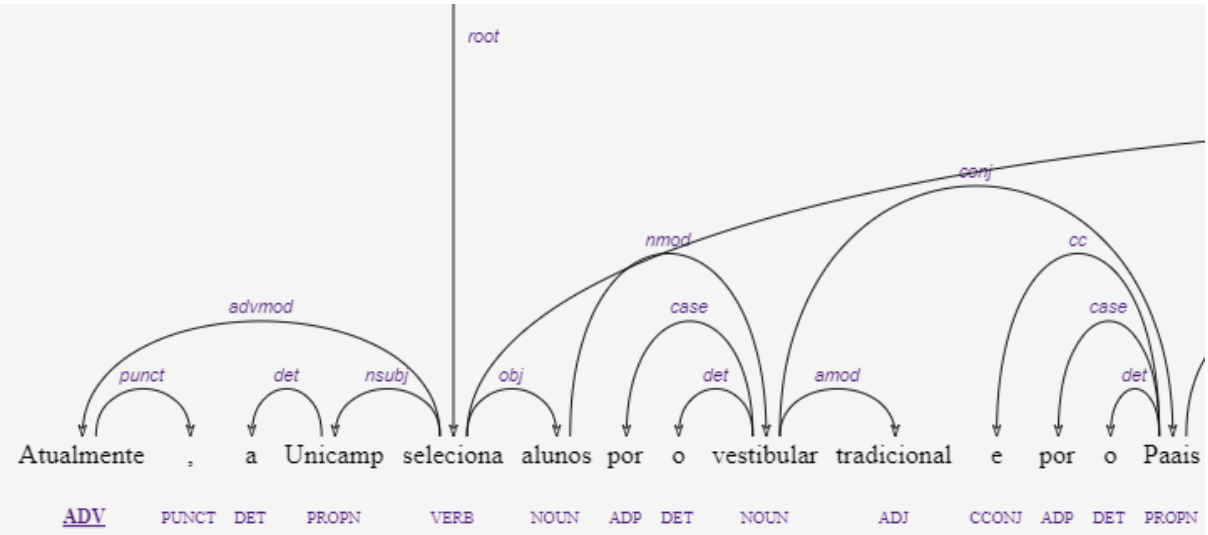


Figura 10a. FOLHA_DOC005015_SENT006 - anotada pelo parser

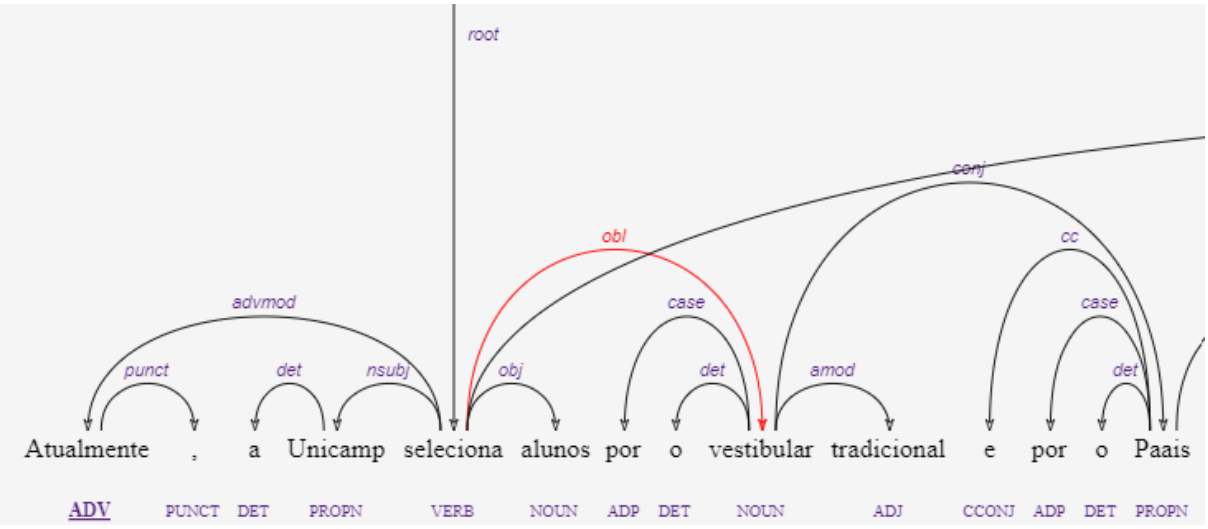


Figura 10b. FOLHA_DOC005015_SENT006 - corrigida manualmente

3.2.2 Erro de head de **obl**

A deprel **obl** é muito frequente em córpus de português. Somente na amostra avaliada, foram 437 **obl** atribuídos. Erros de head de **obl** ocorreram 44 vezes na amostra. Eles podem ser corrigidos apenas mudando o VERB modificado (como o PP "em passeio", nas Figuras 11a e 11b), situação que ocorreu 12 vezes ou, quando o head correto é um

NOUN, a própria deprel é corrigida, passando de **obl** a **nmod** (como o PP "no alto do Corcovado", nas Figuras 11a e 11b), situação que ocorreu 32 vezes na amostra.

Na sentença ilustrada nas Figuras 11a e 11b, o substantivo "passeio" é predicador ("[dar/fazer]passeio em/por" = "passear em/por") e "pede" um complemento de lugar. Se a frequência com que ocorre num córpus for alta, o aprendizado da estrutura argumental é facilitado.

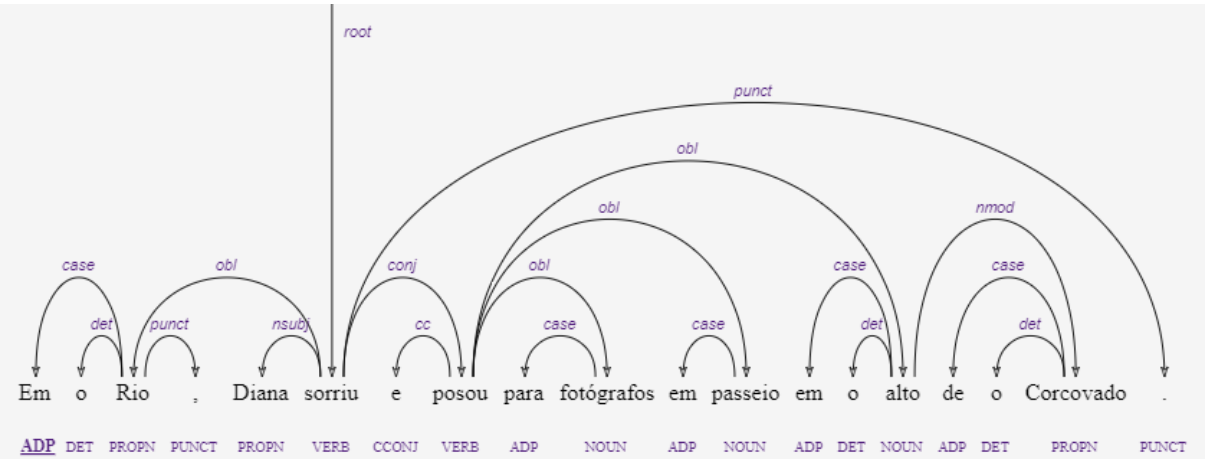


Figura 11a. FOLHA_DOC005068_SENT012 - anotada pelo parser

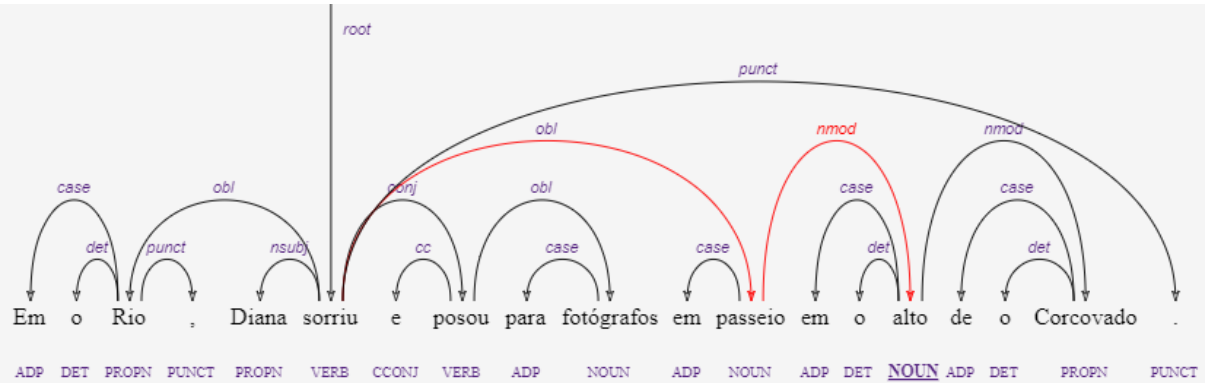


Figura 11b. FOLHA_DOC005068_SENT012 - corrigida manualmente

Quando um substantivo predicador[6] ocorre acompanhado de um verbo suporte[7], sua identificação como possuidor de uma estrutura argumental poderia ser facilitada se a anotação do córpus contivesse, no campo de *features* da anotação, uma indicação de que se trata de verbo suporte e/ou nome predicador[8]. As Figuras 12a e 12b trazem uma construção de verbo suporte: "ter impacto", e pode-se observar que o substantivo é o predicador ("ter impacto em" = "impactar em").

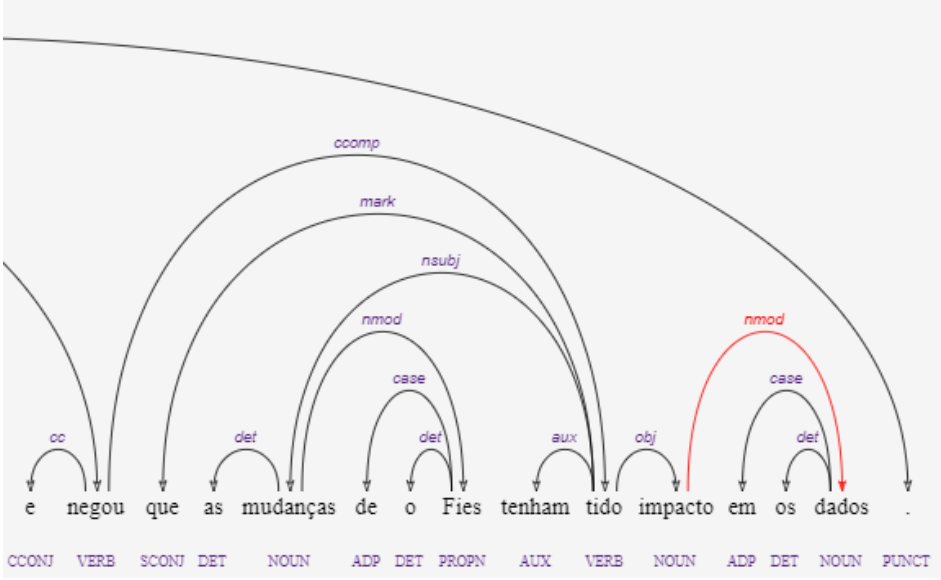


Figura 12a. FOLHA_DOC005027_SENT016 - anotada pelo parser

[6] Um substantivo predicador é aquele que prevê complementos, como é o caso do substantivo "impacto", que prevê um complemento introduzido pelas preposições "em" ou "sobre": "impacto em algo" ou "impacto sobre algo"

[7] Associado a um verbo suporte, um substantivo predicador pode atuar como um predicado verbal. A UD, contudo, não prevê a anotação de verbos suporte e recomenda que sejam anotados como verbos plenos, ao mesmo tempo que o nome predicador é anotado como objeto direto desses verbos. Os verbos suporte mais produtivos em português são: dar, fazer e ter (por exemplo: dar queixa de, fazer aplicação em, ter consideração por).

[8] Enquanto o número de verbos suporte é limitado, pois são palavras semigramaticalizadas, o número de nomes predicadores é da ordem de milhares, pois são palavras de classe aberta.

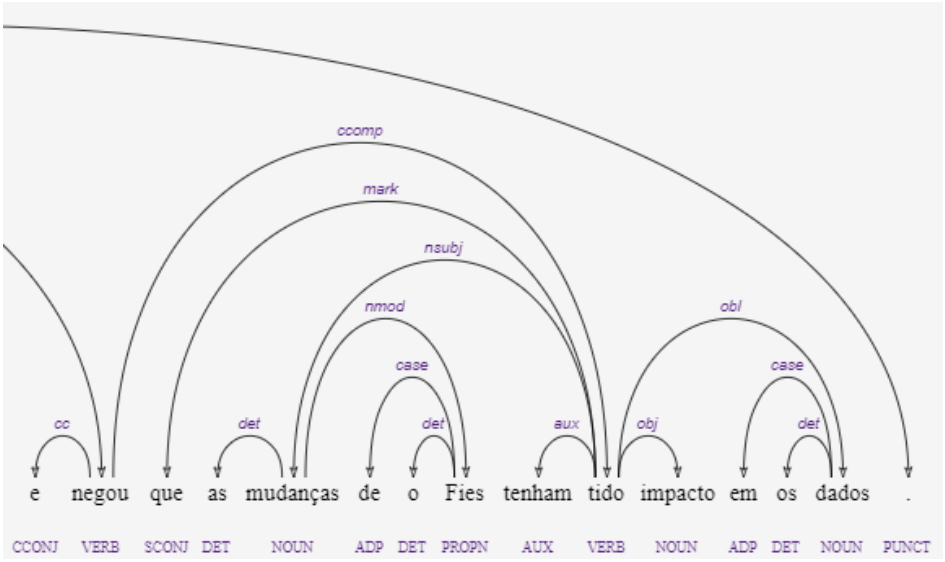


Figura 12b. FOLHA_DOC005027_SENT016 - corrigida manualmente

3.2.3 Erro de head de **acl**

Erros de head de **acl** ocorreram 27 vezes na amostra, sendo 10 casos de **acl** simples, 13 casos de **acl:relcl** (oração relativa) e 4 casos em que o head correto era um verbo, obrigando a mudança de deprel de **acl** para **advcl**. Muitas vezes, o head de uma **acl** pode ser ambíguo e o fato de um anotador humano optar por um head não significa necessariamente que um outro seja incorreto. Nas figuras a seguir, a oração que expressa finalidade poderia ter como head "busca" (Figura 13a) ou "empréstimo" (Figura 13b) ou, até mesmo, poderia ter o verbo "prever" como head, caso em que seria alterada para **advcl** (Figura 13c). Nesse caso específico, optou-se por atribuir o head ao substantivo "empréstimo" (como na Figura 13b) pelo fato de haver grande frequência de uso da expressão de finalidade de empréstimos.

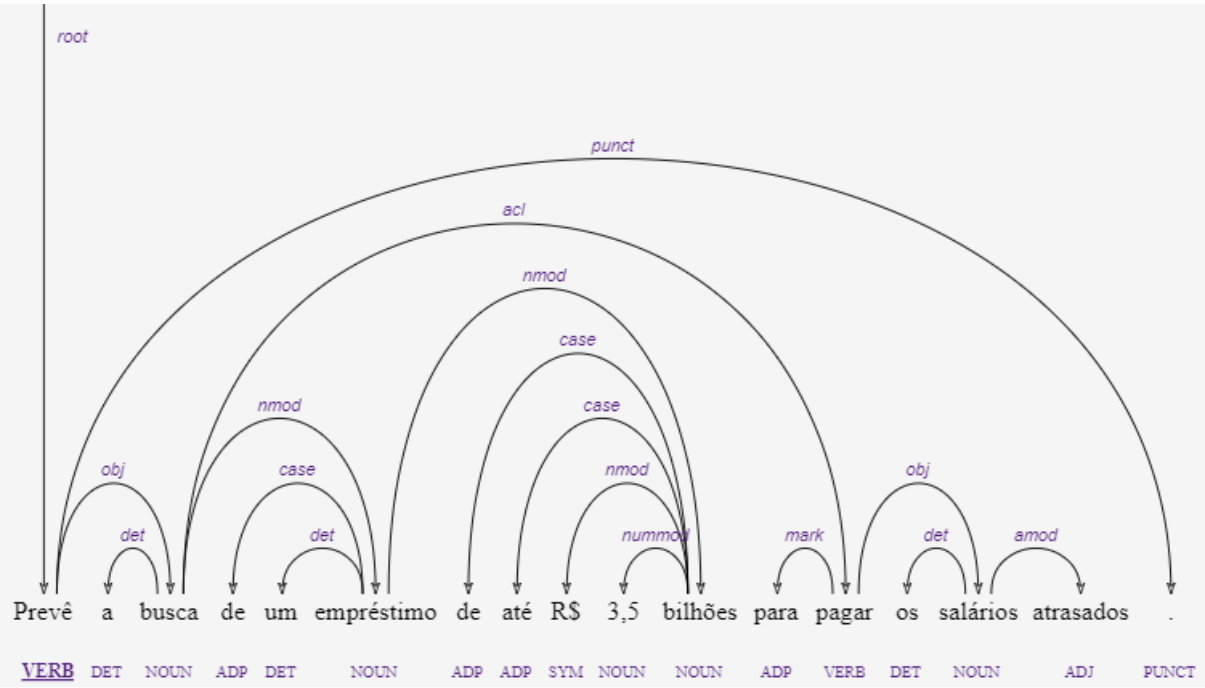


Figura 13a. FOLHA_DOC005007_SENT006 - anotada pelo parser

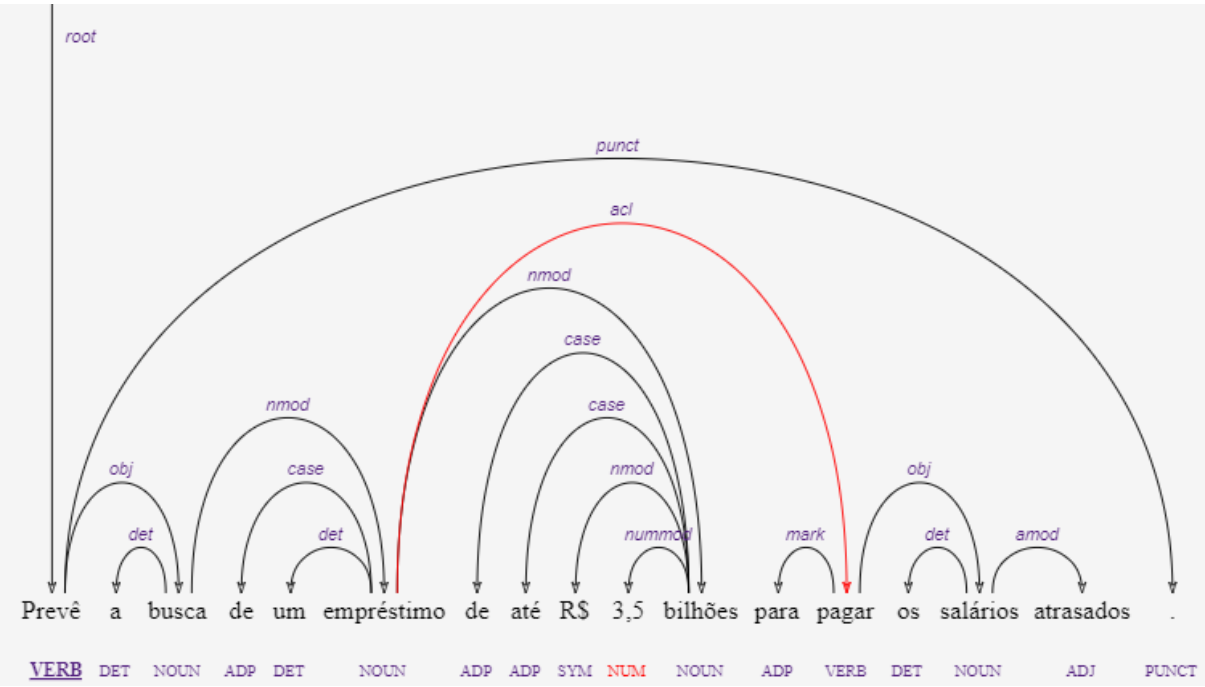


Figura 13b. FOLHA_DOC005007_SENT006 - corrigida manualmente

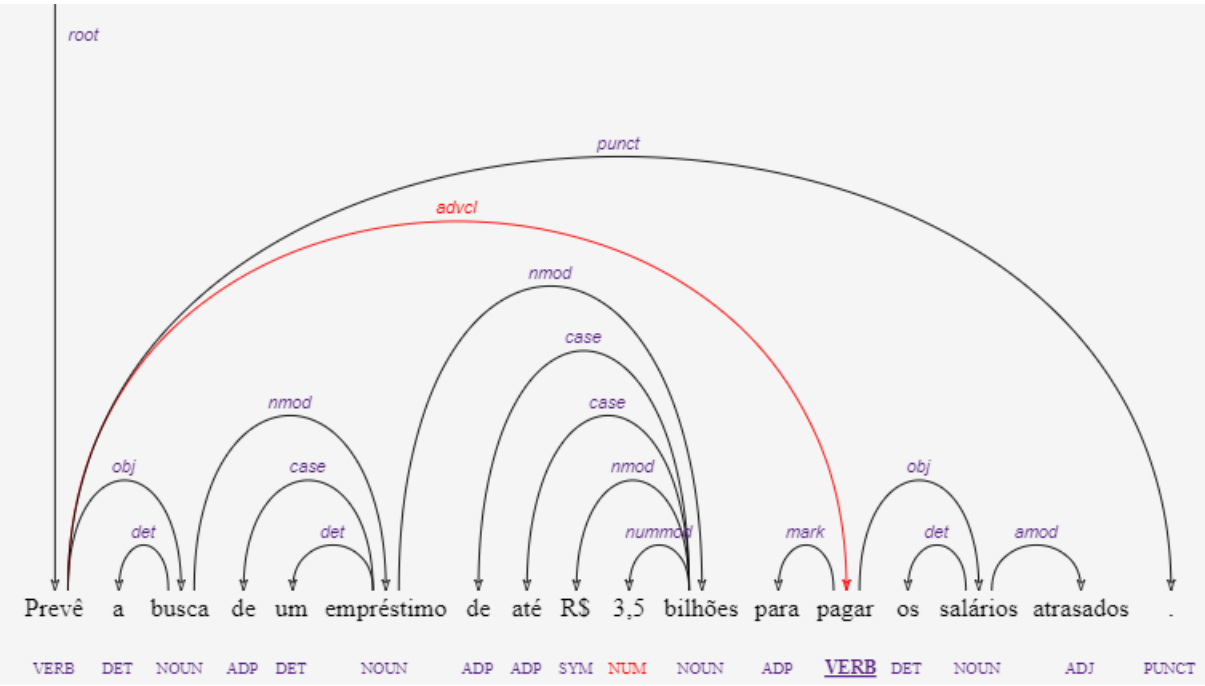


Figura 13c. FOLHA_DOC005007_SENT006 - corrigida manualmente

3.2.4 Erro de head de **advcl**

Erros de head de **advcl** ocorreram 21 vezes na amostra, sendo que em 12 deles o head correto era um substantivo, exigindo a mudança da deprel de **advcl** para **acl**, como no exemplo ilustrado nas Figuras 14a e 14b.

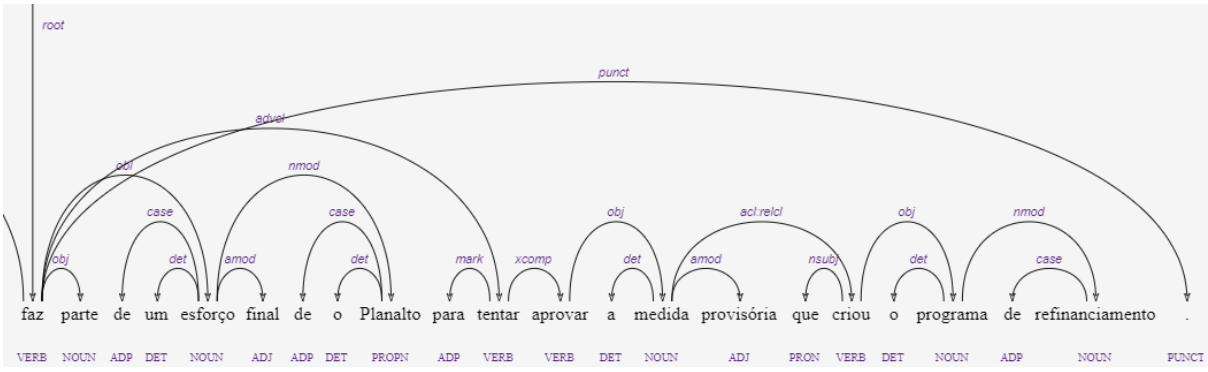


Figura 14a. FOLHA_DOC005033_SENT006 - anotada pelo parser

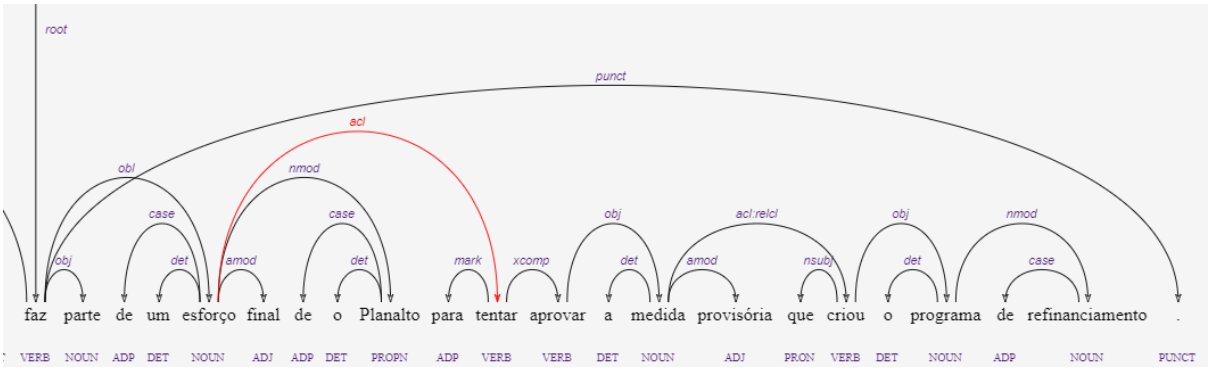


Figura 14b. FOLHA_DOC005033_SENT006 - corrigida manualmente

3.2.5 Erro de head de **conj**

Erros de head de **conj** ocorreram 40 vezes na amostra. A decisão de qual é o head de uma relação de coordenação quase sempre é facilitada pela escolha de elementos de mesma natureza morfossintática: dois ou mais verbos com o mesmo modo, tempo e pessoa; dois ou mais substantivos introduzidos por uma mesma preposição; dois ou mais adjetivos, etc.
Porém, há casos em que esse critério é prejudicado pela presença de mais de uma opção de head, como mostrado nas Figuras 15a e 15b.

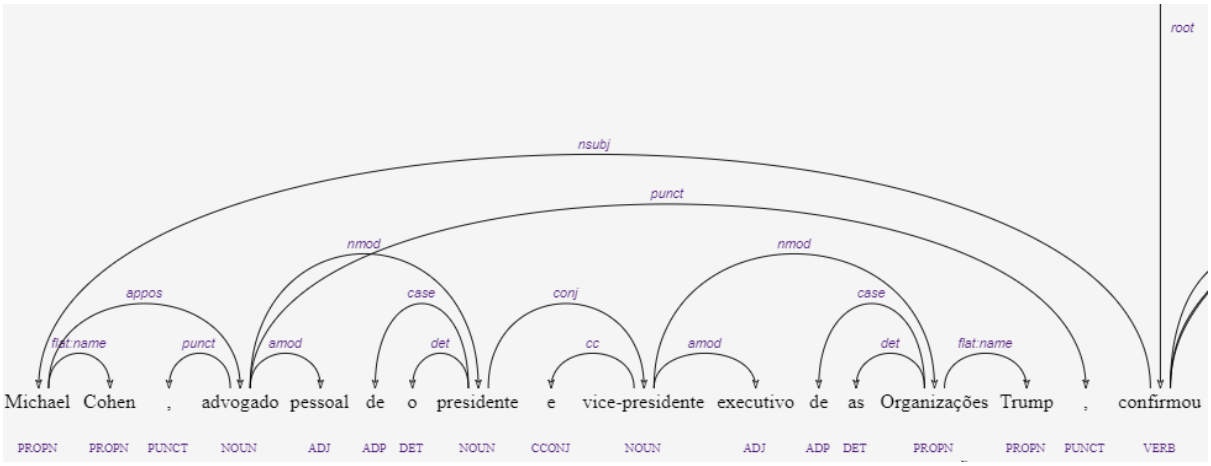


Figura 15a. FOLHA_DOC005004_SENT012 - anotada pelo parser

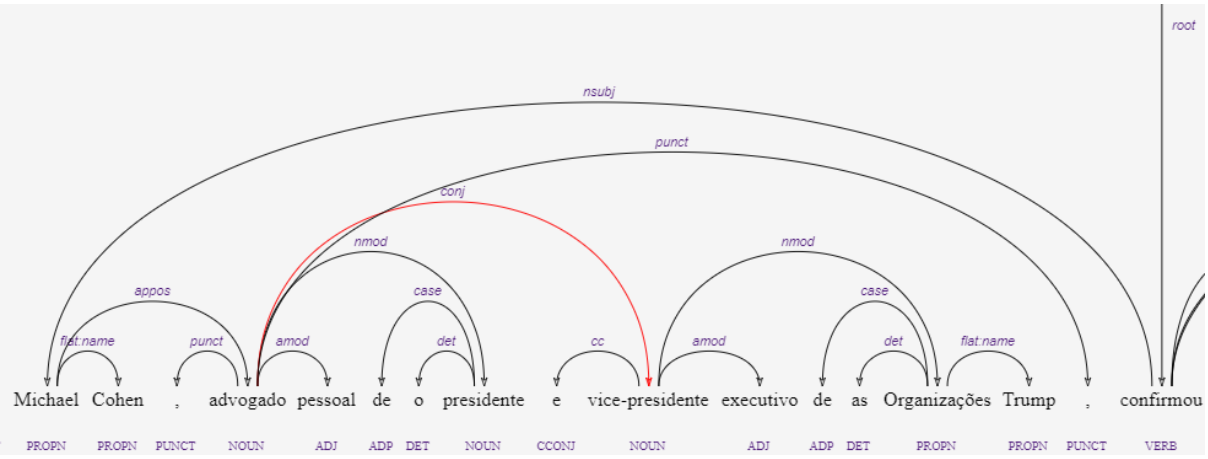


Figura 15b. FOLHA_DOC005004_SENT012 - corrigida manualmente

Outro fato que dificulta a coordenação ocorre quando os elementos têm formas diferentes, como nas Figuras 16a e 16b, em que os elementos coordenados são um PP ("de ameaça") e dois NP ("violência doméstica" e "furtos"). Nessa figura, a preposição "de" ocorre sem acompanhamento de um determinante e está, portanto, introduzindo os três tipos de "crime", ou seja, servindo aos três NPs. A anotação UD, porém, define que a preposição deve acompanhar um único token e sempre o token introduzido por ela, o que torna os três elementos coordenados assimétricos.

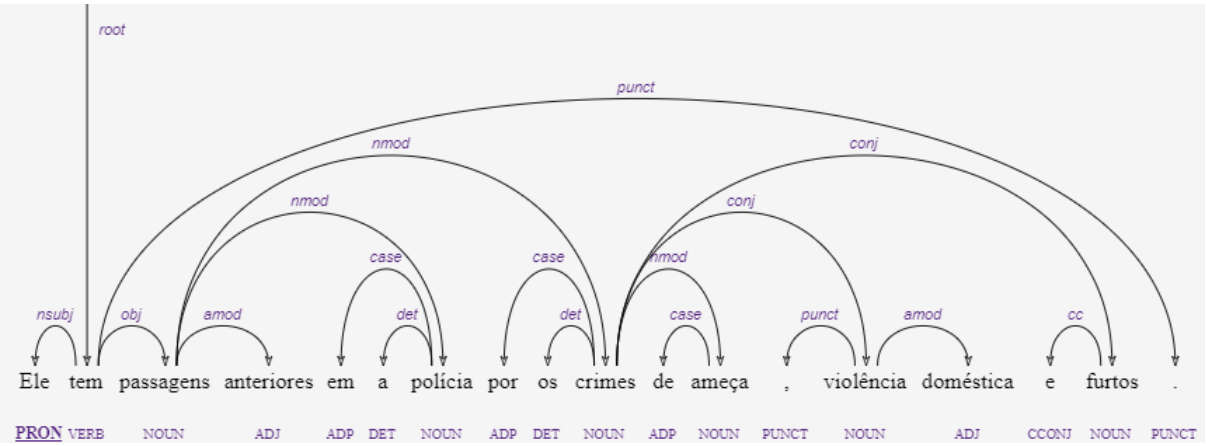


Figura 16a. FOLHA_DOC005023_SENT012 - anotada pelo parser

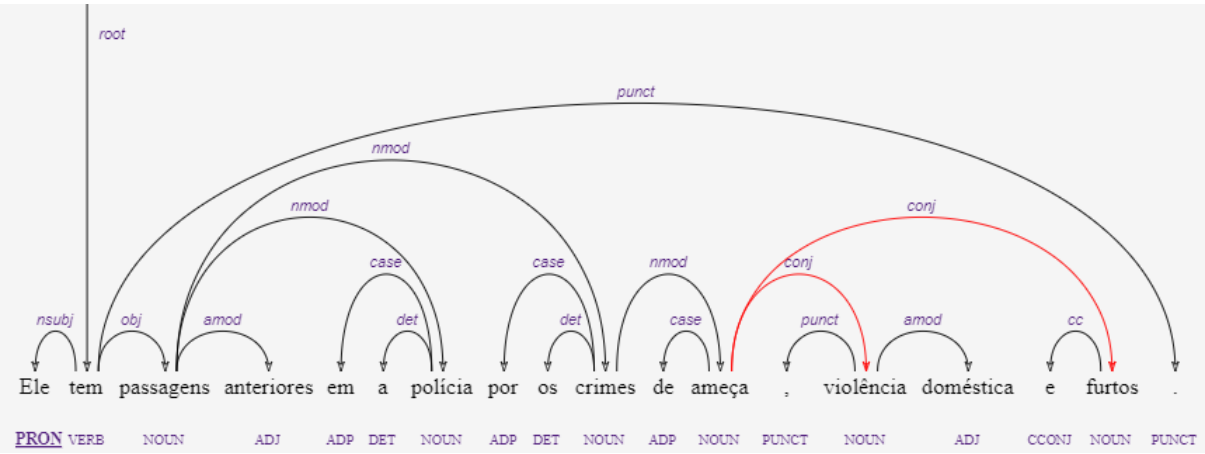


Figura 16b. FOLHA_DOC005023_SENT012 - corrigida manualmente

A mesma assimetria entre elementos coordenados ocorre na sentença ilustrada nas Figuras 17a e 17b, na qual uma deprel **amod** (adjetivo) está coordenada ao que seria, isoladamente, anotado como **nmod**.

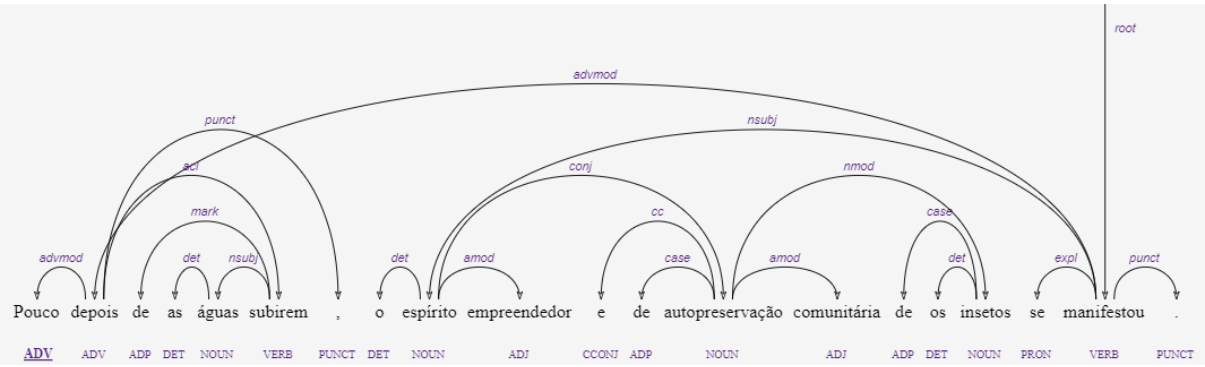


Figura 17a. FOLHA_DOC005022_SENT008 - anotada pelo parser

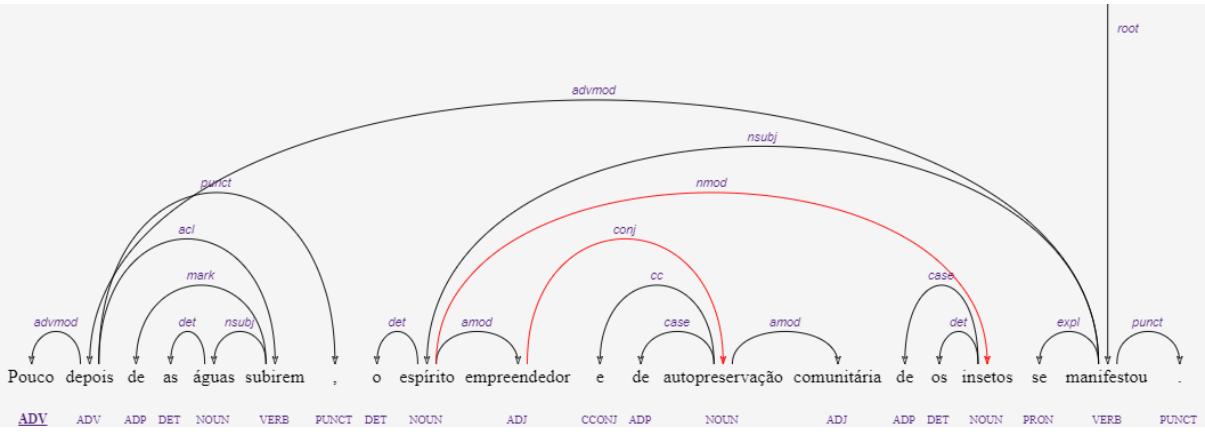


Figura 17b. FOLHA_DOC005022_SENT008 - corrigida manualmente

3.2.6 Erro de head de modificadores de elementos coordenados

Muitas vezes, dois ou mais elementos coordenados possuem um modificador ou um complemento em comum. Porém, não há marca sintática que indique se o modificador modifica apenas o elemento coordenado contíguo a ele ou todos os elementos coordenados. O parser, nessas situações, anotou como head do modificador o elemento mais próximo (Figura 18a). A forma mais lógica de indicar que o modificador se refere a todos os elementos coordenados é atribuir o head do modificador ao elemento que também é head da coordenação (Figura 18b).

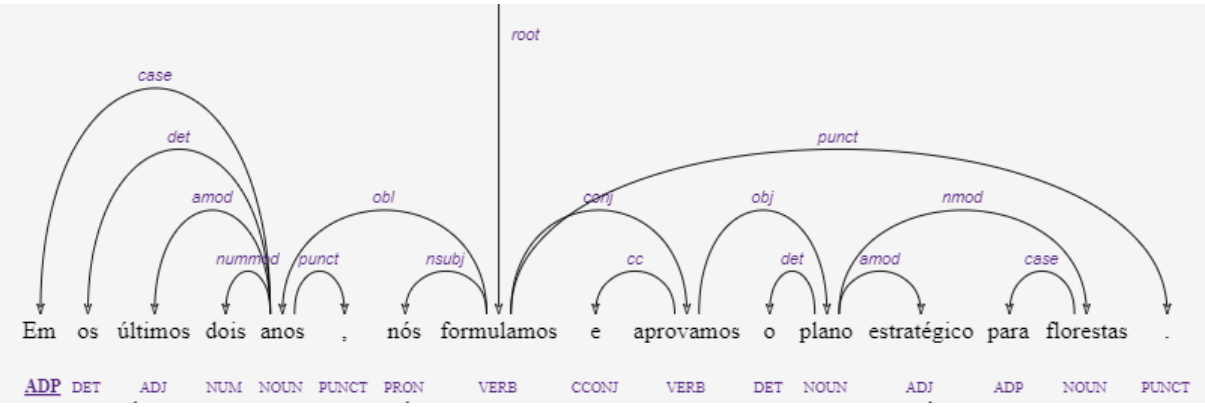


Figura 18a. FOLHA_DOC005216_SENT041 - anotada pelo parser

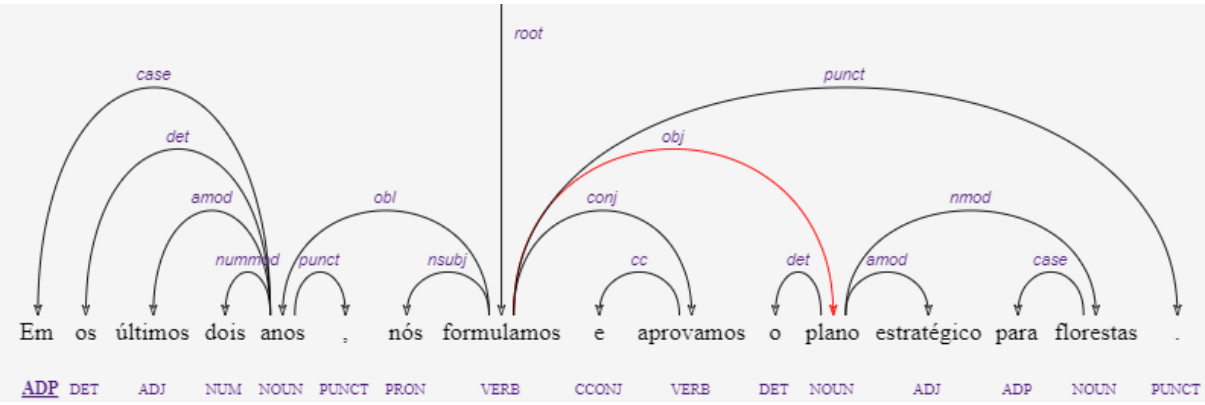


Figura 18b. FOLHA_DOC005216_SENT041 - corrigida manualmente

As Figuras 19a e 19b ilustram o caso de um modificador circunstancial de dois verbos intransitivos coordenados. O importante é notar que nem sempre o modificador diz respeito aos dois elementos coordenados, por isso o modificador do segundo elemento é sintaticamente ambíguo. Por exemplo, em uma sentença como “Ele dormiu e morreu de fome”, o PP “de fome” só modifica “morreu” e não “dormiu e morreu”. Portanto, é uma questão de interpretação que leva à decisão de qual deve ser o head do modificador do segundo elemento de uma coordenação. Esse tipo de erro só ocorreu 1 vez na amostra com a deprel **obj** e 3 vezes com a deprel **amod**. Nas demais vezes, ocorreu com modificadores do tipo **obl** (como exemplificado nas Figuras 19a e 19b) e **nmod** e foram computados nos erros de head dessas deprels.

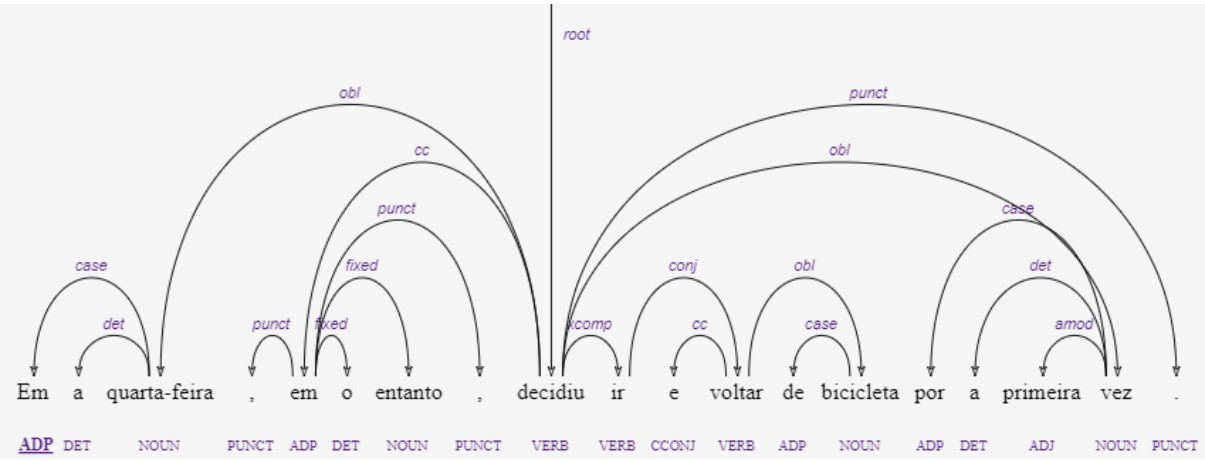


Figura 19a. FOLHA_DOC005050_SENT021 - anotada pelo parser

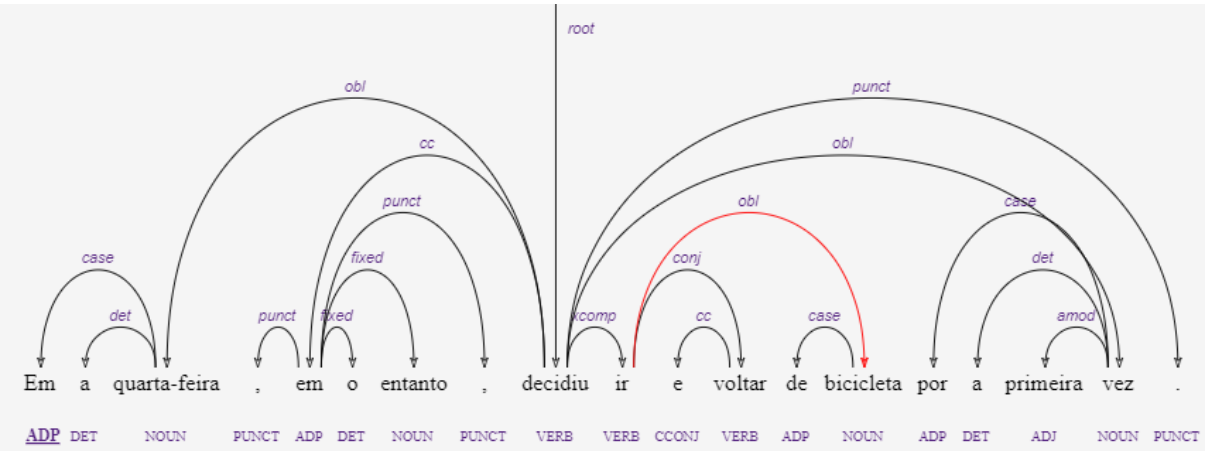


Figura 19b. FOLHA_DOC005050_SENT021 - corrigida manualmente

3.3 Erros de natureza sintática

Consideraram-se erros de natureza genuinamente sintática aqueles que não são motivados por erros de pré-processamento ou que não são dependentes de informações semânticas.

Nessa categoria, foram agrupados: erro de identificação de sujeito; erro de **amod** (adjetivo) anteposto, erro de **det** (determinante) posposto, erro de reconhecimento de **fixed** (expressões fixas) e erro de reconhecimento de **flat:name** (nomes próprios compostos) e erro de **root**.

Antes de comentar esses tipos de erro, é importante destacar alguns pontos em que o parser teve excelente desempenho. Por exemplo, praticamente não houve confusão entre **xcomp** e **ccomp** nos pacotes analisados. Houve apenas 2 erros desse tipo nas 600 sentenças e ambos com o mesmo tipo de construção "ter + PP[9] + VINF" ("ter como objetivo aprovar" e "tem como objetivo reduzir"). As deprels **ccomp** e **xcomp** são atribuídas a dois tipos de orações que constituem complementos verbais. Os próprios anotadores humanos tiveram muita dificuldade para aprender a diferenciar as duas relações no início da atividade de anotação do córpus, mas o resultado do

[9] PP - *Prepositional Phrase* (sintagma preposicionado)

aprendizado mostra que, apesar dessa dificuldade, houve consistência no resultado final da anotação.

As relações funcionais (**case**, **det**, **mark**, **cop, cc** e **aux**) também parecem ter sido muito bem aprendidas pelo parser. Os dependentes dessas relações são corretamente identificados e erros de head dessas relações só ocorrem como acarretamento de outros erros, como erro de **root** e erro na identificação de núcleo de um NP (um adjetivo confundido com um substantivo, por exemplo).

De modo geral, observa-se que sentenças 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.

Os erros de natureza sintática sistemáticos observados na análise são comentados e exemplificados a seguir.

3.3.1 Erro na atribuição do **root** da árvore sintática

O parser treinado errou 12 vezes a determinação do **root** nas 600 sentenças. Como cada sentença tem um **root**, foram 12 erros em 600 ocorrências (2%). O erro de **root** é grave por acarretar erros na atribuição das demais relações, pois é a partir da raiz da árvore que saem os galhos principais: ao mudar-se a raiz, movem-se todos os galhos da árvore que se ligam a ela. Erros desse tipo só ocorreram em situações em que até mesmo um anotador humano teria dificuldades em fazer a anotação.

As Figuras 20a e 20b ilustram um erro de **root** em uma sentença que apresenta estratégia de focalização do objeto direto, na qual “é” e “que”

deveriam ser anotados com a deprel **discourse**, pois exercem função puramente pragmática[10].

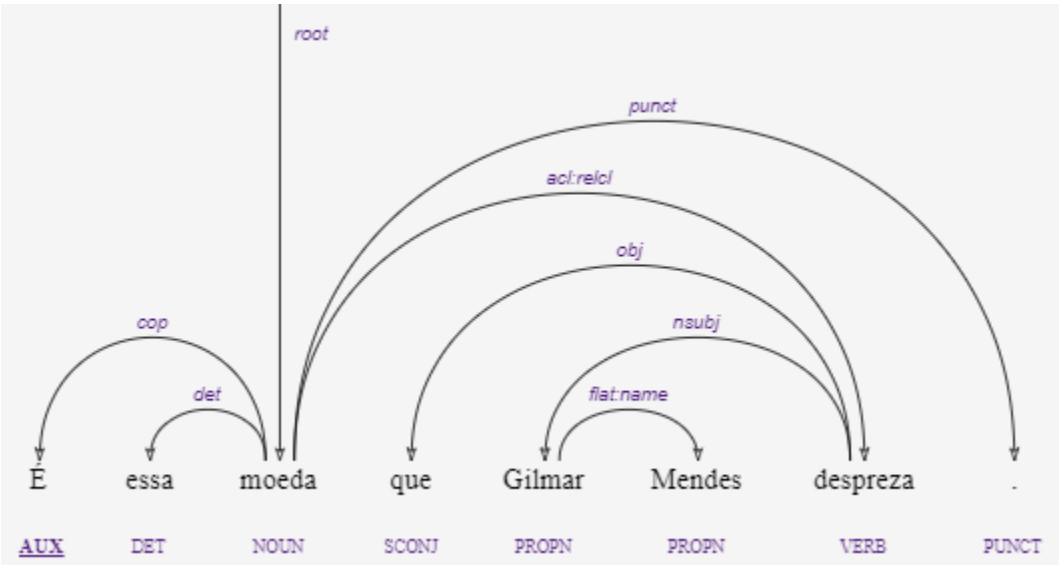


Figura 20a. FOLHA_DOC005065_SENT010 - anotada pelo parser

Figura 20b. FOLHA_DOC005065_SENT010 - corrigida manualmente

[10] A ordem canônica dos elementos seria: “Gilmar Mendes despreza essa moeda”. A estratégia de focalização traz para o início da sentença o objeto direto e “abraça-o” com duas palavras que lhe dão destaque: “é” e “que”, cuja função é discursiva, pois podem ser suprimidas sintaticamente: “Essa moeda Gilmar Mendes despreza”, resultando numa construção OSV (objeto, sujeito, verbo).

3.3.2 Erro na identificação de **nsubj, nsubj:pass e csubj**

O problema na identificação do sujeito ocorreu 38 vezes na amostra analisada e compreende 3 deprels: **nsubj** (sujeito da voz ativa), **nsubj:pass** (sujeito da voz 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 e, dos 38 erros dessa categoria, 11 foram sujeitos atribuídos incorretamente e 27 foram sujeitos não identificados pelo parser, ou seja, 588 sujeitos foram atribuídos corretamente.

Esse tipo de erro está relacionado principalmente à ocorrência do sujeito posposto ao verbo, ou seja, em construções do tipo VS (Verbo Sujeito), que, embora seja uma ordem não canônica do português, apresenta frequência relativamente alta.

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”, “não importa X”, em que X tem papel de sujeito.

As Figuras 21a e 21b trazem o exemplo de um verbo intransitivo, “estrear”, e as Figuras 22a e 22b trazem 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.

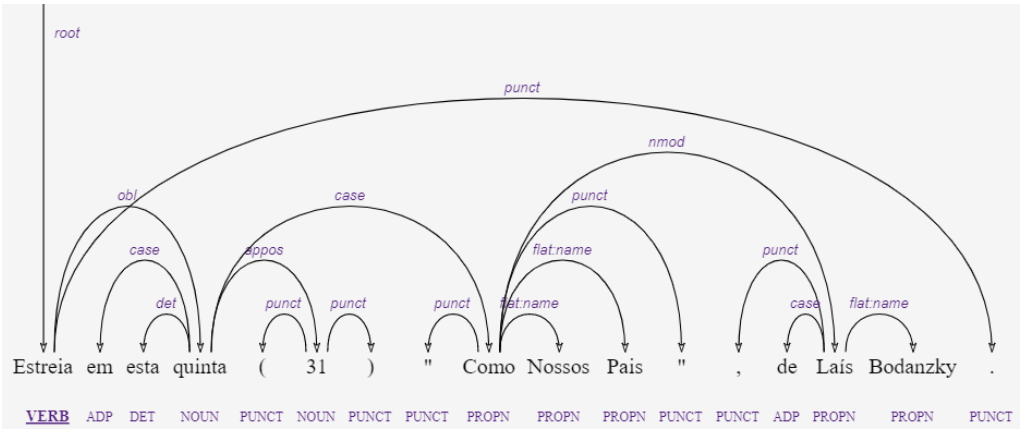


Figura 21a. FOLHA_DOC005081_SENT014 - anotada pelo parser

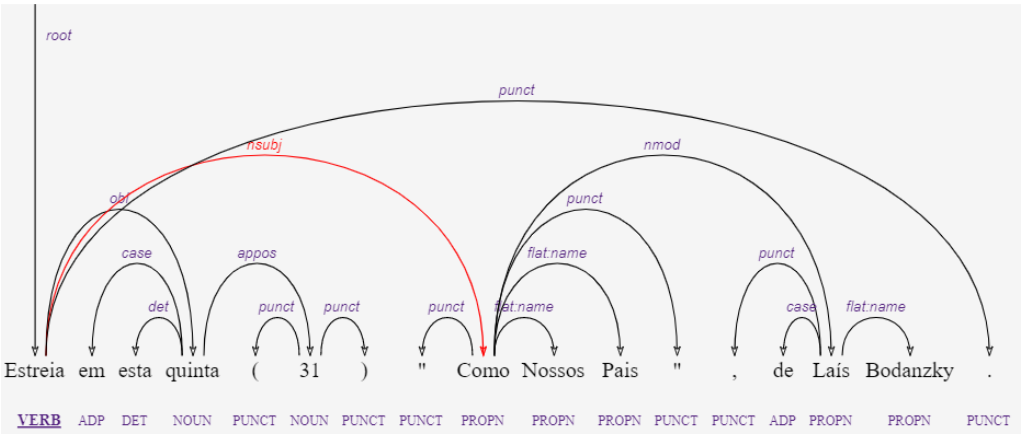


Figura 21b. FOLHA_DOC005081_SENT014 - corrigida manualmente

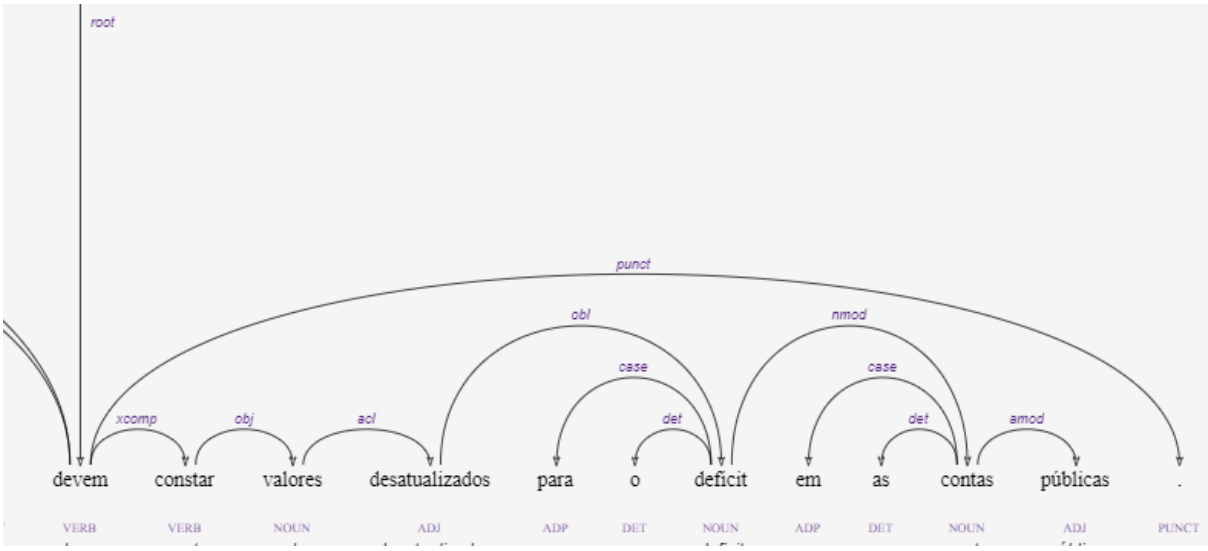


Figura 22a. FOLHA_DOC005040_SENT002 - anotada pelo parser

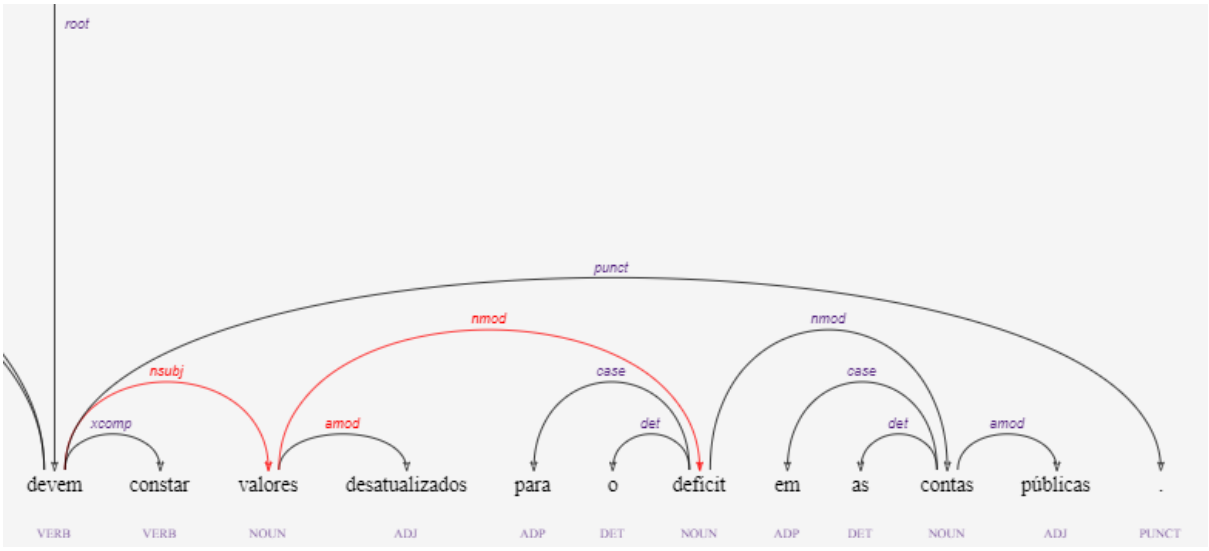
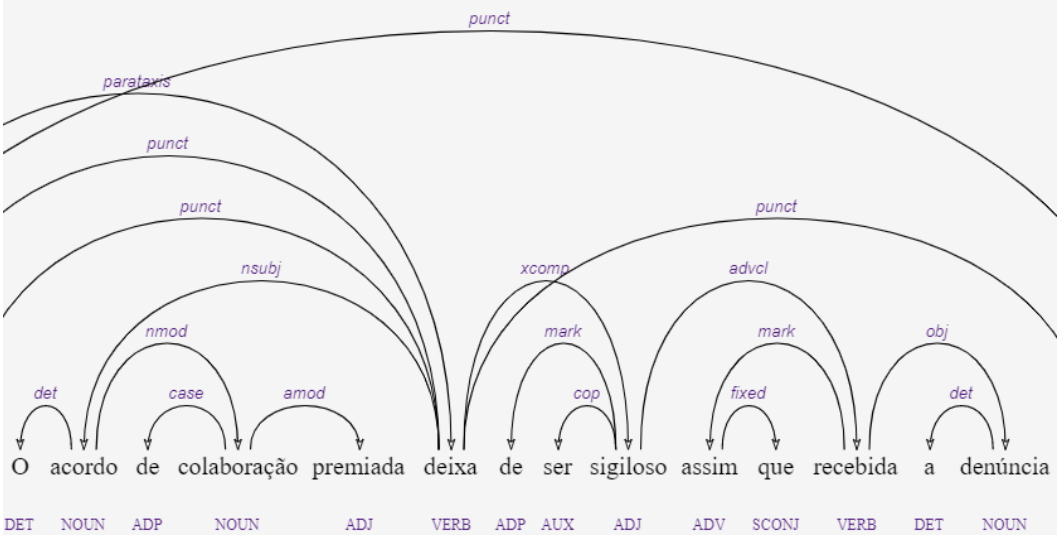


Figura 22b. FOLHA_DOC005040_SENT002 - corrigida manualmente

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 "assim que [for] recebida a denúncia" (Figuras 23a e 23b)[11]. A anotação da *feature* Voice=Pass nesse tipo de passiva pode contribuir para o reconhecimento do sujeito da passiva.



23a. FOLHA_DOC005039_SENT008 - anotada pelo parser

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

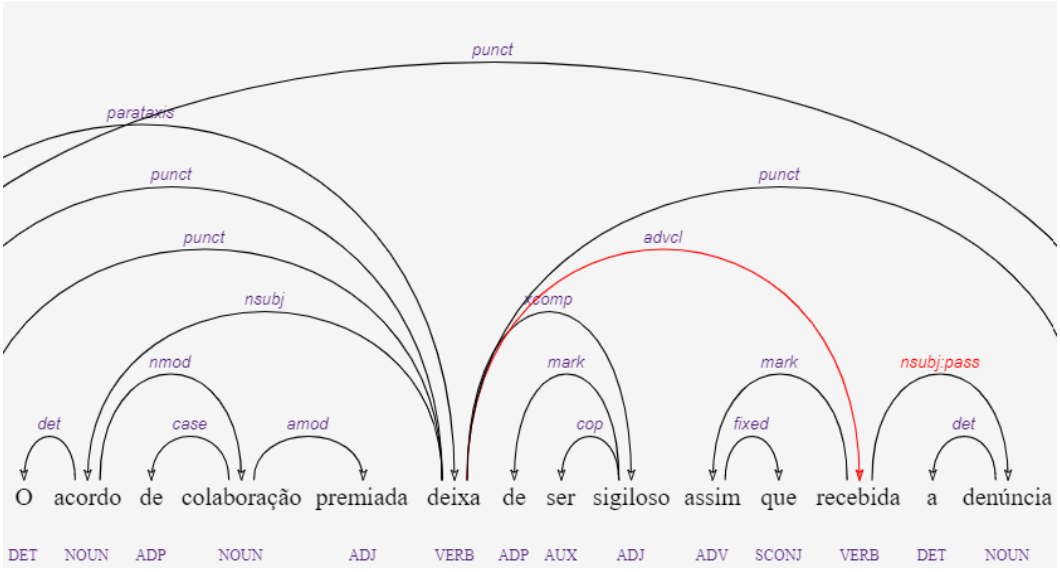


Figura 23b. FOLHA_DOC005039_SENT008 - corrigida manualmente

Também é frequente o sujeito posposto com verbos na voz passiva sintética (Figuras 24a e 24b).

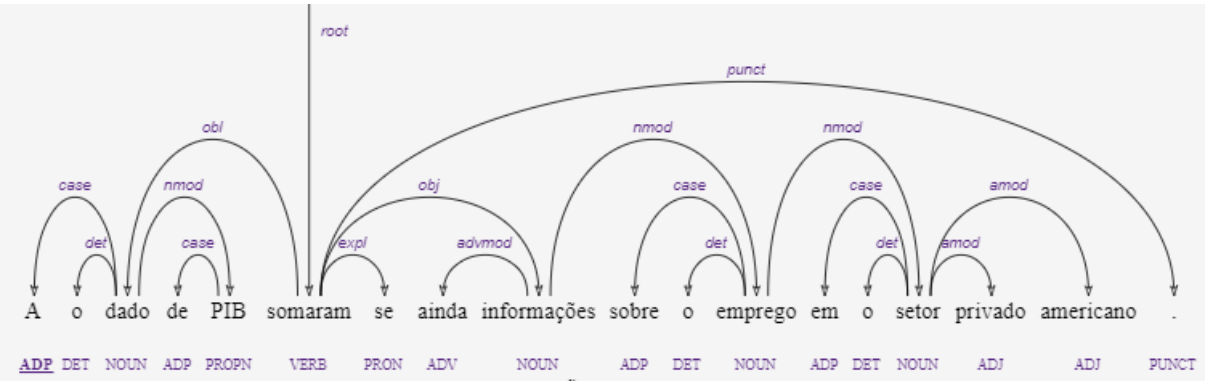


Figura 24a. FOLHA_DOC005139_SENT012 - anotada pelo parser

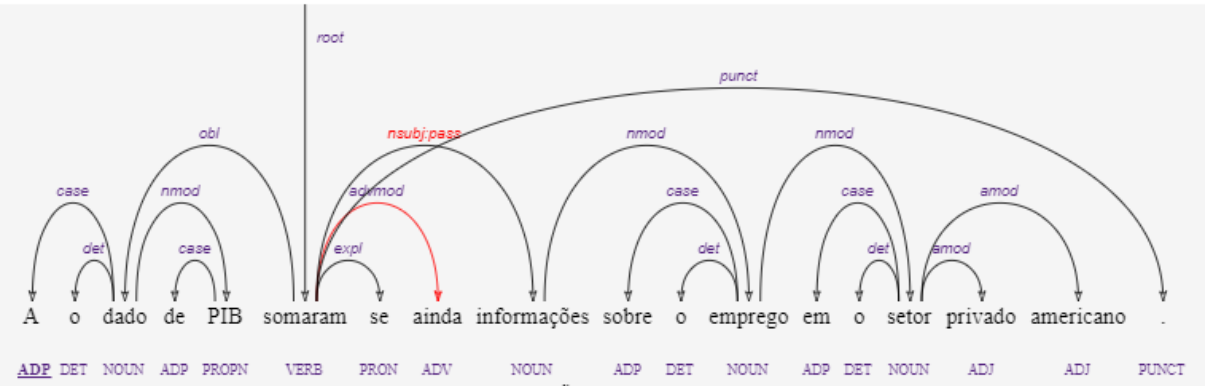


Figura 24b. FOLHA_DOC005139_SENT012 - corrigida manualmente

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 25a e corrigido na Figura 25b.

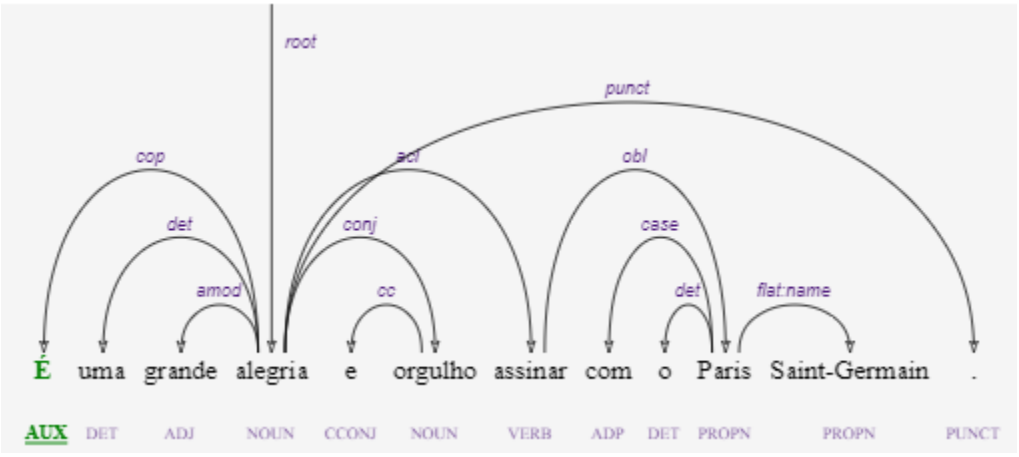


Figura 25a. FOLHA_DOC005008_SENT006 anotado pelo parser

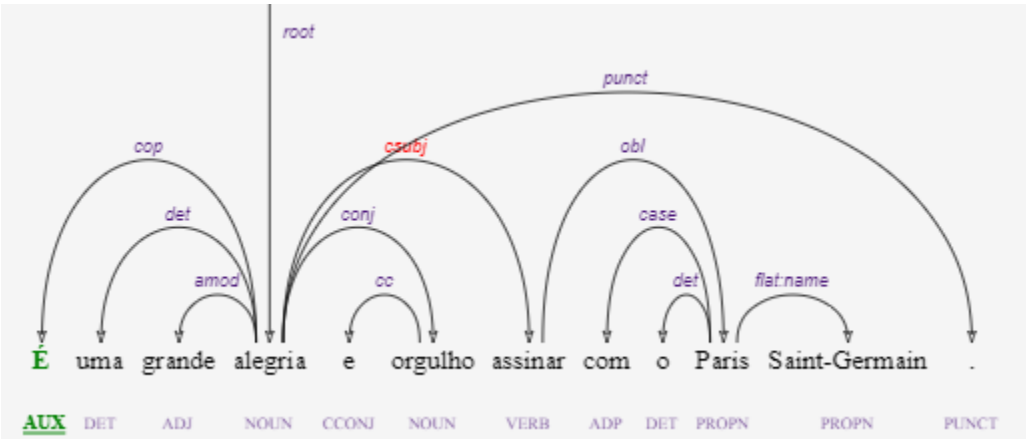


Figura 25b. FOLHA_DOC005008_SENT006 - corrigido manualmente

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 NPs à direita podem ser objeto ou sujeito posposto. As Figuras 26a e 26b ilustram uma confusão desse tipo, em que o objeto direto é oracional e está à esquerda do verbo, ao passo que o sujeito está à direita do verbo, na última posição na sentença.

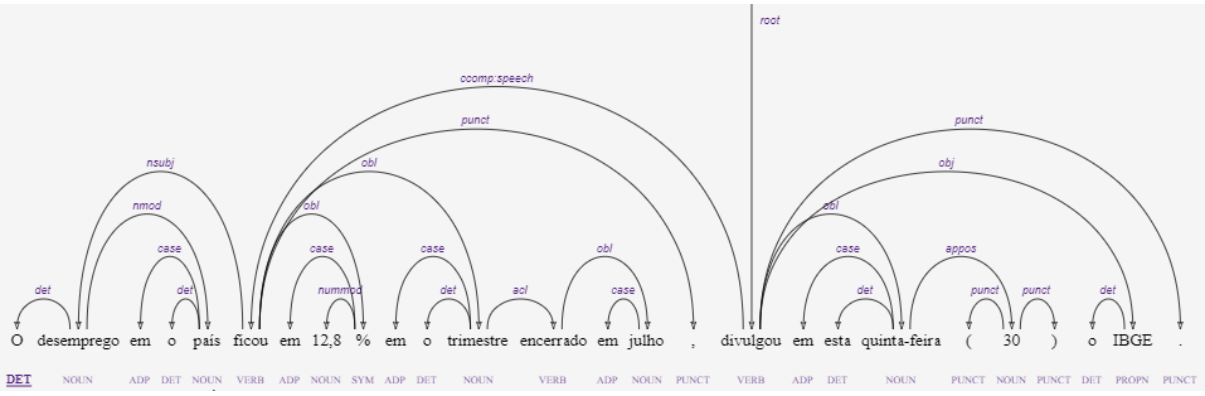


Figura 26a. FOLHA_DOC005011_SENT015 - anotado pelo parser

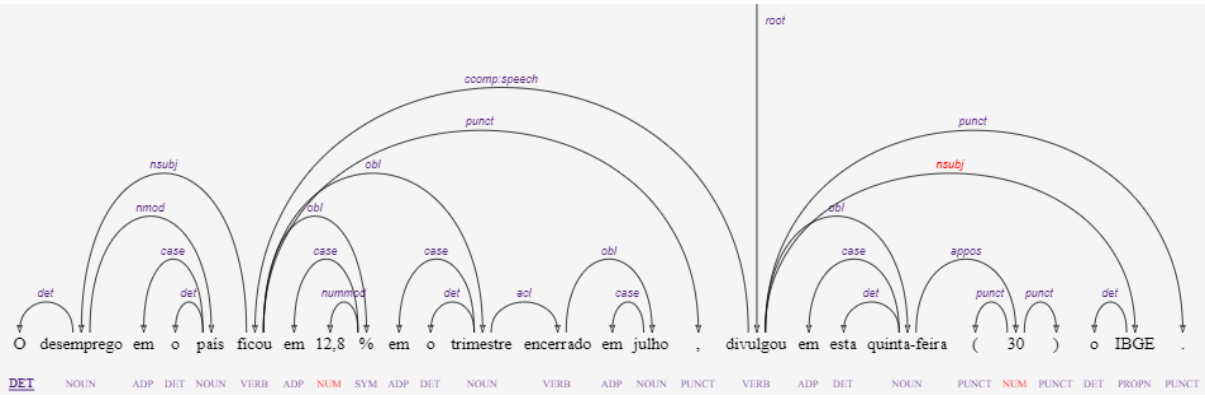


Figura 26b. FOLHA_DOC005011_SENT015 - corrigido manualmente

Um erro de sujeito que ocorreu 4 vezes na amostra foi a atribuição de dois **nsubj** a um mesmo head. Esse problema surpreende, pois no córpus de treinamento não há ocorrência desse tipo. Aliás, **nsubj** é uma das relações que só podem ocorrer uma vez para um mesmo head[12]. Esse erro é ilustrado nas Figuras 27a e 27b. Na sentença anotada, o sujeito é composto ("Andrade Gutierrez, Camargo Correa, Odebrecht"), porém, a falta de um "e" no último elemento coordenado parece ter impedido sua identificação pelo parser. Para o anotador humano é simples inferir a coordenação dos três nomes, independentemente da falta de um conectivo.

[12] As outras deprels que só podem ocorrer uma vez para um mesmo head são: **nsubj:pass, csubj**, **obj, ccomp e xcomp**.

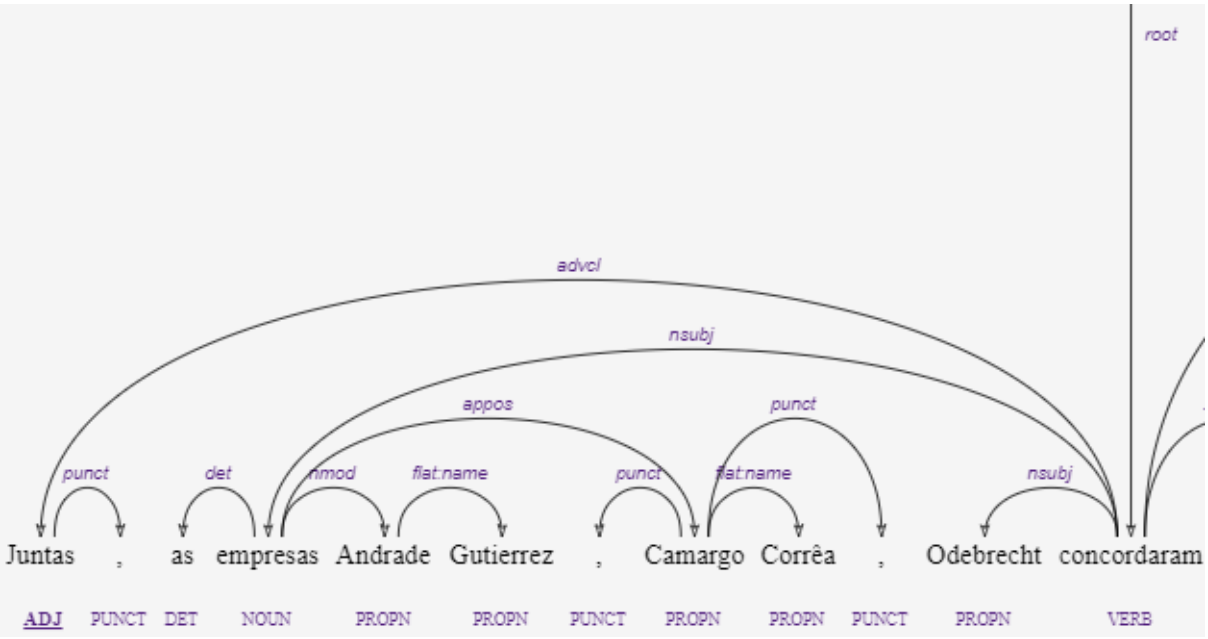


Figura 27a. FOLHA_DOC005041_SENT025 - anotada pelo parser

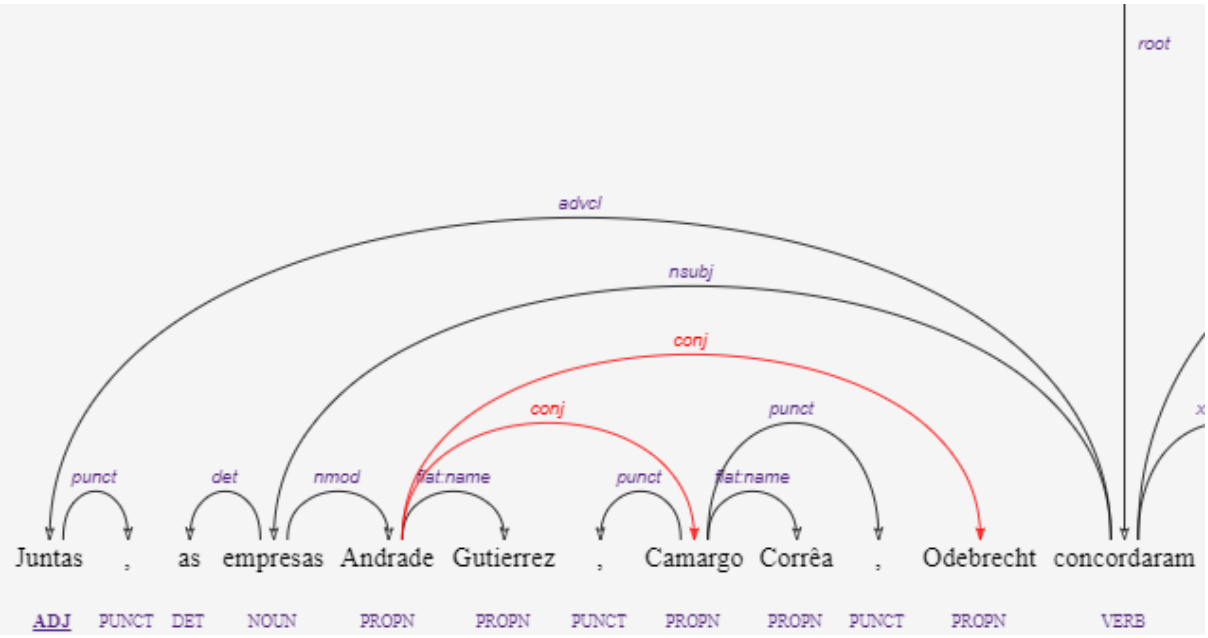


Figura 27b. FOLHA_DOC005041_SENT025 - corrigida manualmente

Formas de melhorar o aprendizado da atribuição das deprels de sujeito incluem:

1) anotação de sub-relações da deprel **expl**, utilizando **expl:impers** para sujeitos indeterminados com "se" (o que aponta o fato de o sujeito estar oculto) e **expl:pass** para voz passiva sintética, ou seja, com partícula apassivadora "se" (o que indica que o sujeito não é elíptico e deve estar à direita do verbo). Embora a anotação das sub-relações de **expl** não faça parte do que a UD exige para todas as línguas (é específica para línguas românicas), esse aprimoramento está previsto para o córpus Porttinari-base em trabalhos futuros. A anotação não foi feita inicialmente porque é uma tarefa complexa, que exige estudos preliminares, elaboração de diretrizes consistentes e treinamento especial de anotadores para a tarefa.
2) aumento de casos de sujeito posposto (**nsubj**, **nsubj:pass** e **csubj** à direita do head) com verbos que apresentaram esse comportamento no córpus de treinamento (por exemplo, pelo uso de técnicas de *data augmentation*), posto que o português é uma língua prototipicamente SVO e casos de VS são relativamente menos frequentes.

3.3.3 Erro na anotação de modificador **amod** anteposto

Em português, os adjetivos que modificam substantivos (deprel **amod**) podem ocorrer tanto à esquerda quanto à direita (anteposto ou posposto). 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 nas Figuras 28a e 28b, que apresenta um substantivo com terminação típica de adjetivo: "contencioso".

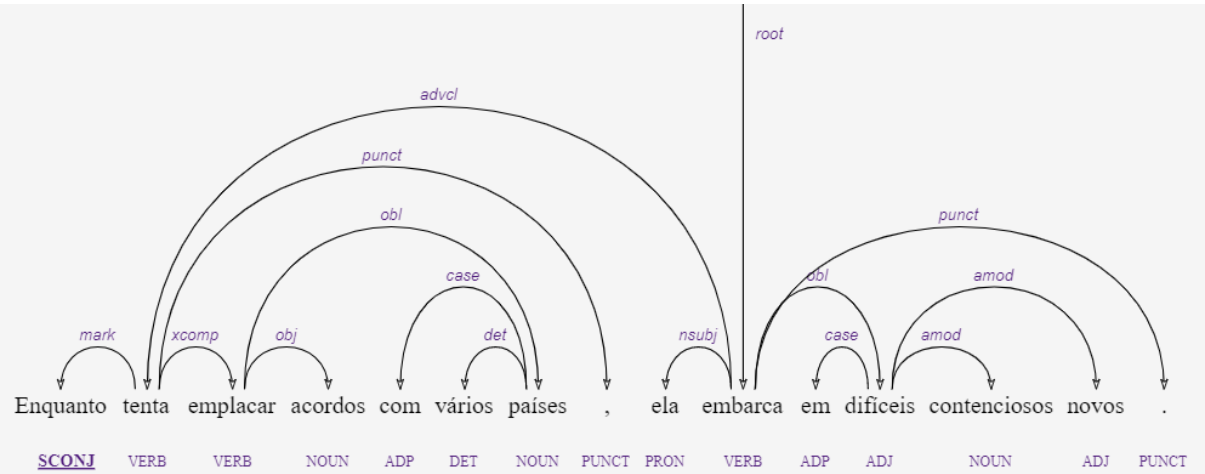


Figura 28a. FOLHA_DOC005102_SENT002 - anotada pelo parser

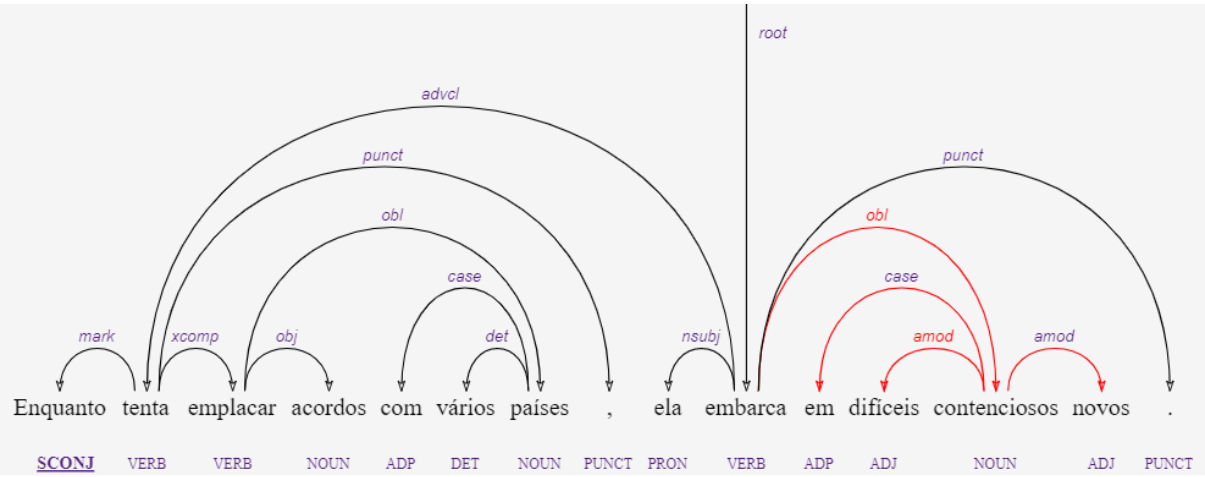


Figura 28b. FOLHA_DOC005102_SENT002 - corrigida manualmente

O problema ocorreu pouco na amostra avaliada: 6 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”, “legítimo interesse”. Seria interessante fazer *data augmentation* com dados de **amod** anteposto (deprel **amod** com sentido da direita para a esquerda), para que o parser aprenda a anotá-lo.

3.3.4 Erro na anotação de modificador **det** posposto

Esse tipo de erro é muito raro (ocorreu uma única vez na amostra), pois em português os determinantes (**det**) ocorrem majoritariamente antes do substantivo. Mas seria interessante fazer *data augmentation* com dados de **det** posposto (deprel **det** com sentido da esquerda para a direita), como o caso ilustrado nas Figuras 29a e 29b, para que o parser aprenda a anotá-lo. Embora pouco frequente no córpus Porttinari-base, é muito normal um **det** posposto: "Espelho *meu*", "proposta *esta* que...", "ele *próprio*", "eu *mesmo*".

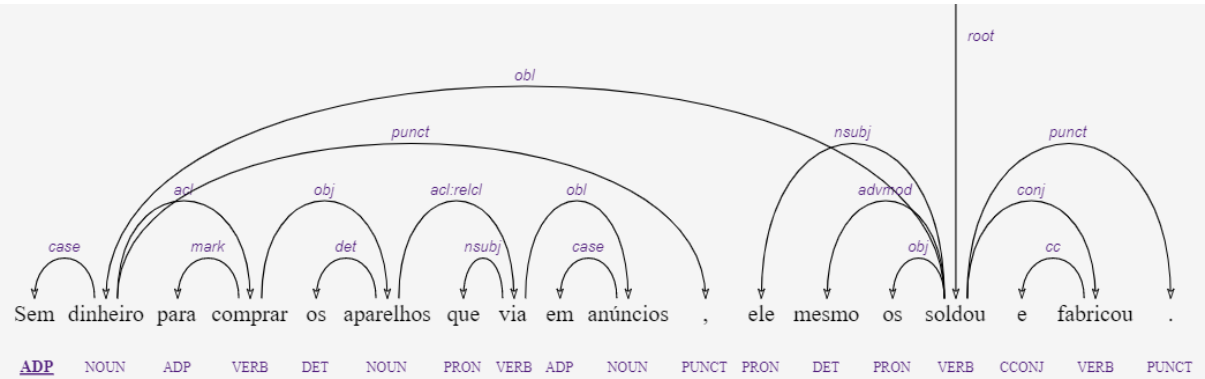


Figura 29a. FOLHA_DOC005108_SENT015 - anotada pelo parser

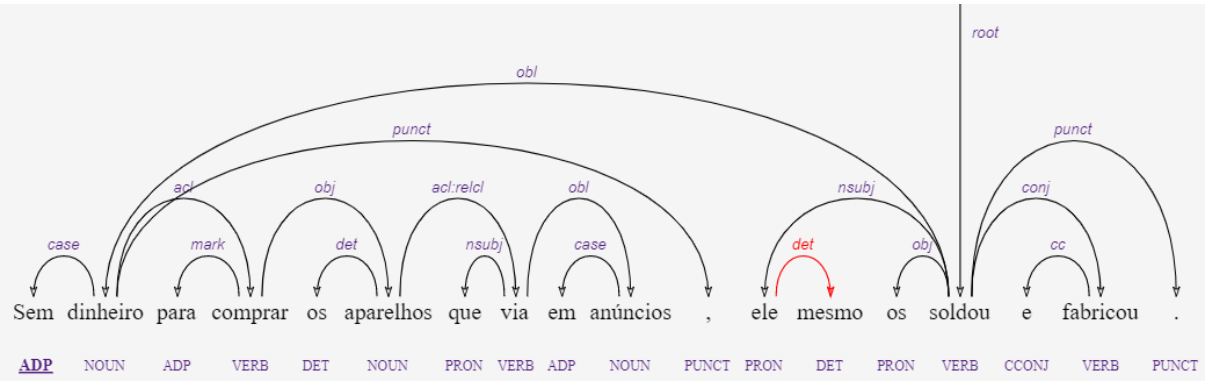


Figura 29b. FOLHA_DOC005108_SENT015 - corrigida manualmente

3.3.5 Erro na identificação de expressões **fixed**

A deprel **fixed** é utilizada para anotar multipalavras funcionais, pois entre palavras funcionais não há relação possível, uma vez que uma palavra funcional só pode ser dependente (nunca head) de uma relação sintática. Essa deprel também é utilizada para pronomes compostos, como "o qual" e algumas locuções adverbiais que não têm sentido composicional, como "mais de o que" (Figuras 30a e 30b). A lista de expressões **fixed** proposta no projeto POeTiSA está disponível no site do projeto[13].

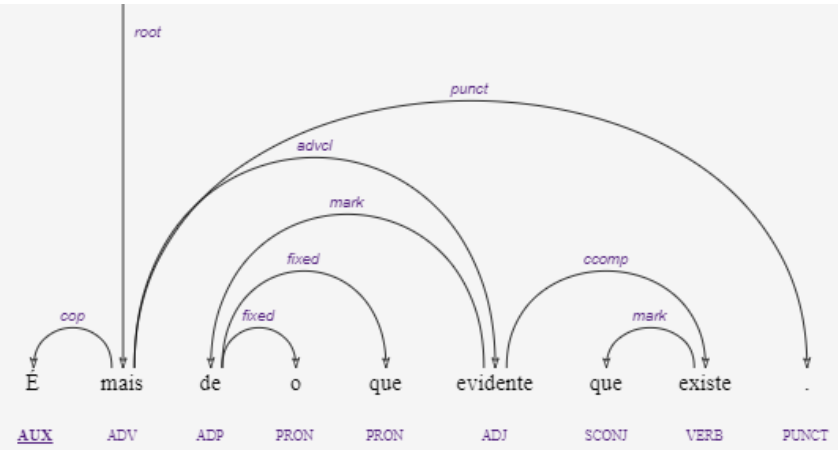


Figura 30a. FOLHA_DOC005036_SENT011 - anotada pelo parser

13 https://docs.google.com/spreadsheets/d/1gR5qlR3PVZ4I6KJKHAZE1Kyunf6rcLuije5jwkszn78/edit#gid=0

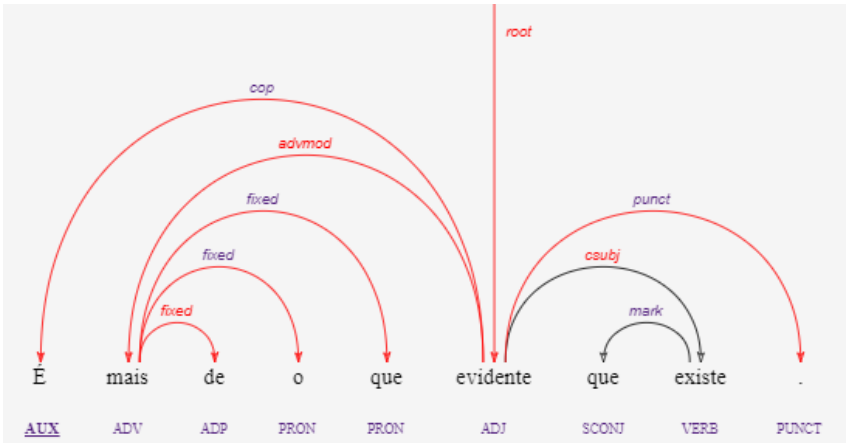


Figura 30b. FOLHA_DOC005036_SENT011 - corrigida manualmente

O problema da não identificação dessas expressões parece estar relacionado somente à esparsidade de dados, pois ocorreu 22 vezes na amostra analisada, sempre com expressões que apresentam baixa frequência no córpus de treinamento. Como expressões **fixed** com cinco ocorrências no córpus de treinamento foram reconhecidas pelo parser, acredita-se que uma estratégia de *data augmentation* dessas expressões menos frequentes possa melhorar o aprendizado automático.

Em alguns casos, observou-se que o parser analisou de forma alternativa (e sintaticamente plausível), expressões que foram definidas como **fixed** nas diretrizes de anotação. As Figuras 31a e 31b ilustram um caso desse tipo. A vantagem de anotar algumas dessas expressões como **fixed** tem, portanto, mais amparo na semântica, pelo fato de não terem sentido composicional, do que na sintaxe. Por isso, se o parser não reconhece uma expressão **fixed**, mas anota o segmento de forma sintaticamente aceitável, não deveria ter seu desempenho penalizado.

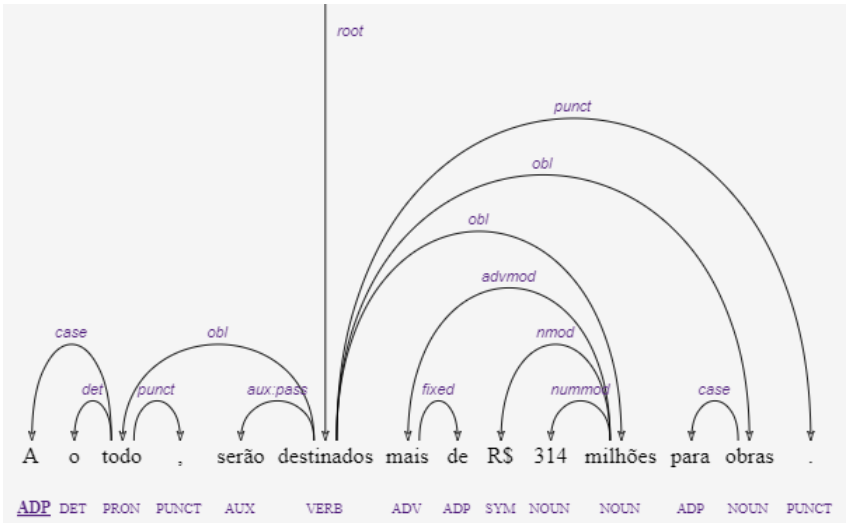


Figura 31a. FOLHA_DOC005010_SENT008 - anotada pelo parser

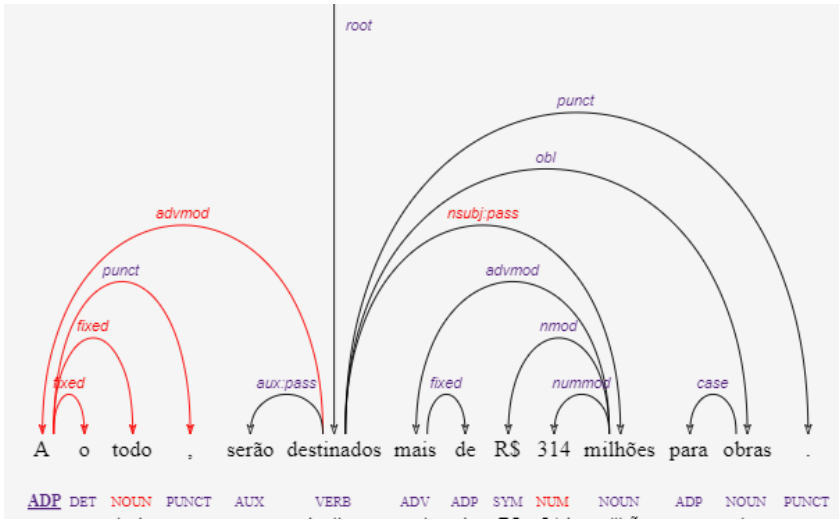


Figura 31b. FOLHA_DOC005010_SENT008 - corrigida manualmente

3.3.6 Erro na identificação de **flat:name**

A deprel **flat:name** serve para unir vários nomes próprios que constituem um nome próprio composto. Erros na identificação de **flat:name** ocorreram 7 vezes na amostra: nomes próprios contíguos ora foram anotados como um único nome composto, ora como dois nomes próprios. Este é um problema inerente às decisões de anotação adotadas no projeto POeTiSA: com o objetivo de não perder a identidade de entidades mencionadas, o referido projeto optou por anotar como nome

próprio (PROPN) todas as palavras com inicial maiúscula que fazem parte de um nome próprio, independentemente de serem palavras comuns da língua. Da mesma forma, usou-se a deprel **flat:name** para reunir todas as palavras anotadas com PROPN que fazem parte de um nome próprio composto. As palavras funcionais em minúsculas, internas ao nome próprio composto, são anotadas com PoS tags comuns e ligadas aos nomes próprios com relações de dependência convencionais.

Essa decisão, contudo, não está em conformidade com as orientações da UD, que recomenda anotar toda palavra comum da língua, independentemente de ser parte de nome próprio, com PoS tags comuns e relações convencionais de dependência. As Figuras 32a e 32b ilustram, respectivamente, a opção adotada no POeTiSA e a anotação recomendada pela UD.

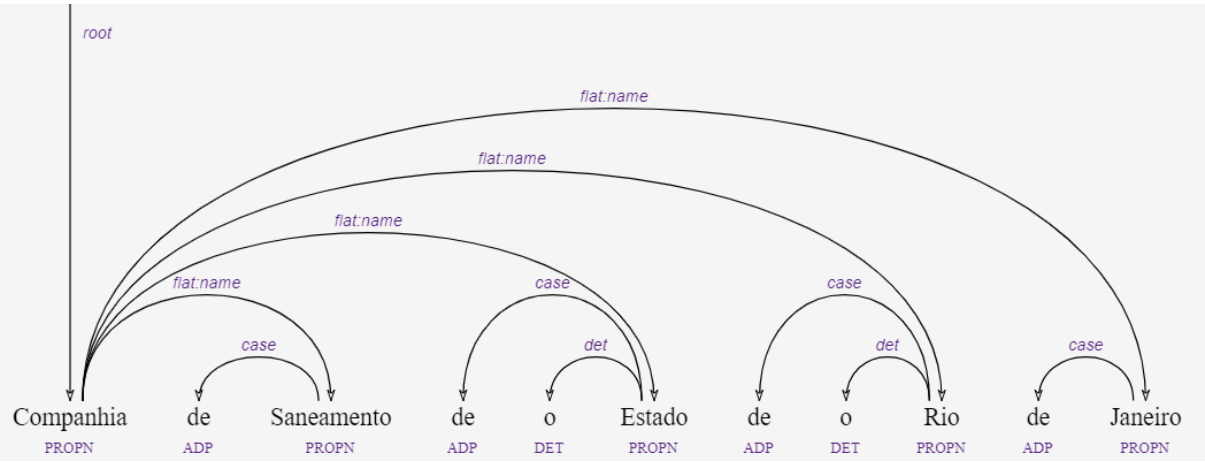


Figura 32a. Anotação de nome próprio composto segundo diretrizes do projeto POeTiSA

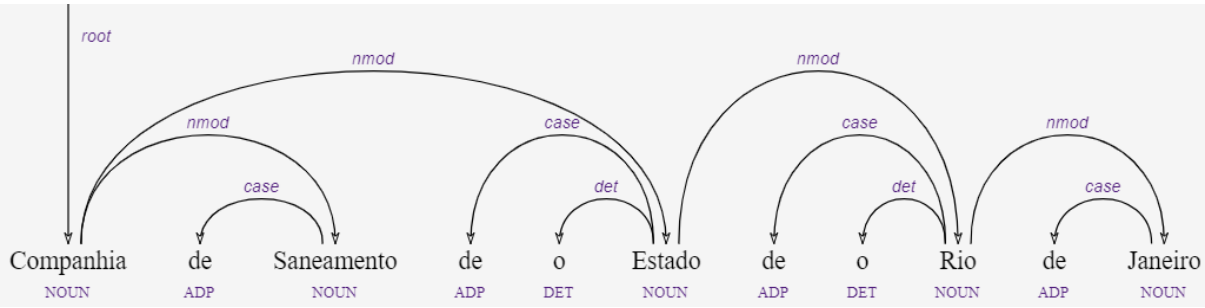


Figura 32b. Anotação de um nome próprio composto segundo diretrizes da UD

A opção de anotação adotada, contudo, traz a possibilidade de um erro: quando dois nomes próprios diferentes ocorrem contiguamente, ambos poderão ser anotados como se fossem parte de um mesmo **flat:name**, como mostram as Figuras 33a e 33b.

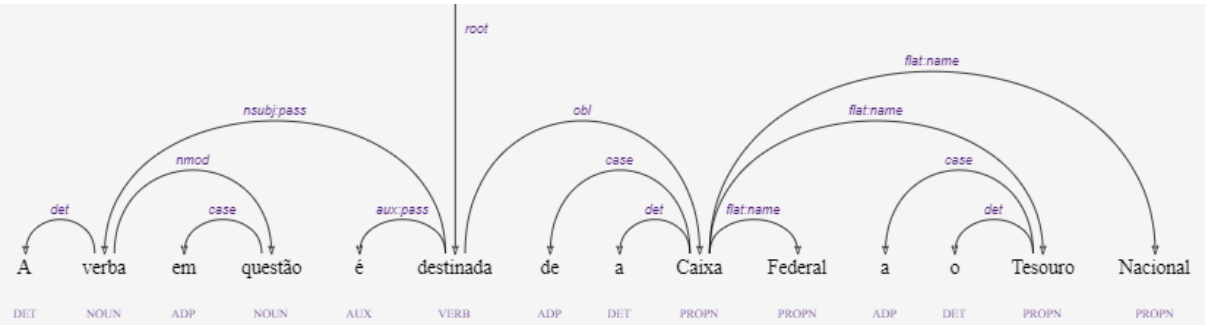


Figura 33a. FOLHA_DOC005005_SENT006 - anotada pelo parser

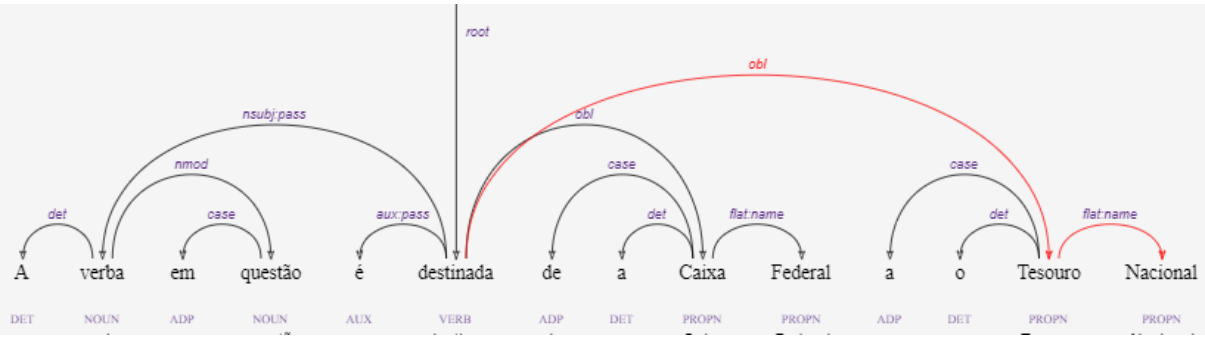


Figura 33b. FOLHA_DOC005005_SENT006 - corrigida manualmente

4. Conclusões

No geral, o parser avaliado apresentou um ótimo desempenho, pois 50% das 600 sentenças não apresentaram nenhum erro. Nas sentenças que apresentaram erros, 23% tinham um único erro e 11% tinham dois erros, ou seja, eram erros pontuais, que não acarretavam outros erros.

As poucas sentenças que tiveram muitos erros apresentaram erro de identificação de expressões **fixed**, de **root** ou algum erro motivado por erro de pré-processamento, como uma PoS tag incorreta. Nesses casos, não se trata de diferentes erros de *parsing*, mas de erros acarretados por um único erro. O erro de **root**, por exemplo, obriga todos os sintagmas ligados ao **root** a "mudarem o endereço" de seus heads.

Como pode ser observado no gráfico da Figura 34, os erros mais frequentes no conjunto de 600 sentenças avaliadas são mais de natureza semântica (head de PPs, head de orações modificadoras e head de coordenações) do que de natureza sintática (cruzamento de arcos, **flat:name**, **root**, sujeito e **fixed**).

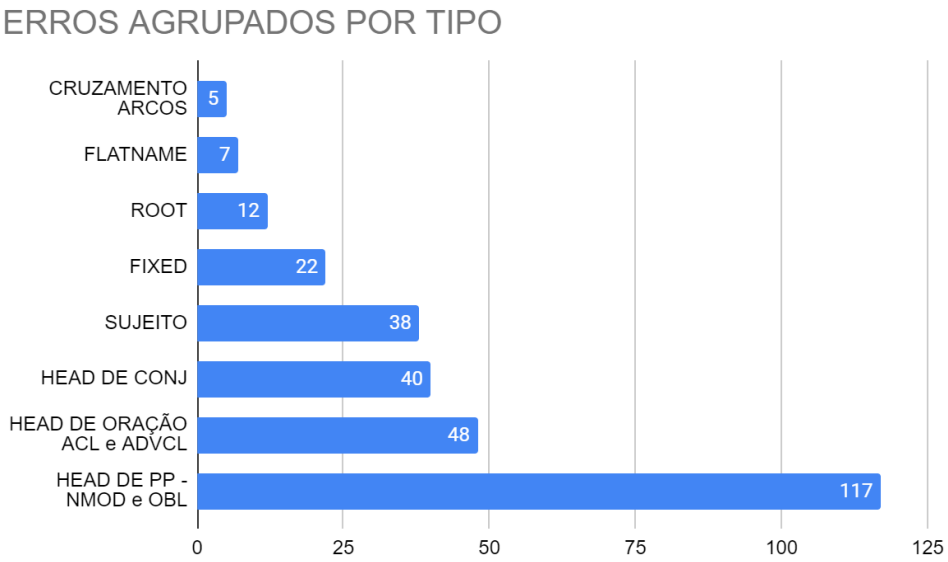


Figura 34. Gráfico por tipos de erro mais frequentes nos 30 pacotes

O gráfico exibido na Figura 35 mostra a quantidade de erros de natureza sintática em cada um dos 30 pacotes (contendo 20 sentenças cada um, como comentado anteriormente). Como as sentenças que integram os pacotes foram escolhidas de forma aleatória, era de se esperar que erros recorrentes aparecessem com frequência semelhante em todos os pacotes. Como a frequência não é semelhante, é possível concluir que os erros não são recorrentes. Isso reflete o que foi observado na análise qualitativa: erros de natureza sintática, quando ocorrem, parecem justificar-se principalmente pelo fato de as sentenças analisadas apresentarem um padrão menos comum na língua, como é o caso dos sujeitos pospostos, ou pela ocorrência de um léxico menos frequente, como é o caso de algumas expressões fixed.

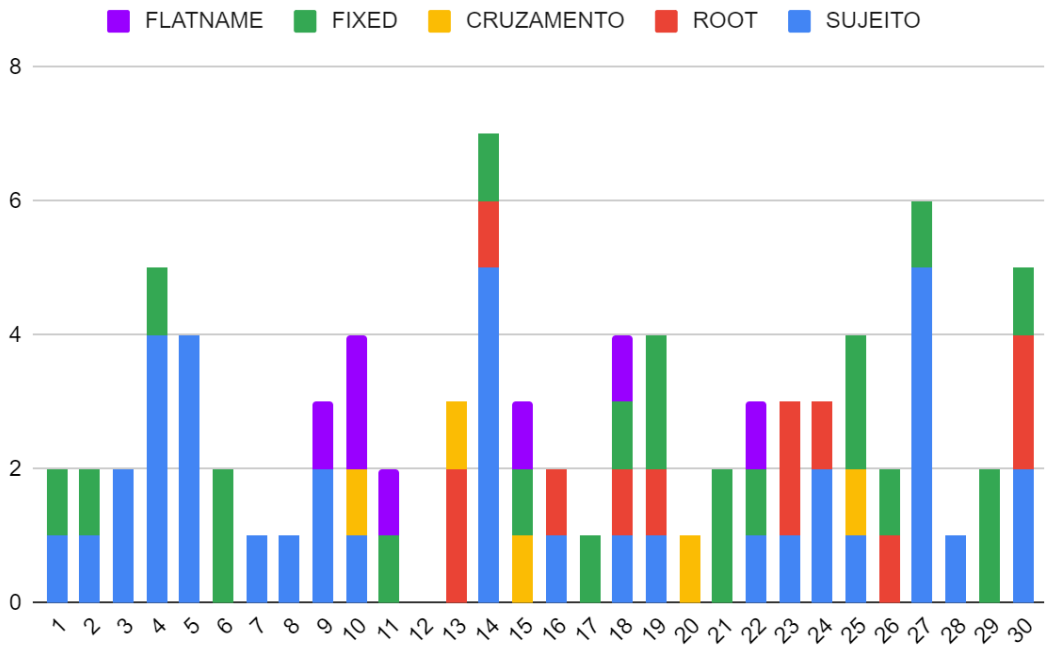


Figura 35. Gráfico por tipo de erro sintático nos 30 pacotes analisados

A média e o desvio padrão por pacote para cada tipo de erro sintático são mostrados a seguir, na Tabela 2, dando uma ideia de com que frequência um usuário do parser pode esperar encontrar esses erros em seus dados. Por exemplo, erros de sujeito e de **fixed** podem ser mais esperados. É fato que os dados analisados qualitativamente são limitados (dado o esforço humano requerido para tal tarefa), mas acredita-se que esses números possam auxiliar o usuário em eventuais esforços de revisão humana dos dados anotados automaticamente.

Também é interessante notar a variação da incidência dos erros nos pacotes. Pode-se ver, por exemplo, que o erro de sujeito se destaca em alguns poucos pacotes, mas ocorre relativamente menos nos outros pacotes (como demonstra o desvio padrão mais alto para esse tipo de erro). Os erros de **flat:name**, por sua vez, ocorrem pouco e em poucos pacotes, daí sua média e desvio padrão pequenos.

Tabela 2. Média e desvio padrão de ocorrência dos erros sintáticos por pacote

Erros de natureza sintática	média nos pacotes	desvio-padrão nos pacotes
SUJEITO	1,27	1,46
ROOT	0,40	0,67
CRUZAMENTO	0,17	0,38
FIXED	0,73	0,74
FLATNAME	0,23	0,50

Como comentado anteriormente, acredita-se que os problemas de natureza sintática possam ser minimizados usando técnicas de *data augmentation* para diminuir a esparsidade de alguns dados, como sujeitos pospostos (relações **nsubj** da esquerda para a direita), adjetivos antepostos (relações **amod** da direita para a esquerda), determinantes pospostos (relações **det** da esquerda para a direita) e expressões **fixed** (constantes de uma tabela disponível no site do projeto).

Observou-se que o parser não aprendeu totalmente algumas restrições sintáticas, embora sejam poucas as ocasiões em que as violou:

- um head não pode ter mais de um **nsubj** (4 violações em 600 sentenças);
- um head não pode ter mais de um **obj** (2 violações em 600 sentenças);
- head e dependente de **amod**, **acl** e **appos** não podem ter número e gênero diferentes (6 violações em 600 sentenças);
- arcos não podem se cruzar (5 violações em 600 sentenças).

Ao contrário do que foi observado nos erros de natureza puramente sintática, os erros de natureza semântica (head de PPs, head de orações

modificadoras, head de coordenação) apresentam uma distribuição mais regular entre os pacotes, como pode ser observado no gráfico da Figura 36. Esses problemas dependem de interpretação semântica e conhecimento de mundo para serem resolvidos, ou seja, não existe pista sintática para decidir qual o head correto de um modificador.

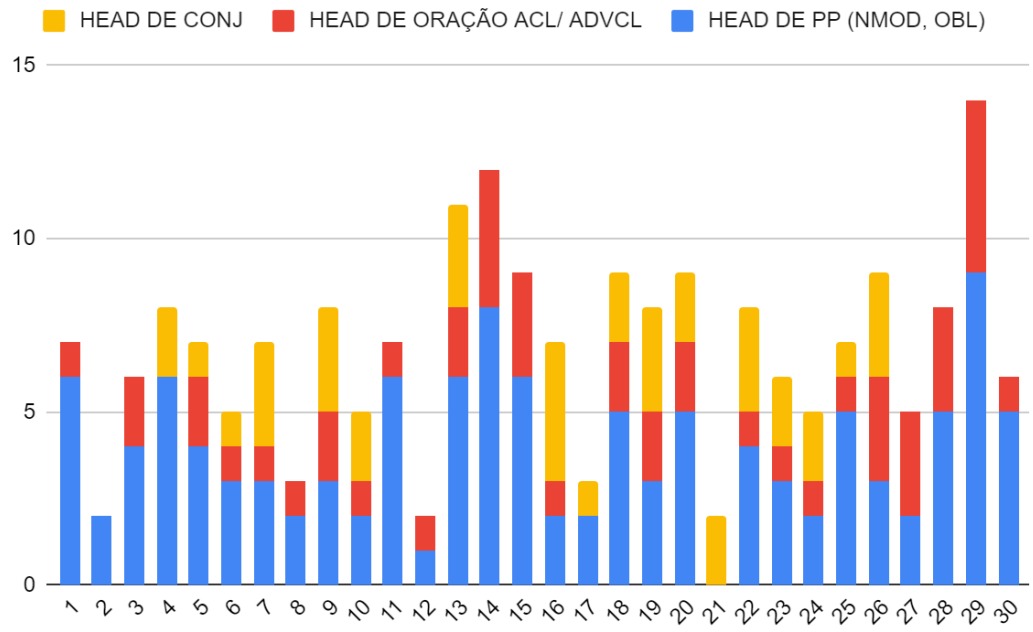


Figura 36. Gráfico por tipo de erro semântico nos 30 pacotes analisados

Novamente, são exibidos na Tabela 3 a média e o desvio padrão por pacote para cada um dos casos citados.

Tabela 3. Média e desvio padrão de ocorrência dos erros semânticos por pacote

Erros de natureza semântica	média nos pacotes	desvio-padrão nos pacotes
HEAD DE PP (NMOD, OBL)	3,90	2,07
HEAD DE ORAÇÃO ACL/ ADVCL	1,60	1,19
HEAD DE CONJ	1,33	1,30

No que diz respeito à ambiguidade do head de PPs, especificamente, é importante ressaltar que o parser anotou ao todo 453 **nmods** e 437 **obl**, num total de 890 PPs. Como 112 PPs estavam com head incorreto, o parser acertou o head de 778 PPs, o que representa 87,41% de precisão.

Acredita-se que a anotação de sub-relações, com informações semânticas de modificadores circunstanciais (tempo, local, causa, finalidade, conformidade, etc.) possa contribuir para o aprendizado, pois os modificadores restantes serão, muito provavelmente, complementos dos verbos e substantivos a que se ligam. Acredita-se também que anotar verbos suporte, bem como verbos aspectuais[14] e modais[15], no nível de *features*, possa ser igualmente benéfico para o aprendizado, pois esses verbos assumem o head dos modificadores circunstanciais, porém nunca são head de argumentos próprios sob forma de PPs. Por fim, é provável que uma camada de anotação de papéis semânticos sobre as árvores sintáticas possa dar um *feedback* para o parser, melhorando o aprendizado da estrutura argumental de verbos e nomes predicadores.

[14] Verbos aspectuais são verbos que atuam como semiauxiliares em locuções verbais, acrescentando um aspecto ao verbo pleno que o sucede. Exemplos clássicos de verbos aspectuais são os frequentativos (por exemplo: viver fazendo, andar fazendo), os incoativos (passar a fazer, começar a fazer) e os terminativos (parar de fazer, deixar de fazer).

[15] Verbos modais são verbos que atuam como semiauxiliares em locuções verbais, modalizando o verbo pleno que os sucede. Os mais comuns são os deônticos, que exprimem obrigatoriedade (dever), necessidade (precisar, necessitar) ou permissão (poder). Por exemplo: deve fazer, precisa fazer, pode fazer.

Agradecimentos

Este trabalho foi executado no Centro de Inteligência Artificial (C4AI-USP) com apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation. Este projeto também é 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 Residência em TIC 13, DOU 01245.010222/2022-44.

Agradecemos também à Lucelene Lopes pela montagem do córpus que constitui o *testbed* para avaliações do analisador sintático e pela constituição dos pacotes de sentenças, extraídas do mesmo córpus, cuja análise é objeto deste relatório.

Referências Bibliográficas

DURAN, M.S. 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. São Carlos-SP, Setembro, 55p. (2021). Disponível em: https://drive.google.com/file/d/1BddPswn-_Ioo-A5GsldA1cO1kqbcCahb/view?usp=sharing

DURAN, M.S. 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. (2022). Disponível em: https://drive.google.com/file/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view?usp=sharing

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

MIRANDA, L.G.M.; PARDO, T.A.S. An Improved and Extended Annotation Tool for Universal Dependencies-based Treebank Construction. In: **Proceedings of the PROPOR Demonstrations Workshop**, 2022, p. 1-3. Disponível em: https://drive.google.com/file/d/1Gz9k3-SU72zXx6v2a0lTrutHVAtpUMUX/view?usp=sharing

NIVRE, J.; MARNEFFE, M.; GINTER, F.; HAJIČ, J.; MANNING, C.; 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), 2020, p. 4034-4043.

PARDO, T.A.S.; DURAN, M.S.; LOPES, L.; DI FELIPPO, A.; ROMAN, N.T.; NUNES, M.G.V. Porttinari - A large multi-genre treebank for Brazilian Portuguese. In the **Proceedings of the XIV Symposium in Information and Human Language** (STIL 2021), 2021, p. 1-10.

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

STRAKA, M. 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, 2018, p. 197-207.