Programa de Pós-Graduação em

LINGUÍSTICA

DESCRIÇÃO SINTÁTICO-SEMÂNTICA DE NOMES PREDICADORES EM TWEETS DO MERCADO FINANCEIRO EM PORTUGUÊS

SÃO CARLOS

2024

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE EDUCAÇÃO E CIÊNCIAS HUMANAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA

DESCRIÇÃO SINTÁTICO-SEMÂNTICA DE NOMES PREDICADORES EM TWEETS DO MERCADO FINANCEIRO EM PORTUGUÊS

BRYAN KHELVEN DA SILVA BARBOSA
BOLSISTA SOFTEX-INOVAUSP-MCTI

Dissertação apresentada ao Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos como parte dos requisitos para a obtenção do título de Mestre em Linguística.

Orientadora: Profa. Dra. Ariani Di-Felippo

São Carlos – São Paulo – Brasil

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Educação e Ciências Humanas
Programa de Pós-Graduação em Linguística

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Bryan Khelven da Silva Barbosa, realizada em 12/08/2024.

Comissão Julgadora:

Profa. Dra. Ariani Di Felippo (UFSCar)

Prof. Dr. Thiago Alexandre Salgueiro Pardo (USP)

Prof. Dr. Otto Araujo Vale (UFSCar)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Linguística.

Agradecimentos

É com profundo apreço que expresso minha gratidão à Profa. Ariani Di-Felippo, minha orientadora, por sua excepcional orientação, sabedoria e apoio constante durante o desenvolvimento deste projeto. Sua expertise e percepção foram pilares para a excelência aqui alcançada.

Dirijo um agradecimento especial à Profa. Maria das Graças Volpe Nunes, cujo comprometimento inestimável com a anotação sintática do córpus DANTEStocks, bem como com a elaboração do "Manual de Diretrizes de Anotação de Relações de Dependência em Tweets do Mercado Financeiro", Apêndice A do presente trabalho, provou-se fundamental para o sucesso deste estudo. Estendo minha gratidão ao Prof. Norton Trevisan Roman, pela sua orientação e liderança no projeto DANTE, que foram essenciais para o meu entendimento dos termos do mercado financeiro e para a progressão geral deste trabalho. Agradeço igualmente ao Prof. Thiago Alexandre Salgueiro Pardo, cuja liderança no projeto POeTiSA, ao qual o DANTE está vinculado, e suas recomendações sobre os processos de Aprendizado de Máquina empregados, foram indispensáveis para a evolução deste trabalho e sua contribuição significativa para a pesquisa em Processamento Automático de Línguas Naturais. Agradeço ainda a Isabela e Breno pelo auxílio quanto à tarefa de avaliação da concordância da anotação semântica, bem com as contribuições acerca de todo o processo de anotação de argumentos.

Não posso deixar de agradecer à minha família e amigos, pelo apoio incondicional, paciência e incentivo ao longo desta jornada acadêmica. Sua força foi o meu alicerce nos momentos mais desafiadores.

O caminho percorrido no mestrado representou uma jornada enriquecedora, permitindo-me explorar cada vez mais as nuances da área de PLN, sob perspectivas tanto linguísticas quanto computacionais. Participar do Núcleo Interinstitucional de Linguística Computacional (NILC) e interagir e conhecer os pioneiros e entusiastas do PLN no Brasil foi e é uma honra verdadeiramente inspiradora!

Neste momento de profunda reflexão e gratidão, volto meus pensamentos ao vasto e misterioso Universo que nos abriga. Com humildade e reverência, agradeço ao Cósmico pelas diversas possibilidades de realização que me foram concedidas até aqui.

Reconheço a mão invisível do Grande Arquiteto do Universo, guiando-me silencio-

samente através dos desafios e aprendizados desta jornada. A sabedoria eterna, que ecoa através das eras, foi uma lanterna a iluminar meu caminho, permitindo-me vislumbrar a perfeição e a justiça na regência do Todo.

Aos Mestres invisíveis, que com sua presença sutil, inspiraram e fortaleceram minha mente e meu espírito nessa busca pelo conhecimento, ofereço meu mais sincero agradecimento.

Agradeço então ao Cósmico pela oportunidade de crescer e evoluir dentro deste vasto laboratório de vida, onde cada desafio superado e cada descoberta feita são passos em direção à Luz Maior. Que a luz da verdade continue a guiar minha jornada, e que eu possa, de alguma forma, contribuir para a elevação e o bem-estar da humanidade.

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.

A todos que apoiaram este trabalho, seja direta ou indiretamente, meus genuínos agradecimentos!

Com meus sinceros sentimentos,
Muito obrigado!
Bryan Khelven

Resumo

Neste trabalho, descreveu-se a estrutura de argumentos (estrutura-A) dos nomes predicadores (Npred) que ocorrem no córpus de tweets do mercado financeiro DANTEStocks, pois há preferência pelo uso desse tipo de predicador em gêneros digitais desse domínio. O objetivo específico foi verificar: (i) a presença/ausência dos argumentos (Arg) nos tweets, (ii) a realização sintática dos Arg e (iii) a influência dos fenômenos linguísticos dos tweets na realização da estrutura-A dos nomes. Especificamente, 145 Npred e 1.756 instâncias (tweets com ao menos um Npred) do córpus foram descritas em nível sintático-semântico. Quanto à sintaxe, fez-se a anotação semiautomática de todo o DANTEStocks de acordo com o modelo *Universal Dependencies* (UD). Em nível semântico, as árvores de dependência sintática guiaram a anotação manual das instâncias segundo o NomBank. O mapeamento sintático-semântico permitiu observar que: (i) a estrutura-A dos Npred de valência um (V_1) é sempre preenchida na sintaxe, (ii) a estrutura-A dos Npred de $V_{>1}$ apresenta algum Arg ausente, (iii) a maioria dos Npred analisados é de V_3, com apenas 2 Arg na maioria das instâncias, (iv) as *deprel* que mais frequentemente conectam os Npred a seus Arg são **nmod** e **amod**, e (v) a realização sintática da estrutura-A em 24 instâncias foi reduzida pela ocorrência de fenômenos dos tweets (truncamento e justaposição de elementos). Esses resultados enriquecem o arcabouço de estudos descritivos sobre os aspectos lexicais da linguagem dos tweets do mercado financeiro. Aliás, a descrição da valência sintático-semântica dos Npred foi sistematizada no NounBank.DS, um repositório lexical online que pode subsidiar outras pesquisas linguístico-computacionais. Como contribuição para o Processamento das Línguas Naturais (PLN), destaca-se a anotação sintática-UD do DANTEStocks, a qual deu origem ao primeiro *tweebank* em português. Esse recurso permitiu o desenvolvimento do primeiro *parser*-UD para UGC na referida língua. A anotação semântica à la NomBank de uma parcela do córpus também gerou um importante recurso. Assim, este trabalho produziu recursos linguísticos de referência e uma ferramenta (*parser*) para o processamento automático de tweets em português, os quais são esseciais para o desenvolvimento de aplicações de PLN destinadas a esse tipo de CGU.

Palavras-chave: PLN, tweet, nome predicador, estrutura de argumento.

Abstract

In this study, the argument structure (A-structure) of predicative nouns (Npred) occurring in the financial market tweet corpus DANTEStocks was described, given the preference for using this type of predicator in digital genres of this domain. The specific objective was to verify: (i) the presence/absence of arguments (Arg) in the tweets, (ii) the syntactic realization of Args, and (iii) the influence of linguistic phenomena in tweets on the realization of the A-structure of the nouns. Specifically, 145 Npreds and 1,756 instances (tweets with at least one Npred) from the corpus were described at the syntactic-semantic level. Syntactically, semi-automatic annotation of the entire DANTEStocks was carried out according to the *Universal Dependencies* (UD) model. At the semantic level, syntactic dependency trees guided the manual annotation of instances according to NomBank. The syntactic-semantic mapping revealed that: (i) the A-structure of valency one (V_1) Npreds is always filled in syntax, (ii) the A-structure of Npreds with $V_{>1}$ shows some missing Args, (iii) most analyzed Npreds are of V_3, with only 2 Args in most instances, (iv) the *deprels* most frequently connecting Npreds to their Args are **nmod** and **amod**, and (v) the syntactic realization of the A-structure in 24 instances was reduced by tweet-specific phenomena (truncation and juxtaposition of elements). These results enrich the descriptive framework of lexical aspects of the language in financial market tweets. Moreover, the syntactic-semantic valency description of Npreds was systematized in NounBank.DS, an online lexical repository that can support further linguistic-computational research. A contribution to Natural Language Processing (NLP) is the UD-syntactic annotation of DANTEStocks, which led to the creation of the first Portuguese *tweebank*. This resource enabled the development of the first UD-parser of UGC for this language. The NomBank-like semantic annotation of a portion of the corpus also generated a significant resource. Thus, this study produced reference linguistic resources and a tool (*parser*) for the automatic processing of Portuguese tweets, which are essential for developing NLP applications targeting this type of UGC.

Keywords: NLP, tweet, predicative noun, argument structure.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Lista de Abreviaturas e Siglas

AM – Aprendizado de Máquina
AMR – *Abstract Meaning Representation*
Arg – Argumento
Arg0 – Argumento 0 (sujeito)
Arg1 – Argumento 1 (objeto direto)
Arg2 – Argumento 2 (objeto indireto ou outro argumento)
Arg3 – Argumento 3 (outro argumento adicional)
ArgM – Argumento modificador
AS – Análise de Sentimento
B3 – Brasil, Bolsa, Balcão
Bi-LSTM – *Bidirectional Long Short-Term Memory Network*
CGU – Conteúdo Gerado por Usuário
CVS – Construções com Verbo-Suporte
DANTE – *Dependency-ANalised corpora of TwEets*
DJIA – *Dow Jones Industrial Average*
DUPB – Dicionário de Usos do Português do Brasil
Hn – Etiqueta funcional para descrever nomes compostos por hífen
IBOVESPA – Índice da Bolsa de Valores de São Paulo
JSON – *JavaScript Object Notation*
LAS – *Labeled Attachment Score*
MO – Mineração de Opiniões
NLTK – *Natural Language Toolkit*
NOM – Nominalização
Npred – Nome predicador
PLN – Processamento Automático de Línguas Naturais
PoS – *Part-of-Speech*
PTB – *Penn Treebank Corpus*
REL – Relação
SN – Sintagma Nominal
SPrep – Sintagma Preposicional ou Preposicionado
TF-IDF – *Term Frequency – Inverse Document Frequency*
UAS – *Unlabeled Attachment Score*
UD – *Universal Dependencies*
UGC – *User-Generated Content*

Sumário

1 Introdução

1.1 Contexto e Justificativa

Segundo os dados do último *Global Overview Report*[1], o número total de usuários do *Twitter*[2] em 2023 era de aproximadamente 550 milhões, sendo, assim, a 14ª rede social mais usada no mundo no referido ano. O Brasil, em especial, tem a quarta maior base de usuários do *Twitter* do mundo, com quase 25 milhões de usuários.

Como resultado dessa proeminência, a referida plataforma, que oferece um serviço que pode ser visto como uma mistura de *microblog* e rede social, tem despertado, há tempos, o interesse de diversas áreas, incluindo política, saúde, finanças e outras.

No contexto das finanças, reconhece-se que processar as opiniões dos usuários do *Twitter* sobre os mercados financeiros permite reunir informações pertinentes sobre os mercados que podem ser usadas para prever mudanças nos preços das ações. Em outras palavras, os tweets sobre mercados financeiros são reconhecidamente fontes importantes de conteúdo pela comprovada relação entre as mensagens sobre uma ação e o volume movimentado das ações no mercado (Bollen; Mao; Zeng, 2011).

Em decorrência da comprovada relação entre o comportamento das ações e o conteúdo do *Twitter*, pesquisadores do Processamento Automático das Línguas Naturais (PLN) tem desenvolvido aplicações computacionais que visam à análise de sentimentos (AS) e mineração de opiniões (MO), sobretudo com o intuito de predizer os movimentos sobre as ações ou ativos nas bolsas de valores. Isso quer dizer que métodos de predição

[1] <https://datareportal.com/reports/digital-2023-global-overview-report>

[2] Embora a plataforma tenha sido renomeada para "X" e as mensagens nela circulantes para "posts" após a aquisição da plataforma por Elon Musk e conseguinte reestruturação ocorrida em 2022, optou-se por utilizar, neste trabalho, as denominações originais (ou seja, "Twitter" para a plataforma e "tweets" para as mensagens) em concordância com a época em que o córpus aqui utilizado foi compilado.

estão sendo baseados na extração de sentimentos e opiniões dos tweets, como pode ser visto, por exemplo, em trabalho os de Bollen, Mao e Zeng (2011), Carosia, Coelho e Silva (2019), Dhabe *et al.* (2023) e em outros muitos.

Para o desenvolvimento de tais aplicações, os córpus de tweets com diferentes tipos de anotação linguística de referência (isto é, manualmente revisadas, sobretudo no que diz respeito ao conhecimento morfo(-sintático)), os chamados *tweebanks*, são essenciais para o treinamento e teste de diversos modelos de PLN.

Enquanto gênero, o tweet parece possuir resquícios de outros (como notícia, propaganda, bilhete, etc.) que foram modificados para atender às necessidades comunicativas na rede (Freitas; Barth, 2015). Assim, a linguagem dos tweets é informal, isto é, não-padrão. As mensagens possuem até 280 caracteres[3] e tendem a conter sequências de sintagmas curtos, orações ou fragmentos, com ou sem problemas de pontuação.

Suas características ortográficas e lexicais incluem fenômenos dependentes da plataforma (como *hashtag*, menções, URL e truncamento), simplificações (p.ex.: acrônimo e inicialismo), empréstimos, inovações, marcas de expressividade (p.ex.: alongamento grafêmico), vocabulário de domínio (como as *cashtags*[4]) e erros de digitação (Sanguinetti *et al.*, 2020). Os fenômenos *hashtag* e URL, por exemplo, ocorrem no exemplo (1), extraído do córpus de tweets do mercado financeiro em português denominado DANTEStocks (Di-Felippo *et al.*, 2021).

Sendo do mercado financeiro, assume-se que os tweets são mais frequentemente compostos por predicadores nominais do que outros tipos de predicadores. Essa afirmação advém trabalho de Voskaki, Tziafa e Annidou (2016), no qual, com base em um córpus multigênero[5] em grego do referido domínio, os autores evidenciaram que os nomes (predicadores) deverbais ocorrem mais frequentemente do que os verbos de que derivam (p.ex.: "securitização" ocorre com frequência maior que securitizar")[6]. Ademais, em comparação a um córpus de língua geral, os autores também verificaram que tais nomes são mais frequentes no discurso especializado. A relevância dos nomes predicadores (deverbais) em

[3]Até 2017, a quantidade máxima era de 140 caracteres.

[4]*Cashtag* é uma cadeia composta pelo cifrão ($), quatro letras e um número (p.ex.: $PETR4). As letras fazem alusão ao nome da empresa e o número se refere ao tipo de ação. A *cashtag* funciona como a *hashtag*, sendo um link para aquilo que é dito no *Twitter* sobre certa ação

[5]Esse córpus é composto por *posts* publicados em fórum do domínio, textos jornalísticos, documentos oficiais da bolsa de valores grega e textos acadêmicos.

[6]Define-se securitização como operação de crédito em que entram títulos como garantia de pagamento (Português, 2024)

relação aos verbos de que derivam também pode ser observada no córpus de tweets do mercado financeiro em português de Silva, Roman e Carvalho (2020), uma vez que nele há 1.237 ocorrências de nomes deverbais e somente 927 verbos primitivos a eles relacionados.

No tweet em (1), extraído do córpus de Silva, Roman e Carvalho (2020), há, por exemplo, dois nomes trivalentes: "acordo" e "compartilhamento". Diz-se isso porque a um predicador (P), está associada uma estrutura de argumentos (estrutura-A) ou valência (Borba, 1996), que especifica o número e tipo (semântico) dos argumentos (Arg), além do modo como os Arg são realizados sintaticamente. O tipo do argumento diz respeito ao papel semântico por ele exercido, como agente, experienciador, etc. (Fillmore, 1968).

Segundo Meyers (2007), os 3 Arg de "acordo" podem ser representados pelos rótulos: Arg0 (*agreer*), Arg1 (*proposition*) e Arg2 (*other entity agreeing*). E os 3 Arg de "compartilhamento" por: Arg0 (*sharer*), Arg1 (*thing shared*) e Arg2 (*shared with*). A partir do tweet (1), pode-se identificar a estrutura-A dos dois nomes (1a,b), as quais compartilham o mesmo Arg0. Nelas, o Arg0 é realizado por um sintagma nominal (SN) externo à valência dos nomes (sendo o sujeito[7] da oração) e os demais (Arg1 e Arg2 no caso de "acordo" e Arg1 no caso de "compartilhamento") são expressos por sintagmas preposicionados (de-SN e com-SN) internos ao SN (isto é, complementos dos nomes).

(1) #BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP. http://t.co/wHGukBg7qp

(a) Arg0[Gol] assina **acordo** Arg1[de compartilhamento de voos] Arg2[com TAP]

(b) Arg0[Gol] assina acordo de **compartilhamento** Arg1[de voos] com TAP.

A estrutura-A, como ilustrada em (1a) e (1b), tem sido muito utilizada há tempos no PLN para a interpretação das línguas naturais em diversas aplicações, como tradução automática, sistemas de perguntas e respostas e outras. Mais recentemente, métodos de AS e MO para os diferentes tipos de "conteúdo gerado por usuário" (CGU)[8] (especialmente tweets) também têm usado essa estrutura para identificar, por exemplo, quem (*who*) emite uma opinião e sobre quem (*towards who*) ela está sendo emitida (Mohammad; Zhu; Martin, 2014; Rudrapal; Das, 2018).

[7] Quando não expressos no SN, um Arg poder ser depreendido no arranjo sintático exterior ao SN. Em (1), por exemplo, o Arg0 (agente) é o sujeito da oração.

[8] Tradução do termo em inglês *"user-generated content"* (UGC), que engloba todo tipo de material (em formato de vídeo, texto, imagem, audio, etc.) postado em plataformas de mídias sociais, como Facebook e Twitter, em *blogs*, *chats*, fóruns de discussão, *wikis*, páginas de avaliação, etc. (Krumm; Davies; Narayanaswami, 2008).

Para ilustrar a relevância do conhecimento sintático-semântico codificado nas estruturas-a para o PLN, destacam-se os projetos PropBank (Palmer; Gildea; Kingsbury, 2005) e NomBank (Meyers, 2004), responsáveis pela associação de papéis semânticos aos Arg, respectivamente, dos verbos e nomes que ocorrem na seção financeira do córpus *Penn Treebank* (PTB) (Marcus, 1993; Taylor; Marcus; Santorini, 2003). Este, aliás, é um dos córpus mais relevantes para o processamento automático da língua inglesa padrão, pois tem mais de 4,5 milhões de palavras e anotação sintática de referência, isto é, manualmente revisada por especialistas.

Diante do exposto, objetivou-se realizar uma descrição sintático-semântica dos nomes valenciais no córpus do projeto DANTE (*Dependency-ANalised corpora of TwEets*)[9], o DANTEStocks, que contém aproximadamente 4 mil tweets do mercado financeiro em português (Di-Felippo *et al.*, 2021). Essa descrição visou gerar conhecimento sobre esses nomes, além de anotações de córpus e repositórios lexicais para o PLN.

1.2 Objetivos e Hipótese

Investigou-se, em outras palavras, a estrutura argumental (estrutura-A) dos nomes em um córpus de tweets do domínio financeiro, assumindo a hipótese de que o tweet, sendo um gênero de linguagem informal e marcado pela brevidade, tende a apresentar nomes com estrutura-A reduzida. Especificamente, objetivou-se:

- Realizar uma descrição detalhada da estrutura-A dos nomes valenciais em tweets do mercado financeiro em português.
- Identificar e analisar a presença ou ausência de argumentos previstos pela semântica do nome, bem como a realização sintática e os papéis semânticos desses argumentos.
- Avançar no entendimento do impacto das características linguísticas específicas dos tweets na análise sintático-semântica no contexto do PLN.
- Contribuir para a construção de recursos linguístico-computacionais, como córpus anotados e repositórios lexicais, que possam auxiliar no processamento de tweets e outros textos relacionados ao mercado financeiro.

Dessa forma, acredita-se que este trabalho pode contribuir para a compreensão mais ampla do uso linguístico em um tipo específico de CGU em português, que é o produzido

[9] <https://sites.google.com/icmc.usp.br/poetisa>

na plataforma Twitter, especificamente sobre o mercado financeiro. A importância deste trabalho reside não apenas na caracterização linguística, mas também no desenvolvimento de recursos e ferramentas que podem ser aplicados no processamento automático desse tipo de CGU.

1.3 Metodologia

O presente trabalho insere-se no campo do PLN, ancorado nos fundamentos da análise linguística baseada em córpus. Visando atingir os objetivos, equacionou-se a pesquisa em 6 etapas ou tarefas, as quais também estruturam este documento. A saber:

- **Revisão da literatura:** essa etapa permeia todo o processo de pesquisa, consistindo na leitura da bibliografia fundamental e demais referências pertinentes, publicadas no decorrer do projeto, sobre (i) nomes predicadores ou valenciais, (ii) tweets e suas características linguísticas, (iii) anotação sintática de córpus e (v) anotação de papéis semânticos em córpus, principalmente da estrutura-A de nomes (como no NomBank).

- **Descrição do córpus e Seleção dos nomes predicadores:** consiste na descrição das características do córpus selecionado para a pesquisa (DANTEStocks) e na seleção do conjunto de nomes predicadores do DANTEStocks para descrição.

- **Anotação sintática dos tweets:** trata-se da anotação sintática dos tweets em que ocorrem as instâncias do nomes que foram selecionados na etapa anterior.

- **Anotação semântica dos tweets:** consiste na identificação dos sentidos dos nomes e na descrição dos papéis semânticos previstos para os seus Arg.

- **Descrição sintático-semântica dos nomes predicadores:** descrição da estrutura-A dos nomes, incluindo aspectos como a presença/apagamento dos argumentos, a realização sintática e os papéis semânticos dos Arg sintaticamente realizados.

- **Criação de repositório lexical:** essa etapa visa à criação de um recurso online no qual a estrutura de argumentos dos nomes predicadores do DANTEStocks e suas respectivas realizações sintáticas possam ser disponibilizadas para as comunidades da Linguística e do PLN.

2 Revisão da literatura

Neste capítulo, abordam-se os conceitos fundamentais relacionados aos tweets e à descrição sintático-semântica da valência nominal, assim como os principais trabalhos da literatura que realizaram tarefas correlatas às aqui propostas.

2.1 Conceitos fundamentais

2.1.1 Gêneros CGU, tweet e mercado financeiro

Segundo Krumm, Davies e Narayanaswami (2008), o termo CGU, como mencionado, engloba todo tipo de conteúdo, seja na forma de imagem, vídeo, áudio ou texto, que é postado por usuários de plataformas online que agregam conteúdo, como as redes sociais, fóruns de discussão, *wikis*, etc. Krumm, Davies e Narayanaswami (2008) destacam ainda que os CGUs são marcados pela acessibilidade e natureza colaborativa, formando um contraponto ao conteúdo produzido por meios de comunicação tradicionais.

De acordo com Sanguinetti *et al.* (2020), o termo CGU recobre, mais especificamente, um contínuo de subgêneros textuais que podem variar conforme (i) convenções e limitações do meio (isto é, *blog*, fórum, chat, etc.), (iii) grau de canononicidade em relação à língua padrão e (iii) mecanismos linguísticos adaptados para veicular a mensagem.

Enquanto gênero, o tweet, particularmente, parece ser uma mistura de outros, como notícia, propaganda e bilhete, os quais foram modificados para atender às necessidades comunicativas da plataforma. Sendo possivelmente uma mistura desses gêneros, pode-se dizer que o tweet constitui um subgênero caracterizado pela informalidade e brevidade, promovendo, assim, uma comunicação concisa e direta. Tal brevidade advém da limitação de caracteres imposta pela plataforma, a qual, atualmente, é de 280 caracteres.

Essa brevidade, aliás, influencia a linguagem do gênero em questão, uma vez que os tweets no geral são compostos por (i) sequências de sintagmas curtos, (ii) sequências de elementos simplesmente justapostos (isto é, sem uma conexão sintática clara entre eles), (iii) orações ou fragmentos de orações, com ou sem problemas de pontuação. Suas características ortográficas e lexicais, em particular incluem fenômenos dependentes da plataforma (como *hashtag*, menções, marcas de *retweet*, URLs e diferentes truncamentos), simplificações (p.ex.: acrônimo e inicialismo), empréstimos, inovações, marcas de expressividade (p.ex.: alongamento grafêmico) e erros de digitação (Sanguinetti *et al.*, 2020).

Sobre a correlação entre o conteúdo do Twitter e a movimentação das ações nas bolsas, destaca-se o trabalho de Bollen, Mao e Zeng (2011), que foi um dos pioneiros a explorar o potencial do conteúdo veiculado pela referida plataforma no cenário das finanças. Nele, os autores encontraram uma forte relação entre as mudanças no estado de ânimo dos usuários do Twitter e as flutuações no *Índice Dow Jones Industrial Average* (DJIA).

Com isso, eles evidenciaram que o humor expresso em mídias sociais tem o potencial de influenciar movimentos de mercado. E, a partir desse trabalho, inúmeros tem sido desenvolvidos no âmbito do PLN com objetivos diversos, mas todos com o fim comum que é o de desenvolver formas mais eficazes de predição dos movimentos do mercado a partir do conteúdo dos tweets (Carosia; Coelho; Silva, 2019; Dhabe *et al.*, 2023).

A respeito desse conteúdo, ressalta-se que, quando se trata do domínio do mercado financeiro, os CGUs caracterizam-se pela ocorrência predominante de nomes predicadores. Essa observação, feita no trabalho de Voskaki, Tziafa e Annidou (2016), pautou-se na exploração de um córpus multigênero em grego do referido domínio de 19 milhões de palavras, que cobre o período de 1999 a 2010. Os autores verificaram que os nomes predicadores, sobretudo os deverbais, ocorrendo com frequência maior que seus verbos de origem em todos os gêneros a que se referem os quatro subcórpus utilizados no trabalho, incluindo o de CGU. No caso, os subcórpus são compostos por: (i) *posts* em fóruns de discussão online sobre o mercado de ações, (ii) notícias (de fontes jornalísticas tradicionais), (iii) relatórios anuais da bolsa de valores de Atenas e (iv) textos acadêmicos. Ademais, os autores verificaram que a frequência dos nomes predicadores é mais elevada nesses subcórpus quando em comparação a um córpus de referência de língua geral.

2.1.2 Nomes predicadores

Com base em Borba (1996) e Neves (2000), o SN, na estrutura de predicado de uma sentença/oração, é um termo, mas o nome, quando predicador, constitui o núcleo de um predicado, selecionando Args, isto é, projetando uma valência[1].

Do ponto de vista morfológico, os nomes predicadores podem ser primitivos (isto é, que não derivam de outra palavra) (p.ex.: "filho" e "vizinho") ou resultantes de nominalizações, podendo ser, portanto, deverbais (p.ex.: "acordo" < "acordar") ou deadjetivais (p.ex.: "beleza" < "belo"). Sendo nominalizações, os nomes deverbais e deadjetivais conservam, em princípio, a estrutura-A dos itens de que deriva (Neves, 2000).

Quanto à subcategorização semântica, os nomes valenciais podem representar conceitos concretos (p.ex.: "filho", "vizinho") ou abstratos, indicando estado (p.ex.: "doença"), propriedade (p.ex.: "temperatura"), qualidade (p.ex.: "beleza"), ação (p.ex.: "intervenção") e processo (p.ex.: "diminuição").

A valência ou estrutura-A, em parcicular, pode ser descrita em 3 níveis: valência lógica, sintática e semântica. No âmbito da língua comum, o nome "alteração", em sua valência lógica, é biargumental; na valência sintática (ou morfossintática), o preenchimento dos Args se dá formalmente por dois sintagmas preposicionados (SPrep); já na valência semântica, os Args estão marcados semanticamente com os traços [-animado] e [+agentivo].

Dessa forma, diz-se que, do ponto de vista lógico, os nomes abstratos podem ser monovalentes (p.ex.: "possibilidade", como em "A **possibilidade** de sucesso é alta"), bivalentes (p.ex.: "confiança", como em "A **confiança** da professora nos alunos é motivadora"), trivalentes (p.ex.: "doação", como em "A **doação** do empresário à cidade foi generosa") ou quadrivalentes (p.ex.: "tradução", como em "A **tradução** do texto do inglês para o português pelo especialistas foi um desafio"). Os concretos, por sua vez, podem ser mono- (p.ex.: "amante", como em "O **amante** da arte visitou o museu"), bi- (p.ex.: "pedinte", como em "O **pedinte** na rua pediu ajuda aos transeuntes") ou trivalentes (p.ex.: "tradução", como em "Uma bela **tradução** do poema do inglês para o francês") (Borba, 1996).

Como ocorre com os verbos, os Args dos nomes nem sempre estão expressos na superfície do texto, sofrendo elipse. Assim, o nome predicador, com dado número de Args

[1] Os nomes avalentes se definem por si mesmo, sendo, portanto, capazes de formar, sozinhos, um SN, como *mar, areia, dragão*, etc.

possíveis, pode ter um ou mais Args expressos (Neves, 2000). Como salienta Neves, no entanto, há situações em que um nome predicador, tomado no geral, deixa de projetar Args, como em "A vida é luta pra triunfo da verdade".

Para a realização dos argumentos dos nomes (valência sintática), é necessária a presença de uma preposição. Isso significa que todo Arg em função de complemento é preposicionado. No entanto, há outras formas correspondentes de realização de alguns Args no interior do SN, como: (i) possessivo (p.ex.: "seu acordo de compartilhamento"> "acordo de compartilhamento [de+ele]" (com y)), (ii) adjetivo classificador (p.ex.: "desenvolvimento econômico" > "desenvolvimento [da economia]" (por y), (iii) pronome pessoal oblíquo (p.ex.: "isso me é de serventia" > "serventia de isso [para mim]") e (iv) pronome relativo "cujo" (p.ex.: "excursão cujo destino" > "destino da excursão") (Neves, 2000). Elementos anafóricos (como os demonstrativos) não representam propriamente Args, mas podem recuperá-los em porção anterior do texto, como relata Bona (2014).

Os nomes predicadores (Npred) tendem a ocorrer em construções com verbo-suporte (Vsup) (isto é, verbos que veiculam os valores gramaticais (flexão) que esses nomes não podem exprimir). Sobre essas construções, destacam-se Barros (2014), Rassi (2023) e Santos (2015), que, com base em córpus, descreveram, respectivamente, as construções compostas pelos verbos "fazer" (2), "dar" (3) e "ter" (4) e um nome predicador (ou predicativo (Npred)). Mais detalhes sobre esses trabalhos são descritos na subseção 2.2.4.

(2) "Zé **fez** um acordo com Ana"
(3) "A prova **deu** calafrio em Ana"
(4) "O aluno **tem** um futuro brilhante"

A estrutura-A ou valência sintático-semântica (1a,b) (página 3) tem sido muito utilizada há tempos para a interpretação das línguas naturais em diversas aplicações de PLN, como a tradicional tradução automática e outras como a de perguntas e respostas.

Com a crescente importâncias das mídias sociais e de seus respectivos conteúdos, destaca-se que, mais recentemente, métodos de AS e MO para os diferentes tipos de CGU (especialmente os tweets) também têm usado essa estrutura para identificar, por exemplo, quem (*who*) emite uma opinião e sobre quem (*towards who*) ela está sendo emitida (Mohammad; Zhu; Martin, 2014; Rudrapal; Das, 2018).

Para desenvolver essas aplicações, voltadas particularmente para o inglês, os

pesquisadores do PLN dispõem dos dados[2] fornecidos pelos projetos PropBank (Palmer; Gildea; Kingsbury, 2005) e NomBank (Meyers, 2007), responsáveis por adicionar uma camada de informação predicado-argumento às estruturas sintáticas do subcórpus financeiro do *Penn Treebank* (PTB) (Marcus, 1993; Taylor; Marcus; Santorini, 2003). Esse subcórpus financeiro em questão é inteiramente composto por artigos (notícias) do *Wall Street Journal*, totalizando aproximadamente 1 milhão de palavras.

Em sua proposta inicial, os pesquisadores do PropBank anotaram as estruturas de argumento sobre as estruturas sintáticas já existentes no córpus PTB, que seguem o paradigma sintagmático de representação (Taylor; Marcus; Santorini, 2003). Em outras palavras, as estruturas sintáticas do referido córpus delimitam e classificam os tipos de sintagmas[3] que compõem uma sentença.

No que diz respeito à estrutura de argumentos propriamente dita, a anotação do PropBank está ancorada em *frames*. Segundo os autores, um *frame* consiste em um arquivo que lista os sentidos de um predicador e, para cada sentido, um conjunto de papéis semânticos previstos (*roleset*). Mais detalhes sobre os pressupostos teórico-metodológicos do PropBank e dos projetos dele derivados são fornecidos na subseção 2.1.2.1.

Embora o modelo sintagmático ainda esteja sendo bastante utilizado na construção de córpus anotados e no desenvolvimento de modelos/sistemas de análise sintática automática (em inglês, *parsing*), o conceito de "dependência" tem ganhado destaque nos últimos anos, sobretudo o modelo *Universal Dependencies* (Nivre *et al.*, 2016). Aliás, esse modelo, devido à sua adaptabilidade a diferentes gêneros e domínios, como salientam Alonso, Seddah e Sagot (2016), tem sido amplamente empregado na construção dos *tweebanks* em diferentes línguas.

Essa predominância, aliás, pode ser verificada no trabalho de Sanguinetti *et al.* (2023), no qual os autores descreveram 30 córpus de CGU construídos desde 2011 para várias línguas distintas. Desse conjunto, 14 *tweebanks* são compostos exclusiva ou parcialmente por tweets, sendo que, 25 deles possuem anotação sintática de dependência. Dos 25, apenas 4 não possuem anotação-UD. Na sequência, fornecem-se mais detalhes

[2]Entende-se aqui que esses dados são os contidos nos bancos de "proposições" dos projetos, sendo que a definição subjacente ao termo "proposiçao" é encontrada na semântica de *frames* de Fillmore (1968). Segundo esse autor, uma "proposição" é a estrutura básica de uma sentença.

[3]De acordo com Koch e Silva (1985), um sintagma consiste em um conjunto de elementos que constituem uma unidade significativa dentro da sentença e que mantêm entre si relações de dependência e de ordem. Organizam-se em torno de um elemento fundamental, denominado núcleo. Quando o núcleo for um nome, tem-se um sintagma nominal.

sobre esse modelo gramatical de dependência em particular.

2.1.3 O modelo *Universal Dependencies*

Em linhas gerais, o modelo UD, segundo Nivre *et al.* (2016) e Nivre *et al.* (2020), é um arcabouço teórico-metodológico para anotação gramatical (que envolve classes de palavras, traços morfológicos e dependências sintáticas) consistente de córpus em diferentes línguas. Em outras palavras, a UD fornece diretrizes e etiquetas para a construção de *treebanks*. A UD enquanto projeto é o resultado do esforço de uma comunidade aberta com mais de 500 colaboradores, as quais já produzindo mais de 200 *treebanks* em mais de 100 línguas.

Para a construção dos córpus anotados, o projeto utiliza um modelo gramatical de dependência, também denominado UD. A partir de uma visão lexicalista da sintaxe, as unidades básicas da anotação de um córpus com base nesse modelo são as palavras sintáticas[4]. Assim, a anotação-UD pressupõe que os clíticos sejam separados de seus hospedeiros ("prepare-se" –> "prepare" "se") e tratados como palavras independentes, assim como as contrações sejam decompostas ("das" –> "de" "as").

Quanto à anotação propriamente dita, a UD prevê 2 níveis. No nível morfológico, especificam-se 3 informações: lema, categoria morfossintática e traços lexicais/gramaticais (*features*). No nível sintático, a anotação se dá por relações de dependência (*deprels*), que são binárias e assimétricas. A representação básica de uma estrutura de dependências é arbórea, na qual uma palavra é o *root* (raiz) da sentença.

Na Figura 2.1, ilustra-se a anotação UD de um tweet (1).

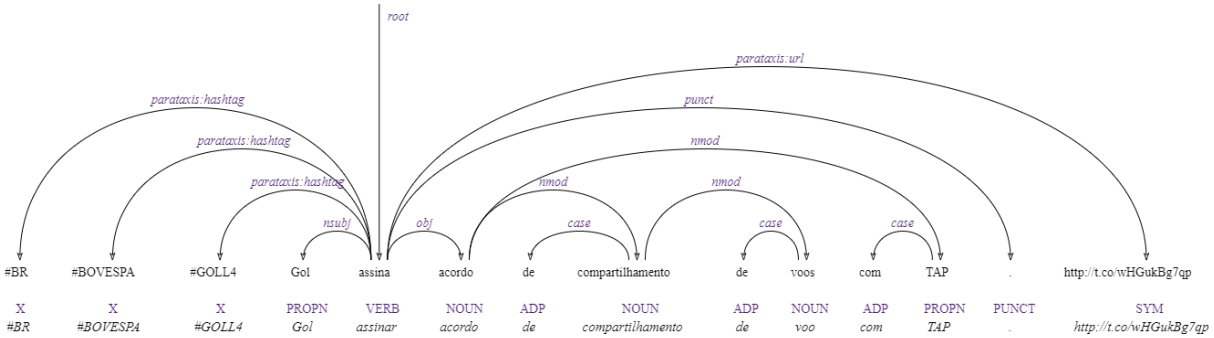
Nela, as etiquetas morfossintáticas ou de partes do discurso (do inglês, *Part-of-Speech* (PoS)) estão em caixa alta, como NOUN para "acordo". Logo acima, estão os lemas, p.ex.: "voo" é o lema de "voos". As *deprels* estão indicadas por setas rotuladas que se originam no *head* (cabeça) e se destinam ao dependente. Na Figira 2.1, "compartilhamento", por exemplo, é dependente de "acordo" pela *deprel* **nmod** (modificação nominal[5]), que é introduzida pela preposição "de". A preposição, por sua vez, é dependente de "com partilhamento" pela *deprel* **case**[6]. A raiz do tweet da Figura 2.1 é verbo "assinou".

[4]Tradução do termo em inglês *syntactic word*, que é definido como a unidade mínima a que corresponde uma função sintática <https://universaldependencies.org/u/overview/tokenization.html>. Na anotação-UD, palavras sintáticas (ou itens lexicais) são sinônimos de *tokens*.

[5]Relação que ocorre entre dois nominais (NOUN, PROPN, PRON), quando um especifica o outro.

[6]Relação que liga uma palavra de conteúdo a uma preposição (ADP) que a introduz. O *head* da *deprel*

Figura 2.1: Exemplo de tweet com anotação-UD no formato de árvore de dependência.



Fonte: O autor, 2024.

A UD também fornece uma lista de atributos[7] que codificam traços (em inglês, *features*) lexicais e gramaticais das palavras. Embora não constem da Figura 2.1, o nome "acordo", por exemplo, está associado aos traços-valores: Gender=Masc e Number=Sing. Quanto à morfossintaxe, a UD possui 17 etiquetas de *PoS* (Quadro 2.1).

Quadro 2.1: As 17 *tags PoS* do modelo UD.

ADJ	adjective	ADJETIVO
ADP	adposition	PREPOSIÇÃO
ADV	adverb	ADVÉRBIO
AUX	auxiliary	AUXILIAR
CCONJ	coordinating conjunction	CONJUNÇÃO COORDENATIVA
DET	determiner	DETERMINANTE
INTJ	interjection	INTERJEIÇÃO
NOUN	noun	SUBSTANTIVO
NUM	numeral	NUMERAL
PART	particle	PARTÍCULA
PRON	pronoun	PRONOME
PROPN	proper noun	NOME PRÓPRIO
PUNCT	punctuation	PONTUAÇÃO
SCONJ	subordinating conjunction	CONJUNÇÃO SUBORDINATIVA
SYM	symbol	SÍMBOLO
VERB	verb	VERBO
X	other	OUTRO

Fonte: Adaptado de Nivre *et al.* (2016).

No Quadro 2.2, tem-se as 37 relações de dependência (*deprels*). Nele, separam-se os argumentos principais dos predicados (*core arguments*) dos demais argumentos (*non-core dependents)*. Além disso, os argumentos e modificadores de predicados estão separados dos

case pode ser NOUN, PROPN, PRON, NUM, ADV, ADJ.

[7] https://universaldependencies.org/u/feat/index.html

modificadores de nominais. No quadro, vê-se também que há etiquetas diferentes quando o dependente da relação está sob forma oracional (isto é, orações subordinadas).[8]

Quadro 2.2: Quadro de relações de dependência sintática do modelo UD.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**Headless**	**Loose**	**Special**	**Other**
conj cc	fixed flat	list parataxis	compound orphan goeswith reparandum	punct root dep

Fonte: https://universaldependencies.org/u/dep/index.html.

No Quadro 2.3, apresentam-se as 37 *tags* das relações de dependência, com seus respectivos significados em inglês e português.

O projeto UD adotou uma versão revisada do formato CoNLL-X chamada CoNLL-U. Com isso, a anotação de córpus é codificada em um arquivo de texto simples (UTF-8) em que cada sentença (ou tweet, no caso) é um bloco de linhas composto por: (i) metainformações (isto é, linhas com o caractere # na primeira posição) e (ii) linhas de *tokens* (isto é, linhas que descrevem as palavras sintáticas da sentença). O bloco correspondente a cada sentença não pode ter linhas em branco, pois uma linha em branco (linha sem caracteres) indica o término desse bloco. Ademais, um bloco de sentença pode ter várias linhas de metainformação, porém, cada sentença deve ter duas informações essenciais: (i) um identificador de sentença definido com a metainformação [# sent_id

[8] O asterisco associado à **advmod** indica que **advmod** é usada para modificadores não só de predicados, mas também para outros tipos de modificadores.

= <*string*>] e (ii) o texto original da sentença definido com a metainformação [# text = <*string*>]. Logo na sequência das linhas de metainformação, tem-se as linhas de *tokens*, sendo que cada uma delas corresponde a cada uma das palavras sintáticas da sentença.

Quadro 2.3: As 37 *tags* das relações de dependência do modelo UD.

acl	adnominal clause	ORAÇÃO ADNOMINAL
advcl	adverbial clause	ORAÇÃO ADVERBIAL
advmod	adverbial modifier	MODIFICADOR ADVERBIAL
amod	adjectival modifier	MODIFICADOR ADJETIVO
appos	appositional modifier	MODIFICADOR APOSITIVO
aux	auxiliary verb	VERBO AUXILIAR
case	case marking	MARCADOR DE CASO
cc	conjunction	CONJUNÇÃO
ccomp	clausal complement	COMPLEMENTO ORACIONAL
clf	classifier	CLASSIFICADOR
compound	compound	COMPOSTO
conj	conjunct	COORDENADO
cop	copula	VERBO DE CÓPULA
csubj	clausal subject	SUJEITO ORACIONAL
det	determiner	DETERMINANTE
discourse	discourse	DISCURSO
dislocated	dislocated	DESLOCADO
expl	expletive	EXPLETIVO
fixed	fixed expression	EXPRESSÃO FIXA
flat	flat structure	RELAÇÃO PLANA
goeswith	goes with	TOKENS QUE VÃO JUNTOS
iobj	indirect object	OBJETO INDIRETO
list	list	LISTA
mark	marker	MARCADOR DE SUBORDINAÇÃO
nmod	nominal modifier	MODIFICADOR NOMINAL
nsubj	nominal subject	SUJEITO
nummod	numeric modifier	MODIFICADOR NUMÉRICO
obj	object	OBJETO OBJETO
obl	oblique nominal	NOMINAL OBLÍQUO
orphan	orphaned dependent	ÓRFÃO
parataxis	parataxis	PARATAXIS
punct	punctuation	PONTUAÇÃO
reparandum	overridden disfluency	DISFLUÊNCIA
root	root	RAIZ
vocative	vocative	VOCATIVO
xcomp	open clausal complement	COMPLEMENTO ORACIONAL ABERTO

Fonte: Adaptado de Nivre *et al.* (2016).

Cada palavra em uma linha de *token* é descrita por 10 campos:

1. ID: o identificador da posição do *token* na sentença (índice numérico a partir de 1);
2. FORM: *token* na forma como ocorre na sentença;
3. LEMMA: lema ou forma canônica da palavra;
4. POS: etiqueta que indica a categoria morfossintática (ou *PoS*) da palavra/*token*.
5. XPOS: etiqueta *PoS* específica da língua em questão;
6. FEAT: atributos morfológicos do *token*;
7. HEAD: ID do *token head* da relação de dependência do *token* que está sendo descrito;
8. DEPREL: relação de dependência do *token* com seu *token head*;
9. DEPS: relação de *enhanced dependency* do *token*;
10. MISC: informações adicionais sobre o *token*.

Na Figura 2.2, tem-se um exemplo do arquivo no formato CoNLL-U correspondente à anotação-UD do tweet cuja árvore de dependência é ilustrada pela Figura 2.1.

Figura 2.2: Exemplo do formato CoNLL-U corresponde à anotação da Figura 2.1.

```
# sent_id = dante_01_454607743392178177I
# text = #BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP . http://t.co/wHGukBg7qp
1   #BR   #BR   X   _   _   5   parataxis:hashtag   _   _
2   #BOVESPA   #BOVESPA   X   _   _   5   parataxis:hashtag   _   _
3   #GOLL4   #GOLL4   X   _   _   5   parataxis:hashtag   _   _
4   Gol   Gol   PROPN   _   _   5   nsubj   _   _
5   assina   assinar   VERB   _   Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin   0   root   _   _
6   acordo   acordo   NOUN   _   Gender=Masc|Number=Sing   5   obj   _   _
7   de   de   ADP   _   _   8   case   _   _
8   compartilhamento   compartilhamento   NOUN   _   Gender=Masc|Number=Sing   6   nmod   _   _
9   de   de   ADP   _   _   10   case   _   _
10   voos   voo   NOUN   _   Gender=Masc|Number=Plur   8   nmod   _   _
11   com   com   ADP   _   _   12   case   _   _
12   TAP   TAP   PROPN   _   _   6   nmod   _   _
13   .   .   PUNCT   _   _   5   punct   _   _
14   http://t.co/wHGukBg7qp   http://t.co/wHGukBg7qp   SYM   _   _   5   parataxis:url   _   SpacesAfter=\n
```

Fonte: O autor, 2024.

Na Figura 2.2, tem-se as duas linhas essenciais de metainformação, referentes ao número de identificação do tweet no DANTEStocks (# sent_id = dante_454607743392178177I) e ao texto do tweet propriamente dito (# text = #BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento de voos com TAP. http://t.co/wHGukBg7qp). Outra linha de metainformação poderia indicar, por exemplo, o especialista responsável pela anotação do referido tweet. Quanto às linhas de *tokens*, destaca-se, como exemplo, a de "Gol", que é o 4° *token* do tweet e tem as seguintes informações a ele associadas: lema "Gol", *PoS* PROPN (nome próprio) e relação de dependência com o *token* 5 por **nsuj**.

A atribuição de relações de dependência deve observar o princípio da projetividade,

isto é, os arcos das relações não devem se cruzar. Ademais, ressalta-se que as diretrizes da UD estão disponíveis no *website*[9] do projeto, assim como os mais de 200 córpus já anotados (*treebanks*). Essas diretrizes já foram instanciadas para a língua portuguesa padrão (Duran, 2021; Duran, 2022) e há alguns córpus de português brasileiro, com revisão manual, disponíveis no site da UD, como o Bosque-UD (Rademaker *et al.*, 2017a) e o PetroGold (Souza *et al.*, 2021).

Para tweets em português (em especial do mercado financeiro), as diretrizes de anotação de *PoS* já foram instanciadas (Di-Felippo *et al.*, 2022). Quanto às dependências, este trabalho, como se verá mais adiante, contribui com uma primeira versão de um manual para anotação de *deprel* em tweets, que busca cobrir somente os fenômenos/estruturas que não estão previstos no manual referente à língua geral de Duran (2022).

2.2 Trabalhos relacionados

2.2.1 PropBank

O projeto PropBank, concebido por Palmer, Gildea e Kingsbury (2005), surgiu com o objetivo de adicionar uma camada de informação predicado-argumento às estruturas sintáticas do subcórpus financeiro do PTB (Marcus, 1993; Taylor; Marcus; Santorini, 2003), que, como mencionado, era inteiramente composto por notícias do *Wall Street Journal*. Tal projeto, com o intuito de construir córpus anotados para classificadores automáticos de papéis semânticos, começou focalizando apenas as estruturas de argumentos projetadas pelos predicadores verbais, mas, como se verá na sequência, esse projeto teve extensões que diferentes tipos, sendo uma delas a inclusão dos nomes entre os predicadores a serem anotados (Meyers, 2007).

A anotação sintática do PTB é sintagmática e, em (5), tem-se um exemplo da primeira versão dessa anotação, elaborado com base em Taylor, Marcus e Santorini (2003) e Jurafsky e Martin (2024). Tal esquema de representação está no formato linear parentético, em que cada sintagma é delimitado por parênteses e começa com uma *tag* que codifica a informação do tipo de sintagma.

Nele, a anotação da sentença *"The mother agrees with the daughter on everything"* ("A mãe concorda com a filha sobre tudo.") engloba as etiquetas de *PoS* (isto é, DT para

[9] <https://universaldependencies.org/>

determinantes, NN para nomes, VB para verbos e IN para preposições) e a delimitação e classificação dos diferentes sintagmas que a constituem, no caso, tem-se que *"the mother"* e *"the daughter"* e *"everything"* são sintagmas nominais (em inglês, *nominal phrase* (NP)), assim como *"with the daughter"* é um sintagma preposicional (em inglês, *prepositional phrase* (PP)) e *agrees* é a verbo/núcleo (em inglês, *verbal base* (VB)) do predicado verbal *"agrees with the daughter on everything"*.

(5)
```
((S
    (NP The/DT mother/NN )
    (VP agrees/VB
        (PP with/IN
            (NP the/DT daughter/NN ))
        (PP on/IN
            (NP everything/NN )))))
```

E foi exatamente sobre representações sintagmáticas desse tipo que se deu a anotação da camada de informação referente à estrutura de argumentos dos verbos no PropBank. A tarefa de anotação, no PropBank, foi precedida pela construção dos arquivos de *frames* (em inglês, *frame files* e pela criação de um diretrizes de anotação (Palmer; Gildea; Kingsbury, 2005). Esses dois recursos permitiram que os pesquisadores distribuíssem a tarefa e reduzissem casos de discordância entre os anotadores.

As diretrizes, em particular, fornecem informações gerais sobre a tarefa. E os *frame files*, como mencionado, descrevem os sentidos dos verbos e, para cada sentido, o conjunto dos papéis semânticos previstos. Os exemplos contidos em um *frame file* servem para orientar os anotadores na escolha do identificador do sentido e das etiquetas de papéis semânticos que serão atribuídas.

Ademais, para melhor garantir o concordância entre anotadores, o PropBank adotou anotação dupla-cega, isto é, cada sentença é anotada por dois anotadores individualmente e sem contato entre eles. Na sequência, essas anotações são confrontadas e, diante de divergências, cada caso é resolvido por um anotador-juiz.

Ainda sobre os papéis semânticos que compõem os *rolesets*, destaca-se que o conjunto de papéis do PropBank pode ser dividido em dois grandes blocos: os papéis semânticos numerados (ArgNs) e os papéis semânticos modificadores (ArgMs).

Os ArgNs (Arg0, Arg1, Arg2, Arg3, Arg4 e Arg5) são previstos pela semântica dos verbos e compõem os *rolesets* armazenados nos *frame files*. Em um *roleset*, os argumentos numerados são "traduzidos" por rótulos mnemônicos[10] que indicam papéis específicos do verbo. Arg0 do verbo *agree* ("concordar") (6), por exemplo, é descrito como *agreer*. O papel Arg0 é geralmente designado para anotar o argumento que exibe características de um agente prototípico, enquanto Arg1 é atribuído a um paciente ou tema prototípico.

O PropBank dispõe de uma série de argumentos não numerados do tipo ArgMs, como LOC (*location*), CAU (*cause*), EXT (*extent*), TMP (*time*), DIS (*discourse connectives*), PNC (*purpose*), ADV (*general-purpose*), MNR (*manner*), NEG (*negation marker*), DIR (*direction*), e MOD (*modal verb*) (Jurafsky; Martin, 2024). Esses modificadores são relativamente estáveis entre os predicadores e, por isso, não são previstos nos *rolesets*.

Para ilustrar a anotação da informação predicado-argumento adicionada às estruturas sintáticas, utiliza-se aqui uma das entradas ou *frame files* do verbo *agree* ("acordar" ou "concordar") (6). Nela, em particular, tem-se o primeiro sentido de *agree* (ou seja, *agree.01*), que é definido por um conjunto de 3 Args com os seguintes papéis semânticos (*roleset*) descritos em (6). Com base nessa entrada, a sentença em inglês do exemplo (5) tem a anotação ilustrada em (7).

(6) *Roleset* do verbo "*agree*" (*agree.01*):

- **Arg0**: *agreer*
- **Arg1**: *proposition*
- **Arg2**: *other entity agreeing*

(7) Exemplo de anotação do PropBank para "*agree*" (*agree.01*):

Sentença: "*The mother* ***agrees*** *with the daughter on everything.*"

Anotação:[*The mother*]$_{\text{Arg0}}$ ***agrees*** [*with the daughter*]$_{\text{Arg2}}$ [*on everything*]$_{\text{Arg1}}$.

Desde que se tornou arcabouço metodológico fundamental para a anotação semântica de córpus e seus dados se tornaram recursos centrais para o desenvolvimento de aplicações como a etiquetação automática de papéis semânticos (em inglês, *sematic*

[10] O uso de argumentos numerados e rótulos mnemônicos foi uma estratégia para equilibrar diferentes pontos de vista teóricos, facilitando o mapeamento consistente entre teorias distintas sobre estrutura-A, como a tradicional teoria papéis *theta* (Kipper; Palmer; Rambow, 2002) e a estrutura lexical-conceitual (Rambow *et al.*, 2003).

role labeling) na primeira metade dos anos 2000, o PropBank tem ganhado extensões importantes, que focam na adição de outros predicados (como os nomes), outros gêneros e domínio e também outras línguas (como o chinês, árabe, etc.) (Pradhan *et al.*, 2022).

Ademais, ressalta-se que, mesmo com a revolução causada pelos métodos de aprendizado profundo (em inglês, *deep learning*), o interesse pela tarefa de etiquetação de papéis semânticos não diminuiu. Uma prova disso foi a incorporação dos arquivos de *frames* do PropBank ao editor de anotação do modelo *Abstract Meaning Representation* (AMR) (Banarescu *et al.*, 2013) para orientar a anotação das estruturas de argumento dos predicados aninhados (isto é, um predicado que é argumento ou modificador de outro predicado) na estrutura AMR (Pradhan *et al.*, 2022). Iniciativas como essa sinalizam a longevidade do projeto PropBank no PLN.

Atualmente, o PropBank possui mais de 110.000 estruturas predicado-argumentos que apontam diretamente para os nós das árvores sintáticas (sintagmáticas) de aproximadamente 50.000 sentenças do PTB. A anotação dessas estruturas foi orientada por um conjunto de aproximadamente 3.300 *frame files* que forneceram um conjunto específico de papéis semânticos para verbos como o argumentos para cada verbo.

Por fim, como salientam Pradhan *et al.* (2022), os *frame files* não estão organizados em nenhuma hierarquia semântica. Na verdade, os *rolesets* são agrupados nos *frame files* basicamente em função da polissemia. No entanto, os autores enfatizam que cada *roleset* potencialmente inclui links para outros recursos lexicais como VerbNet (Schuler; Palmer, 2005), FrameNet (Baker; Fillmore; Lowe, 1998) etc., bem como para os *word senses* da WordNet de Princeton (Fellbaum, 1998). Essa coletividade forma uma rede semântica rica, interconectada e com alta cobertura (de sentidos).

Na esteira do sucesso do PropBank, desenvolveu-se, também para o inglês, o projeto NomBank (Meyers, 2004; Meyers, 2007), que foi o responsável por anotar os argumentos dos nomes que ocorrem no mesmo córpus do PropBank. Para o português, destaca-se o projeto PropBank.Br (Duran; Aluísio, 2012), que foi uma iniciativa centrada na anotação de papéis semânticos em córpus do português brasileiro, aplicando metodologia muito similar à usado no PropBank, e tendo por objetivo final construir córpus anotado para auxiliar no treinamento e desenvolvimento de sistemas de PLN.

Na sequência, fornecem-se mais detalhes sobre ambos, em especial, sobre o NomBank, pois fornece diretrizes para a anotação da estrutura argumental dos nomes.

2.2.2 PropBank.Br

Segundo Duran e Aluísio (2012), o PropBank.Br produziu dois córpus anotados. No primeiro, denominado PropBank.Br v1, as estruturas-a foram anotadas por um único anotador sobre a porção brasileira do córpus Bosque (Afonso *et al.*, 2002), que, por sua vez, é um *treebank* revisado por linguistas. O segundo, denominado PropBank.Br v2, contém 8.350 instâncias[11] anotadas do córpus PLN-Br (Bruckschen *et al.*, 2008) e uma amostra de 840 instâncias do córpus Buscapé (Hartmann *et al.*, 2014).

Ambos os córpus (PropBank.Br v1 e PropBank.Br v2) possuem anotação com rótulos de papéis semânticos realizada sobre as árvores sintáticas geradas pelo *parser* Palavras (Bick, 2000). As árvores sintáticas do PropBank.Br v2, no entanto, não foram revisadas por humanos, diferentemente do PropBank.Br v1.

A anotação do PropBank v1 contou com apenas um anotador, motivo pelo qual não foi possível controlar o nível de concordância na tarefa. Além disso, por restrições de pesquisa, optou-se por uma metodologia um pouco distinta da utilizada no PropBank original: ao invés de primeiro construir os *frame files* e as diretrizes de anotação, os pesquisadores iniciaram a anotação no PropBank.Br v1 usando os *frame files* e o manual de diretrizes do inglês como modelo.

Já para o PropBank v2, os pesquisadores do projeto construíram automaticamente, a partir do PropBank do inglês, 2598 *frame files* em português. Desse total, os 541 *frames* que correspondem aos verbos do córpus PLN-Br com frequência acima de 1000 foram revisados e empregados, em uma anotação duplo-cega, para anotar as 8350 instâncias extraídas do referido córpus em língua portuguesa.

Ressalta-se que a anotação-duplo cega é importe para identificar questões que não são claras para os humanos e que, consequentemente, irão gerar problemas para um possível aprendizado de máquina que venha a ser feito a partir do córpus anotado. Essas questões têm que ser tratadas de maneira a tornar a anotação a mais lógica possível.

Para a tarefa de anotação de ambos os córpus, otimizou-se o editor SALTO (Burchardt *et al.*, 2006), originalmente proposto no projeto FrameNet (Baker; Fillmore; Lowe, 1998). Para os *frame files*, utilizou-se o editor Jubile (Choi; Bonial; Palmer, 2010).

Como resultado do projeto PropBank.Br, destaca-se a elaboração de um manual

[11] Por "instância", entende-se a ocorrência de um nome em uma sentença. Assim, se uma sentença tem 3 nomes, ela é triplicada e cada uma delas usada para anotar um dos nomes e sua correspondente estrutura de argumentos.

de anotação para o português (Duran, 2014), que incorpora a experiência acumulada durante o processo de anotação do PropBank.Br v.1, bem como as alterações mais recentes (à epoca) no conjunto de papéis semânticos promovidas pelo projeto PropBank do inglês.

Além do manual, o PropBank.Br produziu o Verbo-Brasil [12], que é um repositório de 1060 verbos (todos com frequência acima de 1000 no PLN-Br) concebido exatamente para apoiar a tarefa de anotação de papéis semânticos no referido projeto. Na Figura 2.3, tem-se a interface principal da versão do repositório Verbo-Brasil disponível na *web*, na qual é possível observar a função de busca pelos verbos em português.

Figura 2.3: Interface principal da versão online do Verbo-Brasil.

Fonte: http://143.107.183.175:21380/verbobrasil/

Nos moldes do PropBank pioneiro, esse repositório apresenta um arquivo para cada verbo em português, em que estão descritos os sentidos identificados na anotação do português e, para cada sentido, o conjunto dos papéis semânticos previstos. Além disso, cada arquivo apresenta exemplos anotados extraídos do córpus PLN-Br, os quais servem para orientar os anotadores na escolha do identificador do sentido e das etiquetas de papéis semânticos que serão atribuídas. Os sentidos dos verbos foram mapeados, sempre que possível, para os sentidos dos verbos do repositório do PropBank e para as classes verbais[13] da VerbNet[14](Schuler; Palmer, 2005).

Na Figura 2.4, o verbo "acordar" (pleno) tem 2 sentidos numerados (*acordar.01* e

[12]<http://143.107.183.175:21380/verbobrasil/>
[13]A VerbNet é uma extensa base de dados do inglês que contém informações sobre a interface sintaxe-semântica dos verbos de acordo com a teoria das classes de Levin (Levin, 1993)
[14]<https://verbs.colorado.edu/~mpalmer/projects/verbnet.html>

acordar.02)[15], sendo que o sentido *acordar.02* foi mapeado para *agree.01* do PropBank.

Figura 2.4: Sentidos do verbo "acordar" no Verbo-Brasil.

Predicate: *acordar*

Roleset id: acordar.01, *sair do estado de inconsciência; tomar consciência (uso metafórico)*, vncls:, Mapeamento para o inglês: wake.01

Roles:
Arg0: *alarme, agente ou causador do despertar*
Arg1: *aquele que dorme*
Arg2: *sono, sonho ou estado do qual se desperta*
Arg3: *realidade ou estado para o qual se desperta*

Exemplo 1:
O São_ Paulo começou o primeiro tempo apático, mas acordou aos 10 min com uma falta não marcada do lateral Mac Allister em Euller, quando este ia entrar na área.

Arg1: O São Paulo
Rel: acordou
Argm-tmp: aos 10 min
Arg0: com uma falta não marcada do lateral Mac Allister em Euller quando este ia entrar na área

Exemplo 2:
O heptacampeão do mundo de fórmula 1, Michael Schumacher, finalmente acordou do estado de coma, em hospital de Grenoble, na França.

Arg1: O heptacampeão do mundo de fórmula 1, Michael Schumacher,
Argm-adv: finalmente
Rel: acordou
Arg2: do estado de coma
Argm-loc: em hospital de Grenoble
Argm-loc: na França

Exemplo 3:
O governo acordou para as vulnerabilidades na troca de e-Mails sem assinatura e criptografia.

Arg1: O governo
Rel: acordou
Arg: para as vulnerabilidades na troca de e-Mails sem assinatura e criptografia

Roleset id: acordar.02 , *fazer acordo*, vncls: 36.1-1, Mapeamento para o inglês: agree.01

Roles:
Arg0: *entidade que faz o acordo* (vnrole: 36.1-1-agent)
Arg1: *proposta ou objeto do acordo* (vnrole: -theme)
Arg2: *entidade com a qual o acordo é feito*

Exemplo 1:
Os vereadores acordaram que discutirão, em conjunto com a Secretaria de Educação, um trabalho que se desenvolverá nas escolas visando a educação sexual dos nossas crianças.

Arg0: Os vereadores
Rel: acordaram
Arg1: que discutirão, em conjunto com a Secretaria de Educação, um trabalho que se desenvolverá nas escolas visando a educação sexual das nossas crianças.

Fonte: http://143.107.183.175:21380/verbobrasil/textoFrames/acordar-v.html

[15]No Verbo-Brasil, os sentidos dos verbos plenos estão numerados de 01 a 99.

2.2.3 NomBank

O NomBank foi desenvolvido por Meyers (2007) e se baseou e estendeu o PropBank, uma vez que o objetivo foi o de anotar a estrutura-A de todas as instâncias (ou ocorrências) dos nomes predicadores no subcópus financeiro do PTB, isto é, o mesmo no PropBank. Esse projeto foi motivado pelo fato de que, se usado em conjunto com o PropBank, o NomBank tem o potencial de fornecer dados para o reconhecimento de regularidades entre sentenças relacionadas lexical e sintaticamente. Assim, o NomBank e o PropBank podem ser usados para generalizar padrões de modo que um padrão faça o trabalho de vários.

Dado um padrão que indique, por exemplo, que o objeto (Arg1) de *appoint* ("nomeação") seja *John* e o sujeito (Arg0) seja *IBM*, um sistema de PLN enriquecido com esse dado do PropBank/NomBank pode detectar que *IBM* contratou *John* em diferentes sintagmas e sentenças: (i) *IBM appointed John* ("IBM nomeou John"), (ii) *John was appointed by IBM* ("John foi nomeado pela IBM"), (iii) *IBM's appointment of John* ou *the appointment of John by IBM* ("a nomeação de John pela IBM") e (iv) *John is the current IBM appointee* ("John é o atual nomeado pela IBM"). Os sistemas de PLN que não são capazes de comparar predicados distintos (p.ex.: nomes e verbos) e identificar regularidades entre eles precisam de padrões separados para lidar com cada tipo de predicado.

No NomBank, o objetivo foi anotar todos os SNs "marcáveis" (do inglês, *markable*), identificando o seu elemento nuclear (isto é, o nome predicador), seus argumentos e adjuntos permitidos, seguindo o estilo do PropBank. Por SN "marcável", os autores entendem ser todo nome que estiver acompanhado por ao menos um dos seus argumentos (Arg0, Arg1, Arg2, Arg3, Arg4) (Meyers, 2004) ou toda nominalização (palavra similar) que esteja acompanhado por um dos adjuntos permitidos (ArgM-TMP, ArgM-LOC, etc.).

Para tanto, os pesquisadores utilizaram um recurso lexical disponível para o inglês, o NOMLEX-PLUS [16]. Trata-se de um dicionário no formato do COMLEX Syntax (Macleod; Grishman; Meyers, 1998a), enriquecido com as classes verbais de Levin (Levin, 1993). O NOMLEX-PLUS possui 7.000 entradas. Nele, cada entrada de um nome é composta por uma série de informações, incluindo a classe semântica (que indica o sentido de uso do nome), seu esquema de subcategorização (que indica o tipo sintático dos argumentos) e o método de elaboração da entrada do dicionário.

[16] https://nlp.cs.nyu.edu/meyers/nombank/nombank.1.0/NOMLEX-plus.1.0

Em (8), tem-se a entrada do nome *"megawatt"* para o seu uso partitivo (PARTITIVE), como em *"400 megawatts of power"* ("400 megawatts de potência").

(8)
```
(PARTITIVE :ORTH "megawatt"
           :NOM-TYPE ((NOM-REL))
           :FEATURES ((TRANSPARENT))
           :OBJECT ((PP :PVAL ("of")))
           :SEMI-AUTOMATIC T)
```

A entrada está organizada em uma estrutura do tipo atributo-valor. Além da forma da palavra (ORTH) e do tipo de nominalização (NOM-TYPE) (no caso, NOM-REL indica um nome relacional), a entrada especifica que esse nome em questão requer apenas um argumento (OBJECT), o qual é tipicamente realizado em nível sintático como um Sprep introduzido por *of*. Além disso, a entrada fornece a característica (atributo FEATURES) de que o nome nuclear *"megawatt"* é transparente, uma vez que o sintagma inteiro assume a semântica do ARG1 (isto é, *"power"* no exemplo). O atributo-valor SEMI-AUTOMATIC T indica que a entrada foi parcialmente gerada de forma automática.

Especificamente, as entradas do NOMLEX-PLUS estão especificadas em função de uma das 16 classes: (1) RELATIONAL (ACTREL e DEFREL), (2) JOB, (3) HALLMARK, (4) PARTITIVE (PIECE e MERONYMY), (5) SHARE, (6) GROUP, (7) ENVIRONMENT, (8) ABILITY, (9) WORK-OF-ART, (10) VERSION, (11) TYPE, (12) ATTRIBUTE, (13) ISSUE, (14) FIELD, (15) CRISS-CROSS e (16) EVENT.

Ressalta-se que, quando pertinente, os nomes do NOMLEX-PLUS foram organizados em classes a partir da descrição prévia de verbos morfológica e/ou semanticamente relacionados, os quais projetam argumentos semelhantes aos nomes (Meyers, 2004). Esse critério justifica que as classes contenham nomes como os deverbais (p.ex.: *destruction* ("destruição") - cujos sujeitos dos verbos morfologicamente correspondentes são (usualmente) Arg0 e os objetos são (usualmente) Arg1 -, e nomes como *anniversary* ("aniversário"), para o qual se considera o verbo *commemorate* ("comemorar") como base para *picture* em *"Investors celebrated the second anniversary of Black Monday with a buying spree in both stocks and bonds"* ("Os investidores celebraram o segundo aniversário da *Black Monday* com uma onda de compras de ações e títulos").

Cada uma das classes apresenta características específicas quanto à estrutura de argumentos segundo Meyers (2007). Para ilustrar, detalham-se as classes RELATIONAL e PARTITIVE/SHARE/GROUP.

Os nomes da classe RELATIONAL, por exemplo, podem ser ACTREL (ACTion) ou DEFREL (DEFinitional). As diferenças cruciais entre as subclasses são: (i) nomes DEFREL (p.ex.: *father* ("pai"), *capital* ("capital"), *protagonist* e ("protagonista")) aceitam apenas um argumento (Arg1), e (ii) nomes ACTREL (p.ex.: *lawyer* ("advogado"), *president* ("presidente") e *director* ("diretor") aceitam pelo menos um argumento (Arg2) e, em alguns casos, um argumento adicional (Arg1).

Os nomes da classe PARTITIVE e de outras duas classes similares, SHARE e GROUP, requerem um argumento especial B de modo que todo o sintagma nominal represente um múltiplo de B, uma fração de B, uma parte de B, ou qualquer outra quantificação possível sobre uma quantidade de B. B recebe o papel Arg1 (9a,b), (10a), (11b). Os nomes da classe PARTITIVE aceitam por vezes um "tema secundário", que é anotado como Arg3 (9b). Os nomes da classe SHARE aceitam (i) Arg0, que representa a entidade que recebe uma parte (11), (ii) Arg1 que indica a parte em questão (11a), e (iii) Arg2, que indica a porção (ou valor) relativo ao Arg0 (10b). No caso dos nomes da classe GROUP (11), o Arg1 é um nome no plural ou a descrição do conjunto de membros (que ocorre por um adjetivo). Além disso, eles aceitam um Arg2 e um Arg3 (tema secundário).

(9) **(a)** *jillions of dollars* [PARTITIVE]
("zilhões de dólares")
REL = **jillions**, ARG1 = **of dollars**

(b) *40 acres of grapes in California* [PARTITIVE]
("acres de uvas na Califórnia")
REL = **acres**, ARG1 = **of grapes**, ARG3 = **in California**

(10) **(a)** *his stock in a media company* [SHARE]
("suas/dele ações em uma empresa de mídia")
REL = **stock**, ARG0 = **his**, ARG1 = **in a media company**

(b) *Nestle's share of 7%* [SHARE]
("participação de 7% da Nestlé")
REL = **share**, ARG0 = **Nestle's**, ARG2 = **of 7%**

(11) **(a)** *a community of parents* [GROUP]

(“uma comunidade de pais”)

REL = **community**, ARG1 = **of parents**

(b) *ACME’s five-member board of directors* [GROUP]

(“conselho de diretores de cinco membros da ACME”)

REL = **board**, ARG1 = **directors**, ARG2 = **ACME’s**, ARG3 = **five-member**

Seguindo a metodologia do PropBank, o NomBank produziu um inventário de *noun frames* (isto é, os *rolesets* e exemplos) com base principalmente na combinação do NOMLEX-PLUS e dos *frame files* do PropBank. Esse inventário guiou a anotação das estrutura-A dos nomes no córpus. O NomBank possui 4.706 arquivos de *frame*[17], sendo que cada um deles pode ter mais de um sentido ou *roleset*.

Para os nomes deverbais, os *frame files* dos verbos correspondentes do PropBank foram automaticamente convertidos para *frame nouns*. Para os demais casos, os *frames* foram criados de forma manual ou semiautomática. Um exemplo de *noun frame* é descrito em (12), que correspondente ao nome *“megawatt”*, que inclui apenas um *roleset*. Outro exemplo é o *frame noun* em (13) relativo ao nome deverbal (NOM-TYPE:VERB-NOM) *“agreement”* (“acordo”). Esse *noun frame* também descreve apenas um sentido, ilustrado por em duas configurações sintáticas distintas.

(12)
```
<frameset>
  <predicate lemma="megawatt">
    <roleset id="megawatt.01" name="partitive-quant">
      <roles>
        <role descr="quantified" n="1"/>
        <role descr="secondary-theme" n="3"/>
      </roles>
      <example name="autogen1">
        <text> about 500 megawatts of power </text>
        <rel>megawatts</rel>
        <arg n="1">of power</arg>
      </example>
      <example name="autogen2">
        <text> 400 megawatts of power </text>
        <rel>megawatts</rel>
        <arg n="1">of power</arg>
      </example>
    </roleset>
  </predicate>
</frameset>
```

[17] https://nlp.cs.nyu.edu/meyers/nombank/nombank.1.0/frames/

(13)
```
<frameset>
  <predicate lemma="agreement">
    <roleset id="agreement.01" name="agree" source="verb-agree.01">
      <roles>
        <role descr="agreer" n="0"/>
        <role descr="proposition" n="1"/>
        <role descr="other entity agreeing" n="2"/>
      </roles>
      <example name="autogen1">
        <text> any House-Senate agreement on the deficit-reduction
        legislation </text>
        <arg n="0">House-Senate</arg>
        <arg n="2">House-Senate</arg>
        <rel>agreement</rel>
        <arg n="1">on the deficit-reduction legislation</arg>
      </example>
      <example name="autogen2">
        <text> a confidentiality agreement with Dunkin ' Donuts </text>
        <arg n="1">confidentiality</arg>
        <rel>agreement</rel>
        <arg n="2">with Dunkin ' Donuts</arg>
      </example>
    </roleset>
  </predicate>
</frameset>
```

Com base em arquivos de *frames* como os ilustrados em (12) e (13), os pesquisadores realizaram, assim, a anotação dos SNs "marcáveis" no subcórpus financeiro do PTB. A seguir, estão listados alguns exemplos de SNs anotados pelo NomBank.

(14) *her ability to produce higher student-test scores*

("sua/dela capacidade de produzir as mais notas em teste de estudante")

REL = **abiliy**, ARG0 = **her**, ARG1 = **to produce higher student-test scores**

(15) *the absence of patent lawyers on the court*

("a ausência de advogados de patente no tribunal)

REL = **absense**, ARG1 = **of patent lawyer**, ARG-LOC = **on the court**

(16) *order accuracy*

("precisão do pedido")

REL = **acccuracy**, ARG0 = **accuracy**, ARG1 = **order**

(17) *our long term ambition of running a major entertainment company*

("nossa ambição a longo prazo de administrar uma grande companhia de entretenimento")

REL = **ambition**, ARG0 = **our**, ARG1 = **of running a major entertainment company**, ARGM-TMP = **long-term**

(18) *the beauty of a democracy*

("a beleza de uma democracia")

REL = **beauty**, ARG1 = **the democracy**

(19) *offensive capability*

("capacidade ofensiva/de ofensa")

REL = **capability**, ARG1 = **offensive**

(20) *fexibility in regulating pesticides*

("flexibilidade na regulamentação de pesticidas")

REL = **flexibility**, ARG1 = **in regulating pesticides**

(21) *the vulnerability of many small communities to domineering judges*

("a vulnerabilidade de muitas pequenas comunidades a juízes dominadores")

REL = **vulnerability**, ARG1 = **of many small communities**, ARG2 = **to domineering judges**

Para cada instância "marcável" de um nome do PTB anotada no NomBank, criou-se uma "proposição", isto é, um subcojunto das *features* REL, SUPPORT, ARG0, ARG1, ARG2, ARG3, ARG4 e ARGM, emparelhadas com ponteiros para os sintagmas do PTB. Assim, seguindo o PropBank, o NomBank gerou um banco de proposições.

Quanto à anotação da estrutura de argumentos dos nomes no NomBank, Meyers (2007) destaca algumas dificuldades, sobretudo as que se referem à necessidade de repensar ou reinterpretar as árvores sintagmáticas fornecidas pelo PTB. Uma dessas dificuldades, por exemplo, diz respeito aos casos em que os anotadores discordam de uma análise sintática fornecida como ponto de partida para a anotação da estrutura-A.

No exemplo *"approximately 27,500 acres of timberland near Trucke"* ("aproximadamente 27.500 acres de terra florestal perto de Truckee"), a anotação sintática do PTB indica que o sintagma preposicional (SPrep) (locativo) *near Trucker* modifica o SN de núcleo *timberland*. Caso um anotador ache que o SPrep seja um modificar do SN de núcleo *acres*, o projeto adotou a diretriz de preencher o *slot* Arg1 com uma concatenação dos dois constituintes menores representados por (IN *of*) e (NP (NN *timberland*)).

Outro problema diz respeito à hifenização. A anotação do PTB considera elementos hifenizados como um único *token*. A não que se fizesse um ajuste, dois *slots* de Args seriam preenchidos por um mesmo constituinte. Para "remediar" esse problema, a solução adotada foi inclui etiquetas funcionais como H0, H1, H2, H3 e H4, sendo que H0 indica o primeiro

elemento hifenizado, H1, o segundo, e assim sucessivamente. Assim, é possível considerar cada elemento de uma hifenização individualmente. No SN "*English-Spanish translations*", por exemplo, a anotação Arg3-H0=English-Spanish indica que o Arg3 corresponde, na verdade, a *English* e Arg2-H1=English-Spanish indica que o Arg2 corresponde a *Spanish*.

Os autores também destacam a dificuldade de interpretar modificadores sentenciais de aposição. Na sentença "*Down's Syndrome, the leading cause of mental retardation, according to an NIH summary*", a anotação sintagmática do PTB indica que o aposto *"the leading cause of mental retardation"* inclui o modificador *"according to an NIH summary"*. No entanto, os autores alegam que o modificador modifica o aposto inteiro e, como o NomBank não anota aposição, o modificador em questão não é anotado.

Ao final, o projeto NomBank anotou 4.702 nomes distintos e um total de aproximadamente 114.000 instâncias desses nomes no subcórpus financeiro do PTB. Todos os dados, isto é, o NOMLEX-PLUS, o conjunto de nomes, a lista dos arquivos de *frame* e o bando de proposições, estão disponíveis para *dowload* na página do projeto[18].

Com base no descrito, o NomBomk é um projeto que fornece metodologia para a anotação de estruturas de argumento de nomes predicadores em córpus. Para o português, destacam-se algumas iniciativas que envolvem a descrição da estrutura-A desses nomes.

2.2.4 Outros trabalhos e recurso lexicográfio

Os trabalhos de Barros (2014), Rassi (2023) e Santos (2015) focaram na descrição de construções com Vsup (CVS) e um nome predicativo (Npred). Os verbos suportes são aqueles que veiculam os valores gramaticais (flexão) que os nomes não podem exprimir, como pode ser visto nos exemplos (2), (3) e (4). A partir da análise de córpus, Barros (2014) descreveu e classificou 1.815 predicados nominais formados pelo verbo-suporte "fazer", Santos (2015), por sua vez, trabalhou com 2.273 predicados nominais com o verbo-suporte "ter" e Rassi (2023) descreveu 1.489 predicados nominais com o verbo-suporte "dar".

Por se pautarem no referido arcabouço teórico-metodológico, esses trabalhos focaram na inter-relação entre léxico e sintaxe, adotando a frase simples (isto é, estrutura constituída por um predicador e os argumentos essenciais que ele seleciona) como unidade de análise linguística e não um item lexical isolado. Ademais, seguindo os pressupostos de Gross (1968), as propriedades foram formalizadas em uma matriz de dados, que representa,

[18]https://nlp.cs.nyu.edu/meyers/NomBank.html

de maneira concisa, a informação linguística pertinente à descrição das referidas construções.

Especificamente, Barros (2014) destacou 29 propriedades formais (como número de argumentos, tipo de preposição que introduz complemento, etc.), distribucionais (isto é, tipo semântico de sujeito e de complemento) e transformacionais (simetria, conversão, apassivação e nominalização) distintas na análise dos predicados nominais, agrupando-os em 17 classes que demonstram regularidades sintáticas.

Entre as 17 classes, as mais frequentes são: PB-F2HDeNH, PB-FH1 e PB-F2HH. A classe PB-F2HDeNH é representada pelo padrão [*Nhum0* ***fazer*** *Nprep de Nhum1*]; nele, vê-se um nome com tipo semântico (Nhum) na posição de sujeito (N0) e um nome do tipo não-humano (Nnhum) na posição de complemento (N1) (do Npred), como "diagnóstico" em (21) A classe PB-FH1 é representada pelo padrão [*Nhum0* ***fazer*** *Npred*]; nela, também há um nome com tipo semântico (Nhum) na posição de sujeito (N0), mas os Npred não têm complemento, como os nomes de atividades físicas e esportes (22). Já na classe PB-F2HH, cujo padrão é [*Nhum0* ***fazer*** *Npred Prep Nhum1*], também se vê um nome com tipo semântico (Nhum) na posição de sujeito (N0) e um Npred com complemento preposicionado, como "evangelização" em (23). Nos exemplos, os Vsup estão em negrito, os nomes predicadores estão em itálico e os seus argumentos estão sublinhados.

(21) "O médico **faz** o *diagnóstico* da doença"
(22) "Ana **faz** *hidroginástica*"
(23) "Os jesuítas **fizeram** a *evangelização* dos índios"

Santos (2015) realizou um estudo descritivo acerca do léxico e da sintaxe dos predicados nominais com o Vsup "ter" e Npred, verificando a influência desse verbo em alterar aspectos das construções nominais sem afetar a distribuição do Npred e de seus argumentos essenciais. A autora analisou 2.273 predicados nominais distribuídos em 10 classes. Para exemplificação, tem-se PB-TH1, representada pelo padrão [*Nhum0* ***ter*** *Npred NHum1*] (24), como a classe mais frequente. Esse padrão sistematiza as CVS com Npred em que há um Arg na posição de sujeito (N0) obrigatoriamente humano (NHum). A segunda classe mais frequente é PB-TH2, representada por [*Nhum0* ***ter*** *Npred Prep Nhum1*]. Nesse padrão, tem-se também um Arg na posição de sujeito (N0) obrigatoriamente humano (NHum) e um complemento preposicionado introduzido por "por". (25).

(24) "O paciente **teve** *alta* médica."

(25) "Eva **tem** *respeito* por Ivo"

Já Rassi (2023) descreveu construções com o Vsup "dar" e Npred com base não só no Léxico-Gramática (Gross, 1968), mas também na Gramática Transformacional de Operadores (Harris, 1970; Harris, 1978). A autora buscou de forma sistemática recensear o máximo possível de construções em córpus e identificar as principais propriedades formais, distribucionais e transformacionais das CVS. No geral, foram identificadas 15 classes significativas de CVS com o "dar" e Npred, totalizando 1489 ocorrências no córpus. As duas classes mais frequentes incluem (i) DH1, representada pelo padrão [*Nhum0* ***dar*** *Npred*], em que há apenas 1 argumento obrigatoriamente de tipo humano e na posição sujeito (26), (ii) DH2, representada pelo padrão [*Nhum0* ***dar*** *Npred Prep Nhum1*], com 2 argumentos obrigatoriamente de tipo humano em ambas as posições argumentais (N0=:Nhum e N1=:Nhum) (27).

(26) "Ana **deu** uma pirueta."
(27) "Rui **deu** um *castigo* para a Ana"

Rassi também apresenta uma proposta de análise sintática automática das CVS, usando uma abordagem baseada em regras de dependência entre seus constituintes, geradas automaticamente a partir das informações constantes na matriz de dados.

Essas três pesquisas, ao se basearem nos pressupostos de Gross (1968), notadamente produzem descrições muito sistemáticas com base em córpus, reunindo informações importantes sobre as propriedades léxico-sintáticas do português. Em outras palavras, elas geram listas de Npred (que ocorrem em construções com Vsup) e matrizes de dados que incluem padrões valenciais ou de estrutura de argumentos. Devido à adequação linguística e potencial de utilização, ressalta-se que seria fundamental para as comunidades da Linguística e do PLN que essas descrições estivessem disponíveis para consulta e *download* em bases dados online.

Além desses trabalhos, destaca-se aqui um recurso lexicográfico que se relaciona ao tópico deste trabalho. Trata-se do *Dicionário de Usos do Português do Brasil* (Borba, 2002). Diferentemente das obras consideradas cânones do português, o DUPB foi elaborado com base em uma gramática de valência (Borba, 1996), que busca determinar as relações de dependência sintático-semânticas entre um predicador e os itens lexicais que o acompanham na sentença.

Além de seu aporte teórico, o DUPB se destaca por ter sido desenvolvido com base em córpus, isto é, os sentidos das palavras e suas abonações (exemplos de uso) vieram de pesquisas em córpus. No caso, utilizou-se o chamado *Corpus de Araraquara*, que, predominantemente jornalístico, possuía à época da confecção do dicionário cerca de 77 milhões de palavras cujos textos foram escritos no Brasil entre 1950 e 1996. O DUPB, no que diz respeito à sua densidade macroestrutural, engloba 62 mil entradas (Alves, 2014).

Sob a perspectiva teórico mencionada, os sentidos das palavras, quando predicadores, são descritos com base nas relações sintáticas e semânticas elas estabelecem, isto é, de acordo com sua valência sintático-semântica, como se pode ver no verbete de "acordo", apresentado na Figura 2.5.

Figura 2.5: Verbete do nome "acordo" no DUPB.

acordo Nm ★ **[Abstrato de ação] 1** anuência; concordância: Necessitamos do acordo de todos. (AU) **2** entendimento; combinação: Parece que chegaram a um acordo (CI); o caixa quis fazer um acordo: dez mil cruzeiros em dinheiro e dez mil em pêssegos em calda (BH); Vamos entrar num acordo (BO) [±**Compl: com+nome ou entre+nomes coordenados ou nome no plural**] **3** pacto; tratado: com esses objetivos firmamos o Acordo do México (JK-O); Já se começam a ver na frente da Coréia os resultados do acordo concluído com Moscou (CRU) **4** em Direito Trabalhista, concordância de vontades para determinado fim jurídico; ajuste: O acordo, assinado por governo, empresários, camponeses e sindicatos, foi renovado pela 5ª vez (CP); O ano passado entramos em acordo com o parão (EM); Vasconcelos aprovou o acordo salarial que virá beneficiar a categoria (CB); Não houve mais acordo entre as partes (CL) ★ **[Abstrato de estado]** [±**Compl: entre+nomes coordenados ou nome no plural**] **5** concordância; conformidade: Há muitas vezes acordo entre as duas estimativas (NFN); Permanece, em Scoto o acordo entre a fé e a razão (HF) **6** consenso; concordância: Nas ciências naturais há um amplo acordo quanto aos métodos a serem utilizados (IP) ★ **[Núcleo de construção adjetiva]** [**de**+~] **7** bem ajustado à situação; dentro do que convém; sempre descobria barzinho, um bom conjunto com um cantor de acordo (DE); Fontoura não é um bom nome para você. Por que não escolher outro nome mais de acordo? (T) [**Compl: em+oração ou nome abstrato**] **8** concorde; concordante; encontravam-se completamente de acordo em atingir um objetivo (REA); estamos todos de acordo em que o método é engenhoso (AR-O); Se ela não estivesse de acordo, fosse embora (PT) ★ **[Núcleo de construção adverbial]** [**de**+~] **9** convenientemente: ele diz que sempre agiram de acordo [**Compl: com+nome**] **10** em conformidade; em consonância: acréscimos que variavam de acordo com o narrador (AF); O preço varia sempre de acordo com a oferta (REA)

Fonte: Borba (2002, p. 23)

Nesse verbete, por exemplo, o nome predicador “acordo” pode ser de duas classes semânticas (abstrato de ação e abstrato de estado) e ocorrer como núcleo em construções adjetivais e adverbais. Associados às diferentes classes, estão os sentidos numerados em função da frequência de ocorrência dos mesmos em córpus. Se “abstrato de ação”, por exemplo, o nome pode ocorrer com complemento preposicionado introduzido por *com* ou *entre* no sentido de “paco, tratado”.

Diante do exposto, o DUPB é sim um repositório de nomes predicadores, cujos sentidos estão descritos em função da Gramática de Valência de Borba (1996) e do uso em córpus. Aliás, tal gramática é um modelo por dependência, cujos origens estão em Tesnière (1959), assim como a *Univesal Dependencies* (Nivre *et al.*, 2016),

Após a revisão da literatura, fez-se a seleção do córpus de tweets e dos nomes predicadores desse córpus a serem descritos. Essas tarefas são descritas na sequência.

3 Seleção do córpus e dos nomes predicadores

Nesta seção, apresentam-se o córpus de tweets DANTEStocks, assim como o processo de seleção dos NPred nesse córpus. Especificamente sobre o DANTEStocks, apresentam-se as propriedades linguísticas que caracterizam esse recurso, assim como as diferentes anotações já existentes no DANTEStocks.

3.1 O córpus DANTEStocks

O DANTEStocks é um *tweebank* pioneiro no âmbito das pesquisas em PLN para o português, pois é o primeiro córpus de CGU (no caso, tweets) com diferentes tipos de anotação linguística. O ponto de partida para esse recurso foi o córpus de 4.517 tweets de (Silva; Roman; Carvalho, 2020). Os tweets foram coletados automaticamente pelos autores em 2014 e, por isso, cada postagem possui o limite de 140 caracteres. Além disso, a compilação foi feita com base na ocorrência de ao menos um *ticker*[1] de uma das 73 ações do índice IBOVESPA na B3[2]. Esse estratégia de coleta resultou em um córpus que apresenta uma diversidade de conteúdos ligados ao mercado de ações, refletindo várias facetas da comunicação no setor financeiro (Di-Felippo *et al.*, 2022).

Atualmente, o córpus abrange 4.048 tweets, totalizando 81.037 *tokens* (Di-Felippo *et al.*, 2022; Silva *et al.*, 2021). E foi a partir da anotação de *PoS* segundo o modelo UD, descrita mais adiante, que o recurso passou a se chamar DANTEStocks.

[1] Código alfanumérico composto normalmente por 4 letras, que representam a empresa, e 1 número, que indica o tipo da ação; por exemplo, Petr3 é o *ticker* para as ações ordinárias da Petrobras

[2] B3, em referência às letras iniciais de Brasil, Bolsa, Balcão.

3.1.1 Características linguísticas

Diante da relevância dos *tweebanks* para o PLN, vários trabalhos têm descrito as características linguísticas desse CGU, as quais impõem desafios à construção e aplicação desses córpus (Liu *et al.*, 2018; Sanguinetti *et al.*, 2020; Sanguinetti *et al.*, 2023).

Di-Felippo *et al.* (2021) fizeram o primeiro levantamento sobre as características do DANTEStocks, que gerou estratégias de *tokenização* automática para esse córpus (Silva *et al.*, 2021). Segundo os autores, a composição estrutural dos tweets enquanto unidade textual varia bastante no DANTEStocks. Nesse sentido, ele possui tweets formados por uma ou mais sentenças bem delimitadas, como (28), (29) e (30), mas também por tweets que apresentam, considerando a norma padrão da língua, ausência de pontuação (31) ou pontuação equivocada (32), tweets fragmentados (33), assim como com truncamentos (ou quebras) (lexical ou estrutual) (34), também compõem o córpus.

(28) Sera k petr4 já entrou na baixa?
(29) PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(30) #CSNA3: Está em região de suporte que vem resistindo. #Whoknows? http://t.co/UaD0jRffPw http://t.co/ceIbWrRhgF
(31) O #PT conseguiu fazer propaganda eleitoral antecipada O que a @dilmabr tem a dizer sobre isso?
(32) Bom dia Marcos, Alguma previsão para petr4?!
(33) #GGBR4 Suportes e resistências http://t.co/Azw6yIEVI9
(34) Banco do Brasil ON (BBAS3), Gráfico Diário. Ação s... http://t.co/uUcr12d64b

Além das características estruturais, os tweets do DANTEStocks apresentam fenômenos ortográficos e lexicais. Di-Felippo *et al.* (2021) organizaram inicialmente esses fenômenos em 7 classes: (i) simplificação de código: fenômeno que reduz o esforço de escrita de um *token* (p.ex.: ausência de diacrítico); (ii) abreviação: forma reduzida de várias palavras, como contração de elementos gramaticais, acrônimo e inicialismo (isto é, abreviações compostas pelas letras iniciais de palavras comuns); (iii) expressão de sentimento: fenômeno que emula o sentimento (p.ex.: alongamento grafêmico, repetição de pontuação, *emoticons/emojis*); (iv) influência de língua estrangeira: *token* formado com base em outra língua (p.ex.: "estopar", do verbo em inglês "stop" ("parar")); (v) expressão de oralidade: palavra cuja grafia remonta à comunicação informal; (vi) elemento metalinguístico: forma que tipicamente ocorre no Twitter (*hashtag*, menção, marca de

retweet, URL e truncamento lexical), e (vii) fenômeno de domínio: *token* que diferencia os tweets do DANTEStocks dos demais tweets (como *ticker*, *cashtag*, índice de (des)valorização das ações e substituição lexical por símbolo).

Em trabalho mais recente, Scandarolli *et al.* (2023) propuseram uma tipologia para os fenômenos ortográficos e lexicais do DANTEStocks. Essa tipologia é um aprofundamento ou refinamento dos fenômenos descritos em Di-Felippo *et al.* (2021), contribuindo, do ponto de vista linguístico, para a compreensão da linguagem dos tweets. Para propor a tipologia, os autores partiram da análise de 1.363 *tokens* que haviam sido marcados inicialmente como *Typo=Yes* durante o processo de anotação morfossintática semiautomática do córpus. Os *tokens* haviam sido marcados com a referida etiqueta porque a anotação de sua *PoS* requeria tratamento diferenciado, uma vez que eles apresentavam alguma variação ou inovação de forma frente à norma padrão. Nessa tipologia, os fenômenos foram categorizados em duas grandes dimensões: "Norma Padrão" e "Norma Inovadora".

A Norma Padrão engloba as variações gráficas vistas como desvios da norma-padrão por diversos motivos (como desconhecimento da ortografia, influência do meio e dispositivo, influência de novas regras fonéticas, etc.), organizadas em classes, tipos e subtipos. As classes dessa dimensão foram propostas com base no conceito de "caractere" do padrão Unicode. Nessa dimensão, tem-se as classes: (i) substituição, que pode ser de cedilha (p.ex.: "acougue" ao invés de "açougue") ou outro diacrítico, (ii) inserção, que pode ser de diacrítico (p.ex.: Petrobrás), espaço (p.ex.: "a final" ao invés de "afinal") ou outro caractere, (iii) omissão, que pode se de diacrítico (p.ex.: "esta" no lugar de "está"), e (iv) transposição (p.ex.: "acrodo" ao invés de "acordo").

A "Norma Inovadora" captura o uso criativo e inovador para expressar conceitos novos ou concorrentes dentro do domínio de tweets do mercado financeiro. Nessa dimensão, tem-se as seguintes classes: (i) abreviação, que pode ser contração (p.ex.: "enqt" ao invés de "enquanto"), acrônimo (p.ex.: CEMIG) ou inicialismo (p.ex.: "lp" para "longo prazo"), (ii) neologismo, que pode ser aglutinação (p.ex.: "Ibolixo" ("Ibovespa" + "lixo"), derivação (p.ex.: "diretassa" ("direta" + "-assa (-aça)"), ou influência estrangeira (p.ex.: "estopar" criado a partir do verbo em inglês "*stop*" ("parar"), (iii) expressividade, podendo ser prolongamento grafêmico (p.ex.: "noossaaa" ao invés de "nossa"), variação dialetal (p.ex.: "malmita" no lugar de "marmita"), simbolismo (isto é, ocorrência de caracteres simbólicos como *emoticons, emojis* ou outros), capitalização (p.ex.: "FEIO") e disfarce (p.ex.: "m*"

ao invés de “merda”), (iv) reescrita homófona, que pode ser fonetização (p.ex.: “krai” substituindo “caralho”) ou substituição grafêmica (p.ex.: “neh” no lugar de “né”), (v) metalinguagem, podendo ser *hashtag* (p.ex.: #PT), menção (p.ex.: @user), marca de *retweet* (RT) ou URL (p.ex.: http://t.co/LwmlKPqssk) e, por fim, (vi) fenômenos de domínio, como os *tickers* (p.ex.: Petr4) e *cashtags* (p.ex.: $PETR3).

Ressalta-se que, além da contribuição linguística para a compreenção da linguagem CGU dos tweets, Scandarolli *et al.* (2023), a partir da tipologia de fenômenos, propuseram diretrizes para anotá-las em córpus segundo o modelo UD. Com isso, esse trabalho sugere um esquema de anotação dos fenômenos que, uma vez aplicado a um córpus, pode permitir, como apontado pelos autores, que aplicações de PLN levem em conta a distribuição dos fenômenos para, por exemplo, desambiguar termos ou ordenar probabilisticamente sugestões/opções em um corretor ortográfico.

3.1.2 Anotação de emoção

Até o momento em que este projeto teve início, o DANTEStocks possuía dois tipos de anotação disponíveis, de emoção e de *PoS*. Quanto à anotação de emoção, vale destacar que ela já compunha o córpus compilado por Silva, Roman e Carvalho (2020), que deu origem ao DANTEStocks. Aliás, essa anotação foi a primeira dada a importância da relação entre as emoções expressas em tweets e o movimento do mercado financeiro (Zhang; Fuehres; Gloor, 2011; Mao; Counts; Bollen, 2011; Bollen; Mao; Zeng, 2011; Gaskell; McGroarty; Tiropanis, 2013).

A anotação de emoções no córpus DANTEStocks foi baseada na Roda das Emoções (do inglês, *Wheel of Emotions*) de Plutchik e Kellerman (1980), que categoriza emoções em quatro eixos emocionais básicos e proporciona uma ampla cobertura das expressões emocionais. Cada eixo contém pares de emoções opostas que variam em intensidade, incluindo (i) *joy* versus *sadness* (“alegria” *versus* “tristeza”), (ii) *anger* versus *fear* (“raiva” *versus* “medo”), (iii) *trust* versus *disgust* (“confiança” *versus* “nojo”) e (iv) *surprise* versus *antecipation* (“surpresa” *versus* “antecipação”).

O processo de anotação utilizou o método *crowdsourcing* (“contribuição colaborativa” ou “colaboração coletiva”), inspirado em trabalhos anteriores (Suttles; Ide, 2013; Mohammad *et al.*, 2015; Schuff *et al.*, 2017), em que voluntários selecionam uma emoção de cada par oposto. Esse método foi escolhido para assegurar maior diversidade nas anotações,

atenuando potenciais vieses associados ao perfil dos anotadores. Como resultado, Silva, Roman e Carvalho (2020) conseguiram anotar os 4.517 tweets originais do córpus, dos quais 240 foram descartados devido à predominância da marcação de "não sei" em todos os pares de emoção, resultando na primeira composição do córpus DANTEStocks com 4.277 tweets anotados com emoção. Como ilustração, o tweet "PETR4 só me traz alegrias" recebeu os rótulos para 3 dos pares emocionais: *joy*, *trust* e *surprise*.

Para avaliar a confiabilidade das anotações, cada emoção rotulada foi quantificada com base na proporção de anotadores que escolheram aquela emoção. Se a emoção "alegria" tivesse sido atribuída por dois de três anotadores, por exemplo, sua confiabilidade era 2/3. Dos 4.277 tweets anotados, 2.340 (54,71%) receberam um rótulo majoritário em pelo menos um par emocional, enquanto os tweets restantes (18,21%) foram classificados como "neutro" em todos os pares emocionais.

3.1.3 Anotação de *PoS*, lemas e atributos morfológicos

Antes da anotação de *PoS*, os autores, segundo Di-Felippo *et al.* (2021), tomaram algumas decisões de projeto importantes, a saber: (i) definição do tweet como unidade de análise e, com isso, os *posts* não foram segmentadas em unidades menores como sentenças ou sintagmas, e (ii) não normalização da linguagem, aceitando, assim, lidar com tweets que apresentavam os fenômenos estruturais e lexicais já descritos.

Além disso, o conjunto inicial de 4.517 tweets de Silva, Roman e Carvalho (2020) passou por um refinamento para que a camada de anotação de *PoS* (via UD) fosse criada, consistindo na exclusão de 469 tweets distintos repetidos e/ou não pertencentes ao domínio (Gazana; Di-Felippo, 2022). Esse refinamento resultou no conjunto de 4.048 *posts* que foram efetivamente submetidos à anotação de *PoS*.

A anotação de *PoS* foi a primeira realizada no âmbito do projeto DANTE. Seguindo a tendência da literatura atual, utilizou-se o modelo UD para a anotação de *PoS* (Silva *et al.*, 2021). Por conseguinte, a referida anotação se baseou no conjunto de 17 *tags PoS* da UD descritas da Figura 2.1. Tal anotação foi feita por 4 especialistas linguistas.

Para a anotação em questão, os *posts* foram *tokenizados* de forma automática por meio do *tokenizador* simbólico de tweets do pacote NLTK[3], adaptado por Silva *et al.* (2021) ao DANTEStocks por meio da inserção de regras contextuais baseadas na caracterização

[3]Trata-se do NLTK TweetTokenizer (https://www.nltk.org/api/nltk.tokenize.html)

do estatuto de *token* dos fenômenos descritos em Di-Felippo *et al.* (2021). Isso quer dizer que a definição dos fenômenos enquanto "palavras sintáticas" realizada por Di-Felippo *et al.* (2021) subsidiou as regras utilizadas para adaptar o *tokenizador*, permitindo que este reconhecesse os *tokens* válidos para anotação-UD do córpus.

Segundo Silva *et al.* (2021), a anotação semiautomática de *PoS* do DANTEStocks começou com a divisão dos 4.048 tweets em 13 pacotes. Com exceção do 1º pacote, que tinha 147 tweets, utilizado para o treinamento dos anotadores, os outros 12 continham aproximadamente 325 cada. O córpus foi dividido em pacotes para que a anotação automática de *PoS* pudesse ser incremental. Tendo em vista que não havia um *tagger* para CGU em português à epoca, a anotação do 1º pacote foi feita pelo parser UDPipe2[4] (Straka, 2018), que até então havia sido treinado para o português apenas a partir de textos com linguagem formal (no caso, notícias jornalísticas).

Assim, após a revisão manual da anotação automática de *PoS* do 1º pacote, este foi utilizado para retreinar o UDPipe2 de forma a prepará-lo para anotar o 2º pacote, e assim, incrementalmente, até que todos os pacotes tivessem sido anotados e revisados.

A revisão manual da anotação automática de *PoS* de cada pacote de tweets foi guiada pelo manual de Duran (2021a), que contém diretrizes para a anotação UD do português, e pelo manual de Di-Felippo *et al.* (2022), que engloba as diretrizes específicas para a anotação de *PoS* dos fenômenos típicos dos tweets do domínio do mercado financeiro.

Cada um dos 13 pacotes foi submetido à revisão de 3 anotadores humanos diferentes e somente os casos de divergência entre eles foram adjudicados por uma linguista sênior. Em outras palavras, um dos 4 especialistas teve exclusivamente o papel de adjudicador do casos em que houve discordância entre os outros 3 anotadores. Ademais, a revisão manual foi feita por meio de uma versão otimizada, por Miranda e Pardo (2022), da ferramenta Arborator-Grew (Guibon *et al.*, 2020). Em (35), tem-se o exemplo (1) deste documento com a anotação final de *PoS* (pós-revisão manual) segundo o modelo UD.

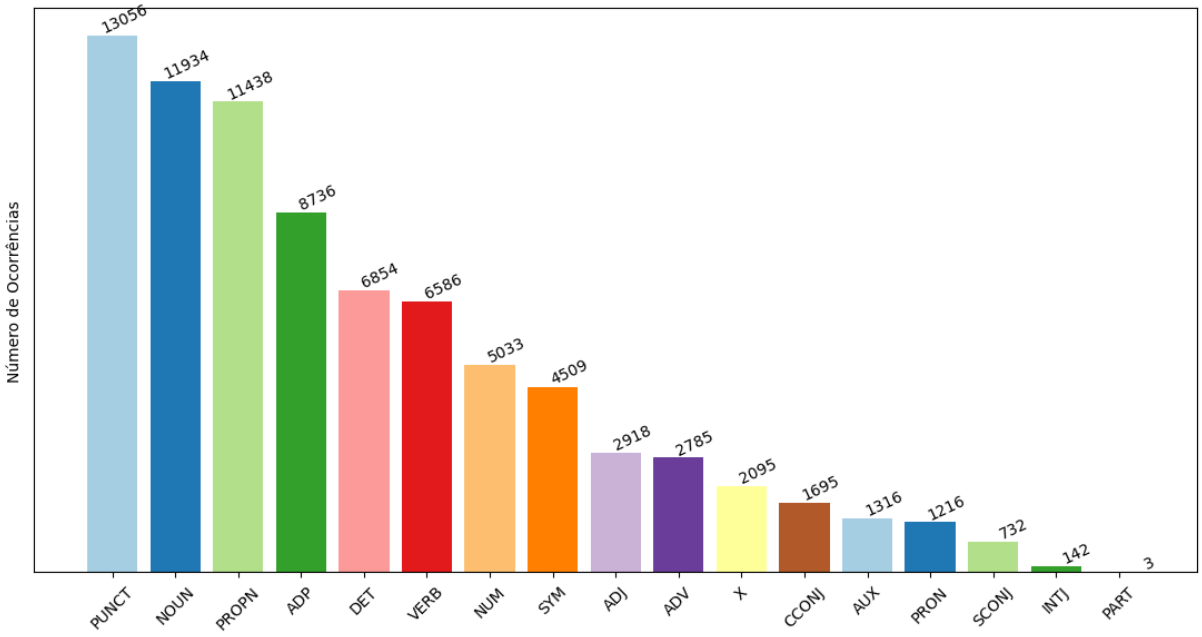
(35) #BR/**X** #BOVESPA/**X** #GOLL4/X Gol/**PROPN** assina/**VERB** acordo/**NOUN** de/**ADP** compartilhamento/**NOUN** de/**ADP** voos/**NOUN** com/**ADP** TAP/**PROPN** ./**PUNCT** http://t.co/wHGukBg7qp/**SYM**

A Figura 3.1 apresenta a distribuição estatística total de *PoS* no córpus. Antes, porém, de se discorrer brevemente sobre essa distribuição, vale ressaltar que os tweets do

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

DANTEStocks têm em média 20 *tokens* e também, em média, têm 8.5 *tags diferentes*.

Figura 3.1: Frequência simples das *tags PoS* no DANTEStocks.



Fonte: O autor, 2024.

Como se vê na Figura 3.1, todas as 17 *tags* propostas pela UD ocorrem no DANTEStocks. Curiosamente, PUNCT é a mais frequente, com aproximadamente 16% de todos os tokens sendo etiquetados com essa *tag*, seguido de NOUN, com cerca de 15%, e de PROPN, com aproximadamente 14% de todas as *tags* do córpus. Juntas, essas três *tags* (PUNCT, NOUN, e PROPN) somam quase metade de todas as *tags* (cerca de 45%).

Sobre as mais frequentes, PUNCT pode ser justificada pelo emprego não convencional dos sinais de pontuação e pela decisão de projeto de tokenizar os sinais de pontuação repetidos em sequência (Di-Felippo *et al.*, 2021). Assim, ocorrências como "!?!?!" originaram 5 *tokens*, sendo cada um deles anotado individualmente com a *tag* PUNCT. Essa decisão certamente afetou o número de PUNCT no total de *PoS* do córpus.

A alta frequência da *tag* NOUN, por sua vez, não é surpresa, uma vez que as palavras dessa classe são as mais frequentes na linguagem (Sardinha, 2000). No entanto, ressalta-se que a proeminência da *tag* NOUN também pode estar ligada à observação de que os nomes (em especial, os predicadores) parecem ser usados com mais frequência em CGU sobre o mercado financeiro, como indicado por Voskaki, Tziafa e Annidou (2016)).

Por fim, pode-se dizer que o fato de a *tag* PROPN figurar entre as mais frequentes revela outra característica do DANTEStocks. Especificamente, a alta frequência desta *tag*

resulta do fato de que os *tickers* utilizados como critério de busca no Twitter para compilar os *posts* foram todos anotados como PROPN, uma vez que são comumente usados como substitutos dos nomes das empresas cujas ações eles codificam.

No outro extremo da escala da Figura 3.1, encontra a *tag* PART, com apenas três ocorrências no córpus. Sobre essa *tag*, é importante mencionar que, de acordo com as diretrizes gerais da UD para o português, PART se refere a uma palavra funcional que deve ser associada a outra palavra ou sintagma para para expressar algum aspecto gramatical e não satisfaz os requisitos de outras categorias gramaticais ((Duran, 2021a)). No DANTEStocks, PART foi aplicada apenas aos prefixos "des-" e "pré-" usados como formas livres em duas ocorrências do mesmo neologismo (e.g. "(des) Graça Foster") e no caso de palavra dividida incorretamente (i.e. "pré abertura" em vez de "pré-abertura").

A *tag* INTJ é também uma das menos frequentes, com 142 casos. A baixa frequência de INTJ pode ser explicada pelo fato de que essa *tag* deve ser usada para anotar uma palavra que estiver sendo utilizada para expressar uma reação emocional. No entanto, em um córpus de CGU, como o DANTEStocks, os usuários geralmente aplicam outras marcas ou dispositivos linguísticos para expressar emoções, como repetição de pontuação, alongamento grafêmico, *emoticons* e *smileys* (Liu *et al.*, 2018; Di-Felippo *et al.*, 2021; Sanguinetti *et al.*, 2023).

Para PRON, uma hipótese para a baixa frequência é a de que os tweets do mercado financeiro tendem a ser factuais, não contendo ocorrência significativa de pronomes como os pessoais, reflexivos, interrogativos ou mesmo demonstrativos. Na verdade, acredita-se que a maior parte das *tags* PRON foi atribuída a pronomes relativos, mas isso não foi verificado porque, à época da escrita desse relatório, os traços morfológicos (que pode, incluir os tipos de pronomes) ainda não estava disponíveis por completo no córpus.

Quanto à SCONJ, AUX e CCONJ, acredita-se que a baixa frequência dessas *tags* pode ter sido causada pelo limite de caracteres imposto pela plataforma aos tweets (que, no caso do DANTEStocks, é de até 140 caracteres). Sobre SCONJ e CCONJ, vale ressaltar que essas (e principalmente SCONJ) são *tags* que devem ser utilizadas em *tokens* que ligam construções (ou sintagmas). Assim, dado o limite de tamanho dos tweets, não é surpreendente encontrar essas duas *tags* com baixa frequência. Podem-se dizer algo semelhante sobre AUX. Uma vez que os auxiliares são utilizados para compor construções de voz passiva, que são mais longas que as construções de voz ativa.

Finalmente, e como decisão de projeto do DANTEStocks, a *tag* X foi atribuída às *hashtags* e *cashtags* (por exemplo, #petr4 e $petr4, respetivamente) quando utilizadas apenas para fins de indexação (ou seja, sem função sintática clara no interior do tweet) e que, por isso, não são tão frequentes no córpus.

Além da anotação de *PoS*, o córpus possui as informações sobre os lemas e os traços morfossintáticos, além de outros adicionais. Essas informações foram inseridas de forma semiautomática (isto é, descrição automática com posterior revisão manual).

Sobre os lemas, registrados no coluna 2 do CoNLL-U, ressalta-se que as variações de forma da dimensão Norma Padrão (p.ex.: "açoes" ao invés de "ações") e da Norma Inovadora de Scandarolli *et al.* (2023) (p.ex.: "malmita" ao invés de "marmita") foram associadas a lemas da língua padrão (p.ex.: lema=ações e lema=marmita) sempre que possível. A não ser casos como os de inicialismos (p.ex.: "lp"), para os quais os lemas são os próprios *tokens* (p.ex.: lema=lp).

No que tange ao traços morfossintáticos ou morfológicos (coluna 6 do CoNLL-U), o DANTEStocks descreve os que são pertinentes ao português. No caso dos nomes, tem-se os traços de *Gender* (gêero) e *Number* (número), com valores binários *Fem/Masc* e *Sing/Plur*. Além desses, o córpus apresenta para alguns casos os atributos *Typo* e *Abrev* na coluna 6, que são para registrar diferentes tipos de variação de grafia.

Na coluna Misc, em especial, encontram-se 3 traços adicionais: *CorrectForm*, *FullForm* e *Trunc*. Esses atributos são responsáveis por explicitar, respectivamente, a forma por extenso de uma abreviação, a forma padrão de um *token* com grafia não-canônica e a forma por extenso de um truncamento.

3.2 Seleção dos nomes predicadores

Para selecionar os nomes predicadores, utilizou-se um critério semelhante ao do projeto NomBank (MEYERS, 2004). No caso, utilizou-se o padrão morfossintático formado pela sequência de *tags* [NOUN+ADP] (isto é, nome seguido imediatamente de preposição), uma vez que os nomes predicadores têm complemento introduzido por preposição (ADP). Por conseguinte, os nomes de interesse eram aqueles que estivessem acompanhados por ao menos um de seus argumentos (Arg0, Arg1, Arg2, Arg3, Arg4). Para identificar o padrão [NOUN+ADP], criou-se-se um script em *Python* que, a partir do arquivo CoNLL-U com

anotação de *PoS*, percorreu cada "tweet" do arquivo, identificou todas as ocorrências de NOUN seguido imediatamente por ADP e as salvou em um *dataframe* estruturado. O *dataframe* contém colunas: (i) a primeira exibe o índice da instância, que é a posição do tweet no córpus (indo de 0 a 4047), (ii) a segunda lista os supostos nomes predicadores, (iii) a terceira exibe as preposições e (iv) a quarta registra o tweet no qual o padrão ocorre. A busca em questão identificou 2.444 instâncias[5].

Na Tabela 3.1, tem-se a ilustração dos resultados e da organização em 3 colunas. Nessa tabela, há repetições de um mesmo NOUN seguido de uma mesma ADP, como #955 e #2292, que contêm "abril de". Para eliminá-las, fez-se um refinamento, excluindo instâncias repetidas de um mesmo NOUN seguido de uma mesma ADP em contexto sintático idêntico. Assim, diante de #955 e #2292, excluiu-se uma das instâncias, uma vez que ambas possuem "abril de" seguido de NUM ("2014" e "2013", respectivamente).

Na Tabela 3.2, para exemplificação, tem-se as instâncias da Tabela 3.1 sem as repetições. Após o refinamento, a lista inicial passou de 2.444 para 1.122 instâncias.

Nesse processo, ressalta-se que os nomes com variação de grafia por capitalização (como "Acordo") ou transposição de caractere (como "acrodo" < "acordo"), seguido de uma mesma preposição, foram considerados instâncias distintas. Isso justifica, por exemplo, a permanência de #1017 e #1237, mesmo que as instâncias "Acordo De" e "acrodo de" sejam seguidas de NOUN em ambos os casos ("Acionistas" e "acionista", respectivamente).

[5]Instância é a ocorrência da sequência NOUN+ADP em um tweet (Meyers, 2004). Se um tweet tem 2 NOUN+ADP, ele é duplicado, consistindo em 2 instâncias, mesmo que se trate do mesmo NOUN.

Tabela 3.1: Exemplos de resultados gerados pela busca do padrão [NOUN+ADP].

Índice	NOUN	ADP	Tweet
3011	abastecimento	de	Sabesp (SBSP3): Responsável por o **abastecimento de** 47% da região metropolitana de São Paulo, o Sistema Cantareira... http://t.co/xs8ji7To5z
955	abril	de	#IBOV #MRFG3 Marfrig realiza assembleia geral ordinária dia 17 de **abril de** 2014, às 10:00h. http://t.co/5BYxsyu6Vm
2203	abril	de	BC eleva juros básicos para 11%. Selic, que em **abril de** 2013 estava em 7,25%, agora já supera o nível do final do governo Lula. ibov petr4
121	Acordo	Com	$BBAS3 - Banco Do Brasil (bbas-nm) - Fato Relevante **Acordo Com** Os Correios http://t.co/kW3xh2fjzU
1017	Acordo	De	$MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo Ao **Acordo De** Acionistas http://t.co/jBJzeY5U5V
1509	acordo	com	Veja as melhores ações para comprar nesta semana, de **acordo com** 8 corretoras: Os papéis da G... http://t.co/L6OsbF6Os6 #infomoney #vale5
1510	acordo	com	Veja as melhores ações para comprar nesta semana, de **acordo com** 8 corretoras: Os papéis da Gerdau (GGBR4) e da ... http://t.co/SK7rRvdHEZ
2076	acordo	com	Confira 5 'top picks' para comprar este mês, de **acordo com** a Socopa: As 'Top Picks' da corre... http://t.co/Zu6BAhev7W #infomoney #vale5
2508	acordo	para	$GOLL4 - GOL e TAP assinam **acordo para** compartilhamento de voos http://t.co/F87EcEzEWK
2511	acordo	de	#BR #BOVESPA #GOLL4 Gol assina **acordo de** compartilhamento de voos com TAP. http://t.co/wHGukBg7qp
1237	Acordo	De	$CCRO3 - Ccr Sa (ccro-nm) - Fato Relevante - Assinatura Do Aditivo Do **Acordo De** Acionista http://t.co/xzQyZPPbBL

Tabela 3.2: Exemplos de instâncias após a exclusão de repetições.

Índice	NOUN	ADP	Tweet
3011	abastecimento	de	Sabesp (SBSP3): Responsável por o **abastecimento de** 47% da região metropolitana de São Paulo, o Sistema Cantareira... http://t.co/xs8ji7To5z
955	abril	de	#IBOV #MRFG3 Marfrig realiza assembleia geral ordinária dia 17 de **abril de** 2014, às 10:00h. http://t.co/5BYxsyu6Vm
121	Acordo	Com	$BBAS3 - Banco Do Brasil (bbas-nm) - Fato Relevante **Acordo Com** Os Correios http://t.co/kW3xh2fjzU
1017	Acordo	De	$MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo Ao **Acordo De** Acionistas http://t.co/jBJzeY5U5V
1509	acordo	com	Veja as melhores ações para comprar nesta semana, de **acordo com** 8 corretoras: Os papéis da G... http://t.co/L6OsbF6Os6 #infomoney #vale5
2508	acordo	para	$GOLL4 - GOL e TAP assinam **acordo para** compartilhamento de voos http://t.co/F87EcEzEWK
2511	acordo	de	#BR #BOVESPA #GOLL4 Gol assina **acordo de** compartilhamento de voos com TAP. http://t.co/wHGukBg7qp
1237	Acrodo	De	$CCRO3 - Ccr Sa (ccro-nm) - Fato Relevante - Assinatura Do Aditivo Do **Acrodo De** Acionista http://t.co/xzQyZPPbBL

Na sequência, realizou-se a efetiva identificação dos nomes predicadores da lista de instâncias. Essa etapa foi relevante porque nem todo NOUN que a busca pelo padrão [NOUN+ADP] retornou era efetivamente um nome predicador.

Essa identificação foi feita manualmente com base em dicionário. No caso, buscou-se por cada um dos nomes distintos constantes na lista exemplificada na Tabela 3.1 no DUPB (Borba, 2002). O DUPB, utilizado aqui como dicionário de referência, foi selecionado por se tratar de um repositório extenso de nomes predicadores, cujos sentidos, como mencionado, já estão descritos em função de sua valência sintático-semântica. O método manual foi necessário porque o DUPB é uma obra lexicográfica impressa. Para tanto, uma quinta coluna foi adicionada ao *dataframe* descrito, na qual fora registrada a análise resultante da consulta ao DUPB (Tabela 3.3).

Tabela 3.3: Exemplos de instâncias/nomes predicadores validados via DUPB.

Índice	NOUN	ADP	tweet	DUPB
3011	abastecimento	de	Sabesp (SBSP3): Responsável por o **abastecimento de** 47% da região metropolitana de São Paulo, o Sistema Cantareira... http://t.co/xs8ji7To5z	S
121	Acordo	Com	$BBAS3 - Banco Do Brasil (bbas-nm) - Fato Relevante **Acordo Com** Os Correios http://t.co/kW3xh2fjzU	S
1017	Acordo	De	$MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo Ao **Acordo De** Acionistas http://t.co/jBJzeY5U5V	S
1509	acordo	com	Veja as melhores ações para comprar nesta semana, de **acordo com** 8 corretoras: Os papéis da G... http://t.co/L6OsbF6Os6 #infomoney #vale5	S
2508	acordo	para	$GOLL4 - GOL e TAP assinam **acordo para** compartilhamento de voos http://t.co/F87EcEzEWK	S
2511	acordo	de	#BR #BOVESPA #GOLL4 Gol assina **acordo de** compartilhamento de voos com TAP. http://t.co/wHGukBg7qp	S
1237	Acrodo	De	$CCRO3 - Ccr Sa (ccro-nm) - Fato Relevante - Assinatura Do Aditivo Do **Acordo De** Acionista http://t.co/xzQyZPPbBL	S
1047	aliança	com	#GOLL4 - GOL negocia **aliança com** aéreas asiáticas - http://t.co/WLYifxc5Cc	S

Assim, dado um nome x da lista de 1.122 instâncias estivesse registrado no DUPB como valencial, assim como o sentido correspondente ao seu uso no DANTEStocks, este foi classificado como S (sim) na quinta coluna. Caso o nome estivesse registrado no DUPB, mas não fosse predicador, ele foi classificado com o valor N (não). Por fim, os casos de nomes não constantes no DUPB foram classificados como NC (não consta).

Para esse processo de validação via dicionário, a verificação dos nomes com variações gráficas foi feita com base na busca pelo nome com grafia padrão. No caso dos *tokens* “Acordo” e “acrodo”, por exemplo, buscou-se por “acordo” no DUPB para verificar se o uso dos mesmos no córpus estava registrado no dicionário como valencial.

Ao final do processo de validação via dicionário, das 1.122 instâncias contendo NOUN+ADP, 269 foram classificadas como S, 789 como N e 64 como NC.

O conjunto de 64 NC englobou (i) nomes em língua inglesa comuns à linguagem CGU (como *post(s)* e *website*) ou termos do domínio do mercado de ações (p.ex.: *"daytrade"*, que pode ser definida como "uma modalidade de *trade*" ("troca")), e "*put(s)"*, que significa "uma opção de venda"), e (ii) siglas e abreviaturas (em português). No caso de (ii), tratam-se especificamente de *tokens* etiquetados como NOUN, a saber: "a.a." ("ao ano")[6], "AGE" ("Assembleia Geral Extraordinária"), "C" (da expressão "série C")[7], "D" (da expressão "série D"), "CP" ("compra") e "CPI" ("Comissão Parlamentar de Inquérito").

Os casos de NC em língua estrangeira foram manualmente traduzidos para o português com vistas a uma segunda consulta ao dicionário. A identificação do sentido dessas palavras no córpus para a tradução contou com o auxílio de especialistas de domínio.

Os *tokens* em inglês que caracterizam o DANTEStocks, como mencionado, são de duas classes, as quais impuseram desafios distintos para a verificação do *status* de predicador. As palavras comuns à linguagem CGU, como *"post"*, têm equivalentes de tradução em português (p.ex.:, "postagem"). Assim, os equivalentes foram utilizados para a consulta ao DUPB. Já as palavras que são específicos de domínio não têm por vezes uma tradução direta e, por isso, fez-se uma definição aproximada que indicou uma tradução possível que pudesse ser consultada no DUPB. Esse foi o caso, por exemplo, de *"daytrade"*. Seguindo os especialistas de domínio, essa palavra foi definida como "uma modalidade específica de *trade*" ("troca") e, por isso, a palavra de busca no DUPB foi "troca".

Depois da verificação feita por meio da tradução, 13 das 64 instâncias classificadas como NC foram validadas como contendo nomes predicadores, o que é indicado, na Tabela 3.3, por meio de NC(S). Entre as não-validadas, estão instâncias compostas, por exemplo, pelos nomes: *videochat, units (unidades", (joint) ventura* e *gap*.

Após a etapa de tradução e segunda rodada de consulta ao DUPB, obteve-se o total de 282 instâncias com nomes predicadores, sendo as 269 já classificadas como S na primeira rodada de consulta e as 13 instâncias que, inicialmente pertencentes à classe NC,

[6]Para a anotação desse caso, os pesquisadores fizeram a *PoS tagging* baseados na diretrizes de Duran (2021a) para as *tokens* híbridos (isto é, constituídos de duas ou mais categorias morfossintáticas), que é a de anotá-los como NOUN. Isso se justifica pela decisão do projeto DANTE em respeitar as idiossincrasias da linguagem CGU sempre que possível.

[7]As expressões "Série C" e Série D" nas opções de ações do mercado financeiro brasileiro indicam vencimentos em março e abril, respectivamente. Acredita-se , alias, que a opção de *PoS tagging* mais adequada seria ambos os *tokens* da expressão por PROPN (série/PROPN C/PROPN).

foram posteriormente também classificadas como S.

A respeito das 282 instâncias contendo nomes que foram validados como predicadores, como as ilustradas na Tabela 3.3, cabe aqui uma observação importante. O padrão [NOUN+ADP], a partir do qual a busca pelos potenciais nomes predicadores foi inicialmente feita no DANTEStocks, identificou apenas os *tokens* etiquetados como NOUN imediatamente seguidos de *tokens* etiquetados com ADP. Por conseguinte, essa busca inicial não recuperou instâncias em que (i) nome predicador ocorresse com complemento preposicionado distante ou mesmo instâncias em que (ii) as outras formas de realização de alguns Args indicadas por (Neves, 2000) ocorressem (isto é, pronome possessivo, adjetivo classificador, pronome pessoal oblíquo e pronome relativo).

Diante disso, fez-se outra busca no córpus com base na ocorrência dos lemas (ou foma canônica) dos nomes que ocorriam nas instâncias validadas. Para tanto, identificou-se como ponto de partida um conjunto de 145 lemas distintos entre os nomes que ocorriam nas 282 instâncias validadas. Na sequência, buscaram-se todas as ocorrências de cada um dos lemas (anotado como NOUN)) no DANTEStocks. Como resultado, foi possível recuperar instâncias com (i) nome predicador e complemento preposicionado distante, como em (36), e instâncias com (ii) outras formas de realização de alguns Args, no caso, pronome possessivo, como em (37), e adjetivo classificador, como em (38).

(36) @ferrisss : sua **compra** dias atras de #JBSS3 fez volume > que a média (rs) . Vc deve ter comprado um fundo . Parabéns + 1x http://t.co/MpQ0Bh0E94

(37) @Live_Trade em as suas **projeções** , CSNA3 mergulha até que faixa de preço?

(38) @gugabianco Até onde sei, 1) não temos cap. instalada para refino, 2) **Acordos** comerciais, Petr4 precisa exp pra gerar receita em dólar.

A busca pelos lemas também levou em conta as informações contidas na coluna MISC do arquivo CoNLL-U do córpus. No caso da DANTEStocks, a coluna MISC registra os atributos adicionais FullForm e CorrectForm, que são os responsáveis por descrever, respectivamente, a forma por extenso de uma abreviação e a forma padrão de um *token* com grafia não-canônica ou truncada. Tais informações foram importantes para recuperar nomes predicadores (i) truncados e/ou (ii) com variação de grafia. Assim, com base no lema=compra e na informação de FullForm=compra para "CP" e de CorrectForm=compra para "co" (truncamento lexical), por exemplo, identificaram-se os tweets contendo "CP" e "co" como instâncias do nome predicador "compra".

Ao final, obteve-se um total de 1.756 instâncias (e 1.218 tweets diferentes). Cada uma delas contém ao menos um dos nomes validados como predicadores. Nesse conjunto de dados, há 145 lemas distintos (ocorrências no córpus) e 336 *tokens* distintos/formas distintas (considerando diferenças de flexão e grafia no geral), como demonstrado na Tabela 3.4.

Sobre os 1.218 tweets nos quais ocorrem nomes predicadores validados, seguem algumas estatísticas. Do total de 1.218, (i) 765 possuem apenas 1 nome predicador e (ii) 453, por sua vez, apresentam 2 ou mais nomes predicadores. Do total de 453 tweets com 2 ou mais nomes predicadores, destaca-se que pode se tratar de (i) nomes distintos (sendo 208 com 2 nomes, 30 com 3 nomes e 3 com 4 nomes distintos) ou (ii) repetições do mesmo nome (sendo 214 tweets com o mesmo nome repetido 2 vezes, 1 tweet com o mesmo nome repetido 3 vezes e 1 tweet com o mesmo nome repetido 4 vezes). Há também 4 tweets com um padrão misto, que combinam nomes repetidos e distintos. Especificamente, dois desses tweets apresentam um nome que se repete duas vezes, além de incluírem mais um nome único. Os outros dois tweets também exibem um nome repetido duas vezes, mas apresentam, além dele, dois outros nomes distintos.

Tabela 3.4: Nomes predicadores: estatística de lemas e *tokens*

N°	Npred/Lema	Qt. ocorrências no córpus	Exemplos de formas distintas	Qt. formas distintas
1	abastecimento	1	abastecimento	1
2	acesso	1	acesso	1
3	acordo	9	Acordo, Acordos, Acrodo, acordo	4
4	aditivo	3	Aditivo, aditivos	2
5	administração	6	Admin, Administracao, Administração, ...	4
6	ajuste	2	Ajuste, ajuste	2
7	alavancagem	2	alavancagem	1
8	aliança	1	aliança	1
9	alienação	1	Alienacao	1
10	alteração	19	Alteracao, a, alteracao, alteração	4
11	amor	1	amor	1
12	antro	1	antro	1
13	apreensão	1	apreensão	1
14	aquisição	10	Aquisicao, Aquisição, aquisicao, aquisição	4
15	assinatura	4	Assinatura, assinatura	2
16	ataque	1	ataque	1

N°	Npred/Lema	Qt. lemas distintos	Exemplos de *tokens* distintos	Qt. *tokens* distintitos
17	atestado	1	atestado	1
18	avaliação	2	avaliação	1
19	briga	9	Briga, briga	2
20	caminho	8	caminho	1
21	candidato	10	Candidato, Candidatos, candidatos	3
22	cara	6	cara, caras	2
23	carteira	47	Carteira, cart, carteir, carteira, carteiras	5
24	causa	9	Causa, causa, causas	3
25	cenário	7	cenário	1
26	chance	8	chance, chances	2
27	comissão	5	Comissão, comissão	2
28	comparação	4	comparação	1
29	compartilhamento	2	compartilhamento	1
30	compra	149	COMPRA, COMPRÃO, CP, comrpa, co, ...	10
31	comprador	4	comprador, compradores	2
32	confirmação	6	confirmação	1
33	construção	3	Construcao, construção	2
34	contrato	7	Contrato, contrato, contratos	3
35	controle	4	controle	1
36	conversa	2	conversas	1
37	conversão	1	conversão	1
38	convocação	2	Convocacao, Convocação	2
39	coordenador	1	Coordenador	1
40	coragem	4	coragem	1
41	corte	4	corte	1
42	cotação	47	Cotações, cotação, cotações	3
43	curiosidade	2	curiosidade	1
44	custódio	1	custódia	1
45	daytrade	3	Daytrade, day-trade, daytrade	3
46	decisão	7	Decisao, decisão	2
47	declaração	3	Declaracao, Declaração, declaração	3
48	deliberação	5	Deliberacoes	1
49	demanda	7	Demanda, demanda	2
50	denúncia	5	Denúncia, Denúncias, denúncias	3
51	descoberta	9	Descoberta, d, descoberta, dscberta, ...	5
52	descolamento	1	descolamento	1
53	desistência	1	desistência	1
54	desova	2	desova	1
55	diretor	6	Diretor, diretor, diretores	3
56	diretoria	4	Diretoria, diretoria	2
57	discussão	2	Discussão	1
58	distribuição	28	Dist., Distr., Distribuicao, distribuição	4
59	divergência	3	divergência	1

N°	Npred/Lema	Qt. lemas distintos	Exemplos de *tokens* distintos	Qt. *tokens* distintos
60	divisão	3	divisão, divisões	2
61	divulgação	11	Divulg., Divulgacao, divulgacao, divulgação	4
62	eleição	14	Eleicao, Eleição, el, eleição, eleições	5
63	encerramento	4	Encermto, Encerramento, encer., ...	4
64	entendimento	3	Entendimentos, entendimento	2
65	entrada	31	Entrada, entra, entrada, entradas	4
66	esclarecimento	8	Esclarecimento, Esclarecimentos, ...	3
67	estimativo	3	estimativa	1
78	exemplo	8	ex, exemplo	2
69	expectativa	9	Expectativa, expectativa	2
70	exploração	4	exploraç, exploração	2
71	extensão	2	Extensao, extensao	2
72	falta	2	fa, falta	2
73	fruto	3	frutos	1
74	fusão	9	FUSÃO, Fusão, fusão	3
75	gestor	12	ge, gestor, gestora, gestores	4
76	homem	1	homens	1
77	hora	11	Hora, hora, horas, hs	4
78	importação	1	importação	1
79	incorporação	5	Incorporacao, incorporacao	2
80	indicador	4	indicador, indicadores	2
81	indicação	320	Indicacao, Indicação, indic., indicacao, ...	6
82	inscrição	2	Inscricoes, inscrição	2
83	instauração	1	instauração	1
84	investimento	33	Investimento, Investimentos, invest., ...	5
85	leilão	33	Leilao, Leilão, leilao, leilão	4
86	licença	2	Licenca, Licensa	2
87	locação	2	locação	1
88	medo	4	Medo, medo	2
99	meio	1	meio	1
90	membro	6	Membro, Membros	2
91	mix	1	mix	1
92	monte	4	monte	1
93	mudança	8	Mudanca, Mudanças, mudan, mudança, ...	5
94	necessidade	1	necessidade	1
95	negociação	4	Negoc., Negociacao, negociação	3
96	notícia	31	Noticia, not., noticias, notícia, notícias	5
97	número	6	Nº, Números, núm̃er, número, números	5
98	obrigação	1	obrigacao	1
99	oferta	14	Oferta, oferta	2
100	olhada	6	olhada, olhadinha	2
101	olho	59	OLHOOOOOOO, Olho, olha, olho, zóio, ...	7
102	outorga	1	outorga	1
103	pagador	1	pagadoras	1

N°	Npred/ Lema	Qt. lemas distintos	Exemplos de *tokens* distintos	Qt. *tokens* distintitos
104	pagamento	35	Pagamento, Pgto, pagamento, pagto, pg	5
105	pedido	4	pedido	1
106	perspectiva	6	perspectiva	1
107	posição	22	POSIÇÃO, Posição, Posições, posição, ...	5
108	post	4	Post, post	2
109	postagem	1	postagem	1
110	procura	1	procura	1
111	projeto	6	Projetos, projeto, projetos	3
112	projeção	10	Projecoes, Projeção, proj, projeção, projeções	5
113	proposta	7	Prop., Proposta, proposta	3
114	racionamento	7	racionamento	1
115	rateio	1	Rateio	1
116	reapresentação	8	Reap, Reapresentacao, reapres	3
117	recomendação	18	Recomendação, recomend, recomenda, ...	5
118	recompra	27	Recompra, recompra, recompras	3
119	redução	2	redução	1
120	reeleição	2	reeleição	1
121	relatório	24	Relat., Relatorio, Relatorios, Relatório, ...	7
122	relação	11	relação	1
123	renúncia	3	Renuncia, renúncia	2
124	resistência	122	RESISTENCIA, Resistencia, resist, ...	6
125	resolução	1	Resolucao	1
126	responsabilidade	1	RESPONSABILIDADE	1
127	retorno	2	retorno, retornos	2
128	reunião	9	Reunião, reunião	2
129	risco	4	risco, riscos	2
130	seller	1	seller	1
131	solicitação	1	solicitação	1
132	spread	4	Spread, spread	2
133	sugestão	2	Sugestao, sugestão	2
134	taxação	6	taxação	1
135	tendência	18	Tend., Tendência, tend., tendência	4
136	teste	9	Teste, teste	2
137	trade	15	TRADE, Trade, Trades, trade	4
138	transferência	1	tranf.	1
139	transporte	2	transporte	1
140	troca	1	trocas	1
141	venda	149	#VENDA, VENDA, Venda, vd, vendas, Vd, ...	9
142	vendedor	3	vendedor, vendedora	2
143	visão	1	visão	1
144	volta	1	volta	1
145	zona	5	zona	1
			Total	336

4 Anotação sintática do DANTEStocks

Para identificar e analisar a presença ou ausência de argumentos semânticos, bem como sua realização sintática, adicionou-se uma camada de anotação sintática ao DANTEStocks por meio das etapas: (i) anotação automática de um conjunto inicial de tweets com vistas à construção de um subcórpus de referência e confecção de um manual de anotação, e (ii) treinamento de um *parser* com o subcórpus de referência e anotação semiautomática dos demais tweets. Em ambas as etapas, a revisão manual da anotação automática foi feita com base em um conjunto de diretrizes, as quais estão descritas e ilustradas em um manual de anotação que compõe o Apêndice A e a publicação de Di-Felippo, Nunes e Barbosa (2024).

4.1 Criação de um subcórpus de referência

Seguindo os pressupostos do projeto DANTE, a anotação sintática do córpus foi feita segundo o modelo UD. Antes, porém, fez-se um estudo linguístico sobre a composição dos tweets do córpus para melhor compreendê-la e propor uma metodologia de anotação. Nesse estudo, observou-se que, embora os tweets apresentassem composição estrutural bastante variada, sobretudo pela não segmentação em unidades menores (sentenças ou sintagmas) (além da fragmentação, pontuação informal e outros fenômenos CGU), uma parcela considerável apresentava linguagem relativamente padrão (isto é, tweets compostos por sentenças relativamente bem delimitadas), como em (39). Observou-se também que muitos tweets apresentavam certas padrões recorrentes, como em (40), e que outros tinham estrutura bastante variada, caracterizando-se principalmente pela fragmentação, como em (41). Por essa variação, aliás, os exemplos em (41) foram classificados como “miscelânea”.

(39) **(a)** A oposição protocolou mais um pedido de criação de CPI para investigar a Petrobras PETR4.SA , desta vez composta por senadores e deputados .
(b) Cada vez que ouço a G. Foster defendendo o plano de investimento da @petrobras , mais me certifico que devemos comprar PETR3 e 4 na BOVESPA
(c) Notas gerais A produção total de petróleo e gás natural da PETROBRAS (PETR4) no Brasil no mês de fevereiro foi ... http://t.co/qEEHshr7en

(40) **(a)** #OIBR4 (mensagem : 956643) http://t.co/VD2ApxqWqR
(b) #USIM5 (mensagem : 956895) http://t.co/Whjn7UVWe2
(c) #BBAS3 Banco da Brasil (mensagem : 956467) http://t.co/75T8wtmEXw

(41) **(a)** Tô de olho no HB esperando o MOMENTO HISTÓRICO de PETR4 na era PT . Falta $0,01 pra 13 . E 13 é ... PT!
(b) #ggbr4 13,33 (região encerro 50%)
(c) R$ 13 ... que ironia hein ? ,) #PETR4

Diante disso, procedeu-se à identificação automática dessas grandes "classes" de tweets, pois a decisão foi iniciar a anotação pelos tweets com linguagem relativamente padrão, seguidos por aqueles com algum tipo de padrão estrutural, e por fim, pelos tweets "miscelânea". Essa decisão se pautou no fato de que, embora os tweets "bem-formados" ainda pudessem apresentar fenômenos CGU (que impõem desafios à tarefa em questão), a sua anotação sintática (via UD) poderia contar com *parsers* já treinados para a linguagem padrão e com o manual de anotação de *deprel* instanciado para a língua portuguesa de Duran (2022). A anotação sintática de tweets com um mesmo padrão estrutural ou semelhante, por sua vez, poderia garantir certa consistência na sua análise/revisão. A decisão por anotar os tweets ditos "miscelânea" ao final justifica-se pela sua complexidade.

4.1.1 Organização dos tweets em blocos para anotação

Para tanto, empregaram-se, em um script Python, duas técnicas de análise de dados textuais consolidadas no âmbito do PLN e do Aprendizado de Máquina (AM), a saber: a vetorização TF-IDF (do inglês, *Term Frequency-Inverse Document Frequency*) e o algoritmo de clusterização *K-means*. Inicialmente, o script começou lendo os tweets registrados em um arquivo no formato xlsx[1] e os converteu em um *DataFrame* Pandas[2],

[1] <https://learn.microsoft.com/en-us/openspecs/office_standards/ms-xlsx/>
[2] <https://pandas.pydata.org>

que é uma biblioteca de manipulação de dados da linguagem de programação Python. Após garantir que todos os tweets estivessem no formato de *string*, o script os transformou em vetores numéricos utilizando o *TfidfVectorizer*, que é uma biblioteca de AM em Python pertencente ao módulo *scikit-learn*[3] (Pedregosa *et al.*, 2011).

Além da vetorização conforme descrita por Ramos (2003), o Tf-Idf determinou a frequência relativa das palavras no córpus de tweets considerando que o peso de uma palavra é inversamente relacionado ao número de tweets em que ela ocorre. Assim, o Tf-Idf é uma medida (de relevância) que considera relevantes as palavras que possuem alta frequência de ocorrência em um número limitado de tweets.

Os tweets vetorizados, juntamente com as medidas de Tf-Idf, foram submetidos ao *K-means* (MacQueen, 1967), que é um algoritmo clássico amplamente empregado em análises de AM para a identificação de clusters (conjuntos) a partir de dados textuais. Cada cluster possui um centroide, que é um conjunto de palavras estatisticamente relevantes que representam o tema ou tópico central do cluster. No caso deste trabalho, tais palavras foram definidas pelo Tf-Idf. Dada a similaridade lexical (isto é, número de palavras em comum) com os centroides, os tweets são alocados nos clusters. Dessa forma, cada cluster é formados por tweets semelhantes entre si e que veiculam um mesmo tópico.

A clusterização pelo *K-means* envolveu as etapas: (i) escolha aleatória de centroides (ou temas centrais de cada conjunto), que serviram de base para agrupar os tweets nos clusters, (ii) atribuição de cada tweet ao cluster de tema mais similar, baseando-se nas palavras relevantes (no caso, mais raras) fornecidos pelo Tf-Idf, (iii) verificação dos temas centrais de todos os clusters baseada nos tweets neles presentes, (iv) atualização dos centroides (isto é, de uma nova atribuição de tema central baseada nos tweets presentes nos clusters), e (v) nova iteração dos passos (ii), (iii) e (iv) até que todos os tweets tivessem sido associados devidamente ao cluster que melhor lhes representava.

O *k-means* dividiu os tweets em um número preestabelecido de 125 clusters. Após a clusterização, o script organizou os clusters em um *DataFrame* e finalizou a gravação dos resultados em um novo arquivo xlsx para gerenciamento dos dados. A Figura 4.1 ilustra essa planilha, em que a primeira coluna registra o índice do tweet (ordem em que aparece no córpus), a segunda exibe o texto no tweet e a terceira indica o cluster.

Na Figura 4.1, vê-se que o algoritmo de clusterização agrupou os tweets em função

[3]<https://scikit-learn.org/stable/>

do tema e das 3 características linguísticas observadas no estudo do córpus. Assim, os tweets do cluster 0 variam quanto à temática e composição estrutural. O cluster 2 engloba *posts* que possuem um padrão estrutura recorrente. E os tweets do cluster 44 possuem linguagem relativamente de acordo com a norma padrão.

Figura 4.1: Exemplo da organização dos clusters em arquivo xlsx.

Índice	Tweet	Cluster
391	Posso estar errado, mas a série D de PETR4 vai mansinha pro meu bolso! E sigo firme tentando lançar as 12 séries de PETR sem ser exercido.	0
520	Gosto de sanb11 nesse patamar! Abaixo d 11 deixo d gostar!	0
742	#petr4 descarrilando... a Pátria descarrilando junto. Lástima. Mas quem opera não faz caridade! Opera para seu bolso. Como deve ser, aliás.	0
1138	série D de a #PETR4 bombando	0
1578	Video - Análise Diária com Predador: Esgotamento de Taxa de a série D (Vale5 e Petr4) - http://t.co/a8XnFGrywg	0
1820	A última vez que fui completamente exercido em PETR4: novembro de 2013, PETRK19. Zerei a carteira e voltei na faixa dos $16.	0
2723	@silviusmille despencam valor d mercado por endividamento para investimento. Então, independente d quem for a faixa,em 2018 petr4 vai bombar	0
3366	Video - Análise Diária com Predador: Operação de Taxa na série F - Petr4 ! - http://t.co/a8XnFGrywg	0
3824	VALE5 continua ladeira abaixo. Sigo 100 % vendido em VALEF30 e F29.	0
3920	A última indicação da #MRVE3 resultou em -0.76 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
3947	A última indicação de a #MRVE3 resultou em -3.25 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
3957	A última indicação da #CSNA3 resultou em -1.53 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
3974	A última indicação da #MRVE3 resultou em -1.18 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
3983	A última indicação da #ELPL4 resultou em -1.44 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
3984	A última indicação da #OIBR4 resultou em -2.56 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
4029	A última indicação da #MRVE3 resultou em 2.06 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7	2
146	Satisfeitíssimo com a recuperação de BVMF3! Dá pra pedalar mais um pouco!	44
441	Já dá p/ acusar de gestão temerária o que está acontecendo com a #Petrobrás? #PETR4 #Bovespa http://t.co/7xx0A1UROG	44
542	Aprovar a investigação na #PETR4 só dá mais poder de barganha por cargos e privilégios ao próprio PMDB. (2)	44
1288	@Live_Trade dá só uma olhadinha no gráfico semanal da vale5 e veja que bonito o padrão de reversão dos candles .	44
1353	@andremassaro vc acha que comprar petr4 pra LP(20-30anos) é bater palma pra louco dançar ou dá pra ter esperança na nossa PDVESA...	44
1817	Mas não dá para ficar sem PETR4 em a carteira, por causa de a liquidez em opções. Essas últimas altas reaqueceram violentamente o mercado.	44
1822	RT @lambari_trader: Mas não dá para ficar sem PETR4 em a carteira, por causa de a liquidez em opções. Essas últimas altas reaqueceram violentame	44
2833	#VALE5 Agora já dá para ter mais confiança na compra, no momento acima do topo de ontem confirmando fundo.	44
3408	#SANB11: O ROE desse banco não reforça o gráfico semanal, mas quem sabe num dá pra sacar um troco? http://t.co/B0ZMd5FXhk	44
3525	Ninguém dá nada por o dia de hoje, mas estou de olho em a #PETRE18. Mais de 100% de alta! #petr4	44

Fonte: O autor, 2024.

A etapa subsequente consistiu em identificar e organizar os 125 clusters em função das 3 características mencionadas, gerando blocos de anotação correspondentes. Uma vez que os clusters foram organizados em blocos, selecionou-se um subconjunto representativo de ~10% do total de tweets de cada um dos 125 clusters. Isso gerou blocos contendo de 15 a 25 tweets e um montante total de aproximadamente 1.000. Essa seleção visou identificar uma amostra relativamente representativa da diversidade do córpus. Cada bloco foi organizado em planilha xlsx (Figura 4.2) com 5 colunas: (i) índice do tweet (ou ordem em que aparece no córpus) para facilitar seu gerenciamento durante o processo de anotação, (ii) ID do tweet no córpus DANTEStocks, (iii) texto do tweet propriamente dito, (iv) identificação numérica do cluster e (v) identificação numérica do bloco de anotação.

Na sequência, os 1.000 tweets selecionados foram anotados de forma automática e posteriormente revisados por especialista humano.

Figura 4.2: Exemplo de um bloco de tweet no formato xlsx.

Índice	dante_id	Tweet	Cluster	N° do Bloco
1646	ante_01_449179780374491137l	INTRADAY PETR4: Suportes 14,14 e 14,27 e resistências 14,61 e 14,82 INTRADAY VALE5: Suportes 27,07 e 27,28 e resistências 27,85 e 28,21	83	116
1803	dante_01_449545153137491968l	INTRADAY PETR4: Suportes 13,90 e 14,74 e resistências 15,99 e 16,40 INTRADAY VALE5: Suportes 27,04 e 27,45 e resistências 28,12 e 28,38	83	116
1898	dante_01_450627129328148480l	INTRADAY PETR4: Suportes 14,97 e 15,32 e resistências 15,89 e 16,11 INTRADAY VALE5: Suportes 27,38 e 27,60 e resistências 28,10 e 28,38	83	116
1986	dante_01_450997756396457984l	INTRADAY PETR4: Suportes 15,41 e 15,59 e resistências 15,87 e 15,97 INTRADAY VALE5: Suportes 27,90 e 28,13 e resistências 28,52 e 28,68	83	116
2035	dante_01_451350272934043649l	INTRADAY PETR4: Suportes 15,36 e 15,58 e resistências 15,93 e 16,06 INTRADAY VALE5: Suportes 27,53 e 27,90 e resistências 28,65 e 29,03	83	116
2181	dante_01_452100882658635776l	INTRADAY PETR4: Suportes 14,86 e 15,13 e resistências 15,66 e 15,92 INTRADAY VALE5: Suportes 28,60 e 29,07 e resistências 29,96 e 30,38	83	116
2261	dante_01_453167412829896704l	INTRADAY PETR4: Suportes 15,14 e 15,29 e resistências 15,65 e 15,86 INTRADAY VALE5: Suportes 28,94 e 29,21 e resistências 29,90 e 30,32	83	116
2352	dante_01_453538156524998656l	INTRADAY PETR4: Suportes 15,45 e 15,95 e resistências 16,71 e 16,97 INTRADAY VALE5: Suportes 29,31 e 29,68 e resistências 30,28 e 30,51	83	116
2441	dante_01_453995260587175936l	INTRADAY PETR4: Suportes 14,91 e 15,45 e resistências 16,80 e 17,61 INTRADAY VALE5: Suportes 29,30 e 29,69 e resistências 30,68 e 31,28	83	116
2461	dante_01_454263228046905344l	INTRADAY PETR4: Suportes 15,22 e 15,53 e resistências 16,05 e 16,26 INTRADAY VALE5: Suportes 29,19 e 29,57 e resistências 30,26 e 30,57	83	116
2541	dante_01_454684052935565312l	INTRADAY PETR4: Suportes 15,26 e 15,47 e resistências 15,97 e 16,26 INTRADAY VALE5: Suportes 29,14 e 29,41 e resistências 29,96 e 30,24	83	116
2601	dante_01_455747320068308992l	INTRADAY PETR4: Suportes 15,15 e 15,67 e resistências 16,45 e 16,71 INTRADAY VALE5: Suportes 28,93 e 29,39 e resistências 30,09 e 30,33	83	116
2650	dante_01_456072860755189761l	INTRADAY PETR4: Suportes 15,51 e 15,72 e resistências 16,25 e 16,57 INTRADAY VALE5: Suportes 28,71 e 28,97 e resistências 29,43 e 29,63	83	116
2750	dante_01_456427787113349120l	INTRADAY PETR4: Suportes 14,55 e 14,94 e resistências 15,90 e 16,47 INTRADAY VALE5: Suportes 26,93 e 27,40 e resistências 28,67 e 29,47	83	116
2822	dante_01_456807140955394048l	INTRADAY PETR4: Suportes 15,06 e 15,42 e resistências 15,96 e 16,14 INTRADAY VALE5: Suportes 27,60 e 27,88 e resistências 28,39 e 28,62	83	116
2925	dante_01_458613005102505984l	INTRADAY PETR4: Suportes 15,11 e 15,75 e resistências 16,81 e 17,23 INTRADAY VALE5: Suportes 27,62 e 28,08 e resistências 28,89 e 29,24	83	116
3082	dante_01_459337336161972224l	INTRADAY PETR4: Suportes 15,63 e 15,83 e resistências 16,15 e 16,27 INTRADAY VALE5: Suportes 27,34 e 27,52 e resistências 27,94 e 28,18	83	116
3165	dante_01_459698699480748032l	INTRADAY PETR4: Suportes 15,65 e 15,89 e resistências 16,35 e 16,57 INTRADAY VALE5: Suportes 27,28 e 27,71 e resistências 28,54 e 28,94	83	116
3264	dante_01_460782961156030464l	INTRADAY PETR4: Suportes 15,50 e 15,77 e resistências 16,18 e 16,32 INTRADAY VALE5: Suportes 27,00 e 27,25 e resistências 27,90 e 28,30	83	116
3402	dante_01_461251700347518977l	INTRADAY PETR4: Suportes 15,51 e 16,03 e resistências 16,82 e 17,09 INTRADAY VALE5: Suportes 26,29 e 26,52 e resistências 27,01 e 27,27	83	116
3747	dante_01_466939517224906752l	INTRADAY PETR4: Suportes 17,67 e 17,98 e resistências 18,47 e 18,65 INTRADAY VALE5: Suportes 27,36 e 27,81 e resistências 28,50 e 28,74	83	116

Fonte: O autor, 2024.

4.1.2 Anotação semiautomática via UDPipe2

Para a anotação sintática-UD, os tweets que compõem os blocos precisavam estar no formato CoNLL-U. Para tanto, empregou-se um algoritmo que leu as planilhas xlsx nas quais os blocos foram organizados (cf. Figura 4.2) e, com base nos IDs dos tweets, extraiu os itens correspondentes do arquivo CoNLL-U que havia sido gerado pela anotação de *PoS* anterior. Isso resultou na criação de novos arquivos CoNLL-U, cada um contendo os clusters referentes a um bloco específico.

Os blocos contendo os 1.000 tweets selecionados no formato CoNLL-U foram submetidos à anotação sintática pelo *parser* UDPipe2 [4] treinado com o córpus *UD Portuguese Bosque*[5] (Rademaker *et al.*, 2017b), que integra o *treebank* Floresta Sintá(c)tica[6]. Embora o UDPipe2 tenha sido treinado a partir do *UD Portuguese Bosque*, que é um *treebank* padrão ouro composto por textos jornalísticos, o seu emprego na anotação da porção inicial do DANTEStocks buscou minimizar o esforço de revisão manual.

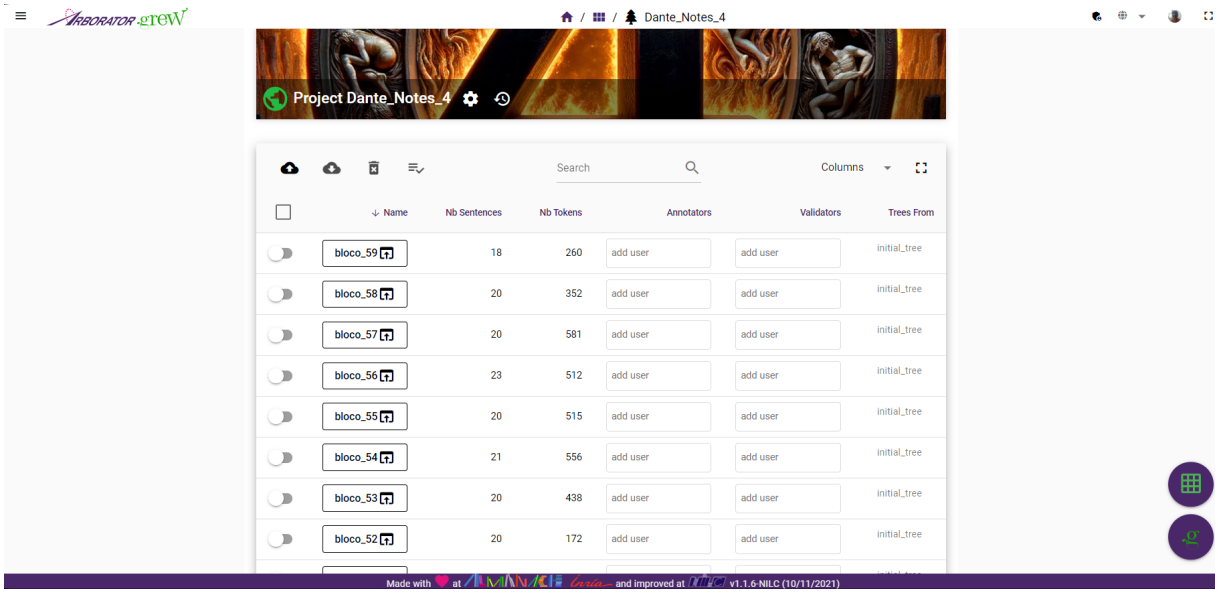
Diz-se isso porque, mesmo sabendo que os tweets com estrutura relativamente bem-formada (os quais foram o foco inicial da anotação) poderiam apresentariam os fenômenos linguísticos CGU que impõem desafios à tarefa de *parsing*, a anotação sintática desses *tweets* poderia se beneficiar com o emprego do UDPipe, que, como mencionado, foi treinado a partir de textos em português de linguagem padrão.

[4] <https://lindat.mff.cuni.cz/services/udpipe/>
[5] <https://universaldependencies.org/treebanks/pt_bosque/index.html>
[6] <https://www.linguateca.pt/Floresta/principal.html>

Após a anotação automática inicial dos ~1.000 com o UDPipe2, procedeu-se à disponibilização dos blocos para revisão manual por um especialista, que foi feita por meio de uma versão otimizada, por Miranda e Pardo (2022), da ferramenta online *Arborator-Grew* (Guibon *et al.*, 2020). A versão otimizada é a *Arborator-Grew-NILC*[7]. Especificamente, trata-se de uma plataforma web para anotações sintáticas do modelo UD.

Os blocos foram agrupados e disponibilizados para revisão manual em 3 projetos, cada um deles contento de 10 a 20 blocos. Grosso modo, o projeto 1 engloba blocos com tweets relativamente bem estruturados (e organizados por tópicos), o projeto 2 engloba os blocos com tweets que possuem padrões estruturais e o projeto 3 reúne os blocos com tweets variados ("miscelânea"). A interface de disponibilização dos blocos no *Arborator-Grew-NILC* está ilustrada pela Figura 4.3.

Figura 4.3: Disponibilização dos blocos/projetos no Arborator-Grew-NILC.



Fonte: https://arborator.icmc.usp.br/#/.

A revisão manual foi realizada por um único especialista com experiência em anotação-UD durante o período de ~4 meses (jul./23 a nov./23). Esse especialista iniciou a revisão da anotação do UDPipe2 pelo projeto 1, seguido pelo projeto 2 e, por fim, pelo projeto 3. Sempre que possível, ele utilizou o manual de *deprels* instanciado para a língua portuguesa padrão de Duran (2022). Mesmo para os pacotes de tweets considerados bem estruturados, a anotação do UDPipe2 necessitou de muitas correções, já que o modelo não havia sido treinado para a linguagem apresentada pelos tweets. E essas correções foram

[7] <https://arborator.icmc.usp>

ainda mais frequentes quando da revisão dos pacotes dos projetos 2 e 3.

As correções da anotação automática dos tweets dos projetos 1, 2 e 3 permitiram a elaboração de diretrizes de anotação, as quais deram origem à primeira versão de um manual de anotação de *deprels* do modelo UD para tweets do mercado financeiro em português (Di-Felippo; Nunes; Barbosa, 2024). Esse manual corresponde ao Apêndice A deste documento. As diretrizes que compõem o manual dizem respeito apenas às características linguísticas dos tweets que não estão previstas no manual de *deprels* instanciado para a língua portuguesa padrão por Duran (2022).

4.2 Treinamento de *parsing* e anotação do córpus

Após a conclusão da revisão manual dos 3 projetos, considerou-se o conjunto inicial de ~1.000 tweets uma parcela relativamente representativa do DANTEStocks a ponto de servir como subcórpus de referência para o treinamento de um modelo de *parsing* específico para os tweets do DANTEStocks. Tal treinamento foi feito com o objetivo de gerar anotações automáticas que precisassem de menos correções manuais. Optou-se por treinar o modelo Stanza[8] (Qi *et al.*, 2020) por ser mais facilmente treinável e por apresentar desempenho similar ao UDPipe2 para o português (Qi *et al.*, 2020).

O treinamento inicial do Stanza foi feito com base na unificação do córpus Porttinari-base (Pardo *et al.*, 2021) e do subcórpus de ~1.000 tweets revisados. O Porttinari-base possui 8.418 sentenças e 168.080 *tokens* (sendo 5.893 para treinamento, 842 para desenvolvimento e 1.683 para teste), extraídas do córpus jornalístico Folha-Kaggle[9], que foram inicialmente anotadas com as *deprels* do modelo UD e posteriormente revisadas com base nos manuais de anotação anotação de *PoS* (Duran, 2021a) e de *deprels* (Duran, 2022). O primeiro treinamento da rede Bi-LSTM (do inglês, "*bidirectional long short-term memory network*") do modelo Stanza usou os cerca de 1.000 tweets iniciais, divididos em 70% para treino, 10% para validação e 20% para teste.

À medida que mais tweets foram anotados com o Stanza pós-treinamento inicial e esses foram manualmente revisados, novas iterações de treinamento do Stanza foram implementadas, integrando progressivamente os tweets anotados. Esses treinamentos subsequentes incorporaram porções adicionais de 203, 300, 400, 400 e 1233 tweets, res-

[8] <https://stanfordnlp.github.io/stanza/>
[9] https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol

pectivamente, seguindo uma nova divisão de 80% para treino, 10% para validação e 10% para teste. Vale destacar que, a cada iteração, a parcela adicional de tweets era unificada ao Porttinari-base. A referida divisão das parcelas de treino, validação e teste também foi aplicada às sentenças do Porttinari-base que compunham o treinamento. Aliás, para assegurar a aleatoriedade e representatividade na seleção das amostras, fatores essenciais ao treinamento do modelo, utilizou-se um script Python.

Os 5 treinamentos posteriores tiveram como objetivo refinar a capacidade do *parser* de compreender mais profundamente os fenômenos inerentes ao córpus de tweets, uma vez que o conjunto inicial (da 1ª iteração) tinha algo em torno de 1.000 tweets. Ao final, realizaram-se 7 iterações de treinamento do modelo Stanza, sendo 6 durante a etapa de anotação e uma posterior, englobando todos os 4.048 tweets do DANTEStocks manualmente revisados.

Conforme os tweets com anotação via Stanza foram sendo revisados, calcularam-se as métricas de desempenho (i) precisão de anotação sintática (UAS) e (ii) precisão de anotação de etiquetas e relações (LAS). Com base nos resultados, o desempenho do modelo foi melhorando ao passo que mais tweets foram sendo incorporados a cada iteração, partindo de uma UAS de 94,46% com LAS de 93,86% iniciais e culminando em uma UAS de 95,96% com LAS de 94,89% após a sétima iteração. Em outras palavras, esses resultados indicam que a capacidade do modelo de capturar as estruturas linguísticas complexas e variadas presentes nos tweets for se aprimorando.

O modelo resultante da 6ª iteração de treinamento do modelo Stanza com o Porttinari-base e os tweets do DANTEStocks foi utilizado para anotar os 512 tweets restantes do córpus, totalizando os 4.048 tweets. Vale lembrar que 1.000 já compunham o subcórpus de referência e 2.536 foram anotados/revisados nas iterações de treinamento.

Embora a quantidade de tweets anotados pelo Stanza (seja durante ou após o treinamento) (isto é, ~3,000) tenha sido maior do que a envolvida na revisão da anotação realizada pelo UDPipe2 (isto é, ~1.000 de referência), o especialista fez a revisão dos ~3,000 em um intervalo aproximado de 3 meses (dez./23 a fev./24), demandando menos tempo que a construção do subcórpus de referência. Isso ocorreu principalmente porque o treinamento do Stanza com os tweets gerou anotações mais precisas/adequadas à linguagem CGU. Com isso, a revisão manual da anotação sintática do Stanza envolveu menos intervenção humana que as anotações geradas pelo UDPipe2.

Quanto ao modelo da 7ª iteração, isto é, contendo a totalidade dos tweets do DANTEStocks (4.048) e o córpus completo do Porttinari-base., seu treinamento visou não apenas melhorar a precisão do *parsing* de tweets do mercado financeiro, mas também explorar a transferibilidade do modelo treinado a outros gêneros, sejam CGU ou não, em português.

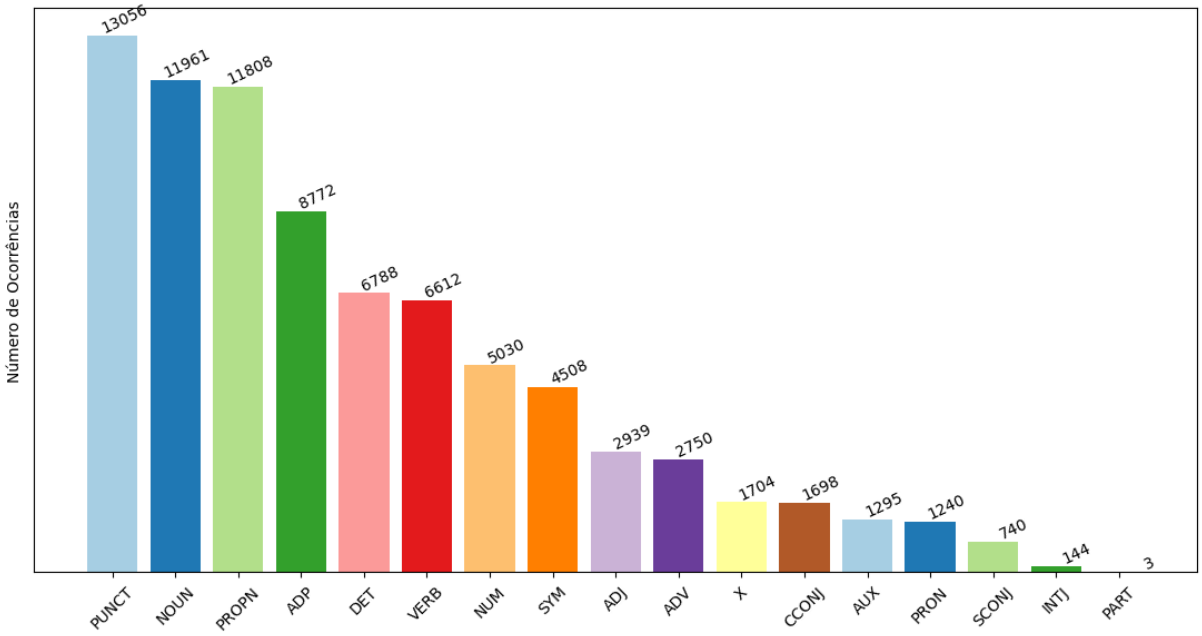
Para avaliar a capacidade de transferência do modelo final para diferentes contextos, utilizou-se o córpus PetroGold como um novo cenário de teste. O PetroGold, conforme descrito por (Souza *et al.*, 2021), é um *treebank* padrão ouro desenvolvido para o domínio de petróleo e gás. Ele é composto por 9.127 sentenças anotadas de acordo com a UD. O córpus, com divisão atualizada em maio de 2023, foi disposto em 80% para treinamento (7.170 sentenças), 12% para testes (1.039 sentenças) e 8% para validação (737 sentenças). O modelo final, quando testado com a porção de teste do PetroGold, obteve uma UAS de 92,33% e uma LAS de 90,25%, demonstrando sua adaptabilidade a um domínio bastante distinto.

4.3 Estatística da anotação das relações sintáticas

Sobre a anotação das *deprels*, ressalta-se que a revisão manual da anotação sintática automática causou a alteração de algumas *tags PoS* que haviam sido anotadas segundo as diretrizes presentes em Di-Felippo *et al.* (2021). Antes, porém, de discorrer sobre as alterações, exibe-se, na Figura 4.4, a distribuição das *PoS* no DANTEStocks após a anotação semiautomática das *deprels*.

Ao comparar as Figuras 3.1 e 4.4, observa-se que as alterações de anotação não influenciaram a distribuição estatística global das *tags*, mas sim a frequência de algumas delas em particular. No intuito de apresentar uma comparação numérica dessas mudanças, a Tabela 4.1 resume as diferenças observadas na distribuição pré- e pós-anotação de *deprels*.

Figura 4.4: Frequência das *tags PoS* no DANTEStocks após a anotação de *deprels*.



Fonte: O autor, 2024.

Tabela 4.1: Comparação da distribuição das *tags PoS* pré- e pós- anotação sintática.

Tag PoS	**Pré-*deprels***	**Pós-*deprels***	**Modificação**
PUNCT	13056	13056	–
NOUN	11934	11961	+27
PROPN	11438	11808	+370
ADP	8736	8772	+36
DET	6854	6788	-66
VERB	6586	6612	+26
NUM	5033	5030	-3
SYM	4509	4508	-1
ADJ	2918	2939	+21
ADV	2785	2750	-33
X	2095	1704	-391
CCONJ	1695	1698	+3
AUX	1316	1295	-21
PRON	1216	1240	+24
SCONJ	732	740	+8
INTJ	142	144	+2
PART	3	3	–

Analisando a Tabela 4.1, a principal mudança está na distribuição de PROPN e X. As ocorrências de PROPN passaram de 11.438 para 11.808, configurando um aumento de 370 ocorrências. Inversamente, *tokens* anotados como X passaram de 2.095 para 1.704, o que indica uma diminuição de 391 casos. Essa diferença reflete o fato de que a interpretação

dos tweets durante a revisão da anotação sintática levou à alteração principalmente das *tags PoS* de *cashtags* e *hashtags* quando ocorreriam no começo ou no final dos *tweets*.

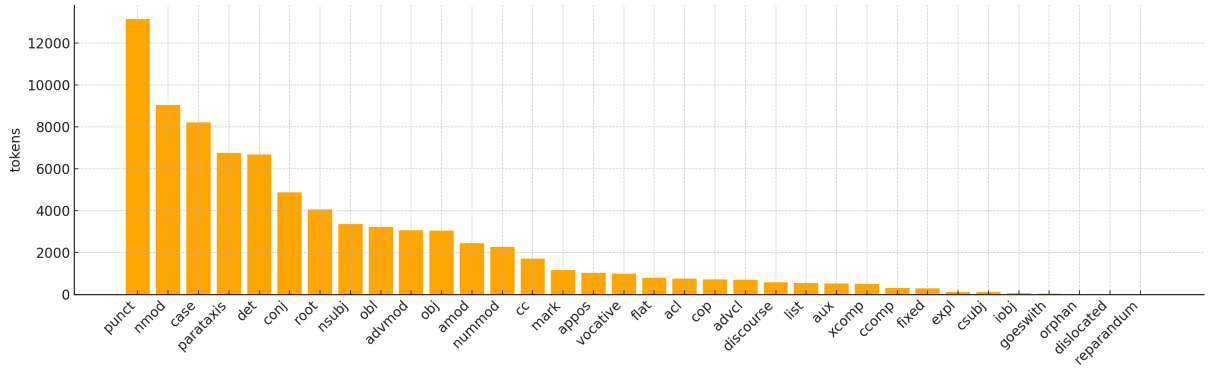
Na anotação morfossintática do córpus, muitos desses casos haviam sido anotados, segundo diretriz de Di-Felippo *et al.* (2021), como X, pois tinham sido interpretados à época como meros indexadores (de assuntos) na plataforma Twitter, sem ter, aparentemente, uma relação sintática clara com a mensagem expressa pelo tweet. No entanto, com o intuito de atribuir uma estruturação sintática possível, buscou-se mais conhecimento de domínio para interpretar os tweets. Essa análise semântica mais apurada dos *posts* levou à integração desses elementos à sintaxe de alguns tweets e, consequentemente, à alteração da *PoS* de X para PROPN, conforme diretriz de Di-Felippo *et al.* (2021). Esse foi o caso das *hashtags* dos exemplos em (42) (descritos na página 53 e listados novamente abaixo), pois elas passaram a ser PROPN, sendo anotadas, inclusive, como **root** da representação por dependência ao se interpretar que a mensagem principal dos tweet é sobre os *tickers* das *hashtags*.

(42) **(a)** #OIBR4 (mensagem : 956643) http://t.co/VD2ApxqWqR
(b) #USIM5 (mensagem : 956895) http://t.co/Whjn7UVWe2
(c) #BBAS3 Banco da Brasil (mensagem : 956467) http://t.co/75T8wtmEXw

Essas alterações, ao atribuírem *tags PoS* mais específicas a fenômenos anteriormente categorizados de forma genérica como X, aprimoram a caracterização linguística de elementos como *hashtags*, *chashtags* e também marcas de *retweet*.

Sobre a distribuição das *tags deprel*, 34 das 37 do modelo foram empregadas na anotação do córpus. As relações não empregadas foram **clf**, **compound** e **dep**. A Figura 4.5 exibe a distribuição das 34 *deprels* no DANTEStocks.

Figura 4.5: Distribuição das *deprels* do modelo UD no DANTEStocks.



Fonte: O autor, 2024.

Especificamente, **goeswith** foi empregada, para além dos casos previstos em Duran (2022) para o português, em uma diretriz específica para anotar repetições em sequência do cifrão (p.ex.: "$ $ $"), nas quais os símbolos foram tokenizados (como em "Long&Short de PETR4 x PETR3 em o lucro forte ! Operação sem usar $ $ $"). Quando integradas à sintaxe, repetições como essa ocorrem em substituição à palavra "dinheiro", funcionando como objeto direto. Assim, optou-se por conectar apenas o primeiro cifrão da sequência por **obj** ao *head* e os outros da sequência por **goeswih** (cf. Apêndice A).

A anotação do DANTEStocks englobou também sub-relações. No caso, 41 combinações de *deprel* e sub-relação foram utilizadas na anotação do DANTEStocks, cuja frequência de ocorrência de cada uma delas consta na Tabela 4.2. Essas combinações se justificam pela especificação das *deprels* em função dos fenômenos CGU e de domínio.

Assim, além das combinações já observadas na anotação-UD de córpus jornalistico em português (**acl:relcl**, **aux:pass**, **flat:foreign**, **flat:name**, **nsubj:outer**, **nsubj:pass**, **obl:agent**) (Duran *et al.*, 2023), as demais combinações do DANTEStocks se referem ao uso das sub-relações: (i) **:strunc** e **:wtrunc** para truncamentos estruturais e lexicais, respectivamente, e (ii) **:url**, **:hashtag**, **:cashtag**, para os fenômenos CGU/domínio.

Tabela 4.2: Quantificação de *deprels* e sub-relações.

Deprel	Quantidade
punct	13155
nmod	9047
nmod:strunc	49
nmod:wtrunc	33
case	8207
case:strunc	6
case:wtrunc	1
parataxis	6760
parataxis:url	1608
parataxis:hashtag	1016
parataxis:strunc	141
parataxis:cashtag	105
parataxis:wtrunc	38
det	6684
conj	4877
conj:strunc	31
conj:wtrunc	6
root	4048
nsubj	3355
nsubj:pass	108
nsubj:outer	10
nsubj:wtrunc	4
nsubj:strunc	1
advmod	3070
advmod:wtrunc	10
advmod:strunc	1
obl	3223
obl:strunc	27
obl:agent	27
obl:wtrunc	26
obj	3044
obj:wtrunc	18
obj:strunc	8
amod	2444
amod:wtrunc	10
amod:strunc	1
nummod	2276
cc	1698
cc:strunc	1
mark	1157
mark:strunc	3
appos	1028
appos:wtrunc	3
appos:strunc	2

Deprel	Quantidade
vocative	978
vocative:wtrunc	1
flat	785
flat:name	711
flat:foreign	30
acl	747
acl:relcl	289
acl:wtrunc	5
acl:strunc	2
cop	712
advcl	699
advcl:strunc	3
advcl:wtrunc	1
discourse	573
list	537
aux	516
aux:pass	101
xcomp	498
xcomp:strunc	4
xcomp:wtrunc	4
ccomp	297
ccomp:speech	30
ccomp:strunc	12
fixed	280
expl	119
csubj	103
iobj	52
goeswith	28
orphan	22
dislocated	16
reparandum	10

4.4 Avaliação da anotação sintática manual

A avaliação da anotação das relações de dependência foi feita pelo cálculo da concordância entre anotadores, o que permite verificar a qualidade e a confiabilidade da anotação. Para tanto, selecionou-se um conjunto aleatório de 100 tweets e a anotação sintática correspondente a cada um (advinda do subcórpus de referência ou de uma das iterações de treinamento do Stanza) foi submetida à revisão manual de um segundo anotador, especialista em UD e CGU. O anotador utilizou o Arborator-Grew-NILC e o manual de anotação de *deprel* para tweet do mercado financeiro de Di-Felippo,

Nunes e Barbosa (2024). Especificamente, empregaram-se duas métricas ou medidas de concordância. Uma delas foi a medida *Kappa* que, conforme a fórmula abaixo, computa o grau de concordância entre os anotadores em determinada tarefa, descontando-se a concordância ao acaso (Carletta, 1996). Na fórmula, P_o é a proporção de concordância observada e P_e é a proporção de concordância esperada pelo acaso. Como resultado, o coeficiente *Kappa* pode ter valores entre 0 e 1, sendo que, em geral, o intervalo 0.0-0.2 indica concordância insignificante e 0.91-1.0 indica concordância (quase) perfeita.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

A outra medida foi a concordância total, que calcula o número de casos em que os anotadores concordam, em relação ao número total de casos anotados, expressando o resultado em termos de porcentagem (Cabezudo *et al.*, 2015). Para calcular a concordância total, seguiram-se os seguintes passos: (i) quantificação do número total de casos anotados, representado por N; (ii) quantificação do número de concordâncias entre anotadores, representado por A (ou seja, o número de casos para os quais todos os anotadores forneceram a mesma anotação) e (iii) cálculo da concordância total utilizando a fórmula:

$$\text{concordância total} = \left(\frac{A}{N}\right) \times 100$$

As duas métricas foram calculadas em 3 cenários distintos, considerando: (i) *head* e *deprel* em conjunto, (ii) *head* e *deprel* separadamente e (iii) somente *deprel*.

No primeiro cenário, que considerou simultaneamente *head* e *deprel*, houve 1.658 casos de concordância e 85 de discordância do total de 1.743 casos. Isso corresponde a uma concordância total de 95,12% e a um *Kappa* de 0.9508. Essa taxa geral de concordância indica consistência significativa entre os anotadores, podendo ser resultante da eficácia do manual de anotação, que torna claro como o conjunto de *deprel* deve ser utilizado, com rica exemplificação de casos comuns e frequentes e casos mais raros e difíceis de anotar.

No segundo cenário, em que a anotação de *head* e *deprel* foram avaliadas separadamente, obteve-se, para *head*, concordância total de 96,67% e *Kappa* 0.9651 e, para *deprel*, concordância de 97,59% e *Kappa* 0.9739. Dos 1.743 casos, a discordância foi de 58 casos para *head* e 42 para *deprel*, havendo uma intersecção de discordância de 15 casos, o que é menos de 1% do total analisado. Essa intersecção indica casos em que

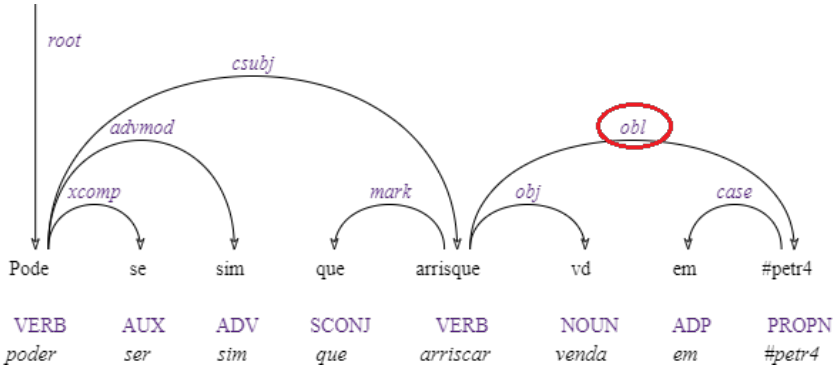
os anotadores discordaram quanto ao *head* e *deprel*. Embora a maior discordância seja referente à anotação do *head* em comparação à da *deprel* (58 e 42, respectivamente), a diferença de 16 casos não chega a 1%, o que a torna estatisticamente insignificante.

No terceiro cenário, calculou-se apenas a concordância total referente a cada relação de dependência. A medida *Kappa* não foi aplicada nesse cenário porque ela não se mostra adequada uma vez que as classes (isto é, os conjuntos de *deprels*) são desbalanceadas e algumas delas apresentam poucos casos ou ocorrências (Delgado; Tibau, 2019). Na Tabela 4.3, exibe-se a concordância total de cada *deprel* nos 100 tweets, incluindo as sub-relações. Os valores dessa concordância estão organizados de forma decrescente. Com base na Tabela, obteve-se 100% de concordância total para mais de 50% das *deprels*, incluindo relações bastante frequentes (como *case*, *nummod* e *cc*) e pouco frequentes, como é o caso das *deprels* com apenas 1 ocorrência (*fixed*, *nmod:wtrunc*, *obl:agent* e *root:hashtag*). A obtenção desse valor de concordância total para *deprels* com frequência de ocorrência opostas no DANTEStocks pode evidenciar uma rigorosa aderência dos anotadores ao manual de anotação de Di-Felippo, Nunes e Barbosa (2024), reafirmando a confiabilidade da anotação sintática para subsidiar pesquisas linguístico-computacionais.

Quanto à discordância, destaca-se que, no geral, foram pouquíssimos os casos em que os anotadores não escolheram a mesma *deprel*. Os 42 casos de discordância equivalem a apenas 2,4% do total de 1.743 ocorrências analisadas. Entre as relações mais frequentes, como *nmod*, com 196 casos anotados pelo primeiro anotador, o segundo anotador discordou em 8 casos. Em um deles, a divergência foi entre *obl* e *nmod*. As Figuras 4.6 e 4.7 ilustram as anotações manuais distintas que causaram essa divergência. Com base nas árvores, vê-se que o primeiro anotador considerou "#petr4" como um complemento nominal do verbo "arrisque" constituído por um nominal preposicionado. O segundo anotador considerou "#petr4" um modificador do nome abreviado "vd" ("venda"). A diferença na anotação reside na definição da estrutura de argumentos em que o PROPN "#petr4" participa, se do verbo ou do nome "vd". As duas interpretações parecem possíveis.

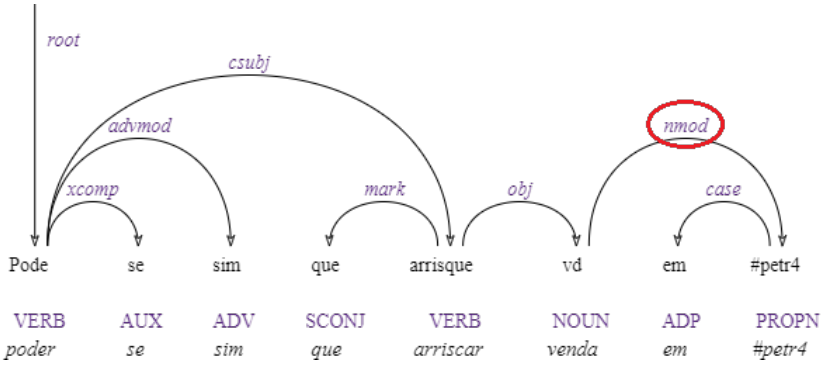
Sobre a discordância das *deprels* menos frequentes, destaca-se que as 3 relações com apenas 1 ocorrência nos 100 tweets da avaliação possuem sub-relações (*punct:strunc*, *nmod:mention* e *ccomp:strunc*). Nesses casos, a discordância diz respeito apenas à especificação das sub-relações, uma vez que apenas um dos anotadores as colocou.

Figura 4.6: Exemplo de discordância entre *nmod* e *obl* (árvore do Anotador 1).



Fonte: O autor, 2024.

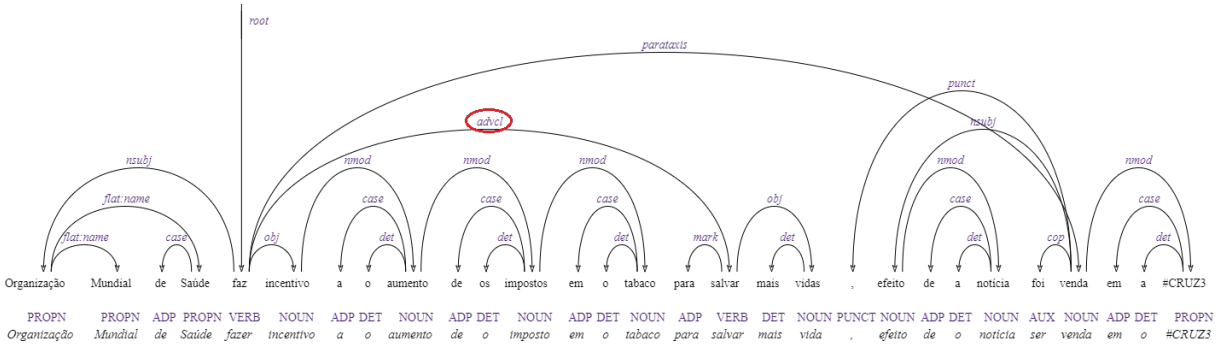
Figura 4.7: Exemplo de discordância entre *nmod* e *obl* (árvore do Anotador 2).



Fonte: O autor, 2024.

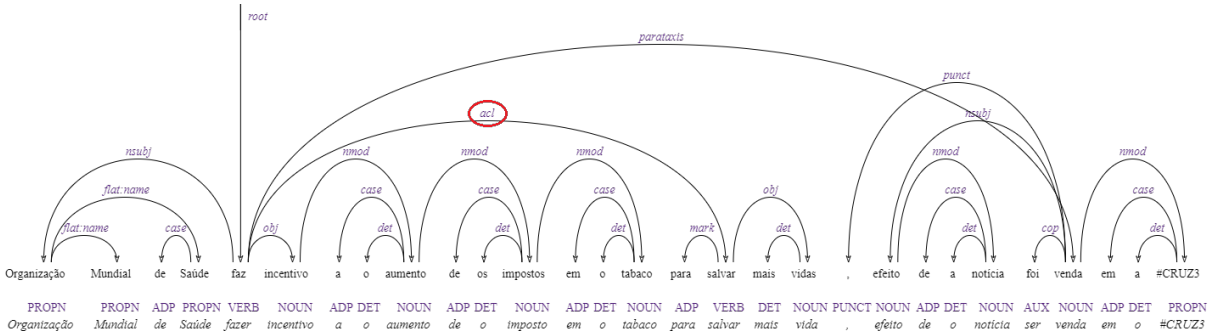
Além dos casos envolvendo sub-relações, ilustra-se aqui uma discordância referente às *deprels* menos frequentes com a única ocorrência de *advcl* (do total de 12 anotadas pela primeiro anotador), que foi anotada como *acl* pelo segundo anotador. As árvores correspondentes a essa divergência entre *advcl* e *acl* estão ilustradas respectivamente pelas Figuras 4.8 e 4.9.

Figura 4.8: Exemplo de discordância entre *advcl* e *acl* (árvore do Anotador 1).



Fonte: O autor, 2024.

Figura 4.9: Exemplo de discordância entre *advcl* e *acl* (árvore do Anotador 2).



Fonte: O autor, 2024.

Com base nas Figuras, observa-se que, ao empregar a *deprel advcl* entre os verbos "faz" e "salvar", o primeiro anotador considerou o trecho "para salvar mais vidas" uma oração que modifica o predicado da oração matriz ("faz"), acrescentando-lhe, no caso, uma informação de finalidade ou razão. O segundo anotador, por outro lado, adotou *acl* para a mesma relação entre "faz" e "salvar". No entanto, segundo o manual de Duran (2022), essa não parece ser a *deprel* adequada, uma vez que *acl* deve ser empregada na conexão de uma palavra de conteúdo, não verbal, e uma oração que a modifica. A escolha do segundo anotador parece ser de fato um equívoco.

Como mencionado anterior, a avaliação da anotação sintática-UD manual revelou, no geral, uma taxa alta de concordância entre os anotadores. Embora as características linguísticas do DANTEStocks impõem desafios a qualquer tipo de anotação, sobretudo de *PoS* e *deprels*, acredita-se que o manual de Di-Felippo, Nunes e Barbosa (2024) se mostrou um material de suporte eficiente, uma vez que fornece diretrizes para anotar a sintaxe com base em padrões estruturais e fenômenos UGC recorrentes no DANTEStocks. E isso parece ter levado a uma anotação sintática consistente.

Atualmente, os 85 casos de discordância referentes à anotação das *deprels* estão sendo analisados conjuntamente pelos dois anotadores envolvidos na avaliação, com o intuito de definir de forma consensual a anotação final ou de referência para esses casos que comporá o DANTEStocks.

Tabela 4.3: Concordância total de anotação por *Deprel.*

deprel	Qt. total	Concordância	Discordância	Concordância total (%)
case	177	177	0	100.00
nummod	51	51	0	100.00
cc	51	51	0	100.00
advmod	46	46	0	100.00
appos	22	22	0	100.00
mark	20	20	0	100.00
flat:name	10	10	0	100.00
nmod:date	9	9	0	100.00
list	9	9	0	100.00
aux	9	9	0	100.00
xcomp	8	8	0	100.00
acl:relcl	7	7	0	100.00
parataxis:strunc	4	4	0	100.00
csubj	3	3	0	100.00
nsubj:pass	3	3	0	100.00
expl	2	2	0	100.00
orphan	2	2	0	100.00
aux:pass	2	2	0	100.00
iobj	2	2	0	100.00
nmod:strunc	2	2	0	100.00
fixed	1	1	0	100.00
nmod:wtrunc	1	1	0	100.00
obl:agent	1	1	0	100.00
root:hashtag	1	1	0	100.00
punct	308	307	1	99.68
det	132	131	1	99.24
root	99	97	2	97.98
amod	42	41	1	97.62
conj	114	111	3	97.37
parataxis:url	38	37	1	97.37
nsubj	72	70	2	97.22
parataxis	67	65	2	97.01
nmod	196	188	8	95.92
cop	20	19	1	95.00
obl	57	54	3	94.74
discourse	14	13	1	92.86
obj	69	64	5	92.75
advcl	12	11	1	91.67
parataxis:hashtag	22	20	2	90.91
vocative	18	16	2	88.89
acl	10	8	2	80.00
ccomp	4	3	1	75.00
parataxis:cashtag	4	3	1	75.00
punct:strunc	1	0	1	0.00
nmod:mention	1	0	1	0.00
ccomp:strunc	1	0	1	0.00

4.5 Considerações sobre a anotação de *deprel*

Sobre a anotação de *deprels*, em particular, seguem aqui algumas observações.

A primeira delas diz respeito à dificuldade da tarefa de anotação/revisão de tweets do mercado financeiro. Embora essas dificuldades já tenham sido mencionadas na literatura geral sobre a construção de *tweebanks*, a anotação de tweets do mercado financeiro impõem obstáculos adicionais, sobretudo pela fragmentação (já que muitos tweets são resultado de copia-e-cola de manchetes e tabelas de outras fontes) e necessidade de conhecimento de domínio para interpretar as mensagens devido à ocorrência de vocabulários terminológicos e de estruturas de linguagem específicas.

Devido a essas características, a linguagem dos tweets do mercado financeiro acarreta elipses e ambiguidades, como mencionado no manual do Apêndice A, e isso dificulta sobremaneira a compreensão do conteúdo e, por conseguinte, a identificação das *deprels*, em especial, do **root**. Dessa forma, um tweet desse domínio pode ter mais de uma interpretação possível e, assim, mais de uma anotação sintática, não havendo, portanto, uma escolha de *deprel* dita "correta". Com isso, as anotações do DANTEStocks, assim como as diretrizes do Apêndice A, refletem a interpretação do anotador responsável pela tarefa. E, consequentemente, a distribuição estatística das *deprels* no DANTEStocks, descrita na 4.2, também é reflexo disso.

Casos paradigmáticos que ilustram isso são os tweets fragmentados do exemplo (40), como o item (b): "*#USIM5 (mensagem: 956895) http://t.co/Whjn7UVWe2*". A interpretação dada com o auxílio de especialistas do domínio do mercado financeiro foi a de que a ação, representada pelo *ticker* "#USIM5", é o conteúdo principal e, por isso, anotada como **root**. Sobre esse **root**, veiculam-se duas informações, isto é, uma mensagem numerada entre parênteses e a fonte de publicação dessa mensagem. Como não há uma ligação sintática clara entre o **root** e os dois blocos de informação, ambos foram conectados ao **root** por *parataxis*, sendo o último deles especificado pela sub-relação *:url*. Em outra interpretação possível, a mensagem numerada publicada sobre a ação poderia ser o **root**. Nesse caso, no entanto, o **root** estaria entre parênteses, o que parece ser menos interessante, pois as informações parentéticas tendem a ser menos importantes sintática e até mesmo discursivamente (segundo teorias como a *Rethorical Structure Theory* (RST) de Mann e Thompson (1988)). Por essa razão, aliás, optou-se pela anotação do *ticker* como **root**.

Aliás, pelas características linguísticas já mencionadas dos dados do córpus, a revisão da anotação automática de *deprels* em tweets do mercado financeiro se configurou uma tarefa extremamente desafiadora, sobretudo na busca por consistência. A construção do manual de anotação-UD no qual as diretrizes foram definidas em função de estruturas e fenômenos recorrentes no córpus e a organização dos tweets, para revisão manual, em blocos/projetos que reunissem essas estruturas e/ou fenômenos foram estratégias que buscaram diminuir a complexidade e contribuir para a homogeneidade/consistência da anotação de *deprels*.

A segunda observação diz respeito à comparação entre as diretrizes de anotação de *deprels* empregadas no DANTEStocks e as propostas por Sanguinetti *et al.* (2023) para a anotação sintática-UD de UGC. Especificamente, Sanguinetti *et al.* (2023) propuseram uma uniformização das diretrizes para a anotação sintática-UD de CGU, fornecendo um arcabouço comum aplicável a diferentes línguas e domínios, contribuindo, com isso, para a construção de *tweebanks* sob um mesmo *framework*. Devido ao rigor metodológico e abrangência dos fenômenos UGC tratados pelos autores, o conjunto de diretrizes de Sanguinetti *et al.* (2023) se tornou a principal referência da literatura internacional. Assim, a comparação mencionada busca evidenciar a adesão da anotação do DANTEStocks a essa principal proposta da literatura, assim como pontuar as principais diferenças entre elas.

A Tabela 4.4 sistematiza apenas os fenômenos CGU que ocorrem no DANTEStocks e que foram abordados por Sanguinetti *et al.* (2023). Optou-se por utilizar a terminologia em inglês para os fenômenos por essa ser amplamente conhecida no cenário internacional de construção de *tweebanks*. As diretrizes empregadas da anotação de deprels do DANTEStocks contidas da Tabela foram sistematizadas a partir de dois trabalhos prévios. Um deles é o próprio manual de anotação sintática-UD para tweets em português de Di-Felippo, Nunes e Barbosa (2024). E o outro é o de Scandarolli *et al.* (2023), já mencionado em 3.1.1, em que os autores caracterizam os fenômenos lexicais do DANTEStocks.

Observa-se com base na Tabela 4.4 que as diretrizes de anotação das *deprels* no DANTEStocks são bastante similares à de Sanguinetti *et al.* (2023). As poucas diferenças residem sobre os fenômenos de *punctuation reduplication* (repetição de pontuação), *truncation* (truncamento) e *at-mention* (menção).

Tabela 4.4: Comparação das diretrizes de anotação sintática-UD do DANTEStocks com as de Sanguinetti *et al.* (2023).

Fenômeno	**Sanguinetti *et al.* (2023)**		**DANTEStocks**	
	Papel sintático padrão	**Outro**	**Papel sintático padrão**	**Outro**
Diacritic omission	✓		✓	
Vowel omission	✓		✓	
Phonetization	✓		✓	
Spelling errors	✓		✓	
Abbreviation	✓		✓	
Contraction	✓		✓	
Punctuation redupl.	✓		✓(`punct`)	
Graphemic stretching	✓		✓	
Disguise	✓		✓	
Neologism	✓		✓	
Truncation	✓		✓	`:wtrunc` ou `:strunc`
Hashtags				
Synt. integrated	✓		✓	
Standalone		`parataxis:hashtag`		`parataxis:hashtag`
At-mentions				
Synt. integrated	✓		✓	`nmod` (de RT precedente)
Standalone		`vocative` (root)		`parataxis`
URLs				
Synt. integrated	✓		✓	
Standalone		`parataxis:url`		`parataxis:url`
Pictograms/emoticons				
Synt. integrated	✓		✓	
Standalone		`discourse`		`discourse`
RTs				
Synt. integrated	✓		✓(`vocative`)	
Standalone		`parataxis`		`parataxis`
Code-switching				
INTRA	✓(se conhecido)	`flat:foreign` (se desconhecido)	✓	`flat:foreign` (se desconhecido)

Quanto à repetição de pontuação, não se trata de fato de uma diferença na seleção da etiqueta para rotular a relação, pois ambas as propostas utilizam *punct*, seguindo, aliás, as diretrizes gerais da UD, mas sim sobre a tokenização desse fenômeno. Enquanto Sanguinetti *et al.* reconhecem a sequência de pontuação como um *token* único, cada sinal de pontuação da sequência foi considerado um *token* individual no DANTEStocks e, por conseguinte, anotado com a relação *punct*. Isso foi indicado na Tabela 4.4 pela explicitação da etiqueta *punct* entre parênteses junto ao sinal de visto (✓).

Sobre os truncamentos, a distinção diz respeito apenas à inserção no DANTEStocks das sub-relações *:wtrunc* e *:strunct* para as casos de truncamento lexical e estrutural, respectivamente, as quais não são propostas por Sanguinetti *et al.* Tais sub-relações foram especificadas nas diretrizes de anotação do DANTEStocks devido ao fato de os truncamentos serem frequentes e, por vezes, os do tipo estrutural afetarem a decisão de anotação sintática. Por essa razão, viu-se a relevância de explicitar, via sub-relação, o tipo específico de ocorrência do fenômeno de truncamento que influenciou a anotação de determinada *deprel*.

Outro caso de diferença é relativo às menções nos tweets. Ambos os trabalhos propõem diretrizes para menções integradas à sintaxe da mensagem e para aquelas que ocorrem em contexto isolado (*standalone*). Além disso, ambos concordam que, quando integradas, elas sejam anotadas pela *deprel* que representa sua função sintática no contexto do tweet, o que é representado pelo sinal de visto (✓) na coluna "papel sintático padrão". A diferença específica entre as proposta aqui se refere ao tratamento dado para as menções precedidas da marca de RT. Enquanto Sanguinetti *et al.* as classifica como *standalone*, conectando-as ao *root* pela relação *vocative*, essas menções no DANTEStocks são dependentes por *nmod* do símbolo de RT que as precede (cf. Figura 3 do Apêndice A).

Embora não haja de fato uma distinção entre as propostas no que se refere à anotação sintática-UD do símbolo de *retweet* (RT), vale observar que essa marca, integrada à sintaxe, ocorreu apenas em 2 tweets do DANTEStocks, tendo sido conectada ao *root* da mensagem por *vocative* (cf. Figura 2 do Apêndice A). Isso foi indicado na Tabela 4.4 pela explicitação da etiqueta *vocative* entre parênteses junto ao sinal de visto (✓).

Diante do fato de as diferenças entre as propostas serem poucas, sendo o tratamento dado às menções precedidas de RT o ponto de maior divergência, pode-se dizer o DANTEStocks foi anotado sintaticamente com base em diretrizes que são em sua ampla maioria similares às de Sanguinetti *et al.*, dando, assim, a esse *tweebank*, o potencial de ser utilizado em pesquisas multilíngue (ou *cross-linguistic*) baseadas em UD.

No próximo capítulo (5), apresenta-se a especificação (ou anotação) dos papéis semânticos projetados pelos 145 nomes predicadores (lemas distintos) nas 1.756 instâncias (e 1.218 tweets distintos) sintaticamente anotadas e revisadas. Tal anotação, como descrita a seguir, foi feita de acordo com o arcabouço teórico-metodológico do NomBank. Os resultados da investigação sobre a valência sintático-semântica dos nomes que ocorrem no DANTEStocks estão descritos no Capítulo 7.

5 Descrição semântica com base no NomBank

Diante do sucesso do PropBank e projetos correlatos devido ao rigor teórico-metodológico e potencial de aplicação no PLN, a descrição da valência semântica, necessária para a investigação da estrutura-A dos Npred, foi feita com base no NomBank. Especificamente, a identificação dos papéis semânticos dos argumentos projetados pelos 145 Npred nas 1.756 instâncias do DANTEStocks se baseou nos 4.706 arquivos de *frames* do NomBank. Para essa tarefa, contou-se com apenas um especialista/anotador. Dessa forma, adotou-se uma metodologia similar à empregada no PropBank.Br v.1, principalmente porque se optou-se por usar os *frame files* do NomBank em inglês como referência.

5.1 Metodologia de descrição semântica

A tarefa de descrição semântica dos nomes foi equacionada em 4 etapas metodológica, todas elas manuais. Assim, dado um Npred x do conjunto de 145 nomes, fez-se: (i) definição do sentido de x e tradução de x para um equivalente em inglês y, (ii) busca pelo *frame* correspondente a y no repositório do NomBank e seleção do *roleset* que codifica o sentido de x no córpus (instância(s)), (iii) classificação semântica do Npred para definição de diretrizes de identificação dos Arg, e (iv) identificação efetiva dos Arg previstos pelo *roleset* selecionado nas instâncias de x e associação de seus papéis semânticos com base nas árvores de dependência.

Para ilustrar as etapas (i-vi), discorre-se sobre o processo de descrição da estrutura-A do Npred "compartilhamento" (definido como "ato ou efeito de compartilhar; partilha de algo entre várias pessoas"), um dos Npred da Tabela 3.4.

5.1.1 Definição do sentido do Npred e tradução

Seguindo a metodologia, a descrição das estruturas-a dos Npred nas instâncias teve início com a definição do sentido do Npred no córpus e tradução dos nomes para o inglês. A identificação do sentido do Npred foi feita com base na interpretação do mesmo em contexto (ou seja, no tweet em que o Npred ocorre) e a tradução, em particular, foi feita com base nos 3 recursos online principais, a saber: *Google Tradutor*[1], *Linguee*[2] e *Cambridge Dictionary*[3].

No caso de "compartilhamento", por exemplo, a interpretação das instâncias ilustradas em (43) levou à definição de "compartilhamento" como sendo "ato ou efeito de compartilhar; partilha de algo entre várias pessoas", sendo que, a partir da consulta aos recursos online mencionados, optou-se por *"sharing"* como equivalente de tradução mais apropriado.

(43) **(a)** BR BOVESPA GOLL4 Gol assina acordo de **compartilhamento** de voos com TAP . http://t.co/wHGukBg7qp

(b) $GOLL4 - GOL e TAP assinam acordo para **compartilhamento** de voos http://t.co/F87EcEzEWK

Os processos de identificação ou definição do sentido e tradução não foram tarefas simples para alguns Npred. Essa dificuldade foi gerada principalmente pela aparente inexistência de um equivalente de tradução direto para o sentidos no Npred no córpus e pela necessidade de conhecimento de domínio.

Um exemplo de Npred para o qual parece não haver uma tradução direta é "antro", que signica "local asqueroso, propenso à corrupção, degeneração moral". A tradução selecionada, *hideaway* ("lugar geralmente remoto usado por bandidos"), foi a que mais se aproximou do sentido em uso no DANTEStocks.

Já um exemplo de Npred cuja definição do sentido demandou conhecimento de domínio é "desova". No mercado financeiro, esse nome tem sentido de "grande volume de venda de títulos em período de tempo curto, resultando na queda dos preços dos papéis". Assim, ele foi traduzido para *sell-off*.

[1] https://translate.google.com/?hl=pt-BR
[2] https://www.linguee.com.br/
[3] https://dictionary.cambridge.org/

Além disso, vale ressaltar que foram poucos os casos que envolveram polissemia, isto é, nomes que ocorrem com mais de um sentido no córpus. Um desses casos é o nome "acordo". Para as 9 instâncias do nome "acordo", as quais correspondem a 9 tweets diferentes, identificaram-se três sentidos distintos, os quais correspondem aos seguintes equivalentes de tradução: *agreement, accord* e *accordance*. Cada um desses nomes está representado no NomBank por um *frame* específico, composto por apenas um *roleset*, a saber: "*agreement.01*" (5 instâncias) , "*accord.01*" (1 instância) e "*accordance.01*" (3 instâncias). As 9 instâncias do Npred "acordo" e seus respectivos *frames/rolesets* estão listados na Figura 5.1.

Quadro 5.1: Descrição da estrutura-A de um Npred por *frames/rolesets* distintos.

Texto	Frame
$BBAS3 - Banco Do Brasil (bbas-nm) - Fato Relevante **Acordo** Com Os Correios http://t.co/kW3xh2fjzU	**agreement**
$MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo Ao **Acordo** De Acionistas http://t.co/jBJzeY5U5V	**agreement**
$CCRO3 - Ccr Sa (ccro-nm) - Fato Relevante - Assinatura Do Aditivo Do **Acrodo** De Acionista http://t.co/xzQyZPPbBL	**agreement**
$GOLL4 - GOL e TAP assinam **acordo** para compartilhamento de voos http://t.co/F87EcEzEWK	**agreement**
#BR #BOVESPA #GOLL4 Gol assina **acordo** de compartilhamento de voos com TAP. http://t.co/wHGukBg7qp	**agreement**
@gugabianco Até onde sei, 1) não temos cap. instalada para refino, 2) **Acordos** comerciais, Petr4 precisa exp pra gerar receita em dólar	**accord**
Veja as melhores ações para comprar nesta semana, de **acordo** com 8 corretoras: Os papéis da G... http://t.co/L6OsbF6Os6 #infomoney #vale5	**accordance**
Veja as melhores ações para comprar em esta semana , de **acordo** com 8 corretoras : Os papéis de a Gerdau (GGBR4) e de a ... http://t.co/SK7rRvdHEZ	**accordance**
Confira 5 'top picks' para comprar este mês, de **acordo** com a Socopa: As 'Top Picks' da corre... http://t.co/Zu6BAhev7W #infomoney #vale5	**accordance**

Fonte: O Autor, 2024.

5.1.2 Identificação do *frame file* e seleção do *roleset*

Uma vez que o equivalente de tradução y tenha sido selecionado, passou-se para a busca pelo *frame file* correspondente a y no repositório do NomBank, análise dos *rolesets* constitutivos do *frame* (caso haja mais de um) e seleção do *roleset* adequado

No que diz respeito ao Npred "compartilhamento", por exemplo, *sharing* foi usado como termo de busca no repositório. Nele, identificou-se o *frame* correspondente *"sharing"* (44). Esse *frame* é composto por apenas um sentido de "*sharing*" e, portanto, apenas um *roleset*. Especificamente, esse *roleset*, herdado do *frame* do verbo/sentido *"share.01"* do PropBank, prevê 3 Arg, associados aos seguintes papéis semânticos e glosas: Arg0 (*sharer*), Arg1 (*thing shared*) e Arg2 (*shared with, if separate from Arg0*).

Na grande maioria das vezes, os *frame files* eram compostos por apenas um *roleset*, o que facilitou o processo de descrição semântica. Ao final, dentificou-se um *frame/roleset* correspondente para cada um dos 145 nomes, indicando que o NomBank teve uma ótima cobertura dos sentidos dos nomes que ocorrem no DANTEStocks (e, por conseguinte, no domínio do mercado financeiro).

```
(44) <frameset>
       <predicate lemma="sharing">
         <roleset id="sharing.01" name="share" source="verb-share.01">
           <roles>
             <role descr="shearer" n="0"/>
             <role descr="thing share" n="1"/>
             <role descr="shared with, if separate from arg0" n="2"/>
           </roles>
           <example name="autogen1">
             <text> a hard-liner who quickly ruled out any sharing of
             power with pro-democracy groups </text>
             <arg n="0">a hard-liner --> who --> *T*-2</arg>
             <arg n="Support">ruled</arg>
             <rel>sharing</rel>
             <arg n="1">of power</arg>
             <arg n="2">with pro-democracy groups</arg>
           </example>
           <example name="autogen2">
             <text> employee profit-sharing </text>
             <arg n="0">employee</arg>
             <arg n="1">profit-sharing</arg>
             <rel>profit-sharing</rel>
           </example>
         </roleset>
       </predicate>
     </frameset>
```

Ainda sobre a seleção dos *frame/rolesets*, salienta-se que, ao final, 6 *frames* foram empregados para descrever mais de um Npred, os quais, por isso, podem ser vistos como sinônimos ou ao menos semanticamente similares. Tais Npred e seus respectivos *frames* estão sistematizados na Tabela 5.1.

Tabela 5.1: Lexicalizações distintas de um mesmo *frame/roleset.*

Frame	**Nomes**
allocation	rateio, locação
change	trocas, alteração
fusion	incorporação, fusão
trade	daytrade, troca
buy	compra, recompra
return	retorno, volta

5.1.3 Definição da classe/tipo semântico do Npred

Uma vez o *frame/roleset* tenha sido definido, o Npred em português foi classificado segundo as classes e tipos semânticos do NOMLEX-PLUS, que representam padrões sintático-semânticos. Assim, as classes do NOMLEX-PLUS serviram de guia para a descrição da estrutura-A, assim como no NomBank.

No Apêndice B deste documento, está descrita a classificação completa dos 145 Npred analisados no trabalho. O Npred "compartilhamento", como exemplo, é, via *sharing*, da classe NOMING[4] e do tipo VERB-NOM. Tal informação de classe/tipo indica que se trata de uma nominalização de um verbo (no caso, "compartilhar") (ou de um nome deverbal). Isso quer dizer que o verbo e o nome correspondente partilham da mesma estrutura de argumentos (com realizações sintáticas diferentes).

No Verbo-Brasil, o verbo "compartilhar" é descrito pelo mesmo *frame/roleset* apresentado em 44. Nele, há uma instância desse verbo anotada segundo o PropBank, a qual, aliás, possui os mesmos argumentos realizados na sintaxe que "compartilhamento" na Figura 5.1. Na instância "a previsão de saques reduzidos sobre a poupança é **compartilhada** por especialistas" do Verbo-Brasil, tem-se que "A previsão de saques reduzidos sobre a" é o Arg1 e "por especialistas" é o Arg0. Por se tratar da voz passiva, o Arg0 (*sharer*) é um Sprep. Em construções com voz ativa, os Ar0 e Arg1 desse verbo não são preposicionados (p.ex.: "Especialistas **compartilham** previsão de saques reduzidos sobre a poupança"). Assim, uma relação que existe entre as realizações das estruturas-A do verbo e Npred correspondente, a qual auxilia na identificação dos Arg, é a de que, considerando a voz ativa do verbo, seu sujeito é o Arg0 do Npred e o seu objeto é o Arg1 do Npred (no caso, preposicionado, como em (**compartilhamento** "de voos").

Assim, sabendo da classe/tipo do Npred , diretrizes de identificação dos Arg/papéis semânticos puderam ser sistematizadas. No caso dos Npred do tipo VERB-NOM, a diretriz principal foi a de analisar a descrição dos verbos correspondem, quando estes estivessem disponíveis no Verbo-Brasil, a fim de garantir o paralelismo entre as estruturas-A, considerando, no caso, a observação feita aqui sobre a distinção entre a realização sintática do Arg1 na predicação verbal e nominal.

Com a classificação dos 145 Npred, buscou-se garantir que as diretrizes traçadas para uma classe/tipo fossem aplicadas consistentemente a todos os nomes constitutivos.

5.1.4 Identificação dos argumentos e papéis semânticos

Uma vez que o *roleset* que adequadamente representa o sentido de um nome no córpus tenha sido selecionado, realizou-se a identificação dos Arg previstos pelo *roleset* nas instâncias, assim como a associação dos respectivos papéis semânticos. Seguindo a

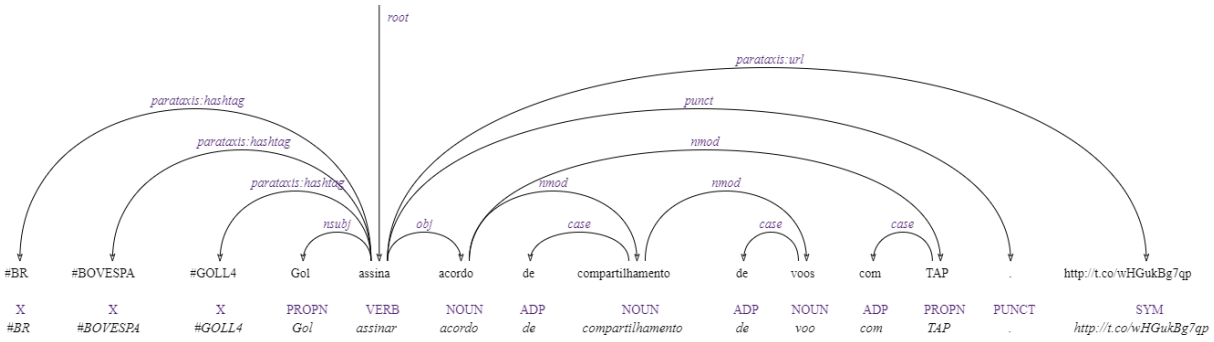
[4]Esse rótulo é usado para nomes em inglês terminados em *-ing*, como *sharing*.

metodologia do NomBank/PropBank, isso foi feito manualmente e com base nas árvores de dependência de cada instância, oriundas da anotação sintática-UD do DANTEStocks.

Ressalta-se, neste ponto, que a tarefa em questão não consistiu efetivamente em uma anotação *à la* NomBank das 1.756 instâncias do córpus, pois os dados resultantes da identificação dos Arg sintaticamente realizados e de seus respectivos papéis semânticos previstos nos *rolesets* foram organizados em um arquivo independente do córpus, no caso, em uma tabela no formato xlsx, como descrito mais adiante. Esses dados, no entanto, podem subsidiar uma futura anotação semântica do DANTEStocks segundo o modelo PropBank/NomBank. Além disso, é importante destacar que este trabalho, por uma questão de delimitação de escopo, focou apenas na identificação dos Arg numerados do modelo NomBank, previstos nos *rolesets* que constituem os *noun frames*. Isso significa que os ArgM (modificadores) não foram descritos.

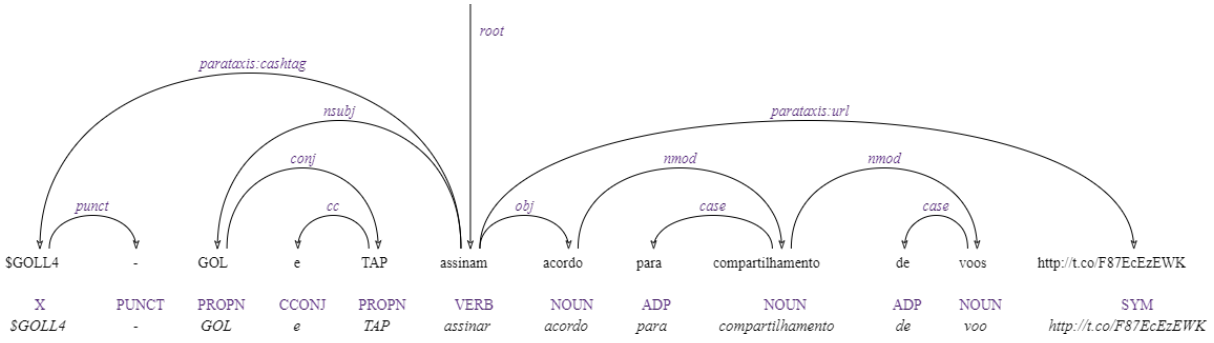
No caso das instâncias de "compartilhamento", as árvores de dependência correspondentes às instâncias descritas em (43) estão ilustradas pelas Figuras 5.1 e 5.2, respectivamente.

Figura 5.1: Árvore de dependência da instância 1 de "compartilhamento".



Fonte: O autor, 2024.

Figura 5.2: Árvore de dependência da instância 2 de "compartilhamento".



Fonte: O autor, 2024.

Com base nas árvores de dependência, identificaram-se os Arg previstos no *roleset* em cada uma das instâncias, organizando-os em um arquivo no formato .xlsx composto por 4 campos (Quadro 5.2). O primeiro se refere à primeira coluna e contém o *sent_ID*, isto é, o identificador único do tweet no DANTEStocks. O campo seguinte, denominado *estrutura-A*, engloba as colunas referentes aos argumentos do *roleset* selecionado. No exemplo, esse campo possui 3 colunas, cada uma delas especifica um dos 3 Arg do nome (isto é, Arg0 (*sharer*), Arg1 (*thing shared*) e Arg2 (*shared with, if separate from Arg0*)). O terceiro campo, intitulado *Texto*, indica o tweet propriamente dito e o quarto campo descreve o *frame/roleset* herdado do NomBank que representa o sentido do nome.

Com base no Quadro 5.2, a estrutura-A projetada pelo nome predicador "compartilhamento" em ambas as instâncias não possui os 3 argumentos realizados sintaticamente, uma vez que o Arg2 (*"shared with, if separate from arg0"*) não ocorre na superfície das instâncias/tweets. O sintagma preposicional "com TAP", que pode ser confundido com o Arg2 de "compartilhamento", é, segundo a árvore da Figura 5.1, o Arg2 de "acordo".

Quadro 5.2: Identificação da *estrutura-A* de "compartilhamento" em suas 2 instâncias.

	Estrutura-a				
sent_ID	**Arg 0**	**Arg 1**	**Arg 2**	**Texto**	**Frame**
dante_01_454607743392178177l	GoI	de voos	-	#BR #BOVESPA #GOLL4 GoI assina acordo de compartilhamento de voos com TAP . http://t.co/wHGukBg7qp	**sharing**
dante_01_454604295078690816l	GOL e TAP	de voos	-	$GOLL4 - GOL e TAP assinam acordo para compartilhamento de voos http://t.co/F87EcEzEWK	**sharing**

Fonte: O autor, 2024.

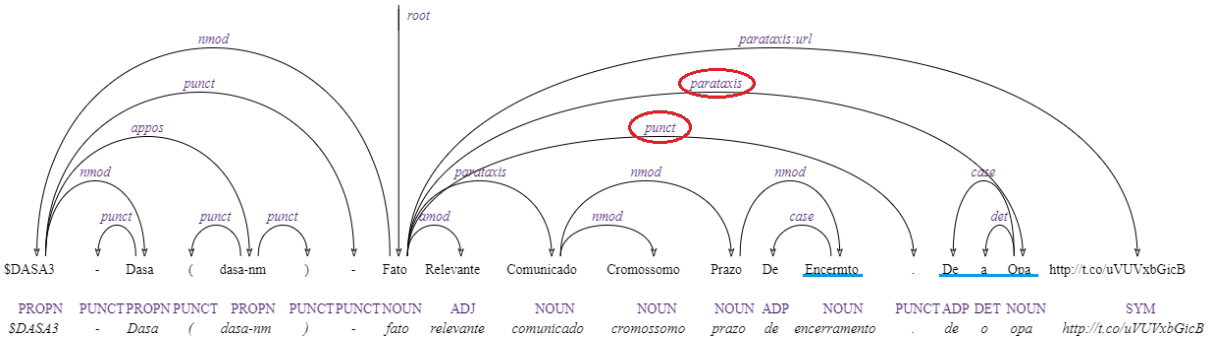
Na descrição de uma instância em particular, referente ao Npred "encermto" (isto é, "encerramento") (Figura 5.3), lidou-se com o que se considerou um equívoco de anotação sintática. Diz-se isso porque, segundo a Figura, o núcleo "Opa"[5] do segmento "Da Opa" (tokenizado para "De a Opa") está conectado por ao *root* ("Fato") por *parataxis*. No entanto, trata-se do Arg1 (*thing closing*) do Npred "encermto" de acordo com o *roleset closure.01* selecionado.

Acredita-se que esse equívoco ocorreu devido à interpretação do sinal de pontuação (.) após "encermto" como um ponto final. Essa pontuação, porém, refere-se ao fenômeno de abreviação (especificamente de *shortening* (ou encurtamento) segundo Scandarolli *et al.* (2023)) que caracteriza o token "encermto". Dessa forma, o Arg1 foi devidamente identificado e o equívoco sintático corrigido.

[5] "Opa" é a sigla para "Oferta Pública de Aquisição".

O equívoco de anotação sintática do DANTEStocks aqui reportado foi o único que afetou diretamente a descrição semântica.

Figura 5.3: Exemplo de anotação sintática-UD com erro.



Fonte: O autor, 2024.

5.2 Diretrizes para identificação dos Arg

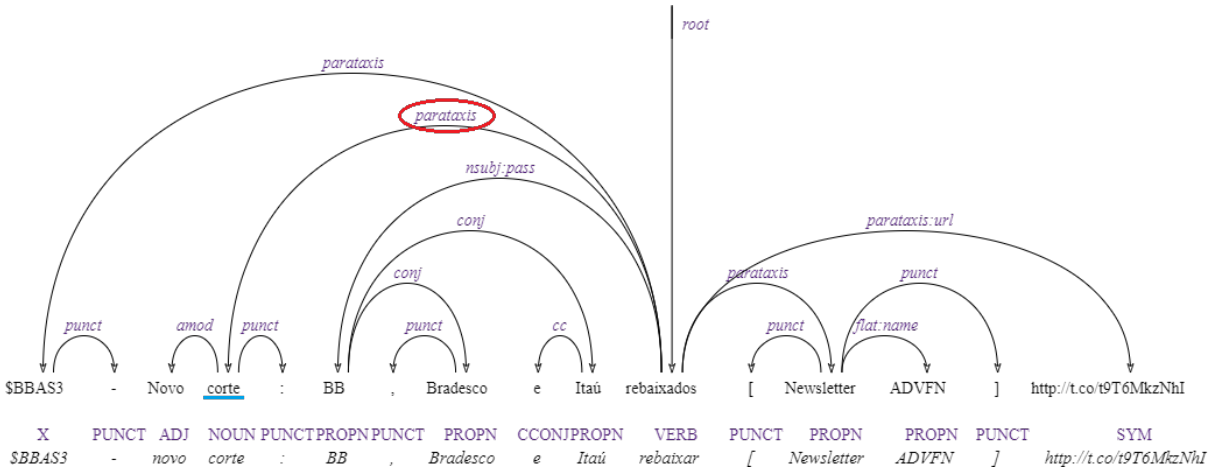
Para a efetiva identificação dos argumentos e conseguinte associação de papéis semânticos, utilizou-se o manual de Duran (2014) proposto no projeto PropBank.Br. Nele, encontram-se as diretrizes principais de uma anotação segundo o modelo PropBank para textos de língua padrão em português. No entanto, a descrição da estrutura-A de nomes predicadores em tweets demandou algumas diretrizes particulares para esse cenário.

A primeira delas é uma diretriz geral que estabelece que somente os Arg diretamente relacionados ao Npred devem ser considerados na descrição. Isso quer dizer que, caso um Arg ocorra em um trecho do tweet relacionado ao Npred por *parataxis*, por exemplo, este não deve ser descrito. A Figura 5.4 ilustra uma instância desse tipo. Especificamente, a instância é referente ao Npred "corte".

Com base na árvore de dependância da Figura 5.4, o Npred "corte" não tem relação sintática com o restante da mensagem, posto que ele está relacionado por *parataxis* ao segmento que contém o *root* do tweet ("rebaixados"). Com isso, o segmento em que os Arg do *roleset cuttback.01* (isto é, *cutter* (Arg0), *thing reduced* (Arg1), *amount reduced by* (Arg2), *start point* (Arg3) e *end point* (Arg4)) pudessem está apenas justaposto ao segmento que contém o Npred. Assim, embora sendo possível inferir que as ações de BB, Bradesco e Itaú (representadas aqui pelo nome das próprias instituições) sejam o Arg1 (*thing reduced*), esse único participante da predicação de "corte" no tweet não foi

considerado na descrição da estrutura-A do Npred. Dessa forma, considerou-se que todos os 4 Arg previstos no *roleset cuttback.01* não estão presentes no tweet em questão.

Figura 5.4: Exemplo de Npred isolado do restante do tweet por *parataxis*.



Fonte: O autor, 2024.

A segunda diretriz é complementar à primeira. Ela estabelece que, caso um Arg não esteja diretamente conectado ao Npred, este somente deve ser incluído na análise se for Arg0 (sujeito) e puder ser "herdado" da predicação verbal principal, garantindo, assim, a identificação do participante da predicação nominal. Esse é o caso, por exemplo, do Arg0 de "compartilhamento" (e também de "acordo") nas Figuras 5.1 e 5.2. Embora ele não esteja diretamente relacionado ao Npred (como o Arg1 "de voos", que está conectado ao Npred por *nmod*), o Arg0 pode ser herdado da predicação principal do tweet, projetada por "assinar" (*root*). Sintaticamente, aliás, é possível percorrer poucas arestas da árvore para restrear a conexão indireta entre o Arg0 e o Npred. Na Figura 5.1, por exemplo, "compartilhamento" é *nmod* de "acordo", "acordo" é *obj* de "assinar" e esse verbo, por fim, tem "Gol" (Arg0) como seu *nsubj*.

A terceira diretriz geral trata da ocorrência de dois argumentos do *roleset* em um único sintagma, como elementos coordenados. Se na instância de "compartilhamento", cuja árvore de dependência está ilustrada na Figura 5.1, os Arg0 (*sharer*) e Arg2 (*shared with*) estão separados, sendo "Gol" identificado como Arg0 e o participante introduzido pela preposição "com" (isto é, "com TAP") como Arg2, o mesmo não ocorre na instância da Figura 5.2. Nela, os Arg0 e Arg2 estão coordenados, compondo o SN "GOL e TAP". Nesse caso, o SN inteiro deve ser rotulado como Arg0, segundo Meyers (2007). E essa foi também a regra adotada no DANTEStocks.

A quarta diretriz geral é relativa à ocorrência de um Npred acompanhado de um único Arg na forma de SPrep, como na instância (45). Diante dessas ocorrências, é preciso decidir se o único participante da predicação nominal expresso na sintaxe é Arg0 ou Arg2. Segundo Meyers (2007), caso esse participante esteja no plural e funcione como complemento de preposições como "*between*"/"*among*" em inglês, esse argumento deve ser identificado como Arg0. Embora não haja em (45) uma preposição como as indicadas por Meyers, optou-se por designar o sintagma introduzido pela preposição "de" (seguido do nome no plural "Acionistas") como Arg0 (e, portanto, "*agreer*") (45a), pois é possível inferir que o "acordo" ocorreu "entre os acionistas".

(45) $MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo Ao **Acordo** De Acionistas http://t.co/jBJzeY5U5V

(a) [...] **Acordo** $_{\text{Arg0}}$[de Acionistas] [...]

Além das diretrizes gerais, definiram-se outras específicas para as classes/tipos.

Uma delas é a já mencionada para os Npred das classes NOM e NOMING que sejam do tipo VERB-NOM. Trata-se da diretriz de analisar a estrutura-A do verbo correspondente, quando este está disponível no Verbo-Brasil, considerando que os sujeitos dos verbos morfologicamente relacionados são (usualmente) Arg0 dos Npred e os objetos são (usualmente) Arg1 dos Npred.

Outras duas diretrizes específicas foram definidas para os Npred da NOM que tenham o tipo SUBJECT (isto é, nominalizações do Arg0=sujeito do verbos correspondentes (p.ex.: "gestor"[6] e "exemplo"[7]) e OBJECT (isto é, nominalizações do Arg1=objeto dos verbos correspondentes) (como "comissão"[8]).

Para os casos de SUBJECT e OBJECT, buscou-se manter o paralelismo entre a descrição dos verbos (ou adjetivos) e dos nomes relacionados, adotando a mesma estratégia de Meyers (2007). Para esses casos, a diretriz é a de descrever/anotar o próprio Npred como argumento "incorporado". Assim, na instância da Figura 5.5, por exemplo, o próprio Npred "gestor" é descrito como Arg0; o mesmo procedimento pode ser observado nas Figuras 5.6

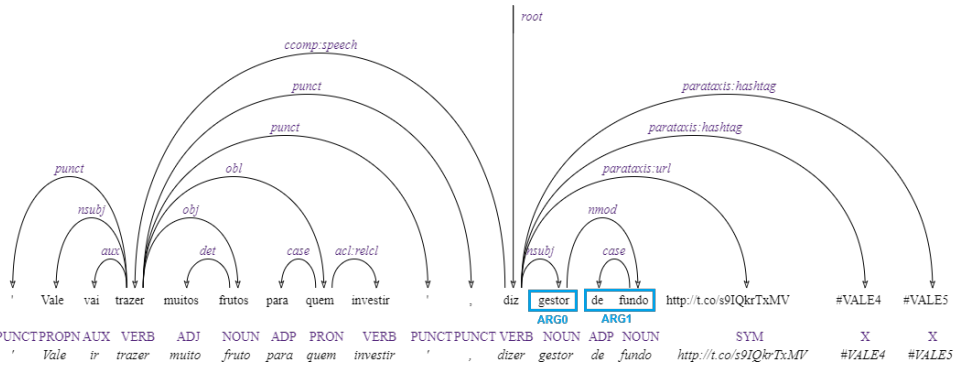
[6]O verbo "gerir" prevê um Arg0 na função de sujeito, como "Maria" em "Maria gere o hotel". A nominalização "gestor" "incorpora" o Arg0 e, por isso, é do tipo SUBJECT (Meyers, 2007).

[7]O verbo "exemplificar" prevê um Arg0, como "a compra da mansão" em "A compra da mansão exemplifica sua ganância". O nome "exemplo" "incorpora" o Arg0 e, por isso, é também do tipo SUBJECT.

[8]O verbo "comissionar" prevê um Arg2 objeto, como "oficiais" em "O general comissionou oficiais para a missão". A nome "comissão" "incorpora" o Arg2 e, por isso, é do tipo OBJECT (Meyers, 2007).

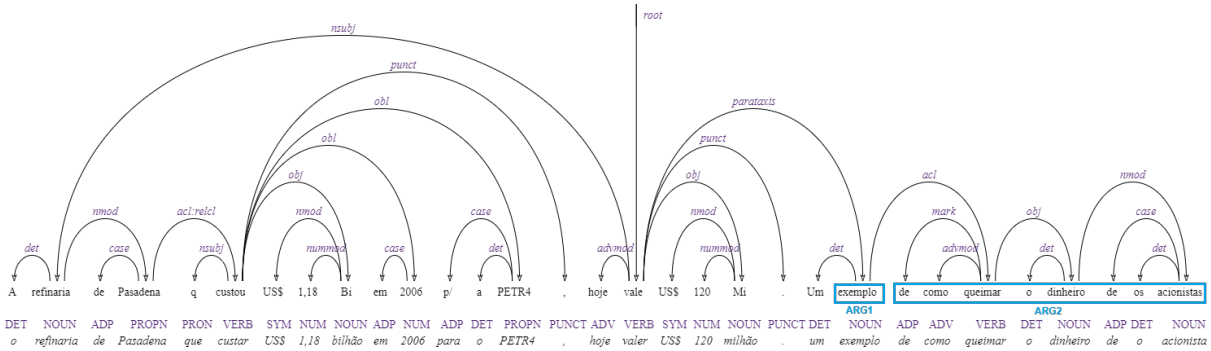
e 5.7 para os demais nomes ilustrados (“exemplo” e “comissão”). A anotação desses Npred como Arg0, Arg1, etc. dependerá de como fazer a descrição da estrutura-A dos nomes ficar compatível com estrutura-A dos verbos correspondentes. Uma das justificativas para a descrição do Npred como Arg “incorporado” é, segundo Meyers (2007), a da possível utilização desse tipo de descrição em trabalhos sobre correferência. Caso se tenha, por exemplo, que uma entidade pessoa (“Maria”) e “gestor de fundos” sejam correferentes em algum contexto, é possível deduzir que “Maria gere/geriu os fundos”.

Figura 5.5: Exemplo de descrição de Npred SUBJECT (“gestor”).



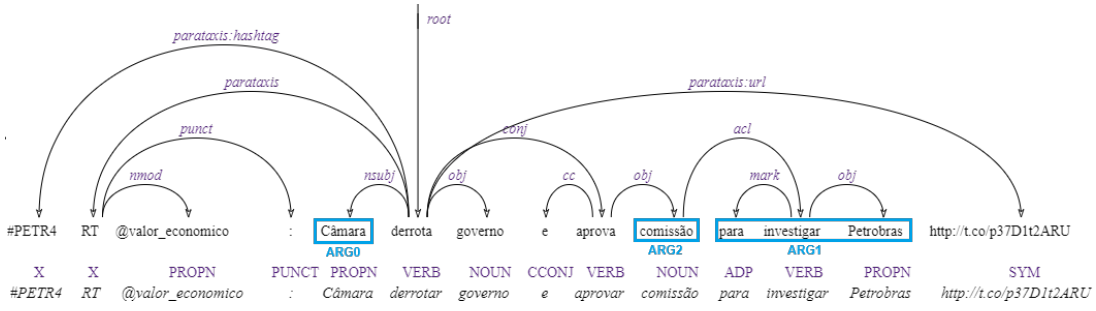
Fonte: O autor, 2024.

Figura 5.6: Exemplo de descrição de Npred SUBJECT (“exemplo”).



Fonte: O autor, 2024.

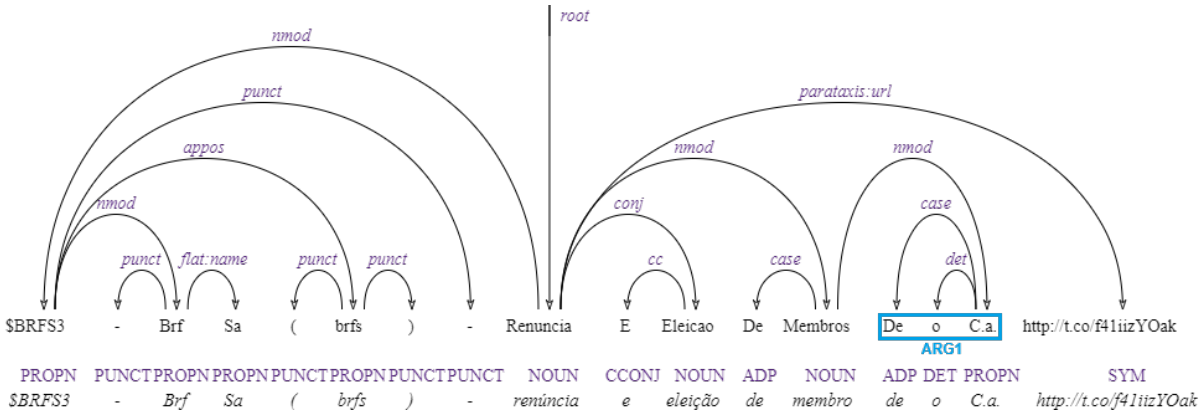
Figura 5.7: Exemplo de descrição de Npred OBJECT (“comissão”).



Fonte: O autor, 2024.

Os Npred da classe PARTITIVE, em particular, representam, no córpus, parte, quantidade ou divisão (seção) de um todo e esse todo é descrito como Arg1. Esse é o caso de "membro" na instância da Figura 5.8, que o trecho "de o C.a. (comitê administrativo)" é Arg1. Os nomes da classe PARTITIVE que expressam quantidade aceitam por vezes um "tema secundário" (que também expressa quantidade), o qual é anotado como Arg3. Os que expressam "divisão", por suz vez, aceitam um "tema" como Arg2, que expressa uma parte do todo. Em "divisão de lácteos da BRF", por exemplo, "de lácteos" é Arg2, ao passo que "de a BRF" é o Arg1.

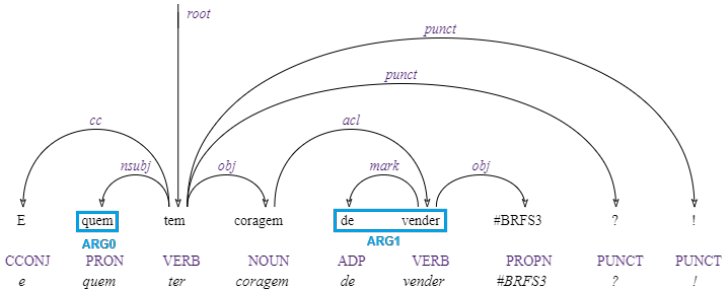
Figura 5.8: Exemplo de descrição de Npred PARTITIVE ("membro").



Fonte: O autor, 2024.

Os Npred da classe NOMADJ compreendem nominalizações derivadas de adjetivos que expressam qualidade ou estado, como "coragem". Seguindo as diretrizes do NomBank, tem-se que o complemento do Npred, quando na forma de uma oração completiva (indicado pela *deprel* **acl** na Figura 5.9), deve ser anotado como Arg1. Caso contrário, o complemento tende a ser descrito como Arg0, como "do rapaz" em "A coragem do rapaz é grande".

Figura 5.9: Exemplo de descrição de Npred NOMAJD ("coragem").

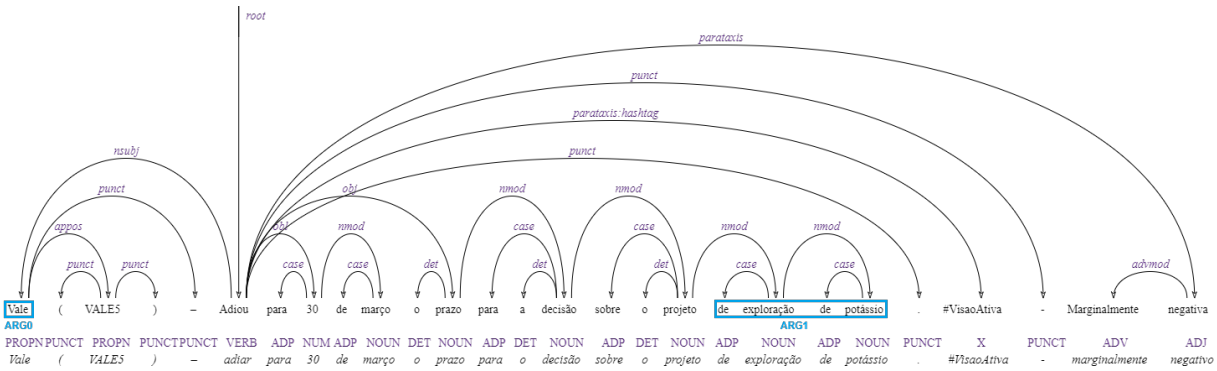


Fonte: O autor, 2024.

Para os Npred da classe ABILITY, como "projeto", a diretriz é a de identificar o argumento na função de complemento (preposicionado), o qual pressupõe uma ação, como

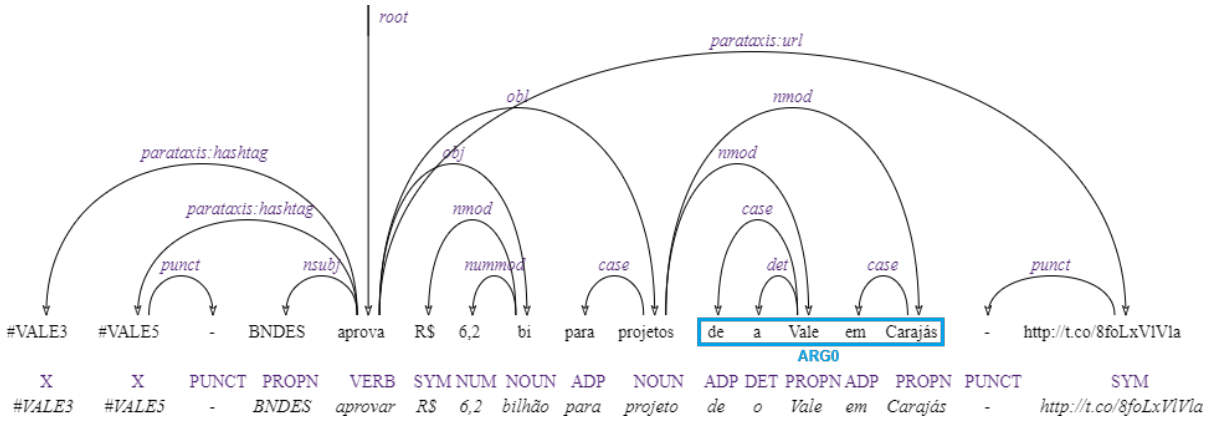
Arg1. Na instância da Figura 5.10, por exemplo, o SN “de exploração de potássio” é o Arg1 do Npred em questão, sendo o Arg0 (“Vale”) herdado da predicação do verbo. Caso haja um SN preposicionado que não expresse ação, como “de a Vale” na Figura 5.11, a diretriz é a de descrever esse SN como o Arg0.

Figura 5.10: Exemplo de descrição de Npred ABILITY (“projeto”) com Arg1.



Fonte: O autor, 2024.

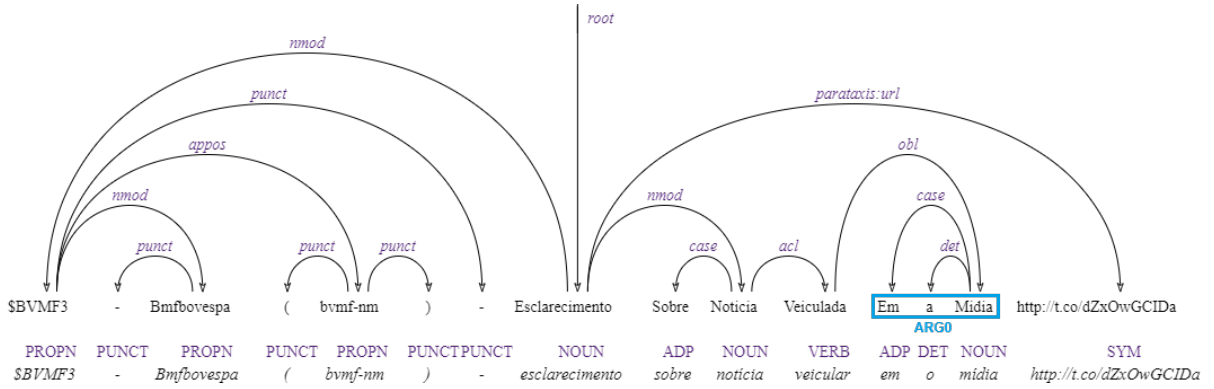
Figura 5.11: Exemplo de descrição de Npred ABILITY (“projeto”) com Arg0.



Fonte: O autor, 2024.

Os nomes da classe WORK-OF-ART projetam semanticamente o “transmissor”/“originador” de um “X” como Arg0 e um “X”, que é uma abstração humana (como ideia, música, notícia, fotografia, pintura, etc.) como Arg1. Na instância da Figura 5.12, o Arg1 do Npred “notícia” não ocorre e o trecho “em a mídia” (relacionado indiretamente à “notícia”) foi interpretado como sendo o Arg0.

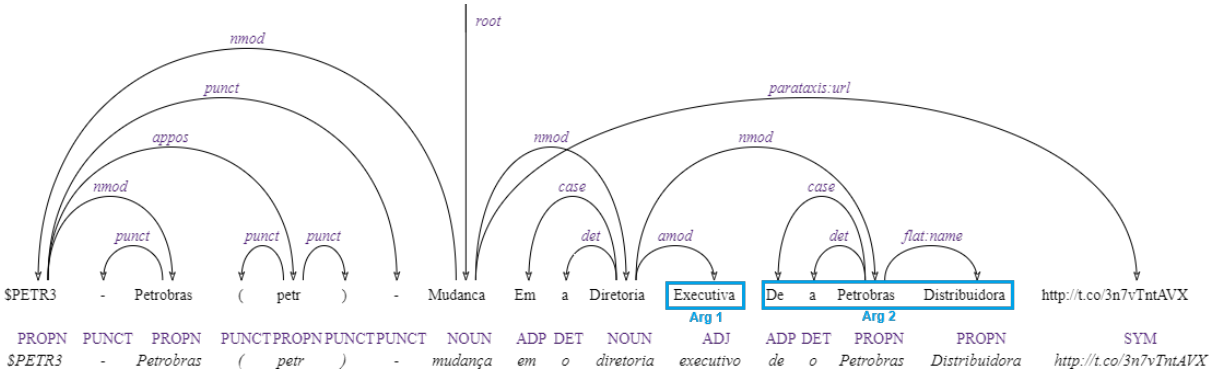
Figura 5.12: Exemplo de descrição de Npred WORK-OF-ART (“notícia”).



Fonte: O autor, 2024.

Nos casos de Npred da classe GROUP, a diretriz é a de descrever o descritor do grupo como Arg1 (que comumente é expresso por um adjetivo) e a entidade a que o grupo pertence como Arg2. No caso do Npred “diretoria” da Figura 5.13, tem-se “executiva” como Arg1 (adjetivo descritor) e “de a Petrobras distribuidora” como Arg2.

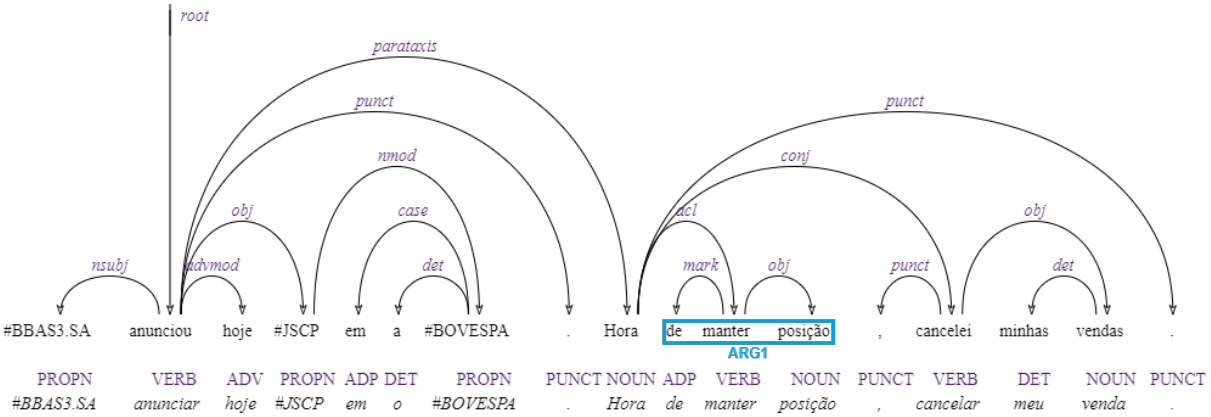
Figura 5.13: Exemplo de descrição de Npred GROUP (“diretoria”).



Fonte: O autor, 2024.

Os Npred da classe ENVIRONMENT, segundo Meyers (2007), podem indicar uma situação, um período de tempo, uma locação, um corpo físico (p.ex.: “a atmosfera da Terra”) ou mesmo um evento ou conjunto de eventos. No DANTEStocks, as ocorrências dos nomes dessa classe indicam “período de tempo”, como “hora”. Para tais Npred da classe ENVIRONMENT, o Arg1 é o único que tende a se realizar na sintaxe com frequência. A diretriz é a de descrever o sintagma preposicionado iniciado por “de” que sucede os Npred como Arg1, o qual indica um evento. Na instância do Npred “hora” na Figura 5.14, tem-se o trecho “de manter posição” como Arg1.

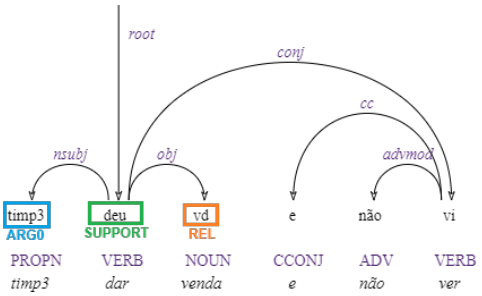
Figura 5.14: Exemplo de descrição de Npred ENVIRONMENT (“hora”).



Fonte: O autor, 2024.

Sobre a descrição das instâncias com Vsup, ressalta-se que a anotação sintática-UD não fornece um rótulo específico para esses verbos. Assim, a identificação dos Vsup nas instâncias foi feita com base na definição de Meyers (2007)[9] e nos trabalhos de Rassi (2023), Barros (2014) e Santos (2015) sobre “ficar”, “dar” e “fazer”, respectivamente. Além desses trabalhos, a classificação do verbo “dar” em expressões compostas pelos Npred “compra” e “venda” foi feita por meio do auxílio de especialista de domínio. Diz-se isso porque o especialista indicou que expressões como “dar venda/compra” podem ser interpretadas como “dar (sinal) de venda/compra” (isto é, “sinalizar venda/compra”) e, com isso, o verbo em questão foi considerado como Vsup. Seguindo a metodologia do NomBank, os Vsup foram descritos com o rótulo SUPPORT, como ilustrado na Figura 5.15. Vale lembrar aqui que, de acordo com a anotação sintática-UD codificada na árvore da referida Figura, o *ticker* “timp3” substitui o nome da empresa (no caso, “TIM”), sendo, portanto, descrito como Arg0 (*buyer.*).

Figura 5.15: Exemplo de descrição de Npred com Vsup.



Fonte: O autor, 2024.

[9]Segundo Meyers (2007), Vsup é um verbo que assume um SN *A* com argumento como um de seus argumentos, e pelo menos outro argumento *B*, tal que *A* também toma *B* como argumento

A seguir, discorre-se sobre a avaliação da tarefa manual de identificação/anotação dos Arg/papéis semânticos.

5.3 Avaliação da descrição/anotação semântica

Para avaliar a qualidade da identificação dos Arg e conseguinte associação de papéis temáticos, a descrição realizada até então foi comparada a de outros dois especialistas, sendo eles estudantes de graduação do curso de Linguística. Para tanto, esses especialistas passaram por um processo de treinamento estruturado em quatro rodadas, cada uma com duração média de uma semana. Durante cada rodada de treinamento, os linguistas receberam 5 Npred distintos e 2 instâncias do DANTEStocks sintaticamente anotadas de cada um dos nomes, totalizando 40 instâncias. Ao final de cada semana, discutiam-se as descrições de cada espacialista para sanar dúvidas. Tanto para o treinamento quanto para a anotação efetiva dos casos selecionados para compor a avaliação, os linguistas utilizaram o manual de Duran (2014) e as diretrizes descritas na seção anterior (5.2).

Vale ressaltar que a avaliação aqui reportada foi feita somente a respeito da tarefa de identificação dos Arg/papéis semânticos. Isso quer dizer que os anotadores já recebiam o *roleset* previamente selecionado para cada um dos Npred envolvidas na avaliação, tendo de verificar de os *roles* estavam ou não sintaticamente realizados nas instâncias. Com isso, cada arquivo para treinamento e para a posterior anotação final era composto por uma tabela, bastante similar à descrita no Quadro 5.2, contendo: (i) o Npred, (ii) as 2 instâncias de Npred no córpus, (iii) a indicação do *roleset* selecionado e (iv) os Arg do *roleset*. Como ilustrado no Quadro 5.2, a tarefa era a de preencher as colunas referentes aos Arg com os respectivos segmentos dos instâncias. Além da tabela, o arquivo continha as árvores de dependência de cada instância e o *roleset* completo do NomBank para consulta.

Após o treinamento, os dois anotadores receberam um mesmo pacote de dados contendo 42 Npred e 180 instâncias (cerca de 10% do total de 1.756). Excluindo os analisados no treinamento, os 42 nomes apresentam diferentes tipos de valência quantitativa (isto é, monovalentes, bivalentes, etc.). Esse critério foi utilizado com vistas a assegurar diversidade quanto à complexidade das valências ou estruturas-a. A distribuição dos nomes para avaliação em função da valência quantitativa está descrita na Tabela 5.2. Vale lembrar que a valência quantitativa foi determinada pela quantidade de papéis previstos

no *roleset* de cada nome já selecionado pelo especialista que gerou a descrição original dos Npred. Aliás, considerando essa descrição como referência, a distribuição dos diferentes tipos de Arg na coleção das 180 instâncias de avaliação está disposta na Tabela 5.3.

Tabela 5.2: Quantidade de Npred na coleção de avaliação em função da valência quantitativa.

Valência quantitativa	Quantidade de Npred
V2	7
V3	16
V4	15
V5	4

Tabela 5.3: Quantidade dos tipos de Arg na coleção de avaliação (anotação original).

Tipos de Arg	Quantidade total
Arg0	90
Arg1	125
Arg2	52
Arg3	14
Arg4	0

A avaliação foi realizada durante um período de 10 dias, de forma individual e sem contato entre os dois linguistas. Para avaliar a consistência da identificação dos Arg e papéis semânticos realizada então pelos três especialistas, utilizou-se o coeficiente *Kappa* de Fleiss. Diferentemente do *Kappa* de Cohen, empregado na seção 4.4, que se limita a dois anotadores, o *Kappa* de Fleiss (1981) estende a medida *Kappa* para três ou mais juízes.

A análise da concordância foi realizada em duas etapas. Na primeira, calculou-se o *Kappa* de Fleiss (i) geral, considerando todos os Arg, e (ii) o específico de cada tipo de Arg (isto é, Arg0, Arg1, Arg2, Arg3 e Arg4). Os valores de (ii) estão dispostos na Tabela 5.4, que também apresenta a classificação dos valores segundo a escala de Fleiss, que estabelece: pobre ($K < 0.4$), satisfatório a bom ($0.4 \leq K < 0.75$) e excelente ($K \geq 0.75$).

O *Kappa* geral obtido foi de 0.7998, sendo um resultado "excelente" segundo a escala descrita. Vale ressaltar que a escala de Fleiss não é consensual, havendo diferentes classificações na literatura. Ademais, a interpretação da escala é dependente da tarefa. No caso, para a identificação de Arg/papéis semânticos, um *Kappa* geral que se aproxima de 0.80 é considerado compatível com a literatura, indicando que, na maioria dos casos, os anotadores concordam sobre a tarefa.

Sobre os resultados na Tabela 5.4, em particular, observa-se que a identificação dos Arg2, Arg3 e Arg4 geraram os maiores valores de concordância, com valores de

Kappa considerados "excelentes" pela escala de Fleiss. Isso quer dizer que, diante de um *frame/roleset* em que o Arg2, o Arg3 ou o Arg4 esteja previsto, os especialistas muito frequentemente concordam quanto à ocorrência ou não deles nas instâncias.

Tabela 5.4: *Kappa* de Fleiss específico de cada tipo de Arg.

***Kappa* de Fleiss**		
Tipo de Arg		**Escala**
Arg0	0.6887	Satisfatório a bom
Arg1	0.7389	Satisfatório a bom
Arg2	0.8043	Excelente
Arg3	0.8557	Excelente
Arg4	0.9066	Excelente

Na segunda etapa da avaliação, calculou-se o *Kappa* de Fleiss em função da valência quantitativa dos Npred descritos na avaliação (isto é, V2, V3, V4 e V5) e dos tipos de Arg previstos nos *frames/rolesets*. Os índices de concordância quanto à valência estão sistematizados em ordem decrescente na Tabela 5.5.

Comparativamente, os índices gerais da Tabela 5.5 indicam que a descrição dos Npred de V5 e V2 gera menos discordância que os demais (V4 e V3), posto que V5 e V2 obtiveram os valores mais altos (0.86 e 0.78, respectivamente). Vale ressaltar, no entanto, que os Npred de V5 e V2 são os menos frequentes na coleção de instâncias avaliadas, com 4 e 7 ocorrências, respectivamente (cf. Tabela 5.2). No que tange aos tipos de Arg, os resultados parecem indicar que a identificação do Arg0 (e de certa forma também do Arg1) gera mais discordância quando a valência envolve vários participantes, posto que os índices de *Kappa* para Arg0 dos nomes de V4 e V5 foram os mais baixos (0.58 e 0.59, respectivamente). No entanto, sabe-se que os Arg0 e Arg1 são o mais frequentes na coleção, com 90 e 125 ocorrências segundo a anotação tomada como referência (cf. Tabela 5.3) e, por isso, passíveis de maior discordância, como salienta Meyers (2007).

Tabela 5.5: *Kappa* de Fleiss por valência e tipo de Arg.

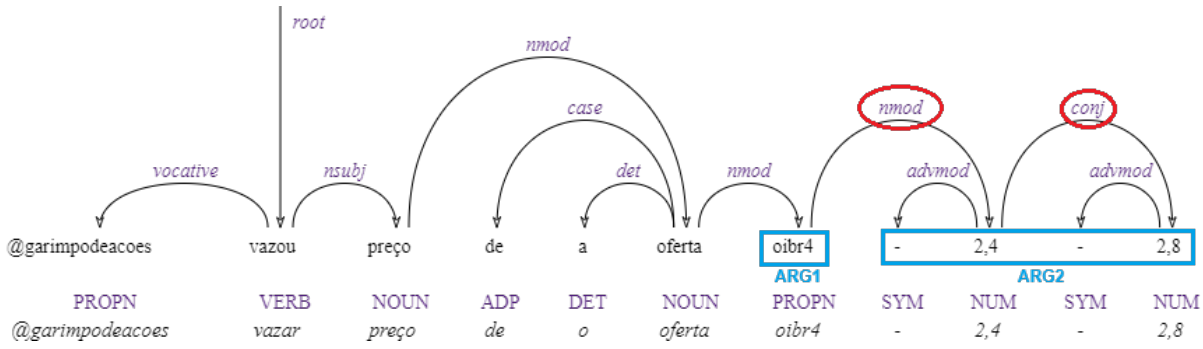
Valência	**Arg0**	**Arg1**	**Arg2**	**Arg3**	**Arg4**	**Kappa/Valência**
V5	0.8434	0.7627	0.8651	0.9519	0.9066	0.8685
V4	0.6687	0.7618	0.7535	0.8372	-	0.7610
V3	0.7165	0.7202	0.8360	-	-	0.7280
V2	0.7855	0.7342	-	-	-	0.7888

No que diz respeito às dificuldades reportadas pelos linguistas que participaram no processo de avaliação, a principal delas é a interpretação do tweet, que, por vezes,

requer conhecimento de domínio para reconhecer os participantes ou argumentos. Essa dificuldade pode ser ilustrada pela descrição do Npred "oferta" na instância representada pela árvore de dependência da Figura 5.16, cujo *roleset* correspondente é *offering.01*, que prevê os seguintes papéis: Arg0 (*entity offering*), Arg1 (*commodity*), Arg2 (*price*) e Arg3 (*beneficiary or entity offered to*). A descrição do Npred nessa instância, aliás, gerou três descrições diferentes. A Figura em questão ilustra os 2 argumentos identificados por um dos 3 especialistas. Nela, o *ticker* "oibr4" corresponde ao Arg1 e os valores da ação ("-2,4 -2,8") ao Arg2. As outras duas descrições indicaram a ocorrência de apenas 1 argumento. Em uma delas, o Arg0 corresponde a "oibr4" e, na outra, o Arg2 foi associado a "-2,4 -2,8".

A indicação de "oibr4" como Arg0 por um especialista e como Arg1 por outro resulta de diferentes interpretações dadas a esse tweet, que é relativamente fragmentado e com elipses. Um especialista entendeu que o *ticker* representa a empresa (ou caso, a Oi) que está fazendo a oferta (Arg0), o que é possível dado que essa é uma das funções dos *tickers* nos tweets. A indicação como Arg1 resulta da interpretação do *ticker* "oibr4" como o *commodity*, isto é, a mercadoria cujo preço é determinado pela oferta e procura. Ambas as interpretações, assim, são possíveis. O não reconhecimento da expressão "-2,4 -2,8" como Arg2 (*price*) parece resultar da fragmentação desse trecho do tweet e da falta de conhecimento sobre o domínio para reconhecê-la como os valores da oferta, mesmo que a árvore indique que se trata de um *nmod* de "oibr4".

Figura 5.16: Identificação dos Arg de "oferta" com base na anotação-UD.



Fonte: O autor, 2024.

6 Descrição sintático-semântica dos Npred

Com a anotação sintática via *deprels* e a identificação dos papéis semânticos nas 1.756 instâncias relativas aos 145 nomes predicadores inicialmente selecionados, fez-se a descrição da valência sintático-semântica desses nomes no DANTEStocks. Para tanto, desenvolveu-se um script Python, responsável por 3 tarefas específicas referentes ao mapeamento sintático-semântico, as quais estão descritas na sequência.

6.1 Etapas do mapeamento sintático-semântico

A primeira tarefa realizada pelo script foi identificar, nas árvores de dependência no formato CoNLL-U do córpus, a realização sintática dos argumentos descritos nos arquivos .xlsx, como ilustrado no Quadro 5.2. Assim, para cada um dos 145 nomes, o script extraiu e salvou em padrão JSON[1] uma série de informações a respeito de cada instância.

O arquivo (45), para o nome "compartilhamento" na instância (42a) (com sentido de *sharing.01*), ilustra essas informações, que são: (i) o identificador único do tweet (*sent_ID*), (ii) o tweet completo no qual o Npred ocorre (campo *Texto*), (iii) o próprio Npred (*rel*), (iv) as relações de dependência em que o Npred é *head* (isto é, 1 **case** e 1 **nmod**) e (v) as *deprel* nas quais o Npred é dependente (no caso, ele é dependente por **nmod** de "acordo" (*head*). Sobre (v), vê-se que, em (45), "acordo" é dependente por **obj** do verbo (*head*) "assina". Por se tratar de um verbo e **root** do tweet, extraiu-se o *token* que está em relação de **nsubj** com esse verbo, pois, com isso, herdou-se o sujeito do arranjo sintático da oração (isto é, "Gol") para a predicação do nome "compartilhamento".

[1] Um formato leve para troca e tratamento de dados que funciona como uma coleção ordenada de chave/valor. <https://ecma-international.org/publications-and-standards/standards/ecma-404/>

O exemplo (46), para o nome "acordo" na instância descrita no exemplo (1) (com sentido de *agreement.01*) (pág. 3), explicita, além das metainformações de *sent_ID*, *Texto* (do tweet) e o próprio Npred (*rel*), que as *deprel* em que o Npred "acordo" é *head* são 2 **nmod** e que esse Npred é dependente por **obj** de "assinar" (*head*). Da mesma forma que em (45), extraiu-se o *token* que está em relação de **nsubj** com esse verbo, pois, com isso, herda-se o sujeito oração (isto é, "Gol") para a predicação do nome.

(45)
```
{"sent_ID": "dante_01_4546077433921781771",
"Texto": "#BR #BOVESPA #GOLL4 Gol assina acordo de
compartilhamento de voos com TAP . http://t.co/wHGukBg7qp",
"rel": "compartilhamento",
"case": "de",
"nmod": "voos",
"nmod (head)": "acordo" --> "obj":"assina" --> "nsubj":"Gol",}
```

(46)
```
{"sent_ID":"dante_01_4546077433921781771",
"Texto": "#BR #BOVESPA #GOLL4 Gol assina acordo de compartilhamento
de voos com TAP . http://t.co/wHGukBg7qp",
"rel": "acordo",
"nmod":["compartilhamento", "TAP"],
"obj (head)": "assina" --> "nsubj":"Gol"}
```

Na sequência, a segunda tarefa do script foi verificar o arquivo .xlsx correspondente ao nome em questão (como o ilustrado pelo Quadro 5.2) e salvar a descrição da estrutura-A em cada uma das instâncias também no formato JSON, conforme ilustrado pelo exemplo (46). Esse processo de conversão do arquivo .xlsx com a descrição semântica para o formato JSON foi realizado com o intuito de facilitar a correlação das informações sintáticas e semânticas, descrita a seguir.

(47)
```
{"sent_ID": "dante_01_4546077433921781771",
"sharer, n=0": "Gol",
"thing shared, n=1": "voos",
"shared with, if separate from arg0, n=2": "--",
"Frame": "sharing.01"}
```

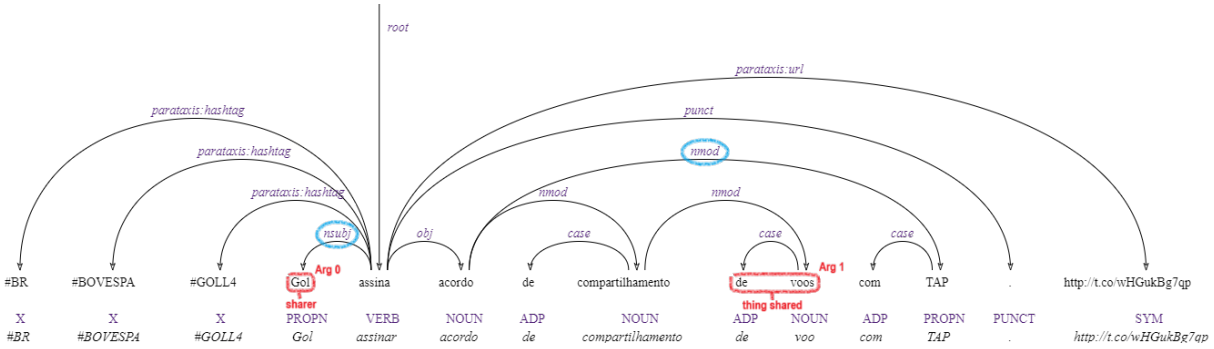
(48)
```
{"sent_ID": "dante_01_4546077433921781771",
"agreer, n=0": "Gol",
"proposition, n=1": "de compartilhamento de voos",
"other entity agreeing, n=2": "com TAP",
"Frame": "agreement.01"}
```

A terceira tarefa realizada pelo script foi a de, a partir do *sent_ID*, mapear as relações de dependência do modelo UD aos papéis semânticos dos *rolesets*, de modo a construir, para cada nome e suas instâncias, uma tabela de correlação, como a ilustrada

na Tabela 6.1 para Npred “acordo”. Mais especificamente, o script comparou a anotação de dependência e a descrição dos Arg, considerando que a palavra-alvo/nome é indicada por *rel*, e contabilizou como se deu a intersecção entre as *deprel* e os Arg.

Comparando as informações sintáticas de (45) e as semânticas de (47) a respeito do Npred “compartilhamento” (42a), por exemplo, foi possível mapear que (i) o Arg0=“Gol” (*sharer*) é dependente de “compartilhamento” pela *deprel* **nsubj**, herdada da predicação verbal (oracional) e (ii) o Arg1=“voos” (*thing shared*) é dependente por **nmod** do Npred em questão. Lembrando aqui que o Arg2 (*shared with, if separeted from Arg0*) não foi sintaticamente realizado na predicação de “compartilhamento”, o que é indicado por “–”. Esse mapeamento pode ser ilustrado com a instância (42a). A Figura 6.1 destaca os papéis semânticos do *frame “sharing.01”* “anotados” sobre a árvore de dependência da Figura 5.1.

Figura 6.1: Valência sintático-semântico de “compartilhamento” (*“sharing.01”*).

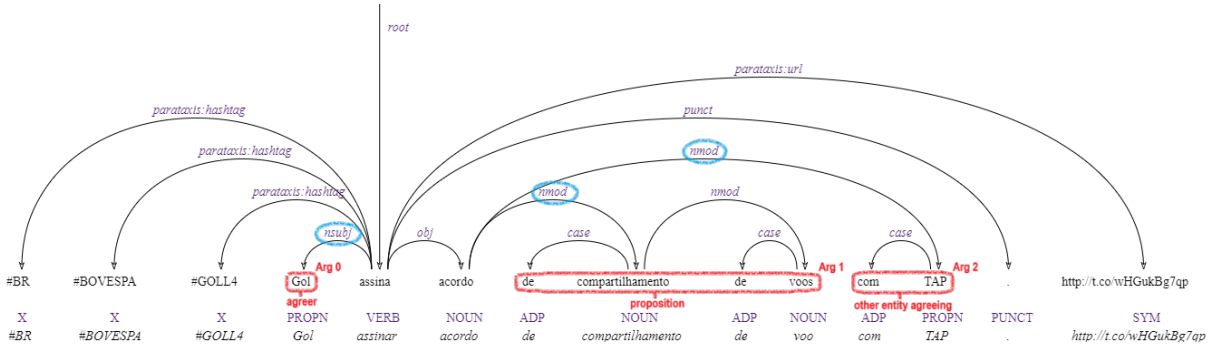


Fonte: O autor, 2024.

No que diz respeito ao mapeamento entre as informações de (46) e (48), tem-se que (i) o Arg0=“Gol” (*agreer*) é dependente de “acordo” pela *deprel* **nsubj**, herdada da predicação projetada pelo verbo **root** do tweet (“assinar”), (ii) o Arg1=“compartilhamento” (*proposition*) é dependente por **nmod** do Npred “acordo” e, por fim, (iii) o Arg2=“TAP” (*other entity agreeing*) é dependente por **nmod** de “acordo”. Esse mapeamento sintático-semântico pode ser ilustrado pela Figura 6.2. Nela, tem-se a instância de “acordo” presente em (1), cuja árvore de dependência é a mesma da Figura 5.1, “anotada” com os papéis semânticos do *frame “agreement.01”*.

O mapeamento aqui ilustrado permitiu investigar/descrever a valência sintático-semântica de cada um dos 145 Npred no córpus. Especificamente, essa descrição focou em verificar (i) se os Arg dos *rolesets* se concretizavam sintaticamente nas instâncias do DANTEStocks e (ii) de que forma essa materialização sintática ocorre por meio das *deprel*.

Figura 6.2: Valência sintático-semântico de "acordo" (*"agreement.01"*).



Fonte: O autor, 2024.

Ressalta-se que tal investigação (assim como as observações decorrentes dela) é totalmente dependente das instâncias dos nomes predicadores no DANTEStocks, pois se buscou verificar exatamente como a estrutura-A dos nomes ocorre nesse tipo de CGU/domínio.

Sobre "acordo", por exemplo, a investigação revelou as características presentes na Tabela 6.1, que sistematiza a valência sintático-semântica desse Npred nas 5 instâncias em que ocorre no córpus somente com o sentido de *"agreement.01"*. Na figura, **nmod** está especificada pelas preposições que introduzem os complementos nominais conectados por essa *deprel*. Segundo os dados, o Arg0 se conecta ao Npred por **nsubj** e, como mencionado, esse argumento é herdado da predição verbal do **root**. Ademais, os Arg1 e Arg2 são dependentes do Npred por **nmod**, sendo, como esperado, introduzidos por preposições. No caso, Arg1 é introduzido pelas preposições "de" ou "para" e Arg2 por "com".

Tabela 6.1: Descrição da estrutura-A de "acordo" (*"agreement.01"*) no DANTEStocks.

	Arg 0 **agreer**	**Arg 1** **proposition**	**Arg 2** **other entity agreeing**
nmod:—|de/para|com	0	2	2
nsubj	4	0	0

Na Tabela 6.1, exibem-se os dados resultantes da descrição da valência sintático-semântica do Npred "acordo" em todas as 9 instâncias do córpus, que foram descritas com base em 3 *frames/rolesets* diferentes. Além das 3 instâncias descritas pelo *frame* "agreement", observa-se na tabela que a única ocorrência de "acordo" com o sentido *"accord.01"* (cf. Quadro 5.1) tem somente o Arg1 preenchido na estrutura sintática do SN/tweet. No caso, trata-se de um ADJ com função de **amod** ("acordos comerciais"). As ocorrências nas instâncias descritas pelo *frame "accordance.01"*, por sua vez, dizem

respeito à expressão “de acordo com”, o que justifica o Arg2 ser dependente do Npred por **nmod** excluvisamente introduzido pela ADP “com”.

Tabela 6.2: Descrição da estrutura-A de “acordo” segundo os 3 *frames/rolesets* distintos.

	Arg0	**Arg1**	**Arg2**	
	agreer	***proposition***	***other entity agreeing***	***Frame***
nmod:—|de/para|com	0	2	2	*agreement.01*
nsubj	4	0	0	
amod	0	1	0	*accord.01*
nmod:—|—|com	0	3	3	*accordance.01*

Realizando o mapeamento sintático-semântico das propriedades dos 145 Npred iniciais nas 1.756 instâncias desses Npred no DANTEStocks, foi possível verificar os aspectos (i) presença/ausência dos Arg definidos nos *rolesets* na estrutura sintática do SN/tweet e (ii) configuração sintática dos Arg via *deprel* no total de nomes investigados.

6.2 Resultados do mapeamento sintático-semântico

Sobre a presença/ausência dos Arg nas instâncias dos 145 Npred, os resultados são os descritos na Tabela 6.3, que sistematiza especialmente a quantidade de Arg sintaticamente realizados dos previstos nos *rolesets*. base nessa tabela, observa-se que as estruturas de argumento dos Npred cujos *rolesets* previram apenas 1 Arg são as únicas realizadas por completa em nível sintático. Especificamente, os 3 Npred para os quais os *rolesets* definiram apenas 1 Arg são “meio” (Arg1=*“whole”*). “membro” (Arg1=*“whole”*) e “deliberação” (Arg0=*“thinker”*). Nas 12 instâncias em que ocorrem, o único argumento previsto se realiza sintaticamente (p.ex.: “Membro $_{\text{Arg1}}$[do Conselho De Adim.] e “Deliberacoes $_{\text{Arg0}}$[Da Ago]”).

Ademais, constata-se que a estrutura-A projetada por todos os demais Npred apresenta algum Arg ausente na sintaxe. Tal observação sobre os Npred de valência (V) maior que 1 ($V_{>1}$), no entanto, não é surpresa, pois essa característica da valência nominal consta dos trabalhos revisados da literatura. No DANTEStocks, a maioria dos Npred analisados (61 de 145) possuem, segundo os *rolesets*, valência três (V_3), sendo que a maioria

das instâncias apresenta apenas 2 argumentos (isto é, 382 das 794). Sobre os Npred de V_4, que são o segundo tipo mais frequente (41 de 145), observa-se que os 41 nomes preenchem no máximo 3 Arg, sendo que a maioria das instâncias distribui-se entre 1 Arg (109 de 282) e 2 Arg (107 de 282). Os nomes de (V_2), que são o terceiro tipo mais frequente na amostra (27 de 145), preenchem a estrutura-A com apenas 1 Arg para a maioria de suas instâncias (188 de um total de 255). Por fim, os 15 Npred de V_5 preenchem apenas com 1, 2 ou 3 argumentos, pois, das 413 instâncias, 179 possuem apenas 1 Arg, 39 possuem 2 Arg e 9 possuem 3 Arg.

Tabela 6.3: Estatística sobre a quantidade de argumentos nas instâncias.

Qt. Npred	**Qt. Instância**	**Qt. Arg previsto**	**Qt. Arg realizado**						**Exemplo Npred**
			0	**1**	**2**	**3**	**4**	**5**	
15	413	5	186	179	39	9	0	0	comprador
41	282	4	38	109	107	28	0	-	conversa
61	794	3	265	381	132	16	-	-	acordo
27	255	2	28	188	39	-	-	-	projeto
3	12	1	-	12	-	-	-	-	membro

Ainda sobre a presença/ausência dos Arg, os dados da Tabela 6.4 especificam a ocorrência dos Arg em função de seus tipos (isto é, Arg0, Arg1, Arg2, Arg3 e Arg4).

Tabela 6.4: Estatística sobre a ocorrência dos diferentes tipos de argumentos.

Nomes	Instâncias	Qt. Arg Previstos	Realização Sintática					
			Nenhuma	Arg0	Arg1	Arg2	Arg3	Arg4
15	413	5	186	37	185	10	46	6
41	282	4	38	133	181	53	35	-
61	794	3	265	116	434	103	-	-
27	255	2	28	88	185	2	4	-
3	12	1	-	5	7	-	-	-

A respeito do preenchimento sintático dos Arg, obtiveram-se os resultados exibidos na Tabela 6.5. Nela, no entanto, as *deprel* não foram especificadas pelas preposições que introduzem os complementos dos nomes por questão de brevidade. Além disso, essa tabela sistematiza somente os Arg que figuram como dependentes das *deprel* que os conectam ao Npred. Embora o foco aqui repouse sobre os Arg dependentes dos Npred, destaca-se que em 33 instâncias (do total de 1.756) o Npred sob análise foi anotado com a *deprel* **root**, sendo, portanto, a informação central do tweet. Ademais, em decorrência da diretriz apresentada na Seção 5.2 e exemplificada nas Figuras 5.5, 5.6 e 5.7, em 36 casos o próprio

Npred foi descrito como Arg, sendo (i) 11 ocorrências Arg0, (ii) 20 como Arg1 e (iii) 5 como Arg2.

Tabela 6.5: Realização sintática dos Arg por *deprel*.

	Arg0	**Arg1**	**Arg2**	**Arg3**	**Arg4**
acl	6	29	3	3	–
acl:relcl	3	3	4	–	–
advcl	–	2	3	–	–
amod	3	58	18	–	–
advmod	1	1	2	2	–
conj:strunc	–	1	–	–	–
nmod	134	580	75	62	3
nmod:hashtag	–	155	–	–	–
nmod:strunc	–	9	–	–	–
nmod:wtrunc	–	4	–	–	–
nsubj	72	15	6	3	–
nsubj:outer	–	2	–	–	–
obj	5	6	1	–	–
obl	5	11	7	5	–
obl:strunc	–	–	1	–	–
vocative	2	–	–	–	–

Sobre os dados da Tabela 6.5, observam-se algumas regularidades.

A primeira delas reside no fato de que Arg0, Arg1, Arg2 e Arg3 são majoritariamente dependentes do Npred por **nmod**, ao passo Arg4 é unicamente dependente via **nmod** em suas 3 ocorrências. Em ambos os casos, os argumentos são preposicionados. Particularmente, as 155 ocorrências de Arg1 como **nmod:hashtag** dizem respeito exclusivamente ao Npred "indicação" em instâncias como a ilustrada na Figura 6.4, em que o Arg1 se realiza como um SPrep cujo núcleo é um *ticker* precedido pelo símbolo de *hashtag*.

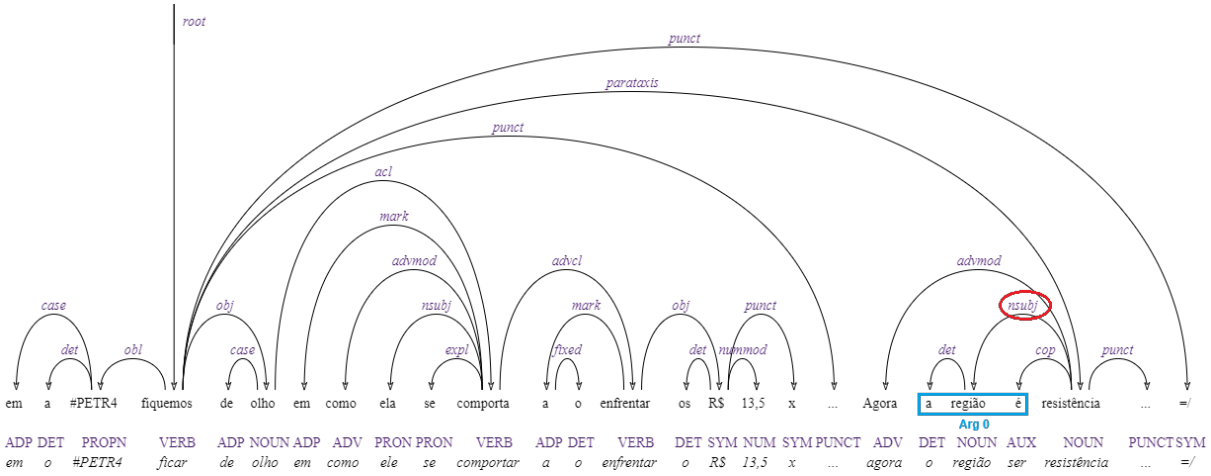
A segunda regularidade diz respeito à herança dos argumentos da predicação verbal para o preenchimento da valência nominal. Dos 72 Arg0 conectados ao Npred por **nsubj**, 71 foram herdados da predicação verbal, como ilustrado em (45), e 1 foi identificado no interior da predicação nominal (cf. Figura 6.3.) Tal como **nsubj**, todos os 12 casos de **obj** e os 28 de **obl** foram herdados da predicação verbal.

A terceira regularidade se refere ao fato de que, depois de **nmod**, **amod** é a segunda *deprel* mais frequente para Arg1 e Arg2 (considerando **nmod:hashtag** como uma subcategoria de **nmod**), uma vez que esses Arg são comumente adjetivos (p.ex.: "acordos/NOUN comerciais/ADJ" (Arg1)).

Ademais, a quarta regularidade observada na Tabela 6.5 é a de que a *deprel* **acl** (oração adjetival), com 29 ocorrências, é a terceira mais frequente como Arg1.

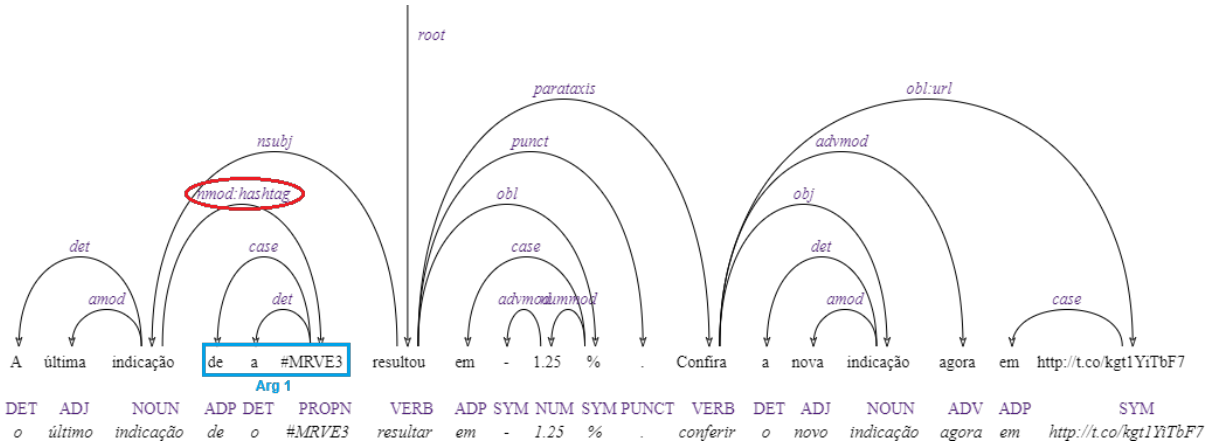
Por fim, os 2 casos de **vocative** são relativos às instâncias únicas dos Npred "olho" e "olhada", nas quais eles ocorrem em expressões com Vsup (sublinhado) ("ficar de olho" e "dá só uma olhada", respectivamente). Nas Figuras 6.5 e 6.6, observa-se que o Arg0 de ambos (com os papéis semânticos de *observer* e *looker*) ocorrem como **vocative**.

Figura 6.3: Exemplo de ocorrência de Arg0=**nsubj** na predicação nominal.



Fonte: O autor, 2024.

Figura 6.4: Exemplo de ocorrência de Arg1 como **nmod:hashtag** de "indicação".

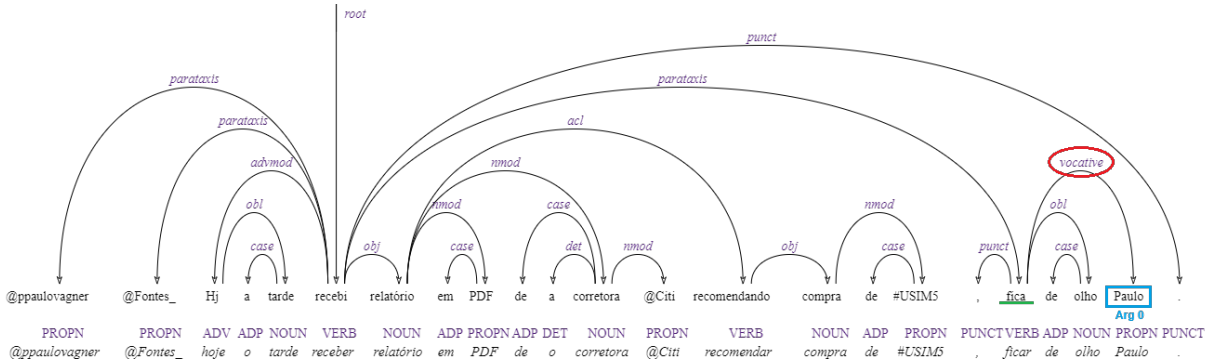


Fonte: O autor, 2024.

Outros aspectos da Tabela 6.5 parecem se relacionar aos fenômenos linguísticos dos tweets/domínio. Um deles é sobre os casos de Arg1 conectados ao Npred por **nmod:strunc**. Tais casos se referem a nomes como "indicadores" e "transporte".

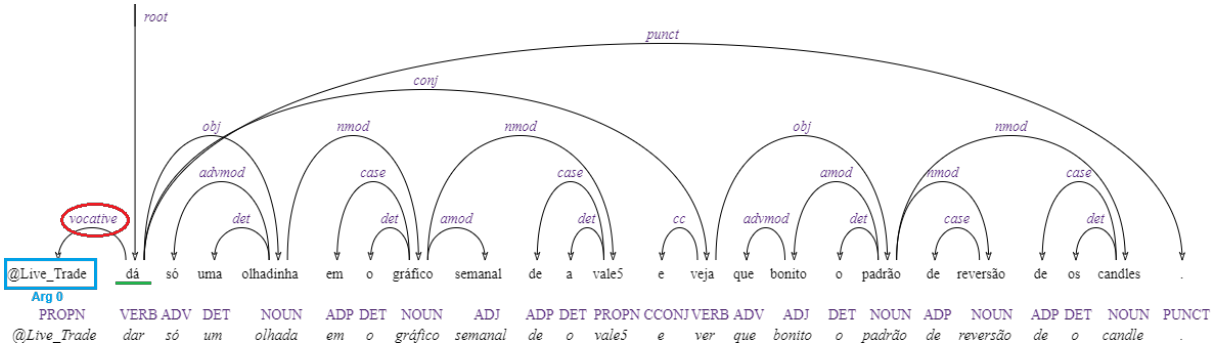
A instância de "indicadores" está descrita em (49) e sua respectiva árvore de dependência está na Figura 6.7. O sentido de "indicadores" foi descrito pelo *frame*

Figura 6.5: Exemplo de Arg0 como *vocative* de "olho" em expressão com Vsup.



Fonte: O autor, 2024.

Figura 6.6: Exemplo de Arg0 como *vocative* de "olhada" em expressão com Vsup.



Fonte: O autor, 2024.

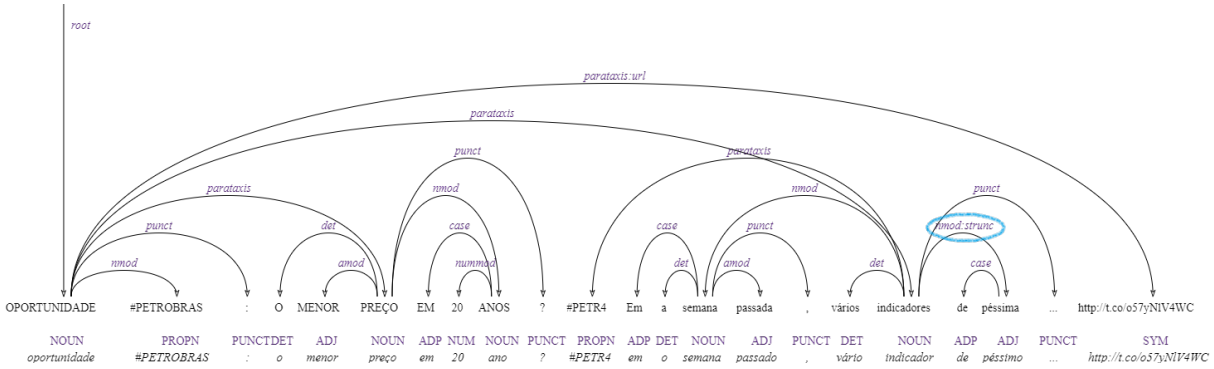
"indication.01", que prevê os seguintes Arg: Arg0 (*"indicator"*), Arg1 (*"thing indicated*) e Arg2 (*"indicated to"*). O Arg1, em particular, que parece ser um sintagma preposicionado composto por DET+ADJ+NOUN, está truncado, omitindo a ocorrência de NOUN. Por isso, o *token* ligado ao Npred como Arg1 é ADJ. Por ser o último *token* antes do truncamento (indicado pelas reticências) e por fazer parte de uma estrutura sintagmática quebrada, a dependência entre o ADJ e o Npred é rotulada por **nmod:strunc**, seguindo as diretrizes de anotação sintática de *deprel* do Apêndice A. Isso quer dizer que, embora exista um Arg1 conectado ao Npred, ele está truncado.

(49) OPORTUNIDADE #PETROBRAS : O MENOR PREÇO EM 20 ANOS ? #PETR4

Em a semana passada , vários **indicadores** de péssima ... http://t.co/o57yNlV4WC

A instância de "transporte" é apresentada em (50) e sua respectiva árvore de dependência está na Figura 6.8. O sentido de "transporte" foi descrito pelo *roleset* composto por 4 Arg: Arg0 (*"causer of motion"*), Arg1 (*"thing in motion"*), Arg2 (*"destination"*), Arg3 (*"source"*). O Arg1, que parece ser um sintagma preposicionado composto por

Figura 6.7: Exemplo da estrutura-A de “indicadores” com truncamento.

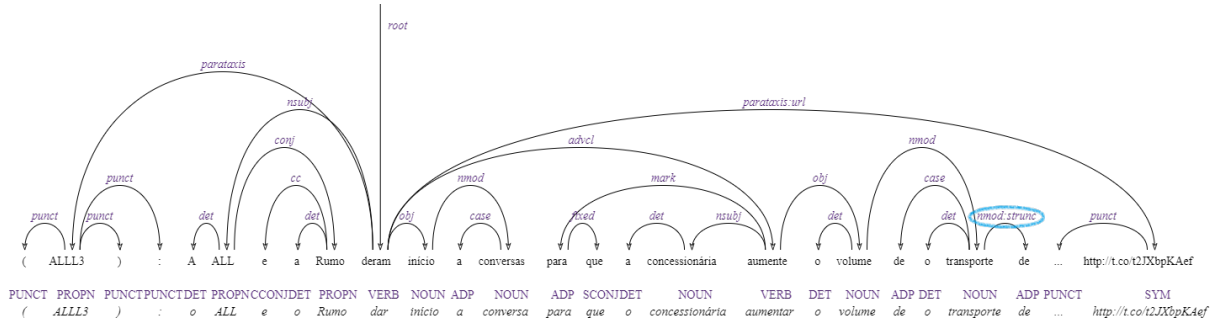


Fonte: O autor, 2024.

ADP+NOUN, está truncado com omissão de NOUN. Consequentemente, o *token* ligado ao Npred como Arg1 é ADP e este é dependente do Npred por **nmod:strunc**. Esse é mais um caso em que, embora exista um Arg1 conectado ao Npred, o truncamento impediu a ocorrência do nome nuclear desse argumento no tweet.

(50) (ALLL3) : A ALL e a Rumo deram início a conversas para que a concessionária aumente o volume de o transporte de ... http://t.co/t2JXbpKAef

Figura 6.8: Exemplo de estrutura-A de “transporte” com truncamento.



Fonte: O autor, 2024.

Além desses 2 casos expostos e mais 13 que aparecem explicitados nas *deprel*, seja com **nmod:wtrunc**, **obl:strunc** ou **conj:strunc**, há mais 9 truncamentos que interferem na realização da estrutura-A dos nomes. Neles, o Arg é truncado por completo, resultando na ausência de todos os Arg. Isso quer dizer que, das 517 instâncias sem a ocorrência de Arg da Tabela 6.4, 9 delas foram causados por truncamento.

Por fim, vale ressaltar que, ao total, foram identificadas apenas 35 instâncias com Npred acompanhados de Vsup.

6.3 Caracterização semântica dos Npred

Com base nos dados que constituem o Apêndice B, fez-se uma caracterização semântica dos 145 Npred que ocorrem no córpus DANTEStocks segundo as classes e tipos do NomLex-Plus. Das 16 classes originais[2], os Npred do DANTEStocks estão distribuídos nas seguintes: NOM, NOMLIKE, NOMING, PARTITIVE, NOMADJ, ABILITY, WORK-OF-ART, GROUP e ENVIRONMENT (Tabela 6.6). Quanto aos tipos (Tabela 6.7), os Npred do DANTEStocks pertencem aos seguintes: VERB-NOM, NOM-REL, SUBJECT, ADJ-NOM, OBJECT, P-OBJ-PART e VERB-PART. A distribuição dos Npred entre classes e tipos está quantificada na Tabela 6.8.

Tabela 6.6: Quantificação de Npred por "classe" do Nomlex-Plus.

Classe	Qt. Npred
NOM	116
NOMLIKE	13
NOMING	6
PARTITIVE	5
NOMADJ	2
ABILITY	2
WORK-OF-ART	1
GROUP	1
ENVIRONMENT	1

Tabela 6.7: Quantidade de Npred por "tipo" do Nomlex-Plus.

Nomlex types	Qt. Npred
VERB-NOM	121
NOM-REL	11
SUBJECT	9
ADJ-NOM	2
OBJECT	2
P-OBJ-PART	1
VERB-PART	1

As classes NOM e NOMING, com 116 e 6 ocorrências, respectivamente, agrupam a maior quantidade de Npred do córpus. Ambas incluem substantivos que, sendo nominalizações de verbos, compartilham a estrutura de argumentos da forma verbal correspondente. A diferença entre elas é a de que o rótulo NOMING é usado para indicar os nomes do inglês terminados em *ing*.

[2]Para definições detalhadas de cada uma das classes e dos tipos, consultar Meyers (2007).

Tabela 6.8: Quantificação de Npred por Classe e Tipo

Classe	VERB-NOM	ADJ-NOM	NOM-REL	OBJECT	SUBJECT	VERB-PART	P-OBJ-PART
ABILITY	0	0	2	0	0	0	0
ENVIRONMENT	0	0	1	0	0	0	0
GROUP	0	0	1	0	0	0	0
NOM	108	0	0	2	8	1	1
NOMADJ	0	2	0	0	0	0	0
NOMING	6	0	0	0	0	0	0
NOMLIKE	13	0	0	1	0	0	0
PARTITIVE	0	0	5	0	0	0	0
WORK-OF-ART	0	0	1	0	0	0	0

O destaque das classes NOM e NOMING no córpus DANTEStocks corrobora a afirmação de Voskaki, Tziafa e Annidou (2016), segundo a qual UGC do mercado financeiro é caracterizado pela prevalência de Npred frente às formas verbais relacionadas.

Com 13 ocorrências, a classe NOMLIKE inclui substantivos que, embora não sejam derivados diretamente de verbos, compartilham características semânticas e estruturais com nominalizações verbais. Esses substantivos comportam-se de maneira semelhante às nominalizações, exigindo complementos.

A classe PARTITIVE engloba 5 Npred que expressam parte/fração de um todo, como “divisão” (“seção”) e “monte” (“quantidade”). No DANTEStocks, a esses Npred descrevem partes de investimentos, segmentos de mercado e distribuições de ativos.

Com 2 Npred, NOMADJ compreende nominalizações derivadas de adjetivos, que expressam qualidade ou estado, como “coragem” e “responsabilidade”. A ocorrência de Npred do tipo NOMADJ indica usos específicos voltados à descrição de características intrínsecas de condições que afetam seus atores.

A classe ABILITY engloba 2 Npred, “projeto” e “caminho”, os quais denotam planos ou estratégias subjacentes a decisões de mercado.

As classes WORK-OF-ART (“criações humanas”), GROUP (“grupo ou coleção”), e ENVIRONMENT (“cenário ou contexto”) englobam somente um Npred em cada uma delas, a saber: “notícia”, “diretoria”, e “hora”, respectivamente.

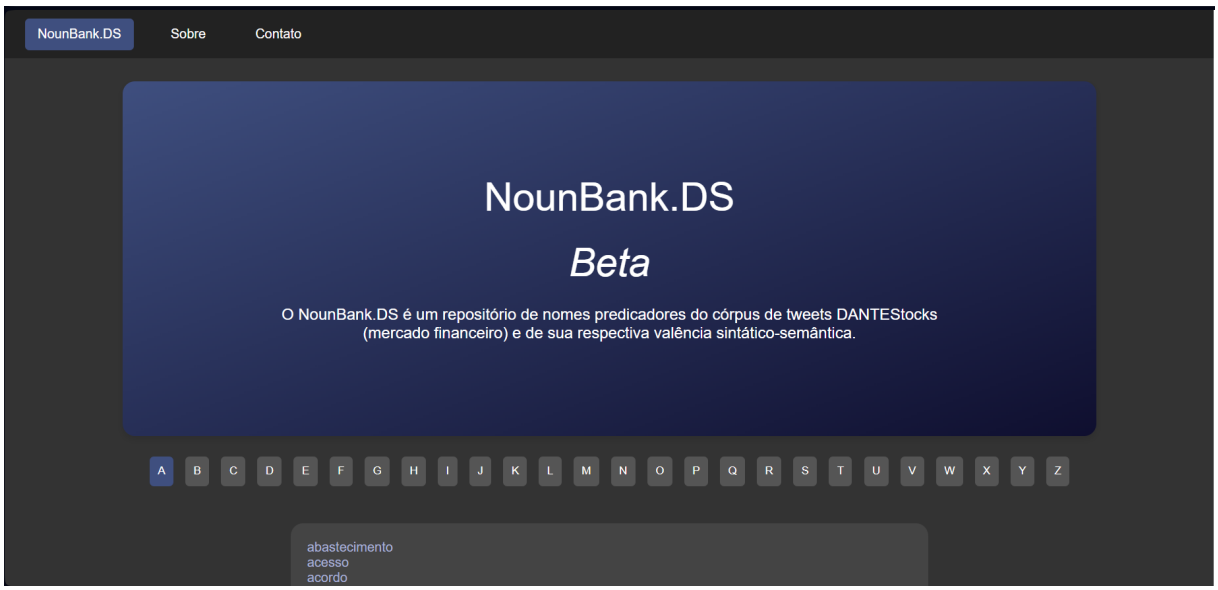
A baixa frequência dessas 5 últimas classes (NOMADJ, ABILITY, WORK-OF-ART, GROUP e ENVIRONMENT) no córpus DANTEStocks é esperada, dado que o foco principal dos tweets financeiros não é em obras de arte, criações culturais, contextos ambientais, entidades coletivas ou títulos de trabalho, mas sim em ações, estados e características financeiras.

6.4 Criação de um repositório lexical online

A partir das descrições sintático-semântica dos 145 Npred distribuídos nas 1.756 instâncias, construiu-se um repositório lexical dos Npred do córpus DANTEStocks nos moldes do Verbo-Brasil. No entanto, ao contrário do Verbo-Brasil, esse repositório não objetiva apoiar a tarefa de anotação de papéis semânticos, no caso, em tweets. Seu objetivo primeiro é o de organizar e disponibilizar as descrições sobre a valência sintático-semântico dos Npred no referido córpus. Naturalmente, por estar disponibilizado online, ele pode ser utilizado como recurso para o PLN e para estudos linguísticos sobre tweets.

O repositório em questão também se difere do Verbo-Brasil por disponibilizar as diferentes realizações sintáticas da estrutura-A dos Npred no DANTEStocks. Nesse sentido, esse recurso se assemelha ao *"FrameNet Index of Lexical Units"*[3] do projeto FrameNet (Baker; Fillmore; Lowe, 1998). Denominado de NounBank.DS ("Banco de Nomes do **D**ANTE**S**tocks), o recurso está em sua primeira versão, contendo os 145 Npred apresentados no presente trabalho. A Figura 6.9 ilustra a interface principal do NounBank.DS[4], disponível na web. Nela, tem-se a lista dos Npred, de modo que cada nome da lista é hiperlinkado a uma página dedicada, como a ilustrada pela Figura 6.10 para o Npred "compartilhamento".

Figura 6.9: Interface principal do NounBank.DS (v.*Beta*).



Fonte: O autor, 2024.

[3] <https://framenet.icsi.berkeley.edu/luIndex>
[4] <https://bryankhelven.github.io/NounBank.DS/>

Figura 6.10: Página do nome "compartilhamento" no NounBank.DS (v.*Beta*).

Fonte: O autor, 2024.

Cada página dedicada disponibiliza as seguintes informações sobre o Npred: (i) o sentido do nome no córpus, numerado de 01 a 99 como no Verbo-Brasil; (ii) *frame* do NomBank que representa o sentido do nome no córpus, (iii) o conjunto dos papéis semânticos do *roleset*, nos moldes do PropBank e NomBank, (vi) instâncias do DANTEStocks com as papéis semânticos "anotados", (v) as diferentes realizações sintáticas dos argumentos nas instâncias do córpus via *deprel* do modelo UD e (vi) a estatística sobre as diferentes realizações sintáticas da estrutura-A, explicitando, assim, a frequência de ocorrência dos padrões sintático-semânticos nos tweets do domínio do mercado financeiro.

Na tela dedicada ao nome predicador "compartilhamento", por exemplo, ilustrada pela Figura 6.10, é possível observar que, de acordo com o NounBank.DS, esse nome possui apenas um sentido no córpus DANTEStocks, isto é, "compartilhamento.01". Esse sentido foi mapeado para o sentido *"sharing.01"* do NomBank, que, por sua vez, está relacionado ao sentido *"share.01"* do verbo *"share"* no PropBank.

Logo abaixo, tem-se o conjunto de papéis semânticos (e suas respectivas glosas) definidos para *"sharing.01"*, que são: (i) Arg0 (*"sharer"*), Arg1 (*"thing shared"*) e Arg2 (*"shared with, if separeted from Arg0"*). Na sequência, ao menos uma instância do DANTEStocks ilustra a ocorrência dos argumentos. Para facilitar a visualização da ocorrência dos argumentos e de seus papéis semânticos, cada um deles está descrito por uma cor diferente. Nos dois exemplos de ocorrência de "compartilhamento", os Arg0 e Arg1 estão preenchidos sintaticamente, ao passo que Arg2 não está. Para "compartilhamento", tem-se então uma proposição descrita nos moldes do NomBank, uma vez que cada *feature* do subconjunto composto por REL, ARG0, ARG1 e ARG2 está associada à sua realização sintática na instância-exemplo, proveniente da árvore UD.

Nos moldes do *"Index of Lexical Units"* do projeto FrameNet, o NounBank.DS exibe as diferentes realizações sintáticas da estrutura de argumentos do nome no córpus de tweets. Essas realizações são explicitadas por meio das relações de dependência ou *deprel* do modelo *Universal Dependencies*, advindas da anotação sintática semiautomática do DANTEStocks. O mapeamento sintaxe-semântica, aliás, é uma característica distintiva do repositório, oferecendo uma visão sobre como os argumentos se materializam na estrutura sintática. Para explicitar as *deprel*, emprega-se o mesmo esquema de cores utilizado na descrição dos Arg previstos no *roleset*. Ao final, a página dedicada ao nome "compartilhamento", exibe, por meio de uma tabela, a frequência de ocorrência da realização sintática dos Arg (via *deprel*) nas instâncias do DANTEStocks. No caso de "compartilhamento", esse nome ocorre em duas instâncias, sendo que em ambas o Arg0 é dependente por **nsubj** e Arg1 é dependente do Npred por **nmod**.

7 Considerações finais e trabalhos futuros

A tarefa de descrever detalhadamente a estrutura-A dos nomes valenciais em tweets do mercado financeiro em português foi o primeiro objetivo traçado. Sobre ele, acredita-se que os seguintes fatores levaram ao seu cumprimento, a saber: (i) identificação manual de 145 Npred distintos com validação em dicionário de valências, (ii) identificação manual de 1.756 instâncias relativas aos Npred, as quais correspondem a 1.218 tweets distintos (isto é, 30% do córpus DANTEStocks), (iii) descrição da valência sintático-semântico (ou estrutura-A) pautada em uma anotação sintática cujo índice de avaliação foi um *Kappa* de Cohen de mais de 95% e em uma descrição semântica com *Kappa* de Fleiss de quase 0.80, os quais indicam a consistência das tarefas.

Sobre o segundo objetivo, que era o de identificar e analisar a presença/ausência dos Arg previstos pela semântica do nome, bem como a sua realização sintática, os resultados indicam que: (i) a estrutura-A dos Npred de V1 é sempre preenchida, (ii) a estrutura-A dos Npred de V>1 apresenta algum Arg ausente, (iii) a maioria dos Npred analisados é de V3, com apenas 1 Arg na maioria das instâncias, e (iv) as *deprel* que mais frequentemente conectam os Npred a seus Arg são nmod (seguido ou não de sub-relação) e amod, e (v) em 24 instâncias, os Npred tiveram sua estrutura-A afetada por fenômenos dos tweets como truncamento de elementos.

Assim, pode-se dizer que parece haver uma relação entre a estrutura-A reduzida dos Npred e a brevidade e fragmentação dos tweets, pois, ao final, das 1.756 instâncias descritas, 517 não manifestaram nenhum Arg previsto nos *rolesets* de seus *frames* correspondentes, o que representa cerca de 30% (29,44%) das instâncias descritas e 12,77% do córpus DANTEStocks. Acredita-se que essas observações avançam no entendimento do impacto das características linguísticas específicas dos tweets na estrutura-A dos nomes.

Por fim, sobre contribuir para a construção de recursos linguístico-computacionais que pudessem auxiliar no processamento de tweets e outros tipos de CGU relacionados ao mercado financeiro, que era outro objetivo, destaca-se que, como resultado do trabalho, os pesquisadores do PLN para a língua portuguesa já contam com o primeiro córpus de tweets com anotação sintática (revisada manualmente) (ou *tweebank*), além de um manual com diretrizes exemplificadas para a anotação de tweets, em especial, do mercado financeiro (Apêndice 1). Assim, pode-se dizer que a inserção de uma camada de anotação sintática no DANTEStocks é uma contribuição significativa deste trabalho. Mesmo que o modelo Stanza treinado com tweets ainda não esteja disponível ao público, ele também é outra contribuição ao PLN, sendo o primeiro para o gênero em português.

Quanto às dificuldades durante o trabalho, a tarefa de anotação sintática dos 4.048 tweets do DANTEStocks foi, de longe, a mais complexa e demorada, dado o desafio que esse tipo de CGU impôs para a anotação segundo modelo UD. Como não havia diretrizes de anotação-UD para CGU em português, a anotação em questão foi pioneira, envolvendo discussões frequentes entre vários especialistas em UD para decidir a melhor maneira de lidar com alguns fenômenos dos tweets do mercado financeiro.

Especificamente sobre a identificação dos sentidos dos nomes no córpus e conseguinte descrição de sua estrutura-A via *rolesets*, destaca-se que essa também foi uma etapa trabalhosa, posto que englobou a identificação manual das classes dos nomes, assim como a definição de diretrizes para a sua descrição semântica. Ademais, salienta-se que, para todos os sentidos dos nomes descritos, foi possível identificar um *roleset* do NomBank correspondente. Por fim, a descrição das 1.756 instâncias segundo a metodologia do NomBank poderá ser usada como base para inserir efetivamente uma camada de informação predicado-argumento ao DANTEStocks, a qual enriquecerá o ainda mais esse córpus enquanto *lingware*.

Como trabalhos futuros, projetam-se as possíveis tarefas de: (i) descrição/anotação dos ArgM (modificadores) nas 1.756 instâncias, enriquecendo a caracterização da estrutura-A dos Npred, (ii) transferência das descrições ora em planilhas para uma ferramenta ou arquivo que permita inseri-las como uma efetiva camada de anotação semântica ao córpus DANTEStocks, (iii) descrição da estrutura-A dos Npred do DANTEStocks em um córpus de língua geral, comparando a ocorrência dos Arg semânticos e as realizações sintáticas desses Arg, (iv) avaliação do modelo de *parsing* treinado no DANTEStocks em

outro córpus de CGU, preferencialmente composto por tweets que não sejam do mercado financeiro, a fim de verificar sua robustez e capacidade de generalização e (v) investigação do impacto que anotações sintáticas e semânticas podem ter no desempenho de LLMs, explorando como essas camadas adicionais de informação podem auxiliar na melhoria da compreensão e geração de texto nesses modelos.

Referências Bibliográficas

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: A treebank for Portuguese. In: GONZÁLEZ RODRÍGUEZ, M.; SUAREZ ARAUJO, C. P. (Ed.). **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**. Las Palmas, Canary Islands - Spain: European Language Resources Association (ELRA), 2002. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2002/pdf/1.pdf>.

ALONSO, H. M.; SEDDAH, D.; SAGOT, B. From noisy questions to Minecraft texts: Annotation challenges in extreme syntax scenario. In: HAN, B.; RITTER, A.; DERCZYNSKI, L.; XU, W.; BALDWIN, T. (Ed.). **Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)**. Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 13–23. Disponível em: <https://aclanthology.org/W16-3905>.

ALVES, C. F. **Tratamento lexicográfico da sintaxe no Dicionário de Usos do Português do Brasil: subsídios da gramática de valências para a descrição do português brasileiro**. 2014.

BAKER, C. F.; FILLMORE, C. J.; LOWE, J. B. The Berkeley FrameNet project. In: **36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1**. Montreal, Quebec, Canada: Association for Computational Linguistics, 1998. p. 86–90. Disponível em: <https://aclanthology.org/P98-1013>.

BANARESCU, L.; BONIAL, C.; CAI, S.; GEORGESCU, M.; GRIFFITT, K.; HERMJAKOB, U.; KNIGHT, K.; KOEHN, P.; PALMER, M.; SCHNEIDER, N. Abstract meaning representation for sembanking. **Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse**, p. 178–186, 01 2013.

BARROS, C. D. d. **Descrição e classificação de predicados nominais com o verbo-suporte fazer no Português do Brasil**. Tese (Doutorado) — Universidade Federal de São Carlos, 2014.

BICK, E. **The parsing system Palavras: automatic grammatical analysis of Portuguese in a constraint grammar framework**. [S.l.]: Aarhus University Press, 2000.

BOLLEN, J.; MAO, H.; ZENG, X. Twitter mood predicts the stock market. **Journal of Computational Science**, Elsevier BV, v. 2, n. 1, p. 1–8, mar. 2011. ISSN 1877-7503. Disponível em: <http://dx.doi.org/10.1016/j.jocs.2010.12.007>.

DE BONA, C. Propriedades valenciais de nomes deverbais: uma reanálise de dados do projeto nurc com base na linguística textual e no estudo dos anafóricos. **Cadernos de Letras da UFF**, v. 49, p. 219–238, 2014.

BORBA, F. **Uma Gramática de Valências para o Português**. São Paulo: Ed. Ática, 1996.

BORBA, F. d. S. **Dicionário de usos do português do Brasil**. [S.l.: s.n.], 2002.

BRUCKSCHEN, M.; MUNIZ, F.; SOUZA, J. G. d.; FUCHS, J. T.; INFANTE, K.; MUNIZ, M.; GONçALVES, P. N.; VIEIRA, R.; ALUíSIO, S. **Anotação Linguística em XML do Corpus PLN-BR**.

BURCHARDT, A.; ERK, K.; FRANK, A.; KOWALSKI, A.; PADO, S. SALTO - a versatile multi-level annotation tool. In: CALZOLARI, N.; CHOUKRI, K.; GANGEMI, A.; MAEGAARD, B.; MARIANI, J.; ODIJK, J.; TAPIAS, D. (Ed.). **Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)**. Genoa, Italy: European Language Resources Association (ELRA), 2006. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2006/pdf/341_pdf.pdf>.

CABEZUDO, M. A.; MAZIERO SOBREVILLA, E. G.; SOUZA, J. W. d. C.; DIAS, M. D. S.; CARDOSO, P. C. F.; BALAGE FILHO, P. P.; AGOSTINI, V.; NóBREGA, F. A. A.; DE BARROS, C. D.; DI FELIPPO, A.; PARDO, T. A. S. Anotação de sentidos de verbos em textos jornalísticos do corpus cstnews. **REVISTA DE ESTUDOS DA LINGUAGEM**, Faculdade de Letras da UFMG, v. 23, n. 3, p. 797, dez. 2015. ISSN 0104-0588. Disponível em: <http://dx.doi.org/10.17851/2237-2083.23.3.797-832>.

CARLETTA, J. Assessing agreement on classification tasks: The kappa statistic. **Computational Linguistics**, MIT Press, Cambridge, MA, v. 22, n. 2, p. 249–254, 1996. Disponível em: <https://aclanthology.org/J96-2004>.

CAROSIA, A. E. O.; COELHO, G. P.; SILVA, A. E. A. Analyzing the brazilian financial market through portuguese sentiment analysis in social media. **Applied Artificial Intelligence**, Informa UK Limited, v. 34, n. 1, p. 1–19, out. 2019. ISSN 1087-6545. Disponível em: <http://dx.doi.org/10.1080/08839514.2019.1673037>.

CHOI, J.; BONIAL, C.; PALMER, M. Multilingual Propbank annotation tools: Cornerstone and jubilee. In: ROSÉ, C. P. (Ed.). **Proceedings of the NAACL HLT 2010 Demonstration Session**. Los Angeles, California: Association for Computational Linguistics, 2010. p. 13–16. Disponível em: <https://aclanthology.org/N10-2004>.

DELGADO, R.; TIBAU, X.-A. Why cohen's kappa should be avoided as performance measure in classification. **PLOS ONE**, Public Library of Science, v. 14, n. 9, p. 1–26, 09 2019. Disponível em: <https://doi.org/10.1371/journal.pone.0222916>.

DHABE, P.; CHANDAK, A.; DESHPANDE, O.; FANDADE, P.; CHANDAK, N.; OSWAL, Y. Stock market trend prediction along with twitter sentiment analysis. In: BALAS, V. E.; SEMWAL, V. B.; KHANDARE, A. (Ed.). **Intelligent Computing and Networking**. Singapore: Springer Nature Singapore, 2023. p. 45–59. ISBN 978-981-99-0071-8.

DI-FELIPPO, A.; NUNES, M. d. G. V.; BARBOSA, B. K. d. S. **Diretrizes de anotação de relações de dependência em tweets do mercado financeiro**. Relatório Técnico – ICMC, USP. n. 446. São Carlos, 70 p.

DI-FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; SILVA, E. H. d.; ROMAN, N. T.; PARDO, T. A. S. Descrição preliminar do corpus dantestocks: diretrizes de segmentação para anotação segundo universal dependencies. In: SBC. **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**. Porto Alegre, 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17813>.

DI-FELIPPO, A.; POSTALI, C.; CEREGATTO, G.; GAZANA, L. S.; ROMAN, N. T. **Diretrizes de Anotação de PoS Tags em Tweets do Mercado Financeiro: Orientações para Anotação em Língua Portuguesa segundo a Abordagem Universal Dependencies**. Relatório Técnico – ICMC, USP. n. 438. São Carlos-SP, 24 p.

DURAN, M. **Guia de Anotação: PropBank.Br (versão 2.0)**. 2014. Acesso em: 03 fev. 2024. Disponível em: <http://www.nilc.icmc.usp.br/semanticnlp/includes/projects/propbankbr/files/manual%20de%20anotacao%20do%20propbank%20v5.pdf>.

DURAN, M.; LOPES, L.; NUNES, M. d. G.; PARDO, T. The dawn of the porttinari multigenre treebank: Introducing its journalistic portion. In: **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: SBC, 2023. p. 115–124. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25443>.

DURAN, M. S. **Manual de anotação de relações de dependência: orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD)**. Relatório Técnico – ICMC, USP. n. 435. São Carlos, 79 p.

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**. Relatório Técnico – ICMC, USP. n. 434. São Carlos, 55 p.

DURAN, M. S. **Manual de anotação de relações de dependência - versão revisada e estendida: orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD)**. Relatório Técnico – ICMC, USP. n. 440. São Carlos, 166 p.

DURAN, M. S.; ALUíSIO, S. M. Propbank-br: a brazilian treebank annotated with semantic role labels. In: **Proceedings of the 8th International Conference on Language Resources and Evaluation**. Istanbul: ELRA, 2012. p. 1862–1867.

FELLBAUM, C. (Ed.). **WordNet: An Electronic Lexical Database**. Cambridge, MA: MIT Press, 1998. (Language, Speech, and Communication).

FILLMORE, C. J. The case for case. In: **Universals in linguistic theory**. [S.l.]: Holt, Rinehart and Winston, Inc., 1968. p. 1–88.

FLEISS, J. L. The measurement of interrater agreement. In: **Statistical Methods for Rates and Proportions**. 2nd. ed. New York: John Wiley, 1981. p. 212–236.

FREITAS, E. C.; BARTH, P. A. Gênero ou suporte? o entrelaçamento de gêneros no twitter. **Revista (Con) Textos Linguísticos**, v. 9, n. 12, p. 8–26, 2015.

GASKELL, P.; MCGROARTY, F.; TIROPANIS, T. An investigation into correlations between financial sentiment and prices in financial markets. In: **Proceedings of the 5th Annual ACM Web Science Conference**. New York, NY, USA: Association for Computing Machinery, 2013. (WebSci '13), p. 99–108. ISBN 9781450318891. Disponível em: <https://doi.org/10.1145/2464464.2464510>.

GAZANA, L. S.; DI-FELIPPO, A. **DANTEStocks: contribuições para seu refinamento e anotação de Part-of-Speech**. São Carlos-SP, 24 p.

GROSS, M. **Grammaire transformationelle du français: 1 - Syntaxe du verbe**. Paris: Cantilène, 1968.

GUIBON, G.; COURTIN, M.; GERDES, K.; GUILLAUME, B. When collaborative treebank curation meets graph grammars. In: **Proceedings of The 12th Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 5293–5302. Disponível em: <https://www.aclweb.org/anthology/2020.lrec-1.651>.

HARRIS, Z. S. Transformations in linguistic structure. In: _____. **Papers in Structural and Transformational Linguistics**. Springer Netherlands, 1970. p. 472–481. ISBN 9789401760591. Disponível em: <http://dx.doi.org/10.1007/978-94-017-6059-1_25>.

HARRIS, Z. S. Operator-grammar of english. **Lingvisticæ Investigationes**, John Benjamins Publishing Company, v. 2, n. 1, p. 55–92, jan. 1978. ISSN 1569-9927. Disponível em: <http://dx.doi.org/10.1075/li.2.1.05har>.

HARTMANN, N.; AVANÇO, L.; BALAGE, P.; DURAN, M.; GRAÇAS VOLPE NUNES, M. das; PARDO, T.; ALUÍSIO, S. A large corpus of product reviews in Portuguese: Tackling out-of-vocabulary words. In: CALZOLARI, N.; CHOUKRI, K.; DECLERCK, T.; LOFTSSON, H.; MAEGAARD, B.; MARIANI, J.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 3865–3871. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/413_Paper.pdf>.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing: an introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. 3rd. ed. [s.n.], 2024. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

KIPPER, K.; PALMER, M.; RAMBOW, O. Extending propbank with verbnet semantic predicates. In: **Workshop on Applied Interlinguas, AMTA-2002**. [S.l.: s.n.], 2002.

KOCH, I. V.; SILVA, M. C. P. d. Souza-e. **Linguística aplicada ao português: sintaxe**. São Paulo: Cortez, 1985.

KRUMM, J.; DAVIES, N.; NARAYANASWAMI, C. User-generated content. **IEEE Pervasive Computing**, v. 7, n. 4, p. 10–11, 2008.

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

LIU, Y.; ZHU, Y.; CHE, W.; QIN, B.; SCHNEIDER, N.; SMITH, N. A. Parsing tweets into Universal Dependencies. In: WALKER, M.; JI, H.; STENT, A. (Ed.). **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**. New Orleans, Louisiana: Association for Computational Linguistics, 2018. p. 965–975. Disponível em: <https://aclanthology.org/N18-1088>.

MACLEOD, C.; GRISHMAN, R.; MEYERS, A. Comlex syntax. **Computers and the Humanities**, v. 31, p. 459–481, 1998a.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: UNIVERSITY OF CALIFORNIA PRESS. **Proceedings of the fifth Berkeley symposium on mathematical statistics and probability**. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MANN, W.; THOMPSON, S. Rethorical structure theory: Toward a functional theory of text organization. **Text**, v. 8, p. 243–281, 01 1988.

MAO, H.; COUNTS, S.; BOLLEN, J. **Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data**. arXiv.org. n. 1112.1051. Disponível em: <https://ideas.repec.org/p/arx/papers/1112.1051.html>.

MARCUS, M. P. e. a. Building a large annotated corpus of english: The penn treebank. **Computational Linguistics**, v. 19, n. 2, p. 313–330, 1993.

MEYERS, A. **Annotation guidelines for NomBank - noun argument structure for PropBank**. Tech Report – New York University.

MEYERS, A. e. a. The nombank project: An interim report. In: **HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation**. [S.l.: s.n.], 2004. p. 24–31.

MIRANDA, L. G. M.; PARDO, T. A. S. An improved and extended annotation tool for universal dependencies-based treebank construction. In: **International Conference on Computational Processing of the Portuguese Language - PROPOR**. [S.l.]: Universidade de Fortaleza, 2022.

MOHAMMAD, S.; ZHU, X.; MARTIN, J. D. Semantic role labeling of emotions in tweets. In: **Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. [S.l.: s.n.], 2014. p. 32–41.

MOHAMMAD, S. M.; ZHU, X.; KIRITCHENKO, S.; MARTIN, J. Sentiment, emotion, purpose, and style in electoral tweets. **Inf. Process. Manage.**, Pergamon Press, Inc., USA, v. 51, n. 4, p. 480–499, jul 2015. ISSN 0306-4573. Disponível em: <https://doi.org/10.1016/j.ipm.2014.09.003>.

NEVES, M. H. M. **Gramática de usos do português**. São Paulo: Editora UNESP, 2000.

NIVRE, J.; MARNEFFE, M.-C. de; GINTER, F.; HAJIČ, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal Dependencies v2: An evergrowing multilingual treebank collection. In: CALZOLARI, N.; BÉCHET, F.; BLACHE, P.; CHOUKRI, K.; CIERI, C.; DECLERCK, T.; GOGGI, S.; ISAHARA, H.; MAEGAARD, B.; MARIANI, J.; MAZO, H.; MORENO, A.; ODIJK, J.; PIPERIDIS, S. (Ed.). **Proceedings of the Twelfth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, 2020. p. 4034–4043. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.497>.

NIVRE, J. *et al.* Universal dependencies v1: a multilingual treebank collection. In: **Proceedings of the 10th International Conference on Language Resources and Evaluation**. [S.l.: s.n.], 2016. p. 1659–1666.

PALMER, M.; GILDEA, D.; KINGSBURY, P. The proposition bank: An annotated corpus of semantic roles. **Computational Linguistics**, v. 31, n. 1, p. 71–106, 2005.

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: **Proceedings of the 14th Symposium in Information and Human Language**. [S.l.: s.n.], 2021. p. 1–10.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PLUTCHIK, R.; KELLERMAN, H. **Theories of Emotion**. [S.l.]: Academic Press, 1980.

PORTUGUêS, A. D. D. C. A. O. de. **Securitização**. 2024. Acesso em: 16/01/2024. Disponível em: <https://aulete.com.br/securitizaÃğÃčo/>.

PRADHAN, S.; BONN, J.; MYERS, S.; CONGER, K.; O'GORMAN, T.; GUNG, J.; WRIGHT-BETTNER, K.; PALMER, M. PropBank comes of Age—Larger, smarter, and more diverse. In: NASTASE, V.; PAVLICK, E.; PILEHVAR, M. T.; CAMACHO-COLLADOS, J.; RAGANATO, A. (Ed.). **Proceedings of the 11th Joint Conference on Lexical and Computational Semantics**. Seattle, Washington: Association for Computational Linguistics, 2022. p. 278–288. Disponível em: <https://aclanthology.org/2022.starsem-1.24>.

QI, P.; ZHANG, Y.; ZHANG, Y.; BOLTON, J.; MANNING, C. D. Stanza: A Python natural language processing toolkit for many human languages. In: **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations**. [s.n.], 2020. Disponível em: <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. Universal dependencies for portuguese. In: **Proceedings of the Fourth International Conference on Dependency Linguistics**. [S.l.: s.n.], 2017.

RADEMAKER, A.; CHALUB, F.; REAL, L.; FREITAS, C.; BICK, E.; PAIVA, V. de. Universal dependencies for portuguese. In: **Proceedings of the Fourth International Conference on Dependency Linguistics (Depling)**. Pisa, Italy: [s.n.], 2017. p. 197–206. Disponível em: <http://aclweb.org/anthology/W17-6523>.

RAMBOW, O.; DORR, B. J.; KIPPER, K.; KUCEROVA, I.; PALMER, M. Automatically deriving tectogrammatical labels from other resources: A comparison of semantic labels across frameworks. **Prague Bulletin of Mathematical Linguistics**, v. 79-80, p. 23–35, 2003.

RAMOS, J. E. Using tf-idf to determine word relevance in document queries. In: . [s.n.], 2003. Disponível em: <https://api.semanticscholar.org/CorpusID:14638345>.

RASSI, A. P. **O verbo dar em português brasileiro: descrição, classificação e processamento automático**. Araraquara, SP: Letraria, 2023. PDF. ISBN 978-65-5434-023-6.

RUDRAPAL, D.; DAS, A. Semantic role labeling of english tweets. **Computación y Sistemas**, v. 22, n. 3, p. 737–746, 2018. ISSN 2007-9737.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; ÇETINOğLU, ; CIGNARELLA, A. T.; LYNN, T.; REHBEIN, I.; RUPPENHOFER, J.; SEDDAH, D.; ZELDES, A. Treebanking user-generated content: a proposal for a unified representation in universal dependencies. In: **Proceedings of the 12th International Language Resources and Evaluation Conference**. [S.l.: s.n.], 2020. p. 5240–5250.

SANGUINETTI, M.; BOSCO, C.; CASSIDY, L.; AL. et. Treebanking user-generated content: a ud based overview of guidelines, corpora and unified recommendations. **Language Resources Evaluation**, v. 57, p. 493–544, 2023.

SANTOS, M. C. A. **Descrição dos predicados nominais com o verbo-suporte ter**. Tese (Doutorado) — Universidade Federal de São Carlos, 2015.

SARDINHA, T. B. Lingüística de corpus: histórico e problemática. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 16, n. 2, p. 323–367, 2000. ISSN 0102-4450.

SCANDAROLLI, C. L.; DI FELIPPO, A.; ROMAN, N. T.; PARDO, T. A. S. Tipologia de fenômenos ortográficos e lexicais em cgu: o caso dos tweets do mercado financeiro. In: **Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2023)**. Sociedade Brasileira de Computação, 2023. (STIL 2023). Disponível em: <http://dx.doi.org/10.5753/stil.2023.233948>.

SCHUFF, H.; BARNES, J.; MOHME, J.; PADÓ, S.; KLINGER, R. Annotation, modelling and analysis of fine-grained emotions on a stance and sentiment detection corpus. In: BALAHUR, A.; MOHAMMAD, S. M.; GOOT, E. van der (Ed.). **Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis**. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 13–23. Disponível em: <https://aclanthology.org/W17-5203>.

SCHULER, K. K.; PALMER, M. S. **Verbnet: a broad-coverage, comprehensive verb lexicon**. Tese (Doutorado), USA, 2005.

SILVA, E. H.; PARDO, T. A. S.; ROMAN, N. T.; DI-FELIPPO, A. Universal dependencies for tweets in brazilian portuguese: tokenization and part of speech tagging. In: **Proceedings of the 18th National Meeting on Artificial and Computational Intelligence**. [S.l.: s.n.], 2021. p. 1–12.

SILVA, F. J. V.; ROMAN, N. T.; CARVALHO, A. M. B. R. Stock market tweets annotated with emotions. **Corpora**, v. 15, n. 3, p. 343–354, 2020. ISSN 1755-1676.

SOUZA, E.; SILVEIRA, A.; CAVALCANTI, T.; CASTRO, M.; FREITAS, C. Petrogold – corpus padrão ouro para o domínio do petróleo. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: SBC, 2021. p. 29–38. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17781>.

STRAKA, M. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: ZEMAN, D.; HAJIČ, J. (Ed.). **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. Disponível em: <https://aclanthology.org/K18-2020>.

SUTTLES, J.; IDE, N. Distant supervision for emotion classification with discrete binary values. In: GELBUKH, A. (Ed.). **Computational Linguistics and Intelligent Text Processing**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 121–136.

TAYLOR, A.; MARCUS, M.; SANTORINI, B. The penn treebank: An overview. In: **Treebanks**. [S.l.]: Springer, Dordrecht, 2003, (Text, Speech and Language Technology, v. 20).

TESNIèRE, L. **Eléments de syntaxe structurale**. [S.l.]: Librarie C. Klincksieck, Paris, 1959.

VOSKAKI, R.; TZIAFA, E.; ANNIDOU, K. Description of predicative nouns in a modern greek financial corpus. In: **Selected Papers of the 21st International Symposium on Theoretical and Applied Linguistics (ISTAL)**. [S.l.: s.n.], 2016. p. 488–503.

ZHANG, X.; FUEHRES, H.; GLOOR, P. A. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". **Procedia - Social and Behavioral Sciences**, v. 26, p. 55–62, 2011. ISSN 1877-0428. The 2nd Collaborative Innovation Networks Conference - COINs2010. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877042811023895>.

Apêndice A – Manual de anotação de relações de dependência em tweets do mercado financeiro

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

DIRETRIZES DE ANOTAÇÃO DE RELAÇÕES DE DEPENDÊNCIA EM *TWEETS* DO MERCADO FINANCEIRO

ARIANI DI-FELIPPO
MARIA DAS GRAÇAS VOLPE NUNES
BRYAN KHELVEN DA SILVA BARBOSA

Nº 446

RELATÓRIOS TÉCNICOS

São Carlos – SP
Abr./2024

Natural Language Processing initiative (NLP2) of the Center for Artificial Intelligence (C4AI) of the University of São Paulo, sponsored by IBM and FAPESP

POeTiSA

POrtuguese processing – Towards Syntactic Analysis and parsing

Diretrizes de Anotação de Relações de Dependência em *Tweets* do Mercado Financeiro

Orientações para anotação de relações de dependência sintática em *tweets* em língua portuguesa segundo a abordagem *Universal Dependencies* (UD)

Ariani Di-Felippo, Maria das Graças Volpe Nunes e Bryan Khelven da Silva Barbosa

Abril/2024

Relatório técnico do
Núcleo Interinstitucional de Linguística Computacional (NILC)

Agradecimentos

À **Magali Sanchez Duran**, por todo o suporte, sempre atencioso, rigoroso e acurado, para a compreensão do nível sintático do modelo *Universal Dependencies* (UD) e para a aplicação dele no tratamento dos *tweets*. Esse suporte foi imperativo para a confecção desta primeira versão de um Manual de anotação das dependências previstas pelo modelo em *tweets* do mercado financeiro em língua português.

Este trabalho 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.

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.

SUMÁRIO

Introdução

Apresenta-se neste relatório o Manual de Anotação de Relações de Dependência Sintática para *Tweets*[1] do Mercado Financeiro em Português desenvolvido no projeto POeTiSA (*POrtuguese processing - Towards Syntactic Analysis and parsing*), que faz parte da iniciativa de Processamento Automático 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 nº. 2019/07665-4).

O POeTiSA é um projeto que visa aumentar os recursos baseados em sintaxe e desenvolver ferramentas e aplicações de PLN relacionadas à sintaxe para o português do Brasil e que alcancem o estado-da-arte. Esse objetivo inclui a produção de um *corpus* multigênero extenso e abrangente anotado segundo o modelo *Universal Dependencies* (UD) (NIVRE, 2015; NIVRE et al., 2020). Para tanto, dá-se continuidade a esforços anteriores, como os de construção da Floresta Sintá(c)tica (AFONSO et al., 2002), do MAC-MORPHO (ALUÍSIO et al., 2003) e do Bosque-UD (RADEMAKER et al., 2017). O *corpus* multigênero, chamado Porttinari[2] (de *PORtuguese Treebank*) (PARDO et al., 2021), deve subsidiar estudos linguísticos e/ou o desenvolvimento de ferramentas de análise textual para o português, como *taggers* e *parsers*.

O modelo UD possui o formato de anotação CoNLL-U, constituído de 10 colunas, algumas das quais pedem decisões de anotação. Este Manual estabelece diretrizes de anotação de relações de dependência em *tweets* do mercado financeiro. Essa tarefa envolve decidir (i) os participantes da relação de dependência, (ii) o *head* e o dependente e (iii) o nome da relação que os liga. As colunas do CoNLL-U envolvidas pelas diretrizes são a 7ª, a 8ª e a 9ª. Este Manual complementa o Manual de Anotação de *PoS* Tags de PoS Tags em *Tweet*, publicado em março de 2022 na série de Relatórios Técnicos do ICMC sob número 438 e disponível na página do POeTiSA[3]. O Manual de *PoS tags* contempla a 4ª coluna do CoNLL-U. A divisão das diretrizes em 2 manuais se deve à decisão de revisar as colunas de anotação por etapas, a fim de que uma etapa pudesse abreviar o esforço requerido na outra, pois a tarefa de revisar *PoS tags* e relações de dependência, separadamente, já é bastante complexa.

Este manual reúne diretrizes para anotação de *tweets* do mercado financeiro, sendo que a anotação das estruturas gerais do português presentes nos exemplos segue o Manual de Duran (2022)[4], proposto para a anotação de textos que possuem linguagem formal (ou padrão). Por isso, aliás, sugere-se o estudo do manual de Duran (2022) antes da leitura deste. Os exemplos deste documento foram extraídos de uma versão revisada do *corpus* de Silva *et al*. (2020) que já possui anotação de PoS segundo a UD (denominada DANTEStocks) e está disponível na página do POeTiSA[5]. O DANTEStocks contém *tweets* que mencionam ao menos um *ticker* de uma das 73 ações que compõem o Ibovespa (principal indicador da B3, que é a bolsa de valores oficial do Brasil), sendo que nele o *tweet* é a unidade básica de análise, uma vez que as postagens não passaram por qualquer segmentação, a não ser a tokenização. Os *tweets* do DANTEStocks também não passaram por um processo de normalização, contendo, por isso, todas as peculiaridades da linguagem informal dos *tweets* e do domínio do mercado financeiro (cf. Di-Felippo *et al.,* 2021).

---

[1] Embora a plataforma tenha sido renomeada para "X" e as mensagens nela circulantes para "*posts*" após a aquisição da marca por Elon Musk e conseguinte reestruturação ocorrida em 2022, optou-se por utilizar as denominações originais ("*Twitter*"' para plataforma e "*tweet*" para mensagem/postagem) em concordância com a época (2014) em que o córpus aqui utilizado foi compilado.

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

[3] https://drive.google.com/file/d/1ka-GVNb8XgEJWmBOrNcd-Grfg10l_QEb/view?usp=sharing

[4] https://drive.google.com/file/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view?usp=sharing

[5] https://drive.google.com/file/d/1wr9M4czkPgkUj1--U9GT9h8ncXc6rzv4/view?usp=sharing

Diretrizes de anotação de dependências para *tweets*

O esquema de anotação do modelo *Universal Dependencies* (UD), em sua versão 2.0, disponibiliza 37 *tags* de relações de dependência ou *deprel* (do inglês, ***dep**endency **rel**ation*).

Na Figura 1, extraída das *Guidelines* da UD[6], exibem-se as 37 *deprel* da UD. A Figura destaca os argumentos principais (ou *core*) dos predicados, separando-os dos demais argumentos não-*core*. Na Figura 1, separam-se também os argumentos e modificadores de predicados dos modificadores de nominais. Ademais, a Figura 1 exibe as etiquetas empregadas quando o dependente da relação está sob forma oracional (coluna *clauses*), que correspondem às orações subordinadas.

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	**MWE**	**Loose**	**Special**	**Other**
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figura 1 - Quadro de relações de dependência da UD.

Há três relações do tipo *core* em que os dependentes têm forma nominal[7] (**subj**, **obj** e **iobj**) e três em que os dependentes têm forma oracional (**csubj**, **ccomp** e **xcomp**). Não são considerados argumentos *core* os complementos verbais preposicionados ou os adjuntos adverbiais.

Nas relações **obl**, **advmod**, **advcl**, **vocative**, **expl**, **dislocated**, **discourse**, **aux**, **cop** e **mark**, o predicado é o *head* e os dependentes são considerados seus modificadores. Apenas uma dessas relações tem forma oracional: **advcl**.

Já nas relações **nmod**, **appos**, **nummod**, **acl**, **amod**, **det** e **case**, o *head* é um nominal e o dependente é um modificador. O dependente nessas relações pode ser um nominal, uma palavra funcional (como as preposições, na *deprel* **case**, e os determinantes, na *deprel* **det**) ou um numeral (**nummod**). A *deprel* **acl** é a única que tem forma oracional, sendo usada para ligar nominais a orações adjetivas e completivas nominais.[8]

As *deprel* **cc** e **conj** são empregadas para tratar da coordenação de elementos da sentença/*tweet*.

---

[6] ttps://universaldependencies.org/u/dep/index.html

[7] Neste manual, o termo “nominal” se refere a palavras que podem exercer as funções típicas de substantivos (substantivos, pronomes, adjetivos e numerais com função substantiva).

[8] Como destacado por Duran (2022), a relação de modificador nominal **clf** não é usada no português e, por isso, não está incluída neste manual.

Por fim, tem-se as 11 relações artificiais criadas pela UD, sendo **root** a mais importante delas. Essa *deprel* foi criada para marcar a raiz da árvore sintática de dependências. Estabelecer o **root** é o primeiro passo para se fazer a anotação de uma sentença ou *tweet*. As demais relações **fixed**, **flat**, **compound**, **parataxis**, **list**, **orphan**, **goeswith**, **reparandum**, **punct** e **dep** são usadas para anotar *tokens* que não apresentam relação sintática com outros *tokens*. Como essas relações não possuem sintaxe, a identificação do *head* da *deprel* é arbitrária, sendo esse papel normalmente atribuído ao primeiro dos dois *tokens* unidos pela relação.

Uma *deprel* liga dois elementos de uma sentença (ou *tweet*, no caso), tal que:

- um deles é denominado ***head*** (núcleo da relação) e o outro é denominado **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 associado à função que o dependente realiza em relação ao *head*;
- a seta que representa uma *deprel* parte sempre do *head* em direção ao dependente da relação;
- o elemento apontado pela seta, no caso de dependente oracional, será o predicado da oração dependente;
- um *head* é sempre uma palavra de conteúdo (verbo, nome/substantivo, adjetivo, pronome, numeral e advérbio); exceções são símbolos que podem ser expressos por palavras, como R$ (reais), % (por cento), etc.
- palavras funcionais (determinantes, preposições, conjunções) e sinais de pontuação, por sua vez, deverão ser sempre dependentes; no caso dos *tweets* do mercado financeiro, no entanto, há exceções, como “>” quando interpretado como verbo e anotado com **root;**

Além disso, a atribuição de *deprel* deve observar o princípio da projetividade, ou seja, os arcos das relações não podem se cruzar. A Figura 2 ilustra um *tweet* anotado com relações de dependência.

Antes de apresentar as diretrizes propriamente ditas, no entanto, cabem algumas observações importantes sobre a linguagem CGU (“conteúdo gerado por usuário”), especificamente dos *tweets* do mercado financeiro, uma vez que o córpus, por não ter sido normalizado, impõe desafios diversos à tarefa de anotação de *deprel*.

Os referidos *tweets* são comumente fragmentados, podendo ser compostos por sequências de sintagmas ou elementos simplesmente justapostos, apresentar truncamentos diversos, fenômenos típicos do *Twitter* (como URL, menção, *hashtag,* marca de *retweet* e outros), além de linguagem informal, com pontuação e ortografia que não seguem a norma padrão. Tudo isso acarreta elipses e ambiguidades, as quais dificultam a compreensão do conteúdo, sobretudo, a identificação do root. Além disso, a interpretação de um *tweet* sobre ações requer conhecimento do domínio financeiro devido à ocorrência de vocabulário terminológico (como as *cashtags*) e de estruturas de linguagem específicas. Dito isso, um *tweet* desse domínio pode ter mais de uma interpretação possível e, por conseguinte, mais de uma anotação sintática, não havendo, assim, uma escolha de *deprel* dita “correta”. Assim, as diretrizes apresentadas aqui refletem a interpretação do anotador especialista.

Diferentemente do Manual de *deprel* para o português de Duran (2022), as diretrizes de anotação que compõem este Manual são apresentadas em duas partes.

Sobre as diretrizes, a primeira parte descreve as diretrizes de anotação das *deprel* do modelo UD para ocorrências idiossincráticas dos *tweets* do mercado financeiro. Tais ocorrências englobam:

a) contextos com sintaxe não-padrão (e, por isso, não previstos no manual de anotação da língua portuguesa de Duran (2022), o que acarreta o emprego particular de certas *deprel*
b) fenômenos que ocorrem nos tweets do referido corpus/domínio (cf. Di-Felippo et al. 2021), a saber: marca de *retweet*, menção, URL, *hashtag*, *cashtag*, truncamento, índice de valor das ações, *emoticon*, *emoji* e repetição de pontuação).

Na segunda parte, descrevem-se as diretrizes que são específicas para certos padrões estruturais que ocorrem com frequência nos *tweets* do mercado financeiro, sendo que algumas estruturas correspondem ao *tweet* como um todo.

Ademais, ressalta-se que, dada a fragmentação e, por vezes, a ocorrência de fenômenos típicos do domínio que dificultam a compreensão dos *tweets,* alguns exemplos estão associados a uma possível interpretação. Essa interpretação, aliás, é fundamental para compreender a anotação sintática (e escolha das *deprel*) ilustrada.

PRIMEIRA PARTE – IDIOSSINCRASIAS

Nesta segunda parte, descrevem-se as diretrizes de anotação das *deprel* do modelo UD para ocorrências idiossincráticas dos *tweets* do mercado financeiro. Tais ocorrências englobam:

i. os fenômenos CGU ("conteúdo gerado por usuário") que ocorrem nos *tweets* do referido *corpus*/domínio, a saber: *marca de retweet, URL, menção, hashtag, cashtag,* truncamento*,* índice de (des-)valorização das ações, *emoticon* e *emoji* e repetição de pontuação
ii. contextos com sintaxe não-padrão, o que acarreta o emprego particular de certas *deprel*

Fenômenos GCU

1. RT: marca de *retweet*

A marca de *retweet* (RT) sempre ocorre seguida de uma menção (*@mention*), sendo o *head* da relação de **nmod** que se estabelece com a menção (Figura 2).

Uma RT pode ocorrer integrada à sintaxe ou de forma isolada (em inglês, *standalone*).
Se integrada, deve ser anotada com a *deprel* relativa à sua função e PoS.
No caso dos *tweets* do DANTEStocks, a única ocorrência integrada à sintaxe foi anotada com **vocative**, como pode ser observado na Figura 2.
Nesse contexto, a RT é dependente do *head* "claro", que é **root** do *tweet.*

Exemplo:

(1) Mas é claro **RT** @garimpodeacoes : FITCH , cia de avaliação de risco diz que seca prolongada deve pressionar Sabesp (SBSP3) de forma mais severa

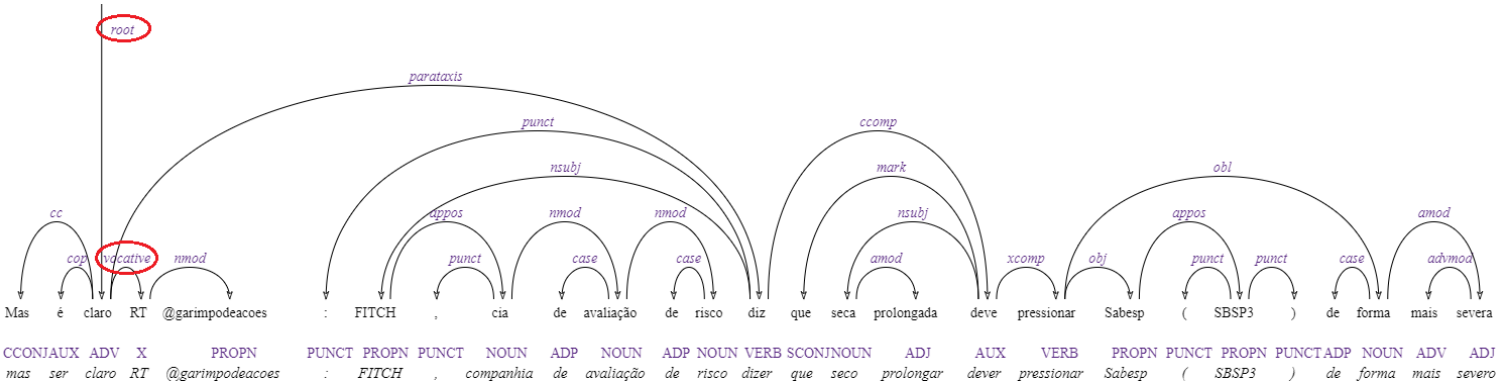


Figura 2 – Anotação de marca de *retweet* (RT) integrada à sintaxe com **vocative.**

Quando *standalone*, as marcas RTs devem ser conectadas ao **root** ou a outro *head* pela relação **parataxis**.

Exemplo:

(2) **RT** @daltonvieira : Ação ex-dividendos hoje : PCAR4 . As cotações históricas foram ajustadas . Saiba mais ! http://t.co/C7k4DuDID2

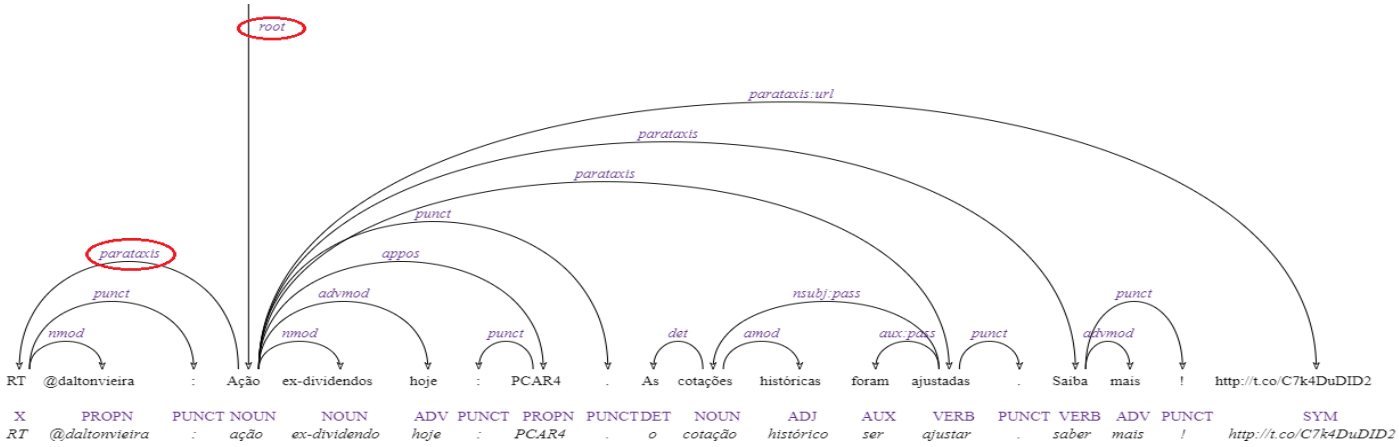


Figura 3 – Marca de retweet (RT) *standalone*, anotada como **parataxis** do **root**.

Um *retweet* (que se inicia com a marca de RT) pode apresentar conteúdo novo, inserido pelo usuário que realizou a repostagem.
Nesses casos, o **root** deve estar no trecho novo e a mensagem retuitada estará conectada ao *root* por **parataxis**.
Esse é o caso do *tweet* da Figura 4, em que o trecho sublinhado é a mensagem original (repostada) e o trecho em negrito foi inserido pelo usuário que retuitou a mensagem original.
Como se vê, o **root** está no trecho novo ("piorando"), sendo que a marca de RT está associada ao *head* do trecho original ("acompanhando") por **parataxis**, e esse head, por sua vez, está ligado ao **root** também por **parataxis**.

Exemplo:

(3) RT @Pepez_Legal : @coroneldoblog Amigo , está acompanhando PETR4 ? ? > http://t.co/WxnN4AOaKT **A Graça está piorando ainda mais o cenário .. rs …**

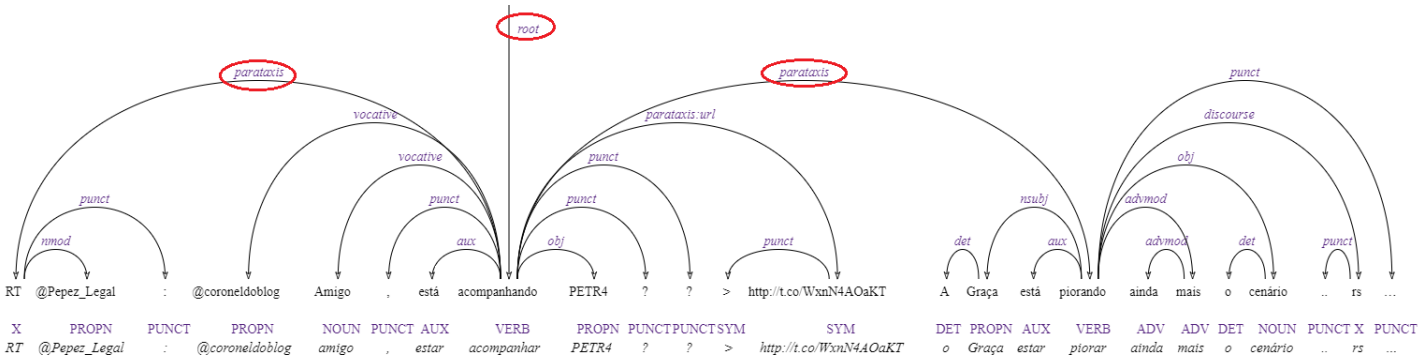


Figura 4 – *Retweet* com conteúdo novo, inserido pelo responsável pela repostagem.

2. Menção

As menções a usuários/perfis do *Twitter* foram anotadas com a PoS *tag* PROPN e podem ocorrer integradas à sintaxe da mensagem ou em contexto isolado (*standalone*).

Quando integradas, devem ser conectadas em nível sintático pela *deprel* que representa sua função/posição na mensagem, podendo ser **root** (Figura 5), **nmod** (Figura 6), **vocative** (Figura 7, 8, 9, 10, 11), **obl** (Figura 12), **obj** (Figura 13) e **nsubj** (Figura 14).

2.1. Root: raiz

A anotação de uma menção com **root** é dependente da interpretação do *tweet.* No caso da Figura 5, por exemplo, *"*@DepEduardoCunha" foi anotado como **root** por ser interpretado como predicado.

Exemplo:

(4) @marisascruz sim , sim . este é @**DepEduardoCunha** , contra o MC . E > @eduardocampos40 que fez PSB votar contra CPI #PETR4 http://t.co/An5WtvvCYj

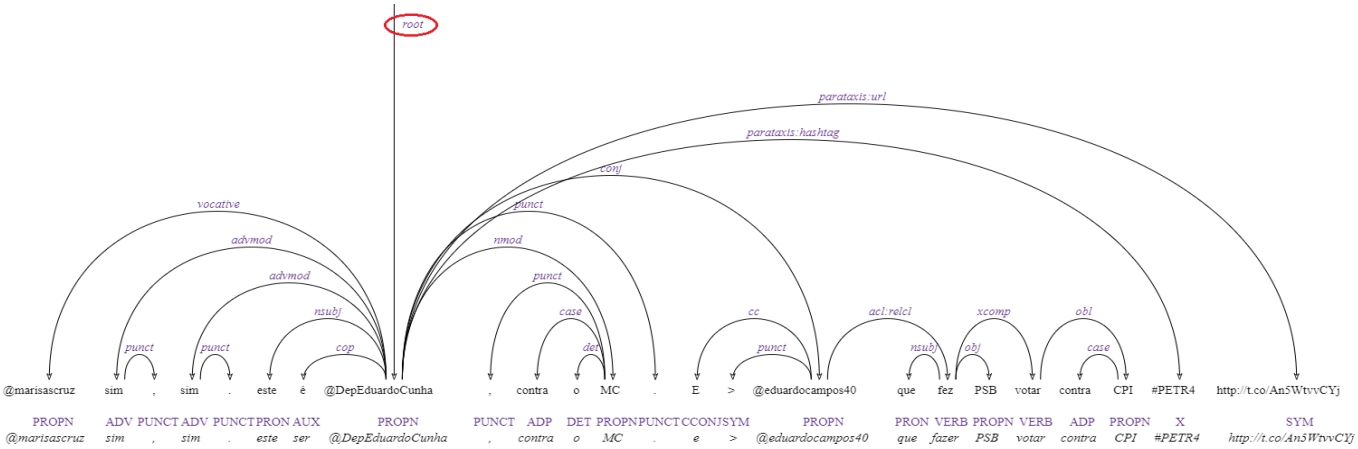


Figura 5 – Menção integrada à sintaxe anotada com **root.**

2.2. Nmod: modificador nominal

Quando uma menção funciona como modificador de outro substantivo, deve-se anotá-la com **nmod,** sendo que tal modificação ocorre primordialmente da esquerda para a direita.
Geralmente, a *deprel* **nmod** é intermediada por alguma preposição que, por sua vez, recebe a *deprel* **case**, a qual partirá do substantivo modificador (no caso, a menção) em direção à preposição.

Exemplo:

(5) Cada vez que ouço a G. Foster defendendo o plano de investimento de a @**petrobras** , mais me certifico que devemos comprar PETR3 e 4 em a BOVESPA

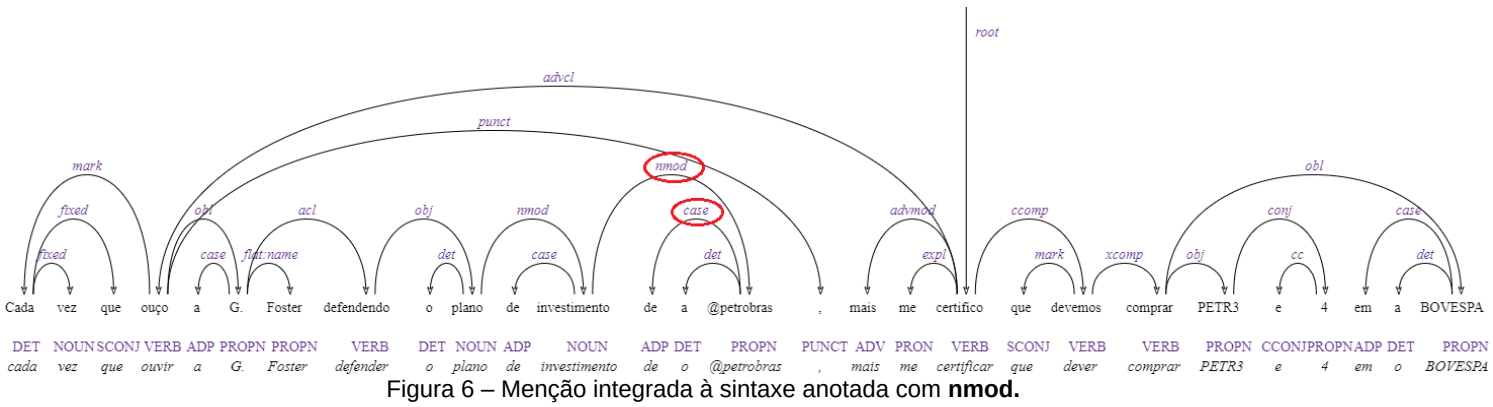


Figura 6 – Menção integrada à sintaxe anotada com **nmod.**

2.3. Vocative: vocativo

Nesses casos, interpreta-se a menção como uma referência ao participante do diálogo a quem se dirige a mensagem, sendo o predicado da oração principal o *head* da *deprel* **vocative.**

Nessa função, uma ou mais menções podem ocorrer no início ou no fim dos *tweets* (ou de uma sentença que os integra). Cada uma das menções deve ser conectada ao *head* por **vocative.** Quando ocorrem no final dos *tweets,* as menções podem ou não ser precedidas de pontuação.

Exemplo:

(6) **@alvarodias**_ tem q ser contundente com essa petista q está ajudando a acabar com a Petr4 ...

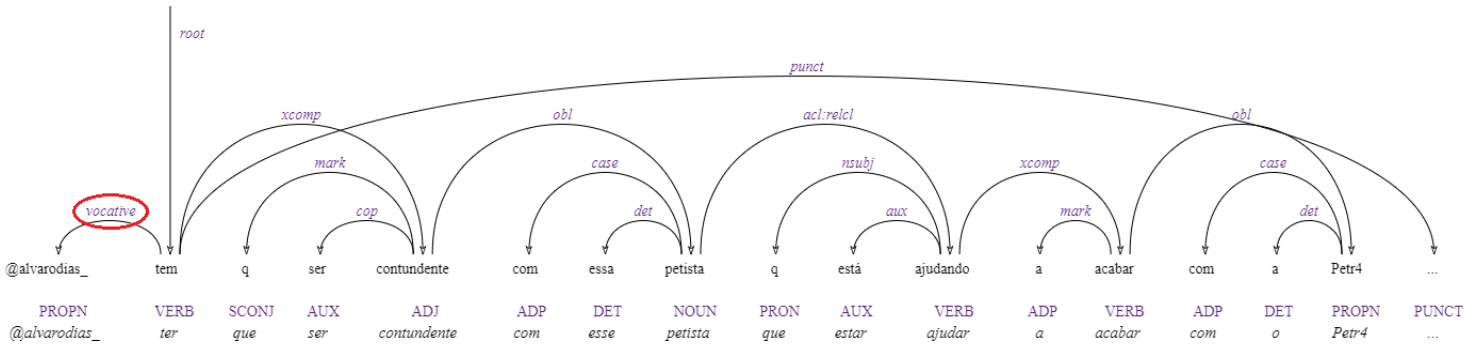


Figura 7 –Menção no início do *tweet* conectada ao **root** por **vocative** (exemplo 6).

Exemplo:

(7) **@ferrisss @dfittarelli** vendinha de itub4 em os 35,10 começando a dar frutos , VAI QUE VAI !

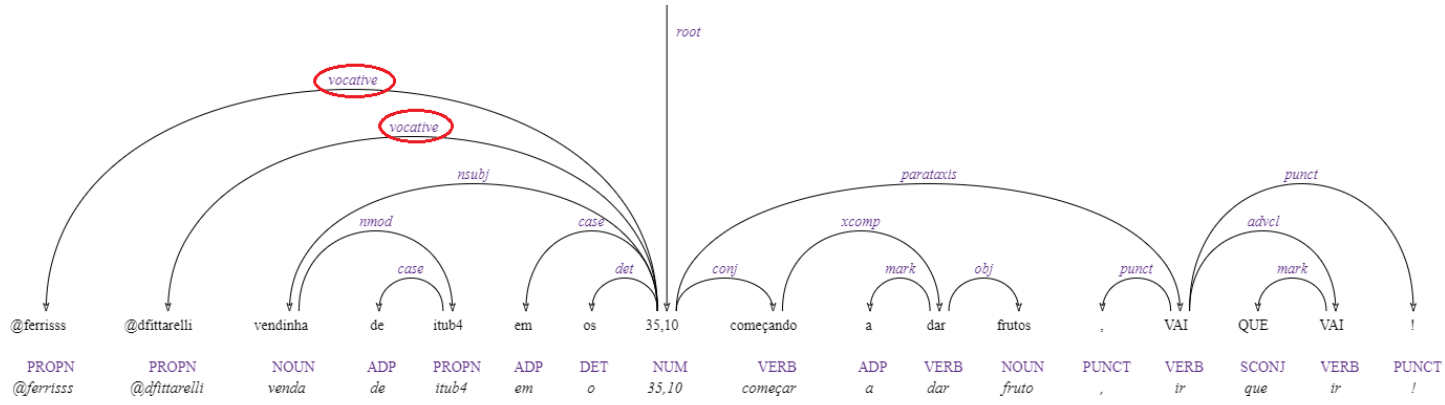


Figura 8 – Duas menções no início do *tweet* conectadas ao **root** por **vocative** (exemplo 7).

Exemplo:

(8) Essa virada final em a #vale5 pagou meu japonês hoje . @**ELISLEITAO**

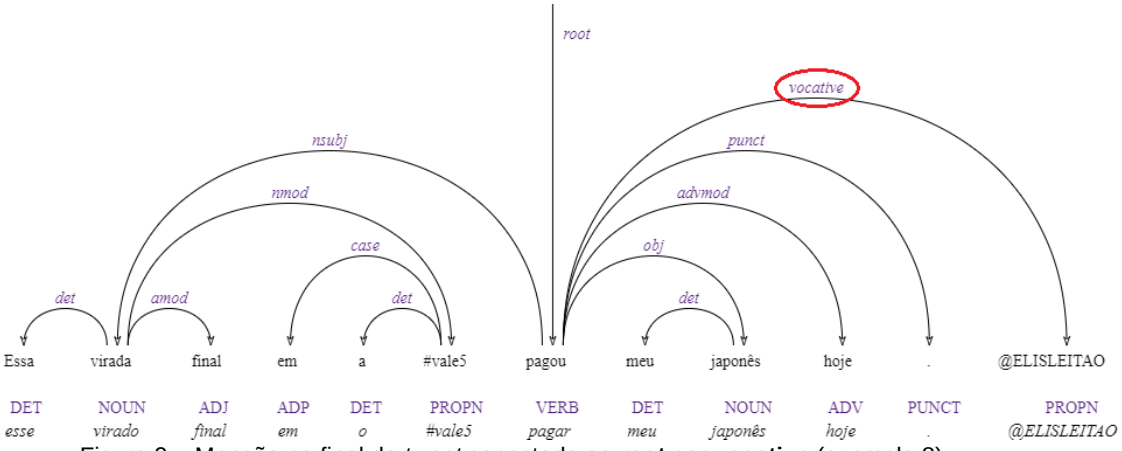


Figura 9 – Menção no final do *tweet* conectada ao **root** por **vocative** (exemplo 8).

Exemplo:

(9) Aliás , essa p* de essa Pasadena tá funcionando ? 03/02/06 #PETR4 aprova compra de refinaria EUA http://t.co/lfZWv799p0 @**ivomarcelino** @**AryAntiPT**

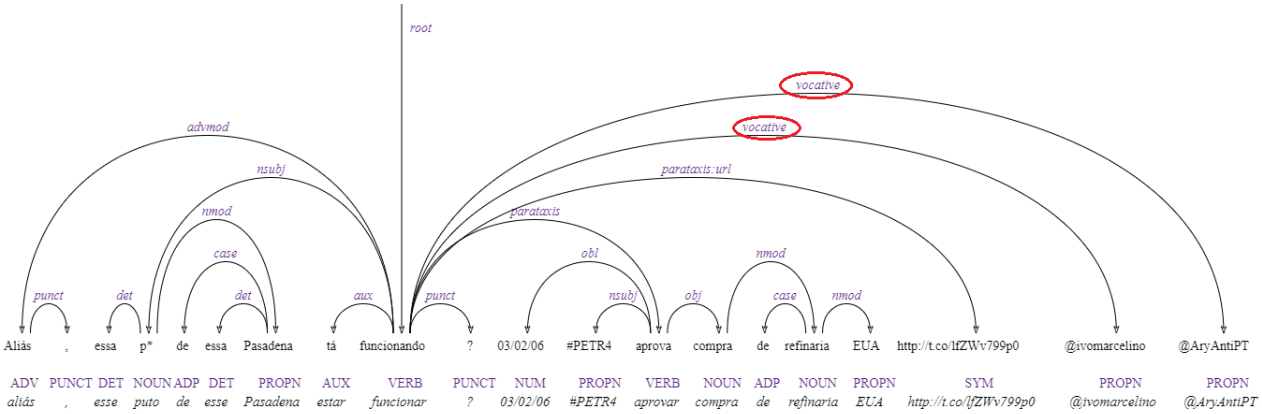


Figura 10 – Duas menções no final do *tweet* conectadas ao **root** por **vocative** (exemplo 9).

Exemplo:

(10) Eita + um rojão em a Petr4 ! @**clubedopairico** Refinaria Pasadena – MP investigará Petrobras por evasão d divisas e peculato http://t.co/egXYdJ7L8u

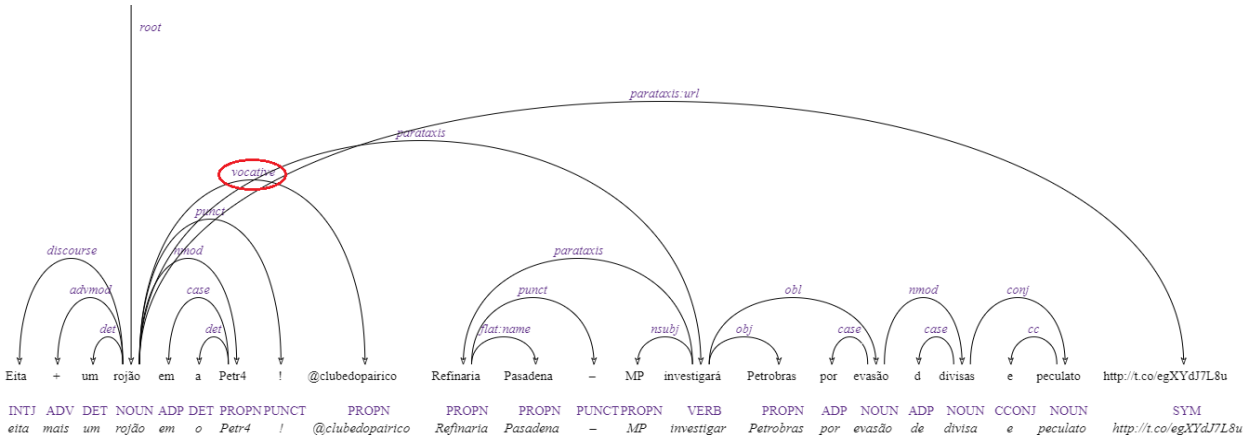


Figura 11 – Menção no final de uma sentença do *tweet* conectada ao **root** por **vocative** (exemplo 10).

2.4. Obl: nominal oblíquo

Uma menção deve ser anotada com a *deprel* **obl** quando for empregada para modificar um verbo (seu *head*). Nesse caso, a menção dependente de **obl** é introduzida por preposição.

Exemplo:

(11) #PETR4 : EXCLUSIVO , **em** @**VEJA** : PF prende ex-diretor de a Petro em operação de lavagem de dinheiro http://t.co/3wGEVANwmT http://t.co/YHocJpiEkv

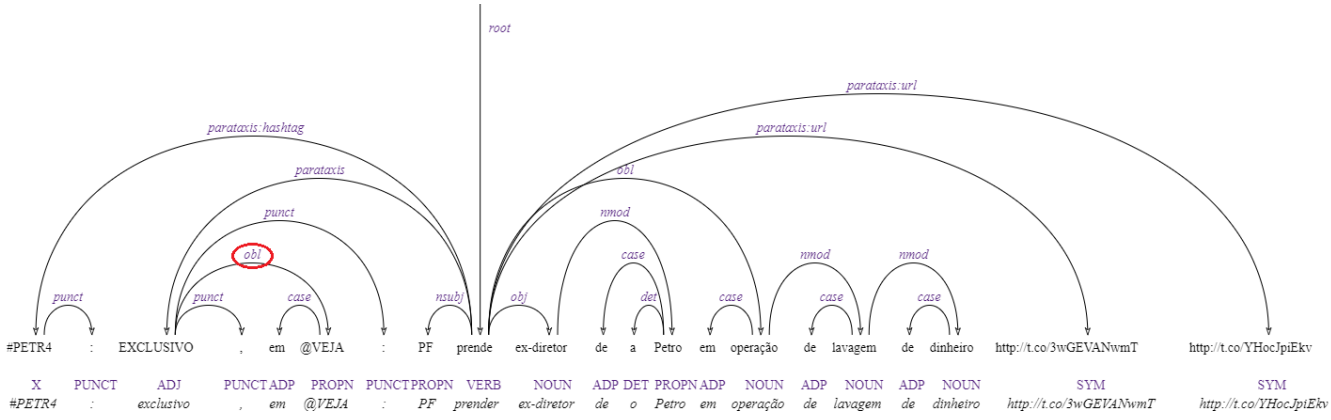


Figura 12 – Menção integrada à sintaxe anotada com **obl**.

2.5. Obj: objeto direto

Uma menção deve ser anotada com a *deprel* **obj** quando for empregada como segundo argumento *core* do verbo, sendo que o primeiro é **nsubj**. No caso da Figura 13, por exemplo, a menção "@petrobras" é dependente por **obj** do verbo "vincular". Este, por sua vez, está conectado por **xcomp** a "tentam", uma vez que o sujeito da oração dependente de **xcomp** ("vincular @petrobras a vergonha") coincide com o sujeito da oração matriz, que é "coxinhas".

Exemplo:

(12) Os coxinhas aderem a o jogo sujo de o PIG e tentam vincular @**petrobras** a vergonha . Mas todo mundo quer compras ações de a PETR4

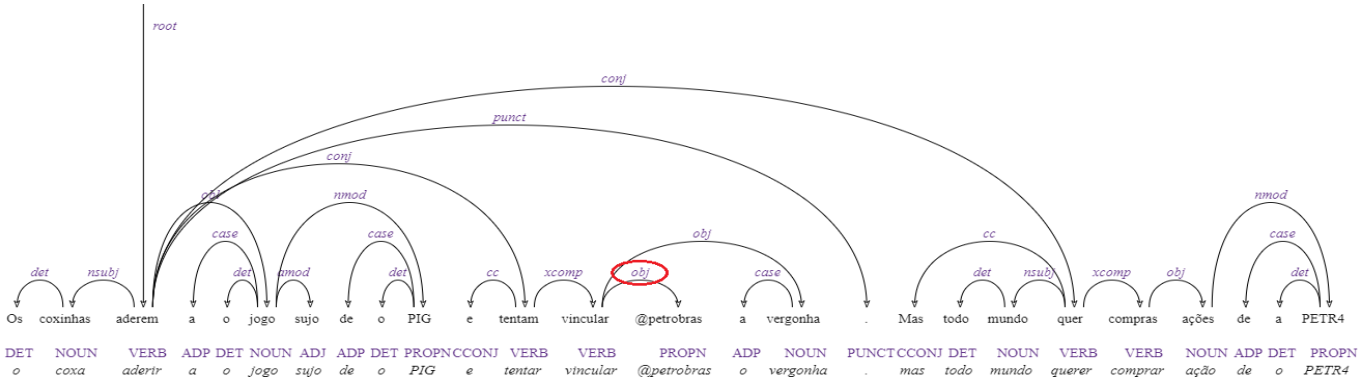


Figura 13 – Menção integrada à sintaxe anotada com **obj**.

2.6. Nsubj: sujeito

Deve-se anotar uma menção com a *deprel* **nsubj** quando ela ocorrer como primeiro complemento *core* de um predicado, isto é, como sujeito. No caso da Figura 14, por exemplo, a menção "@geraldoAlckmin_" foi conectada ao predicado verbal "diz" por meio de **nsubj.**

Exemplo:

(13) último dia de o Gov FHC , a ação Petrobras (PETR3) > R$ 3,3 @**geraldoAlckmin**_ diz q ações de a Petrobras " viraram pó " . http://t.co/vAdFc65rmh

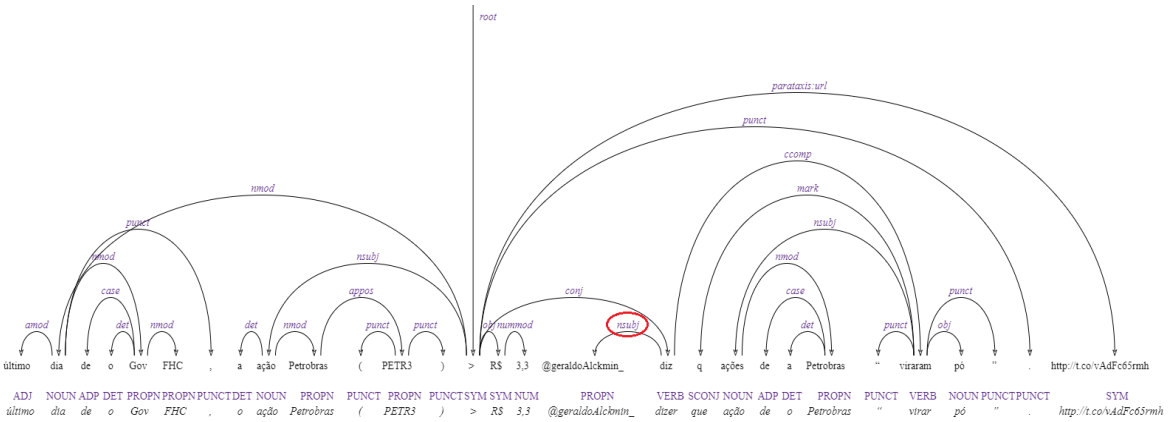


Figura 14 – Menção integrada à sintaxe anotada com **nsubj**.

Quando *standalone,* uma menção não tem relação sintática explicitada com seu *head* e, por isso, deve ser anotada como dependente por **parataxis** do *head* (que pode ser o *root*). Nesses casos, a menção pode ocorrer no começo, no meio ou no final do *tweet.* Quando no começo e no meio, tende a ser seguida por dois pontos, como nas Figuras 15 e 16. E, quando a ocorrência se dá ao final do *tweet,* pode aparecer entre parênteses (Figura 17).
A conexão de uma menção ao *root* ou a outro *head* por **parataxis** depende da interpretação do conteúdo do *tweet.* Nas nas Figuras 15 e 17, por exemplo, as menções "@JornalMercantil" e "@oswaldosena_" conectam-se ao **root** por **parataxis**, uma vez que dizem respeito à informação central da postagem. O mesmo não ocorre com "@Info_BMFBOVESPA" (meio) (Figura 16), que é dependente do *head* "fecha".

Exemplo:

(14) " @**JornalMercantil** **:** China planeja reduzir importação de minério de ferro http://t.co/Yqyjr6Z2fB " #vale5

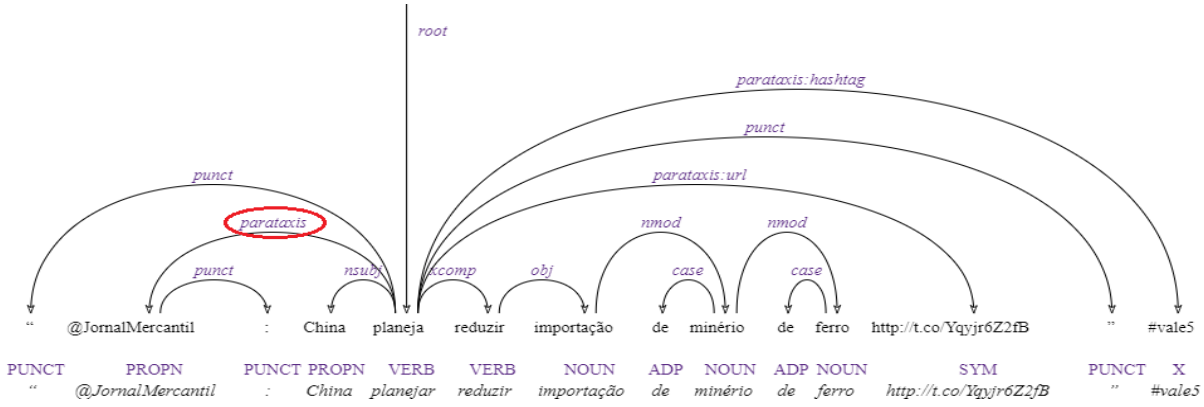


Figura 15 – Menção com **parataxis** no início do *tweet* e seguida de pontuação.

Exemplo:

(15) $PETR4 R$ 13,11 - 1,58 % / $VALE5 R$ 25,9 - 2,12 % , ' **@Info_BMFBOVESPA :** #Ibovespa fecha em baixa de 0,91 % a os 45.443 pontos . '

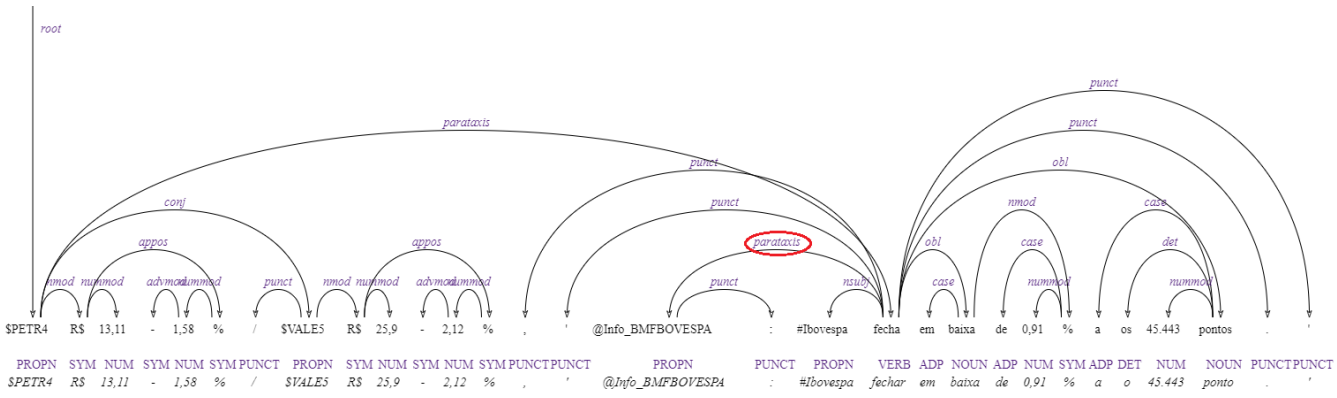


Figura 16 – Menção com **parataxis** no meio do *tweet* e seguida de pontuação.

Exemplo:

(16) Petrobras (#PETR) , Rossi (#RSID3) e BB (#BBAS) batem mínimas de até 10 anos . http://t.co/O0X4s9y6Ab (**@oswaldosena_1**)

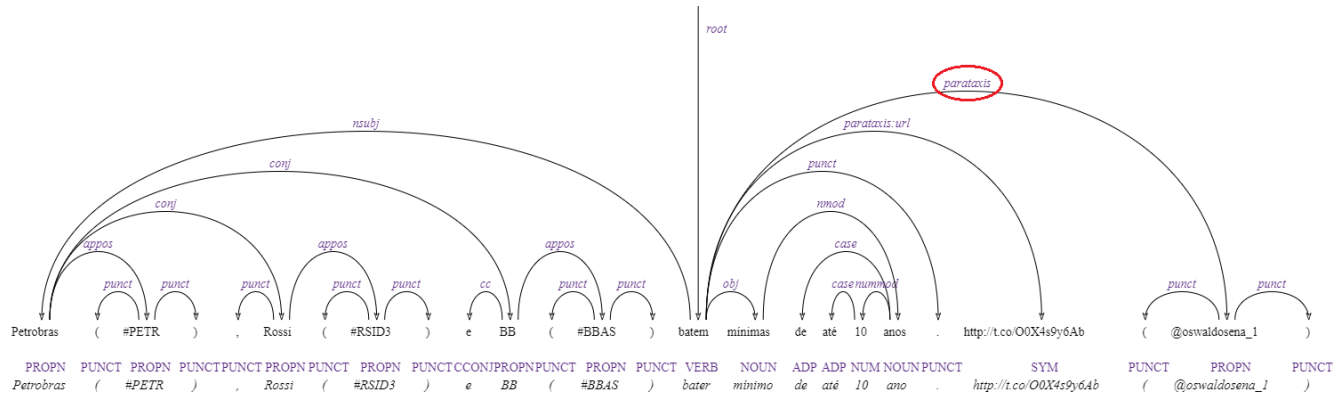


Figura 17 – Menção com **parataxis** no final do *tweet* e entre parênteses.

3. URL: endereço *web*

Uma URL tem sempre PoS SYM e pode ocorrer integrada à sintaxe ou *standalone*.
Se integrada, pode ocorrer precedida de preposição (Figura 18) ou de dois-pontos (Figura 19).
Nesses casos, a URL deve ser conectada a um *head* por **obl**.

Exemplo:

(17) Publiquei estudo de a #HGTX3 em o gráfico diário . Rompendo tendência de baixa ? ? ? Veja em **http://t.co/oRA4bA8Qye**

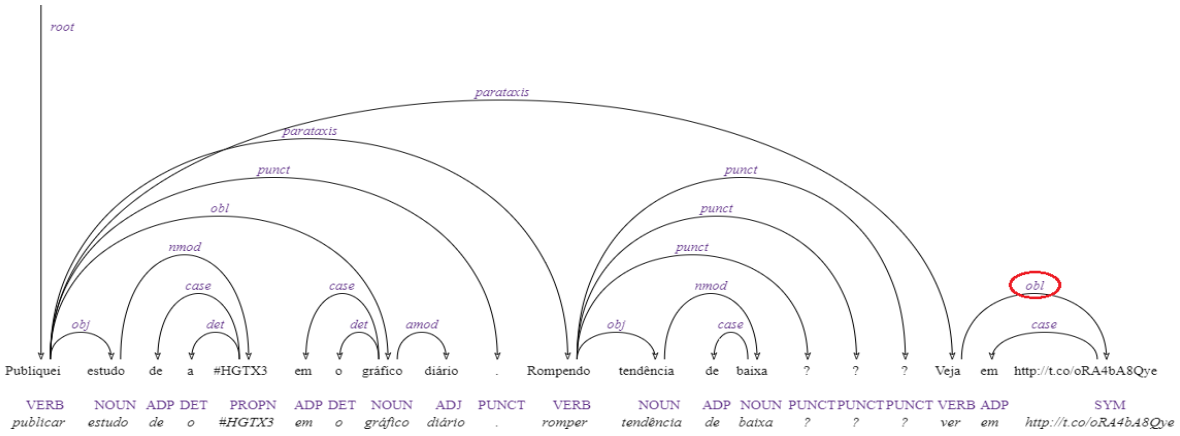


Figura 18 – URL precedida por preposição e conectada a um *head* por **obl**.

Exemplo:

(18) Em os últimos 5 pregões #CSNA3 acumulou uma baixa de 17.1 % enquanto #USIM5 - 14.8 % e #VALE5 - 10 % . Veja o ranking : **http://t.co/XjRazUAN9b**

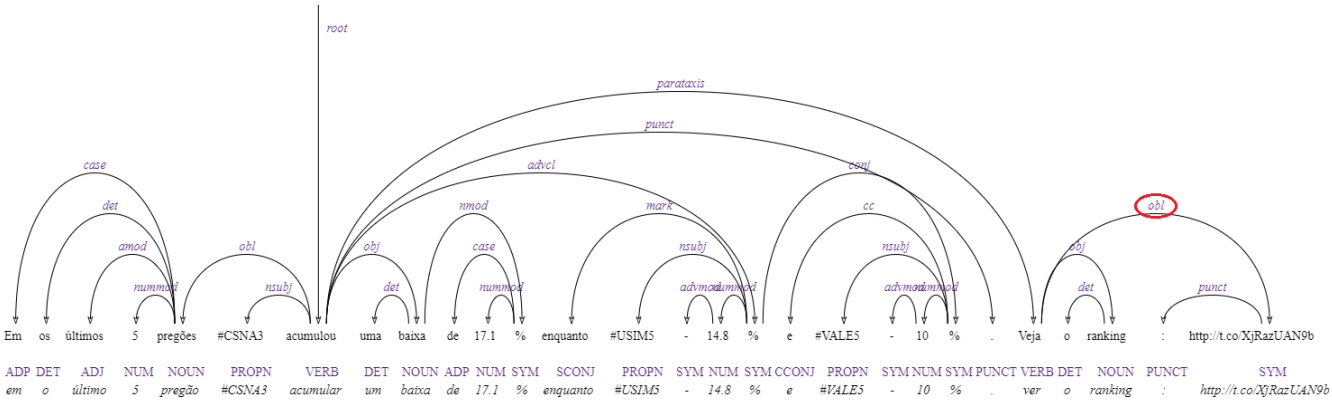


Figura 19 – URL precedida por dois pontos e conectada a um *head* por **obl**.

Ainda quando integrada, uma URL pode completar o sentido de um verbo, comumente no imperativo, ligando-se, portanto, ao seu *head* por meio da *deprel* apropriada.

Exemplo:

(19) Hoje , a as 9h , o Bradesco realizará teleconferência sobre os resultados de o 1T14 . Acesse **http://t.co/FHE37z52Q2** e participe ! #BBDC4

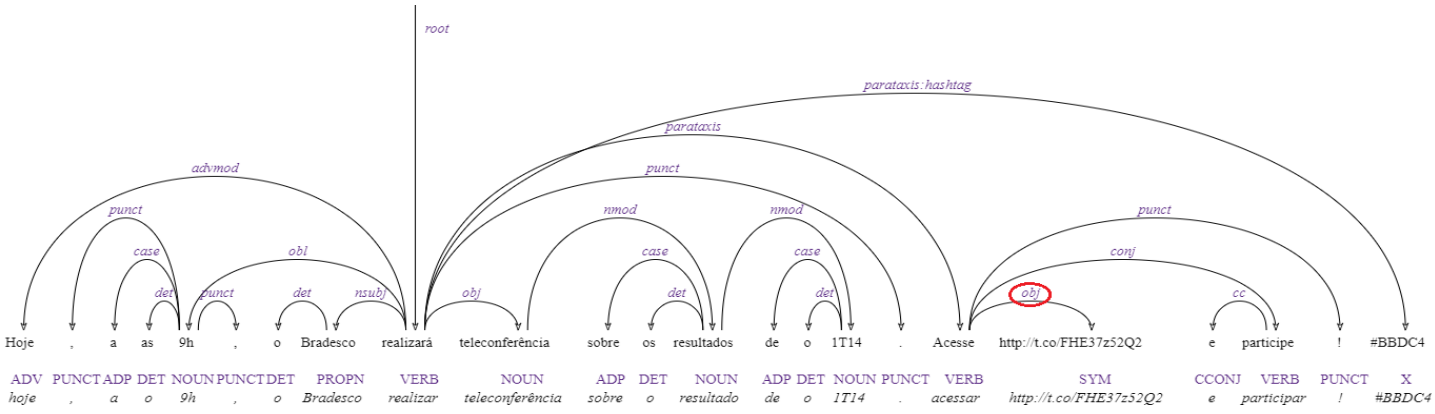


Figura 20 – URL com preposição implícita e conectada a um *head* por **obj**.

Se *standalone,* uma URL ocorre ao final dos *tweets,* podendo ser ou não precedia por um ponto final (Figura 21) ou reticências (Figura 22).
Nesses casos, deve-se conectá-la ao **root** por **parataxis**, acrescida da sub-relação **url**.

Exemplo:

(20) @garimpodeacoes $RSID3 4Q13 (N) Geração Op de Caixa forte , com desalavancagem financeira . Margens se recuperam YoY . **http://t.co/CC8MGagICJ**

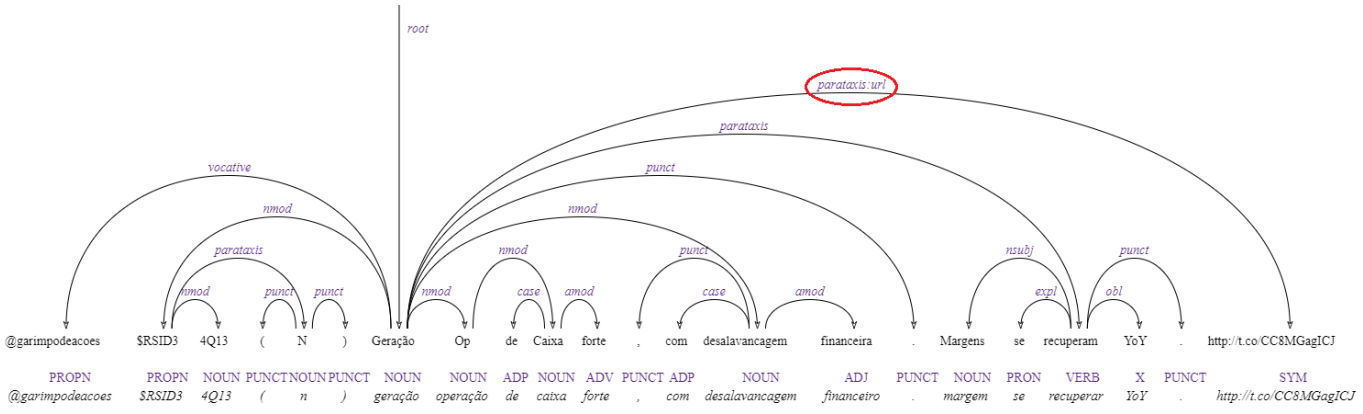


Figura 21 – URL *standalone* precedida de ponto final e conectado por **parataxis:url** (exemplo 20).

Exemplo:

(21) Petrobrás Pn (Petr4) , Gráfico Diário . Ação registr ... **http://t.co/Zsml5piTaT**

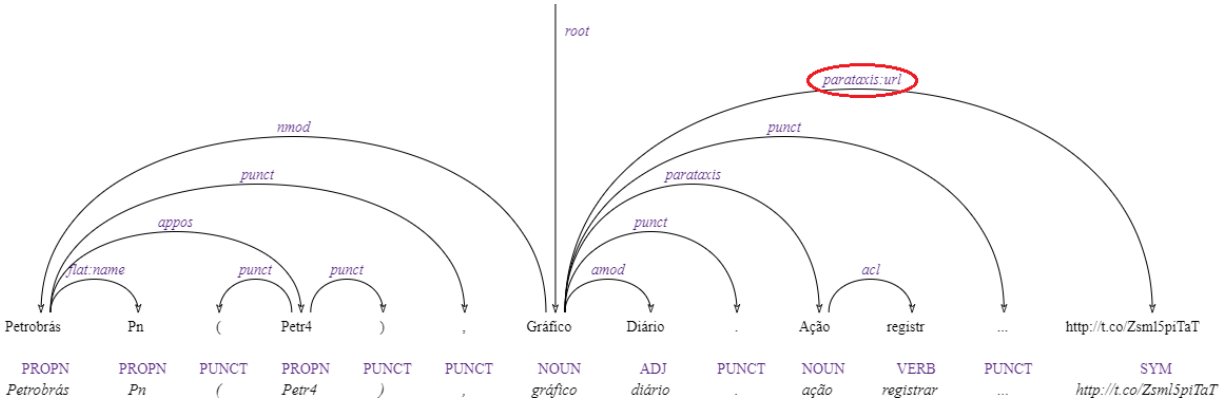


Figura 22 – URL *standalone* precedida de reticências e conectada por **parataxis:url**. (exemplo 21).

Exemplo:

(22) #BBAS3 #BBDC4 RT @passagensaereas : Banco de o Brasil e Bradesco fazem parceria para criar novo programa de fidelidade **http://t.co/zeQF7pMYMo**

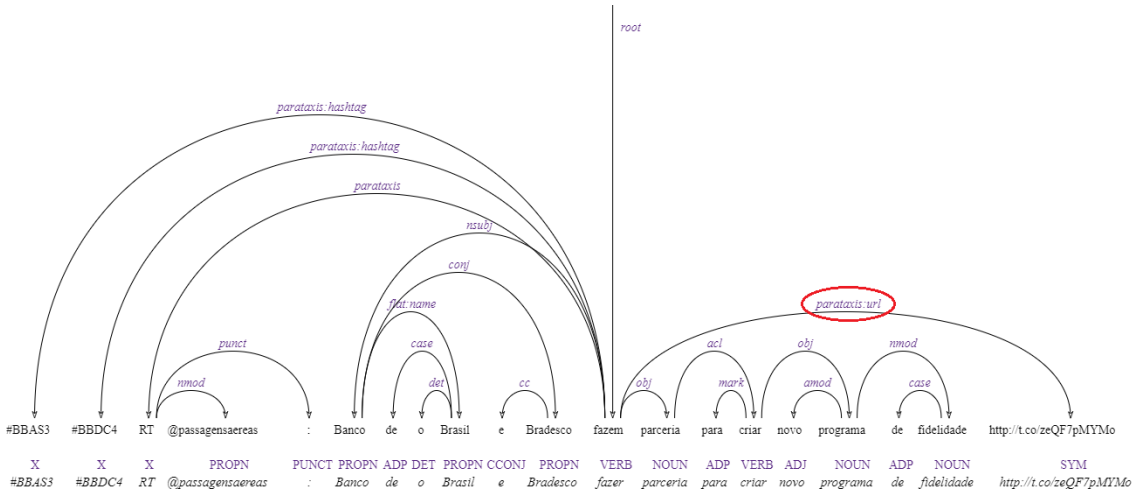


Figura 23 – URL *standalone* não precedida de pontuação e conectada por **parataxis:url** (exemplo 22).

4. Hashtag

As *hashtag* são *tokens* comumente compostos por uma palavra precedida pelo símbolo "#" (como #petr4). Elas são típicas da linguagem das redes sociais, podendo ocorrer integradas ou *standalone*.

Quando integradas (PoS PROPN) devem ser anotadas com base na sua função/posição sintática. A seguir, estão as diretrizes para as diferentes funções sintáticas (*deprel*) que as *hashtags* podem ter quando integradas à sintaxe do *tweet*, a saber: **root, nmod, nsubj, obj, obl, appos** e **vocative.**

4.1. Root: raiz

Exemplo:

(23) **#petr4** stop se fechar acima 17,775 e romper

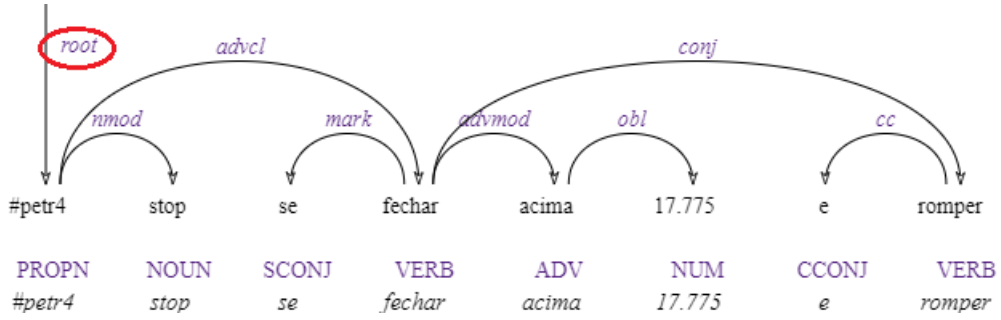


Figura 24 – *Hashtag* anotada com **root**.

4.2. Nmod: modificador nominal

Exemplo:

(24) Lembra de o sinalzinho de fundo de a **#petr4** ?

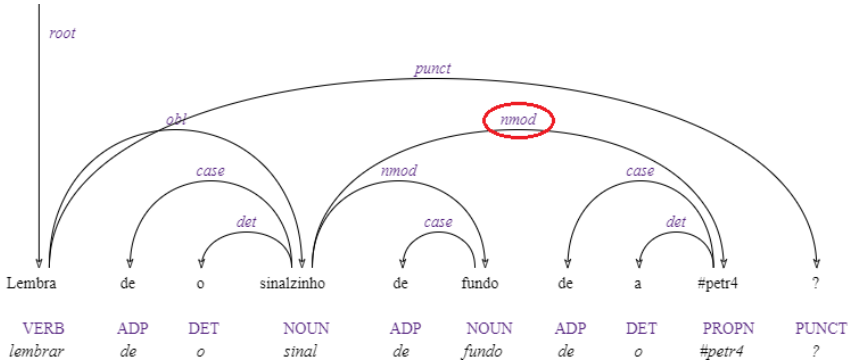


Figura 25 – *Hashtag* anotada com **nmod**.

4.3. Nsubj: sujeito

Exemplo:

(25) RT @Live_Trade **: #petr4** cumpriu certinho e ...

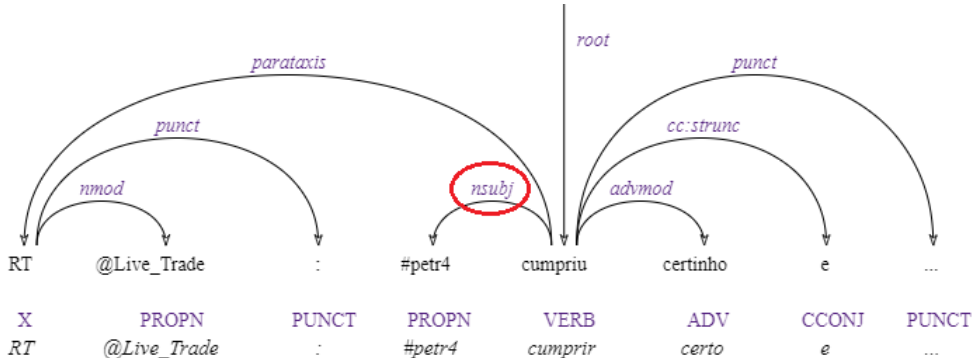


Figura 26 – *Hashtag* anotada com **nsubj**.

4.4. Obj: objeto direto

Exemplo:

(26) Agora entendem pq eu estava tranquilo comprando **#petr4** #BBAS3 e #vale5 Agora ficou desenhado para os q não entendiam .

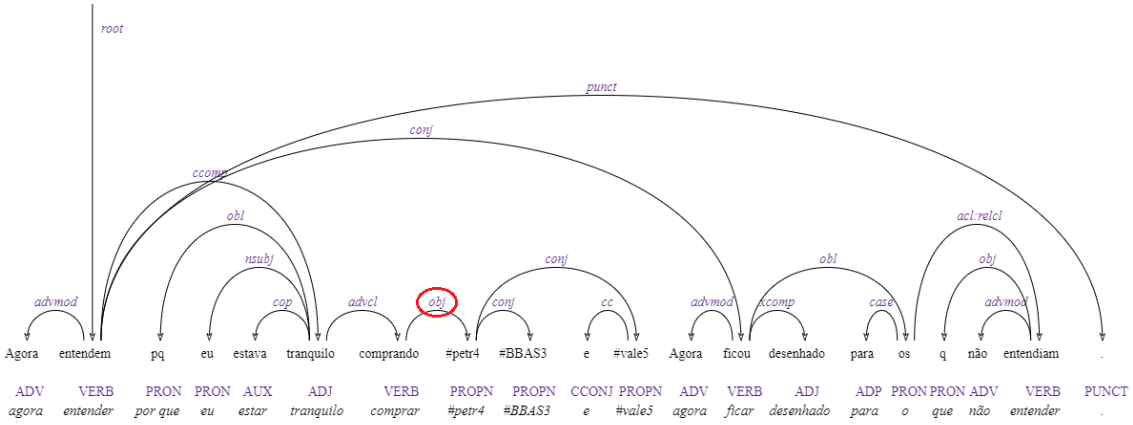


Figura 27 – *Hashtag* anotada com **obj**.

4.5. Obl: nominal oblíquo

Exemplo:

(27) Entrei em **#petr4** , com projeção de crescimento . #hype3 ta maravilhosa , continua em a tendência de alta .

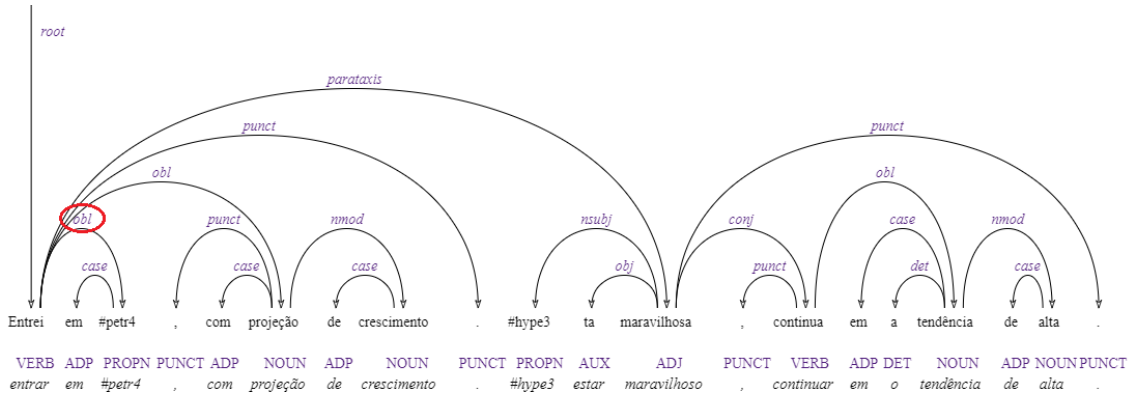


Figura 28 – *Hashtag* anotada com **obl**.

4.6. Appos: modificador apositivo

Exemplo:

(28) @Live_Trade Bom dia meu amigo , e o KING KONG (**#petr4**) , sabe me dizer suporte e resistência ? abraço

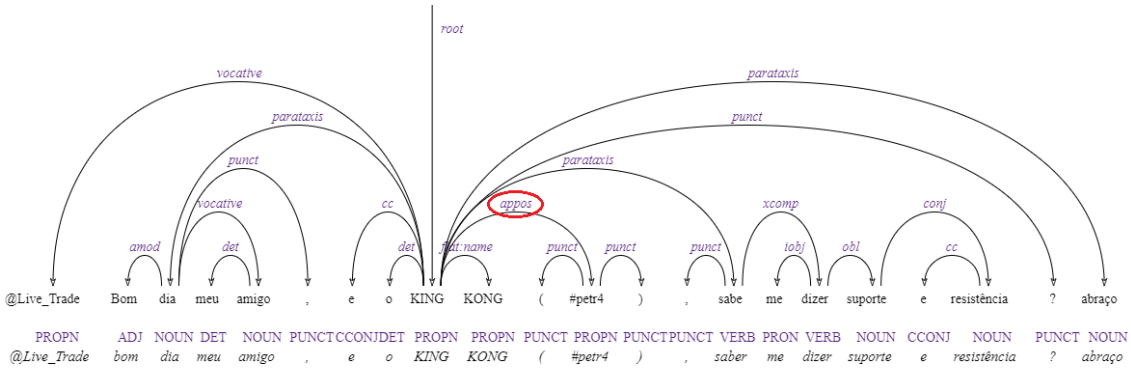


Figura 29 – *Hashtag* anotada com **appos** (exemplo 28).

Exemplo:

(29) @lambari_trader E não é que a Petrobras **#petr4** visitou os R$ 12 … . ! ! !

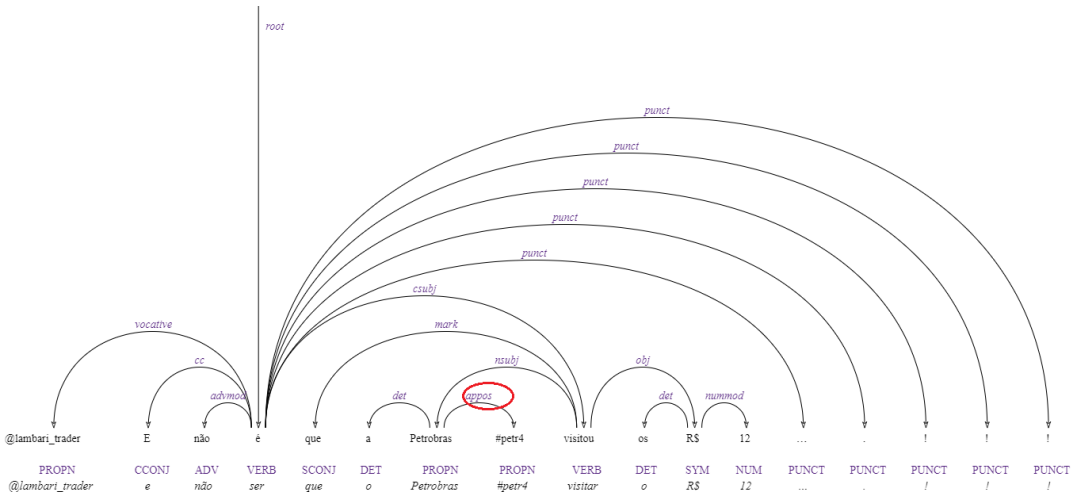


Figura 30 – *Hashtag* anotada com **appos** (exemplo 102).

4.7. Vocative: vocativo

Exemplo:

(30) **#vale5** vc é uma putinha mesmo .. rsrs .. em o finalzinho que vc resolve abrir as pernas .. rsrs

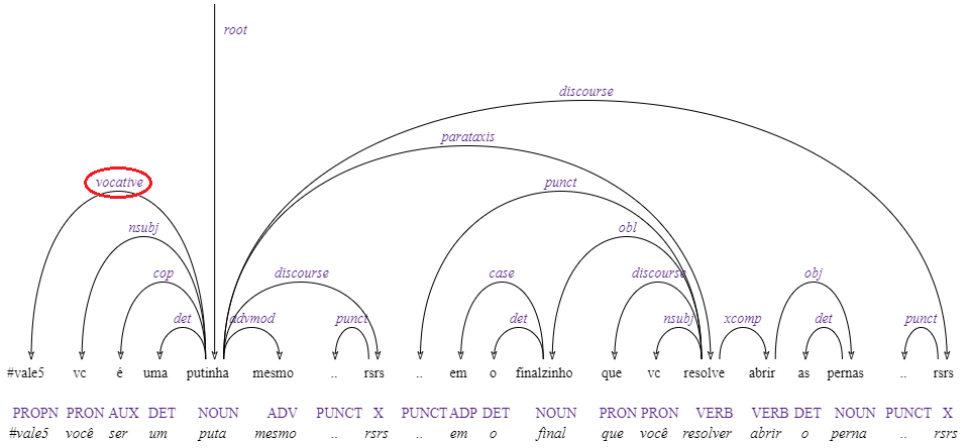


Figura 31 – *Hashtag* anotada com **vocative**.

Quando *standalone*, as *hashtags* podem ocorrer no início ou no final dos *tweets.* Nesses casos, a PoS é **X** e, no nível sintático, devem ser conectadas a um *head* por **parataxis**, acrescida pela sub-relação **:hashtag**. Caso haja mais de uma *hashtag,* cada uma deve ser conectada ao *head*.

Exemplo:

(31) **#PETR4** RT @valor_economico : Câmara derrota governo e aprova comissão para investigar Petrobras http://t.co/p37D1t2ARU

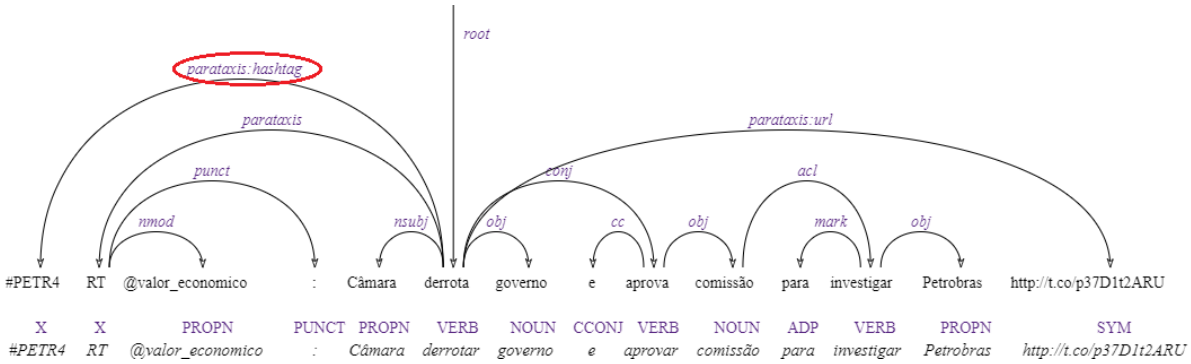


Figura 32 – *Hashtag standalone* no início do *tweet* e anotada com **parataxis:hashtag** (exemplo 31).

Exemplo:

(32) **#IBOV** **#KLBN4** Klabin fecha contrato com a Pöyry para Projeto Puma : http://t.co/WPqvXaEQB2

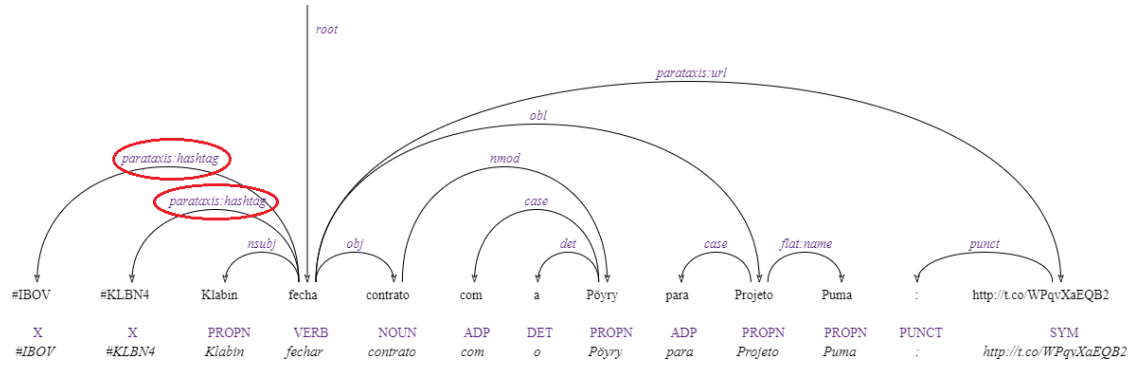


Figura 33 – Duas *hashtags standalone* no início e anotadas com **parataxis:hashtag** (exemplo 32).

Exemplo:

(33) Agora é oficial ... Ministério de a Fazenda publicou portaria que aumenta a tributação de bebidas frias . **#ABEV3**

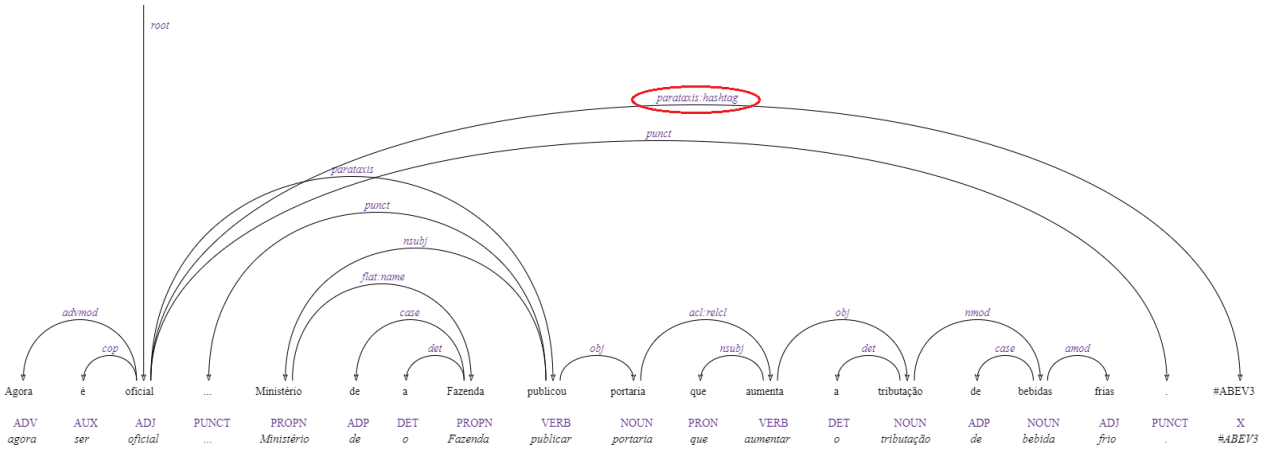


Figura 34 – *Hashtag standalone* no final do *tweet* anotada com **parataxis:hashtag** (exemplo 33).

5. Cashtag

As *cashtags* são *tokens* comumente compostos por uma palavra precedida pelo símbolo “$” (como $petr4). Elas são características da linguagem das redes sociais no domínio financeiro, podendo ocorrer integradas ou *standalone*.

Quando integradas (PoS PROPN) devem ser anotadas com base na sua função/posição sintática. A seguir, estão as diretrizes para as diferentes funções sintáticas (*deprel*) que as *cashtags* podem ter quando integradas à sintaxe do *tweet*, a saber: **root, nmod, nsubj, obj** e **appos.**
Diferentemente das *hashtags,* as *cashtags* <u>não ocorrem</u> como **vocative** e **obl.**

5.1. Root: raiz

Exemplo:

(34) **$PETR4** R$ 13,11 - 1,58 % / $VALE5 R$ 25,9 - 2,12 % , ' @Info_BMFBOVESPA : #Ibovespa fecha em baixa de 0,91 % a os 45.443 pontos . '

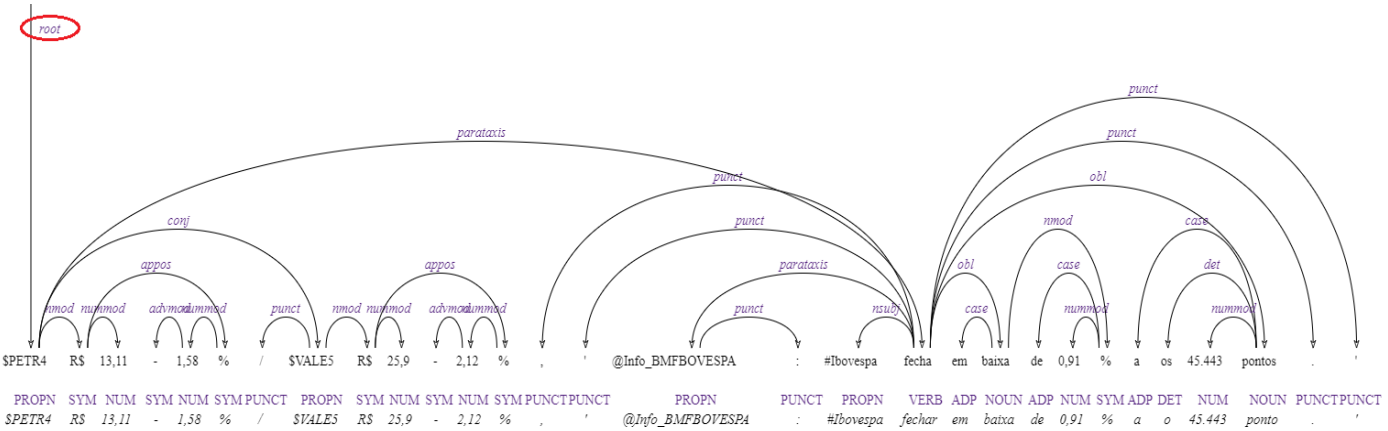


Figura 35 – *Cashtag* anotada com **root**.

5.2. Nmod: modificador nominal

Exemplo:

(35) **$PETR3** - Atencao Para O Preco Petrp73 (petr3) http://t.co/frle2I3TBF

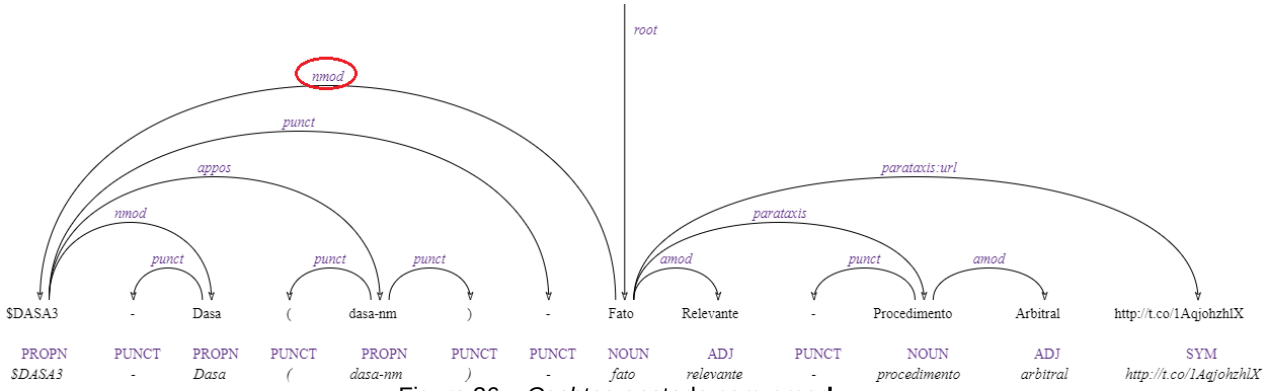


Figura 36 – *Cashtag* anotada com **nmod**.

5.3. Nsubj: sujeito

Exemplo:

(36) **$PETR4** termina o dia com um belo doji , mas acima de a 21EMA ... repique para manhã ?

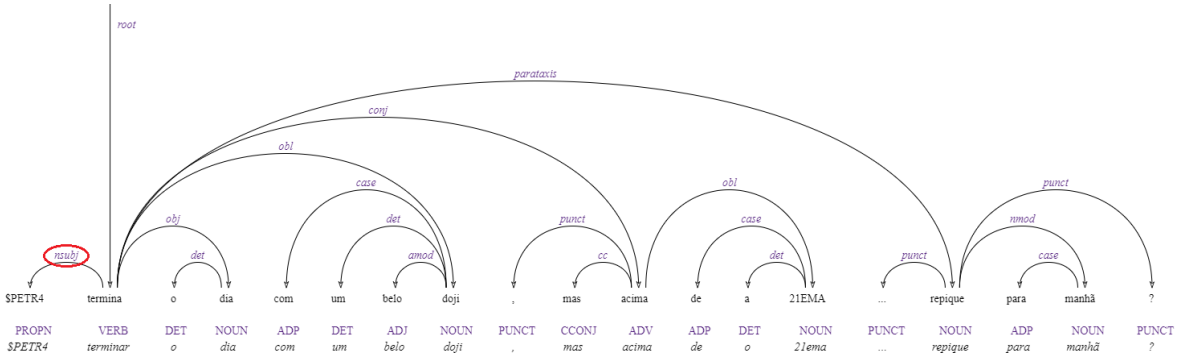


Figura 37 – *Cashtag* anotada com **nsubj**.

5.4. Obj: objeto direto

Exemplo:

(37) Diz[9] bom de comprar **$ABEV3** . Caindo 5 % com notícia de taxação . Eu vou em essa .

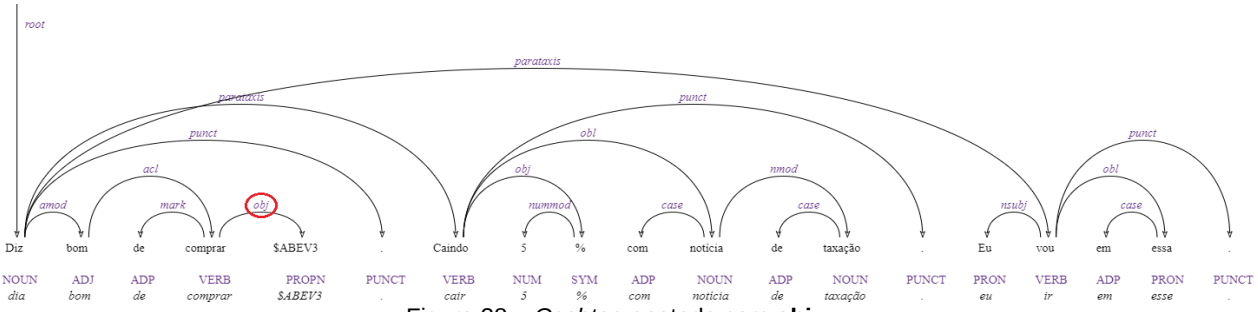


Figura 38 – *Cashtag* anotada com **obj**.

[9] *Tokens* com desvio ortográfico (ou erro de digitação) não foram alterados no córpus, mas seus traços morfossintáticos (incluindo PoS e lema) estão descritos em função de sua forma correta, a qual, aliás, também é indicada na coluna MIRC do CoNLL-U pelo traço adicional CorrectForm. Na anotação sintática, assume a função da forma correta no tweet.

Quando *standalone*, as *cashtags* podem ocorrer no início ou no final dos *tweets.* Nesses casos, a PoS é **X** e, no nível sintático, devem ser conectadas ao **root** com a *deprel* **parataxis:cashtag**.

Exemplo:

(38) **$PETR3** - Atencao Para O Preco De Petrp73 (petr3) http://t.co/Qd9pmmLVNw

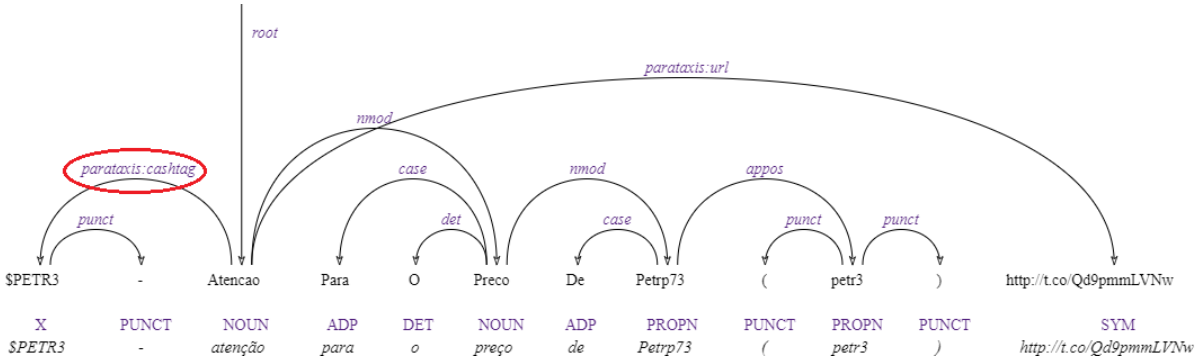


Figura 39 – *Cashtag standalone* anotada com **parataxis:cashtag**.

2.6. Truncamento

Um *tweet* pode apresentar (i) palavras ou (ii) estruturas (sentenças ou sintagmas) quebradas, principalmente devido ao limite de caracteres impostos pela plataforma *Twitter.*
Os truncamentos ocorrem majoritariamente seguidos por reticências ("...").
Para anotar os diferentes tipos de truncamentos, adotam-se as sub-relações :**wtrunc** (do inglês, *word truncation*) e :**strunc** (do inglês, *structure truncation*) para a *deprel* **trunc**.
Nos *tweets* do mercado financeiro, os truncamentos ocorrem em diferentes cenários, para os quais as diretrizes de anotação são as seguintes:

1. A relação que chega no predicado principal do trecho truncado recebe a sub-relação **:strunc**. Se a última palavra do trecho for truncada, ela recebe **:wtrunc**, a menos que ela seja o predicado principal.

   Exemplo:

(39) O que há de melhor em a Bovespa : as ações mais indicadas por os analistas : A Vale ficou em **prim** ... http://t.co/tkdUiSqQUs #infomoney #vale5

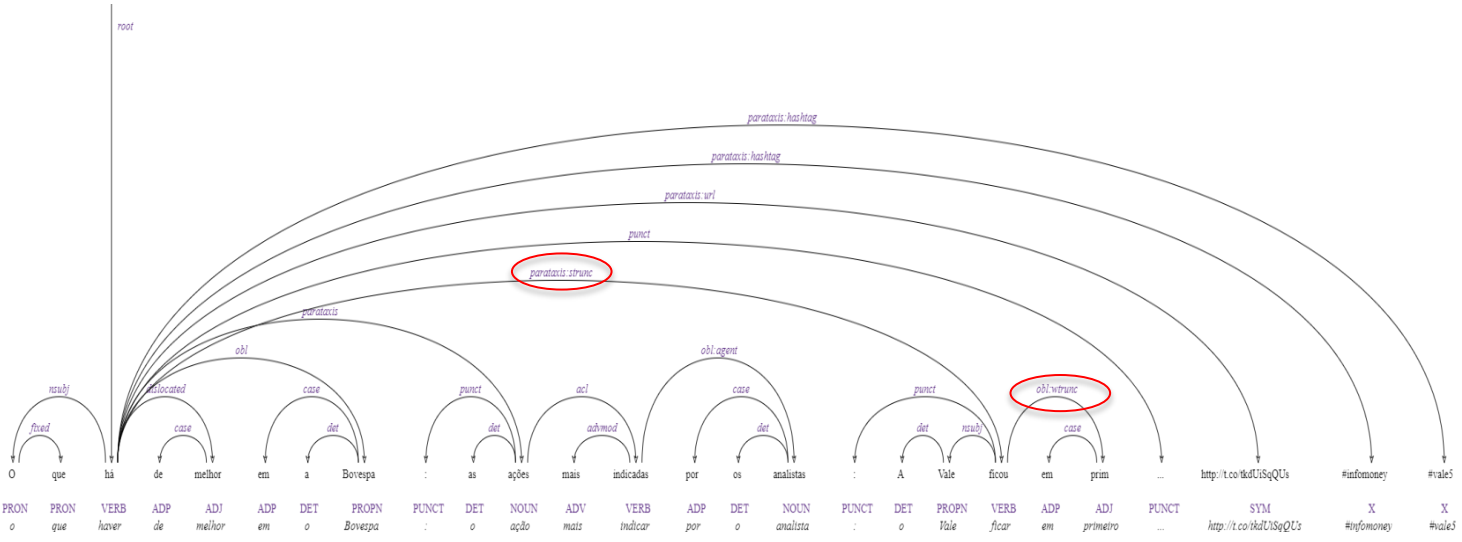


Figura 40 – Relação que chega no predicado principal do trecho truncado com sub-relação **:strunc** e última palavra do trecho truncado com **:wtrunc**.

Exemplo:

(40) Elétricas sofrem ' apagão ' e levam junto siderúrgicas , 11 ações caem mais de 4 % : Apenas 9 **açõ** ... http://t.co/UfrAbJS2fS #infomoney #vale5

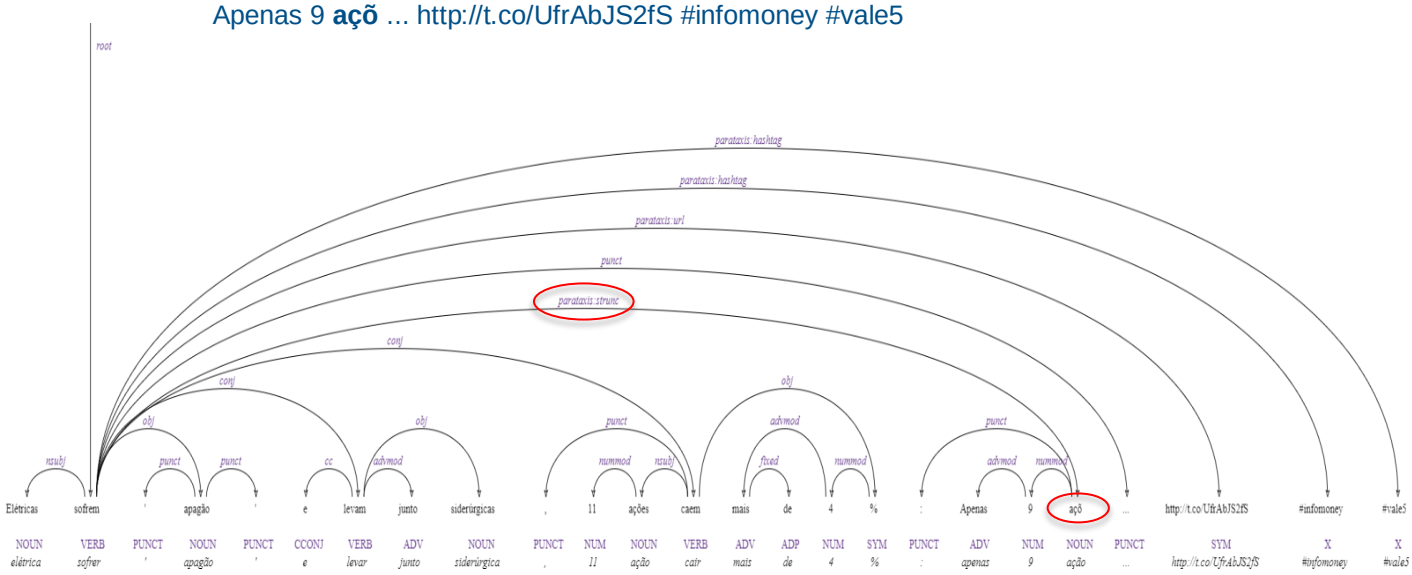


Figura 41 – *Head* truncado de trecho truncado: relação da qual o *head* é dependente tem sub-relação **:strunc**.

2. Se as reticências (“...”) indicarem um truncamento estrutural, mas o texto/trecho estiver sintaticamente completo, sugere-se, nesse caso, não usar a sub-relação **:strunc**.

Exemplo:

(41) Notas gerais A VALE (VALE5) pretende pagar , como primeira parcela de remuneração mínima a os acionistas em 2014 , **...** http://t.co/RB6XViLTFD

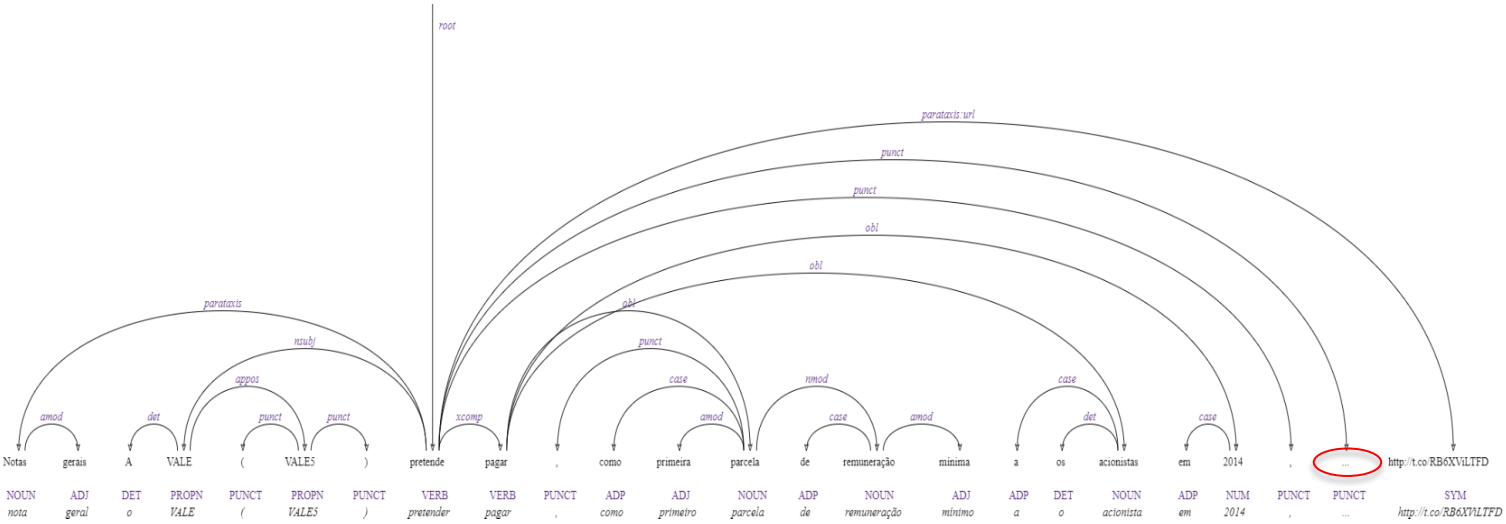


Figura 42 – Truncamento estrutural sem quebra de conteúdo não requer sub-relação **:strunct** (exemplo 41).

Exemplo:

(42) Notas gerais A PETROBRAS (PETR4) demitiu o diretor financeiro de a subsidiária Petrobras Distribuidora , Nestor **...** http://t.co/CMZD46pQOq

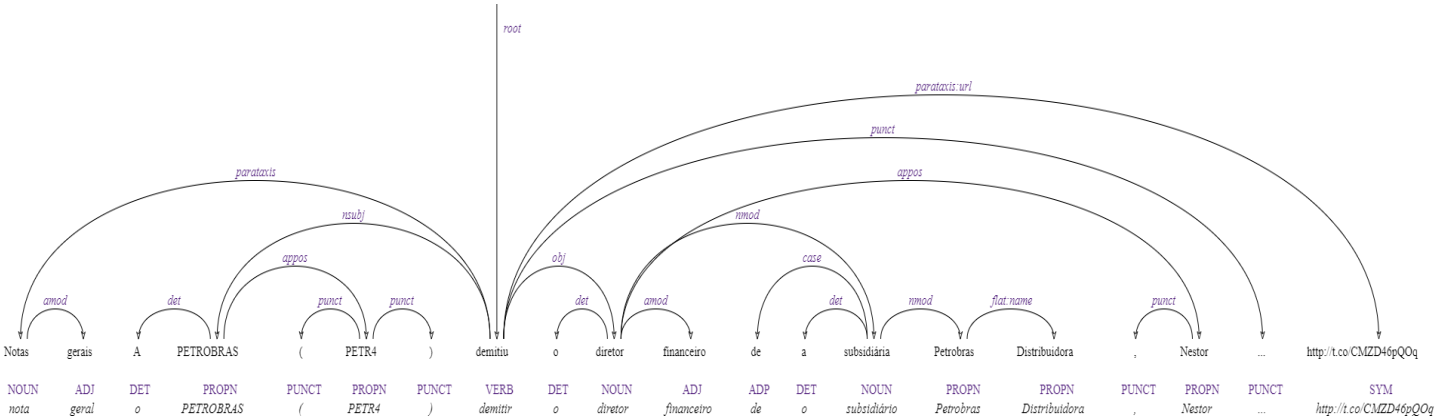


Figura 43 – Truncamento estrutural sem quebra de conteúdo não requer sub-relação **:strunct** (exemplo 116).

2.7. Índice de (des-)valorização das ações

Segundo Di Felippo *et al* (2021), os *tweets* sobre o mercado financeiro apresentam os chamados "índices de (des-)valorização das ações", como **- 1,3 %**, que são expressões sempre compostas pela seguinte sequência de PoS: SYM, NUM, SYM.
A anotação de *deprel* entre esses 3 elementos segue as diretrizes (exemplos 43):

1. símbolo (PoS SYM) "%" é *head*
2. valor numérico (PoS NUM) "1,3" é dependente de "%" por **nummod**
3. símbolo "-" (PoS SYM) é dependente de % por **advmod**

Os índices ocorrem em dois contextos sintáticos distintos, sendo que em cada um deles o *head* ("%") será dependente de outro *token* por meio de *deprels* diferentes.
Um índice pode aparece inserido em contexto nominai, sem que um verbo ocorra ou seja inferido. No *tweet* do exemplo (43) a expressão nominal "maiores altas (seguida de dois pontos)" possui o **root** ("altas") e a sequência de índices ocorre depois dos dois pontos.
Nesses casos, o *head* do índice ("%") é **nmod** do *ticker* que o precede e o primeiro *ticker* da sequência é dependente por **appos** do **root.**

Exemplo:

(43) Maiores Altas : **LLXL3 + 7,79 % | OIBR4 + 6,14 % | JBSS3 + 6,11 % | RSID3 + 5,67 % | ELET3 + 4,83 %** .

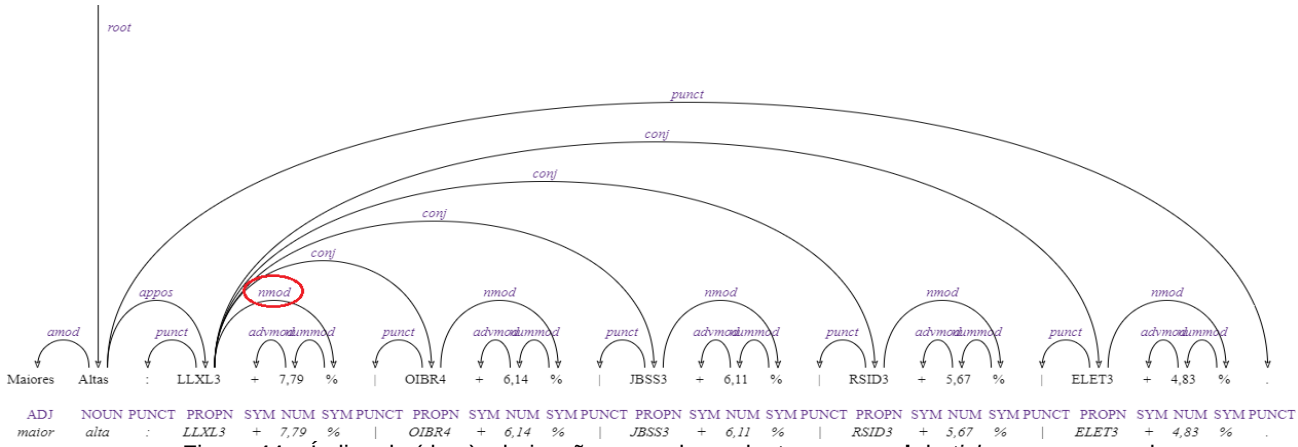


Figura 44 – Índice de (des-)valorização como dependente por **nmod** do *ticker* que o precede.

Em outro contexto, um índice pode aparecer em contexto verbal, com verbo explícito ou implícito. Em (44), por exemplo, pode-se inferir que a sentença inicial seja “Ações da VALE também estão em alta”, sendo que cada índice herda o predicado inferido, como “Vale3 está em alta de...”.
Diante dessa interpretação, o **root** do *tweet* é o *token* “alta”, sendo o *head* do índice (“%”) dependente por **orphan**[10] do *ticker* que o precede, pois o símbolo “%” e o *ticker* ficam “órfãos” de *head* em função do predicado elíptico.
O primeiro *ticker* da sequência, nesse caso, é dependente por **parataxis** do **root.**

Exemplo:

(44) Ações de a VALE também em alta . Vale3 , + 2,1 % e Vale5 , + 1,6 ...

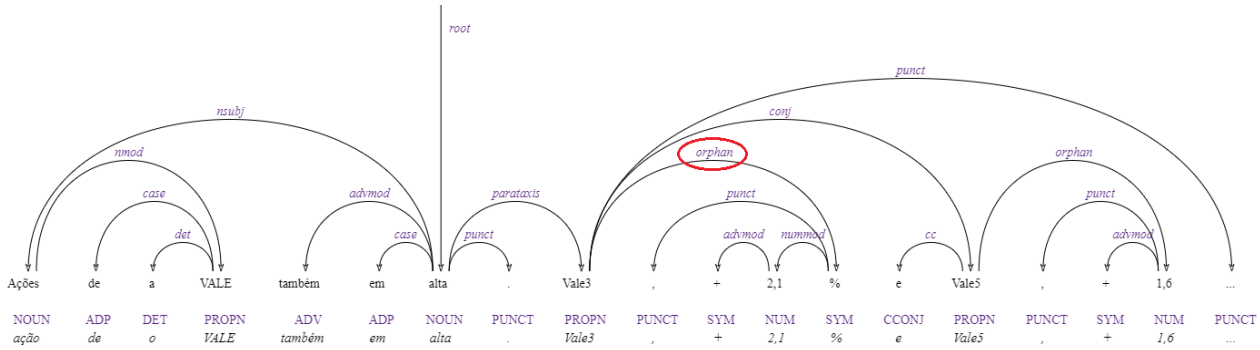


Figura 45 – Índice de (des-)valorização como dependente por **orphan** do *ticker* que o precede.

2.8. Emoticon e emoji

Os *emoticons* (como “:)”, “:|”, “o.O” e “:D”) e *emojis* (como 🔥) só ocorrem *standalone* no *corpus* que deu origem a este manual.
Nesses casos, *emoticons* e *emojis* devem ser conectados ao **root** por **discourse**, pois não têm uma relação clara com a estrutura sintática do *tweet*, exceto de maneira expressiva.
A identificação do *head* adequado, no entanto, é contextual. No caso da Figura 46, por exemplo, a ocorrência das aspas indica que a sequência de *emojis* se refere ao *head* (“enviada”) do trecho entre aspas e não ao **root** do *tweet.*
Para sequências de um mesmo *emoticon* ou *emoji* (exemplo 45), cada um deles deve ser conectado ao seu *head* por **discourse**.

Exemplo:

(45) Vendido ! **:)** “ @felipefdeaguiar : #ELPL4 ordem enviada R$ 9,30 vendo ! 🔥 🔥 🔥 ”

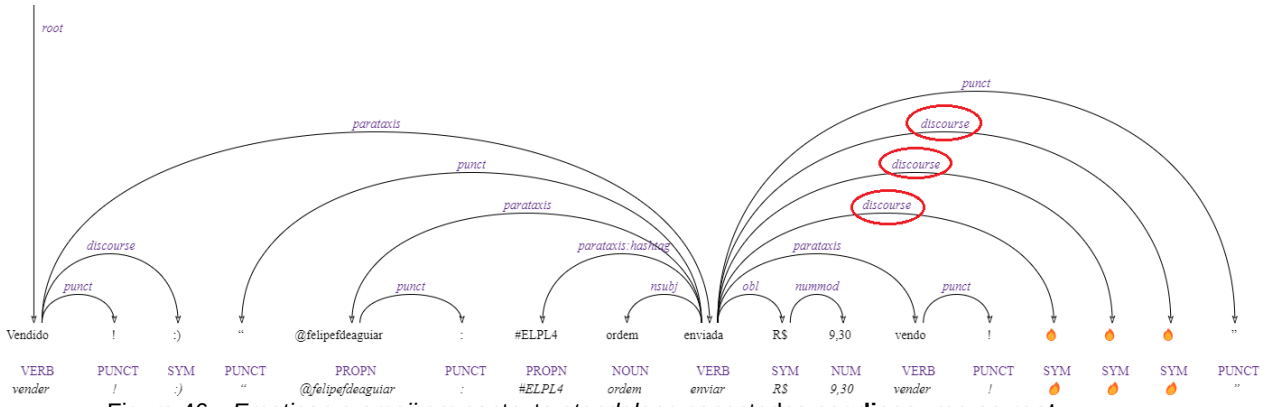


Figura 46 – *Emoticon* e *emoji* em contexto *standalone* conectados por **discourse** ao **root.**

[10] A *deprel* **orphan** (do inglês, *orphaned dependente*) liga dois elementos que ficaram “órfãos” de *head*, em função da elipse do *head* que tinham em comum. Essa relação é usada tipicamente quando há elipse de um predicado e pelo menos duas palavras de conteúdo que se ligariam a esse predicado (DURAN, 2022).

2.9. Onomatopeias

As onomatopeias de risos (como hehehe, kkkkk, haha e outras) (PoS X) nos *tweets* do mercado financeiro ocorrem em contexto *standalone* e, assim como *emojis/emoticons,* relacionam-se à estrutura sintática do *tweet* apenas de maneira expressiva, sendo, portanto, conectadas ao **root** pela *deprel* **discourse**.

Exemplo:

(46) #Vale5 Opções é caça niquel , igual tem em os botecos **kkk**

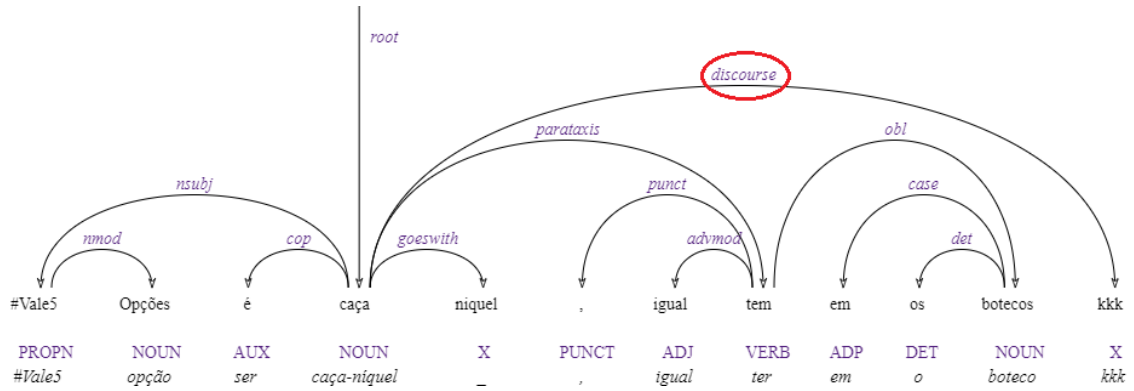


Figura 47 – Onomatopeia *standalone* conectada por **discourse** (exemplo 46).

Exemplo:

(47) #petr4 King Kong me acordem qdo bater em 12,50 que tenho interesse ... **rsrsr**

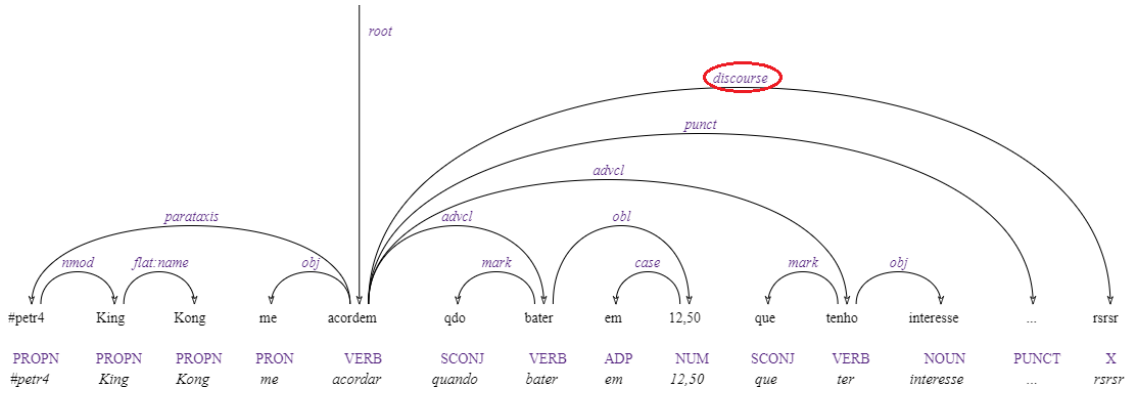


Figura 48 – Onomatopeia *standalone* conectada por **discourse** (exemplo 47).

2.10. Repetição de sinal de pontuação

Para sequências de um mesmo sinal de pontuação (como “? ? ?”), cada um deles deve ser conectado ao seu *head* por meio de **punct**.

Exemplo:

(48) @Live_Trade Marcos , projeta algum fundo para OIBR4 **? ? ?**

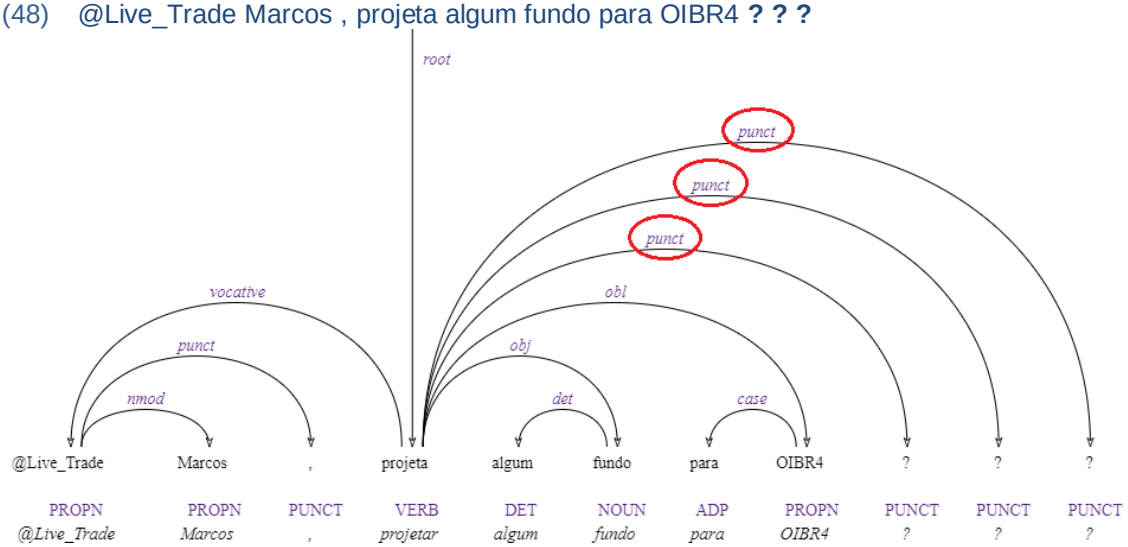


Figura 49 – Cada sinal de pontuação repetido em sequência conectado por **punct** ao *head* (exemplo 48).

2.11. Substituição lexical por símbolo (SYM)

Repetições em sequência do cifrão “$” podem ocorrer integradas à sintaxe ou *standalone.* Quanto integradas, ocorrem em substituição à palavra “dinheiro”, com função de objeto direto. Nesses casos, a opção foi por conectar apenas o primeiro cifrão da sequência por **obj** ao *head* e os outros símbolos da sequência por **goeswih**[11] ao primeiro.

Exemplo:

(49) Quando Lula mandou trabalhador aplicar **$ $** de o FGTS em a #PETR4 , ações valiam R$ 24 . Hoje valem R$ 15,5 e contando para baixo : - 34 % !

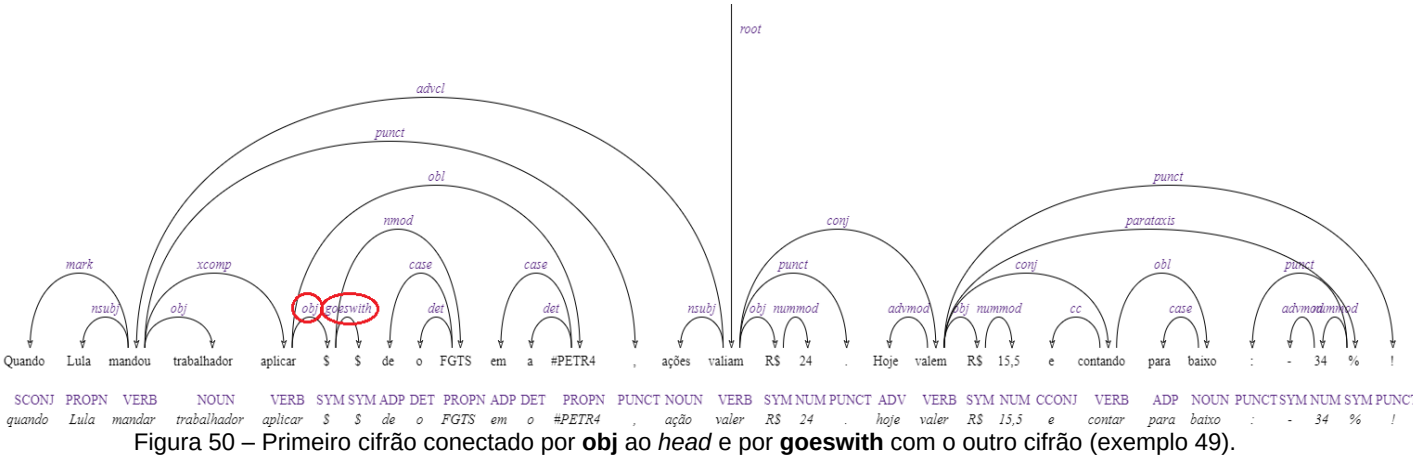


Figura 50 – Primeiro cifrão conectado por **obj** ao *head* e por **goeswith** com o outro cifrão (exemplo 49).

[11] Nos *tweets* originais, os cifrões repetidos em sequêcia formam um único *token*, sem espaço em branco entre eles (p.ex.: “&&”). Assim, optou-se por usar **goeswith**, pois, segundo Duran (2022), essa *deprel* pode ser empregada para conectar elementos que foram *tokenizadas* indevidamente, como parece ser o caso.

Quando *standalone,* as repetições sequenciais do cifrão devem ser conectadas a um *head* por **discourse**, pois não têm uma relação clara com a estrutura sintática do *tweet*, exceto de maneira expressiva. Nesses casos, cada um dos símbolos da sequência deve ser conectado ao *head* por **discourse.**

Exemplo:

(50) #PETR4 indo p picas @SakaSakamori : Denúncia . Estrangulamento **$ $ $** de a Petrobras http://t.co/RMBjR9d80F Beira de o colapso ! #DilmaTemMedoDaCPI

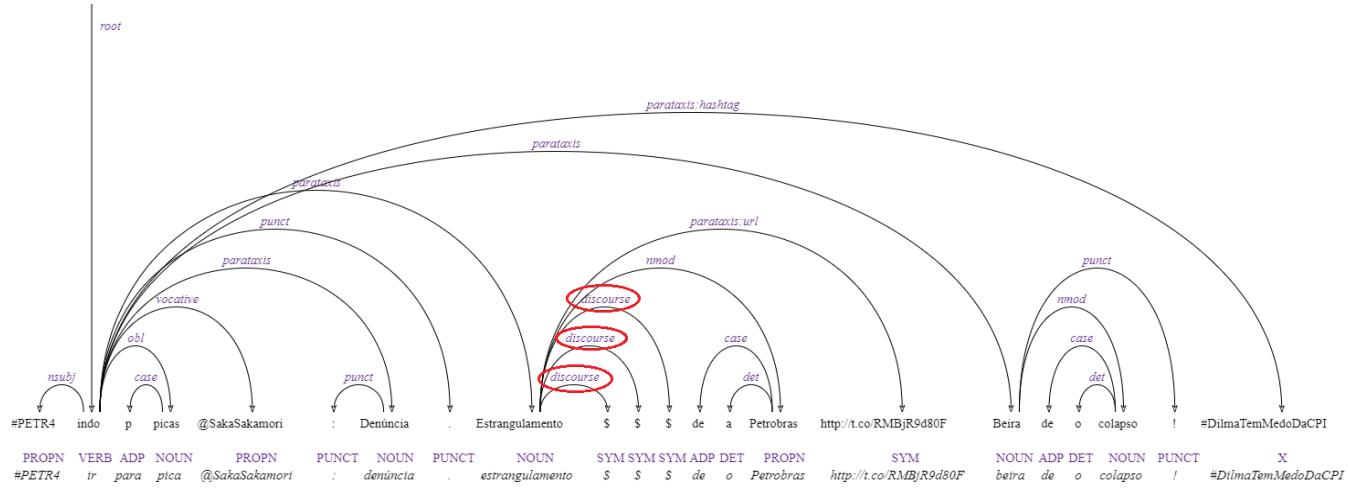


Figura 51 – Cada cifrão repetido em sequência conectado por **discourse** ao *head* (exemplo 50).

Sintaxe não-padrão

Sujeito (nsubj) separado do predicado (*root*) por dois pontos

Exemplo:

(51) @jcvolemos **VALE5 : Sentiu a resistência de 29,90 e lateralizou** . Acima de 29,90 pode chegar a 31,33 e abaixo de 29,02 pode chegar a 27,59 .

(Interpretação: "[...] **VALE5 sentiu a resistência de 29,90 e lateralizou** . [...]".)

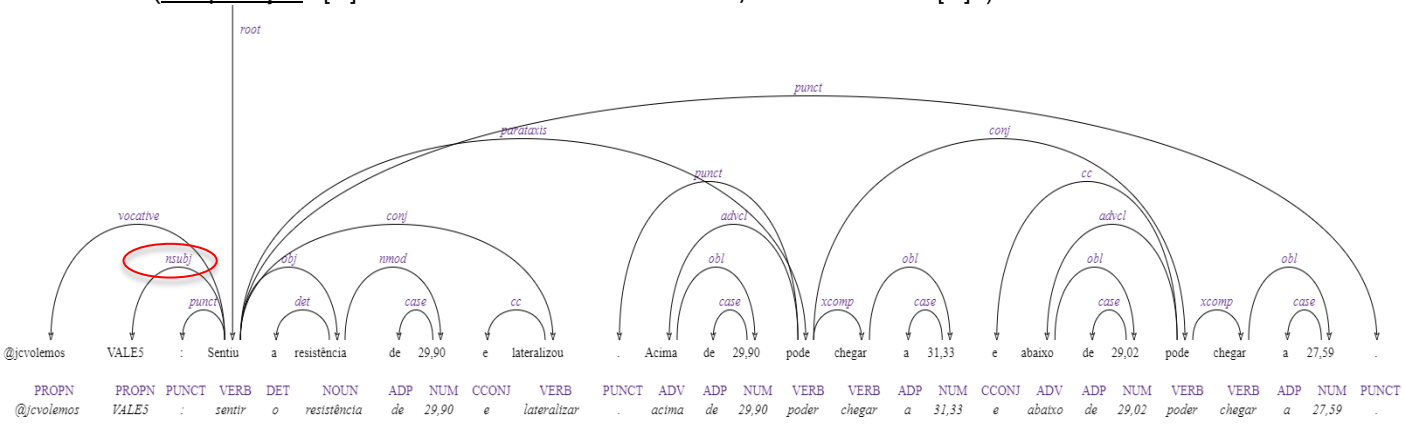


Figura 52 – Atribuição de **nsubj** com sujeito separado do **root** por dois pontos (exemplo 51).

Exemplo 2:

(52) **#CRUZ3 : pode estar revertendo tendência de baixa após romper canal** . Melhor se superar os 23.52 . #Whoknows ? http://t.co/tZVAMpMQVG

(Interpretação: “CRUZ3 pode estar revertendo tendência de baixa após romper canal. [...]”.)

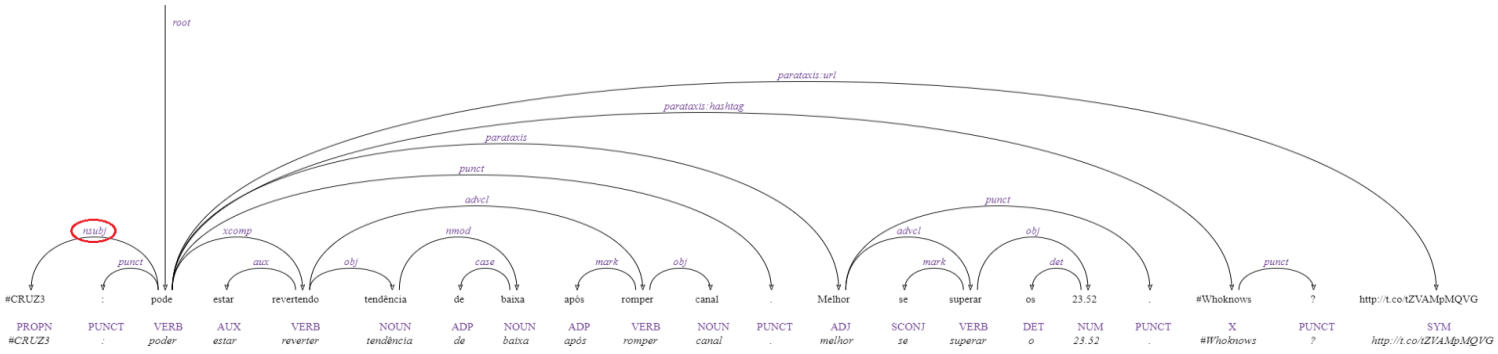


Figura 53 – Atribuição de **nsubj** com sujeito separado do **root** por dois pontos (exemplo 52).

Advcl sem predicado

Como mencionado, para que a anotação sintática (via *deprel*) de um *tweet,* é preciso, por vezes, inferir uma interpretação do *post.* Esse é o caso do trecho em negrito do *tweet* abaixo, do qual se infere a expressão “se estiver” em “(se estiver) acima de 29,90” e “(se estiver) abaixo de 29.02”, o que permite, de acordo com Duran (2022), anotar “acima_ADV (de 29,90)” como dependente por **advcl** de “pode_VERB”, pois **advcl** liga o predicado de uma oração matriz ao predicado de uma oração que a modifica.

Exemplo:

(53) @jcvolemos VALE5 : Sentiu a resistência de 29,90 e lateralizou . **Acima de 29,90 pode chegar a 31,33 e abaixo de 29,02 pode chegar a 27,59 .**

(Interpretação: [@jcvolemos VALE5 sentiu a resistência de 29,90 e lateralizou. Pode chegar a 31,33 se estiver acima de 29,90 e pode chegar a 27,59 se estiver abaixo de 29,02).

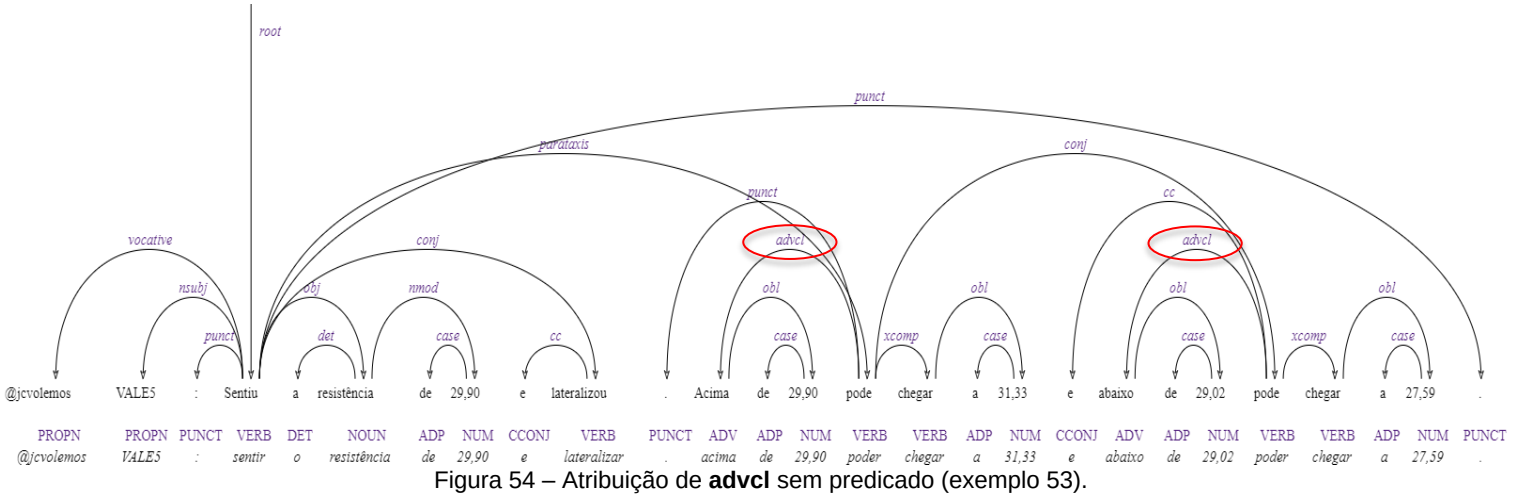


Figura 54 – Atribuição de **advcl** sem predicado (exemplo 53).

Exemplo:

(54) não rompeu, não compro RT @Live_Trade : #vale5 26,28 o de 15 que fechar acima , cp rompimrnto.

(Interpretação: Não rompeu, portanto, não compro. [RT @Live_Trade: #Vale a 26,28. O de 15 que fechar acima. Compra no Rompimento].)

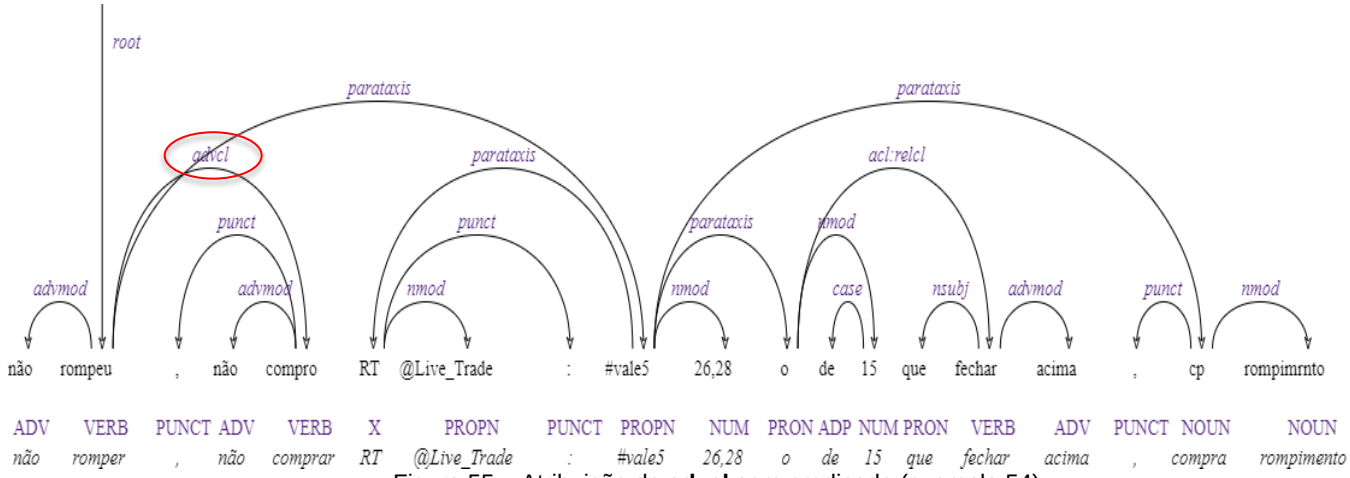


Figura 55 – Atribuição de **advcl** sem predicado (exemplo 54).

Coordenação (conj) introduzida por pontuação ou símbolo

Sinais de pontuação (PoS PUNCT), como barras inclinada (**"/"**) e barra vertical simples (**"|"**) ou dupla ("| |")) e o símbolo (PoS SYM) "X", podem funcionar como conjunção quando a intepretação do *post* sugerir que eles estejam introduzindo uma coordenação.
Nesses casos, a *deprel* entre o elemento coordenado e PUNCT é **punct**, e entre o elemento coordenado e SYM é **cc**.
Em ambos os casos, a *deprel* entre os elementos coordenados, seja por PUNCT ou SYM, é **conj.**

Exemplo (com barra inclinada):

(55) $CIEL3 – Cielo (ciel-nm) – ago/e – 31/03/14 – 10h00 – bonificacao **/** alteracao De o Estatuto Social http://t.co/ZODi2E6KmJ

(Interpretação: [...] Bonificação **e** alteração do estatuto social (decididas) na assembleia geral ordinária ou extraordinária do dia 31/03/14, às 10h00.)

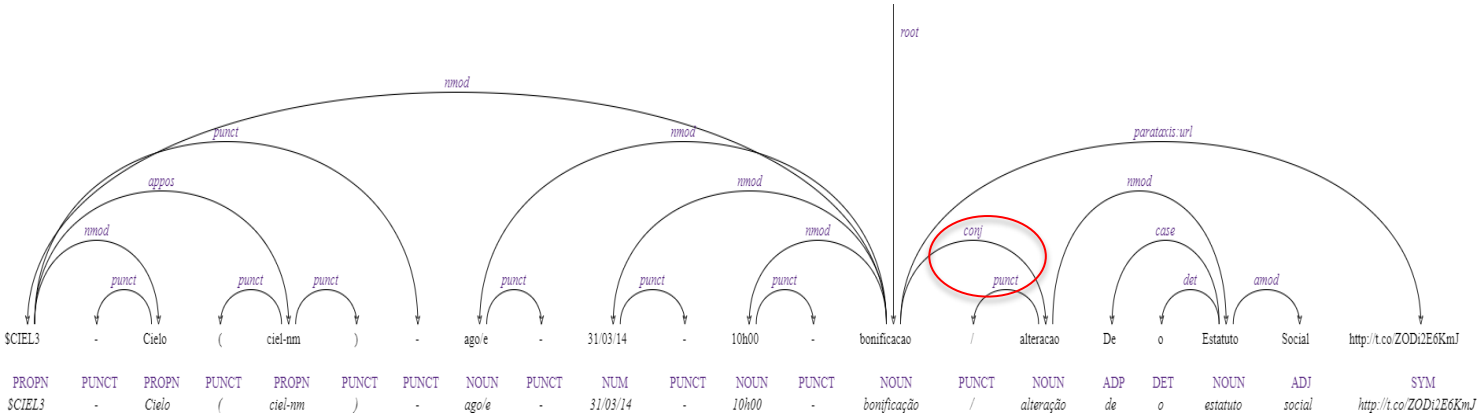


Figura 56 – Atribuição de **conj** introduzida por pontuação (barra inclinada) (exemplo 55).

Exemplo (com barra vertical simples e dupla):

(56) RT @dividendo_br : ELETROBRAS jscp | aprov 30/04/2014 **|** ex 02/05/2014 **| |** pg n/d | ELET3 R$ 0,399210837 **|** ELET5 R$ 2,178256587 **|** ELET6 R$ 1,63369244 htt …

(Interpretação: a respeito dos juros sobre o capital próprio (jscp) da ELETROBRAS : aprovação em 30/04/2014; ex em 02/05/2014; pagamento indefinido, sendo: R$ 0,399210837 para ELET3 e R$ 1,63369244 para ELET6; fonte: http://t.co/IjtHKlQlfr)

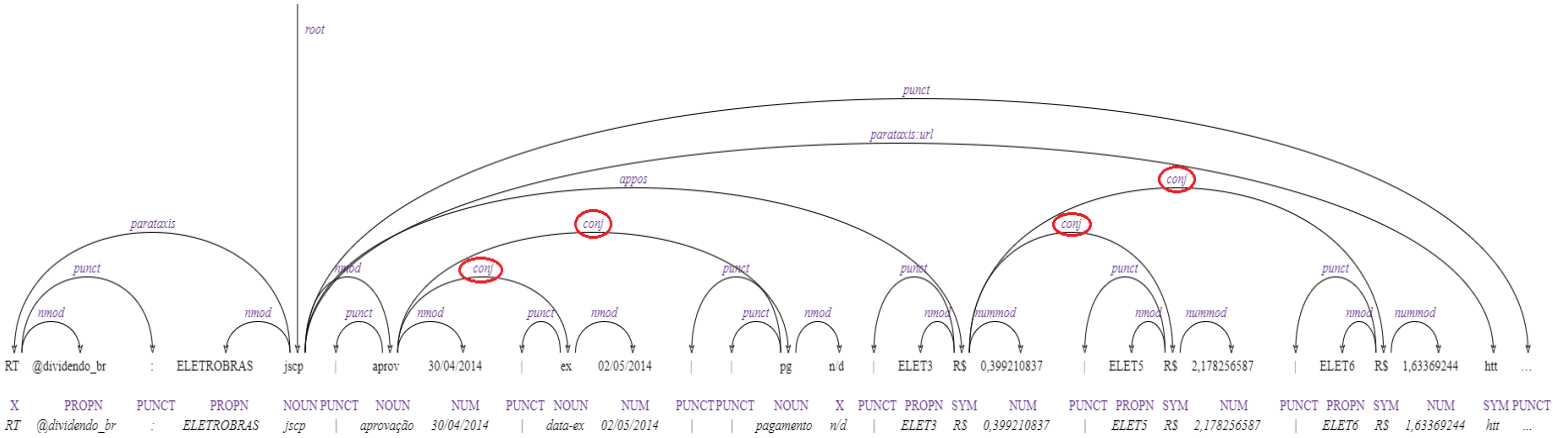


Figura 57 – Atribuição de **conj** introduzida por pontuação (barra vertical simples e dupla) (exemplo 56).

Exemplo (com símbolo):

(57) LONG&SHORT de PETR4 **x** PETR3 indo MTOOOO BEM ! ! !

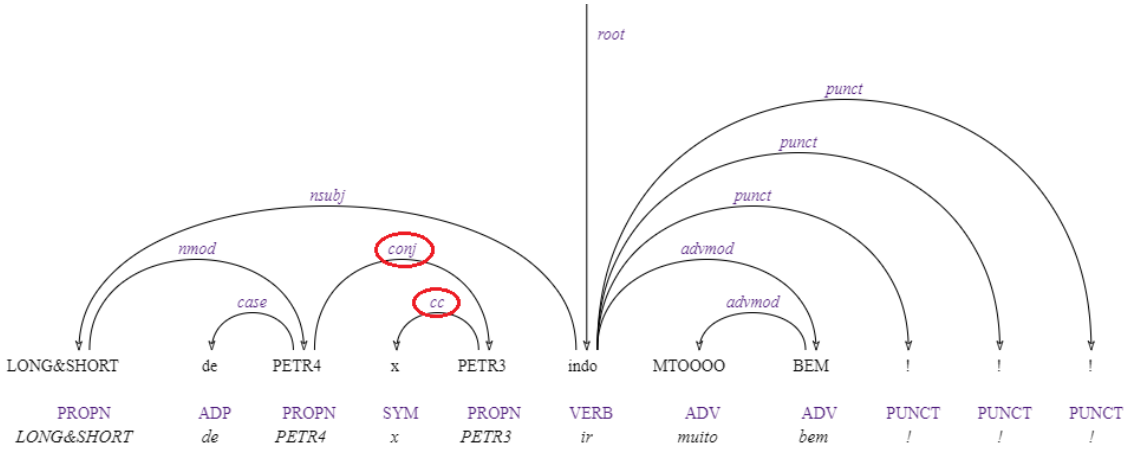


Figura 58 – Atribuição de **conj** introduzida por símbolo (SYM) com **cc** (exemplo 57).

Aposição (appos) sinalizada por PUNCT “/” (barra inclinada)

Padrão: **PUNCT “/” como elemento de aposição (appos)**, quando:

1. O dependente (elemento da direita) especifica ou descreve o ***head*** (elemento da esquerda)

Exemplo:

(58) $LIGT3 - Light S/a (ligt-nm) - Aviso A os Acionistas **/** Distribuicao De Dividendo http://t.co/a8VHtsh8Xw

(Interpretação: Aviso aos acionistas (**:** haverá) distribuição de dividendo [...]”)

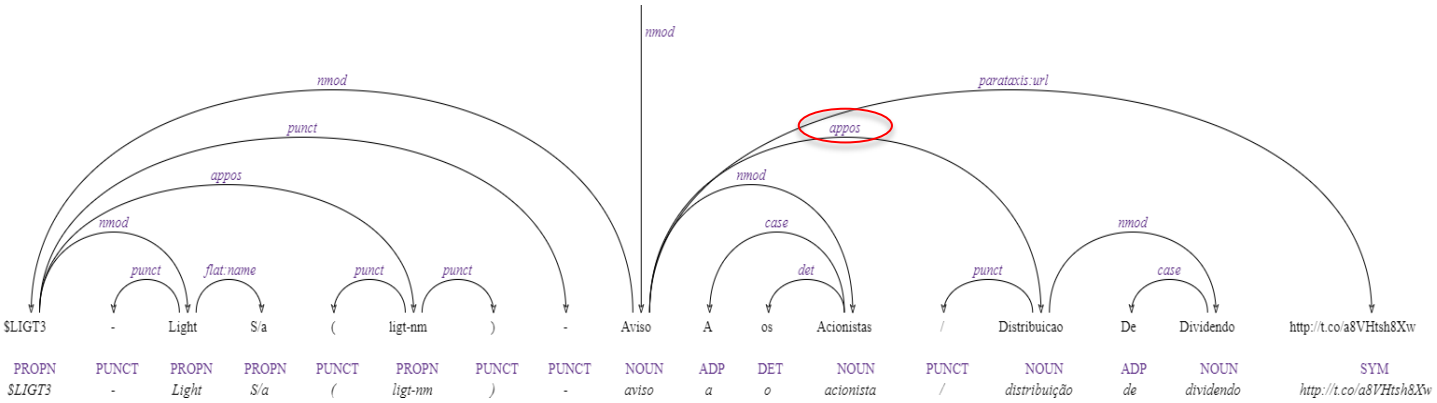


Figura 59 – Atribuição de **appos** introduzida por pontuação (barra inclinada).

Root expresso por símbolo (SYM)

Em certos contextos, o SYM “>” pode ser interpretado como verbo, funcionando como **root**. No exemplo da Figura 60, o referido símbolo foi “traduzido” para “foi (a)” (PoS VERB ADP), ao qual “R$” foi conectado por **obj** (segundo argumento *core* do predicado).

Exemplo:

(59) último dia de o Gov FHC , a ação Petrobras (PETR3) **>** R$ 3,3 @geraldoAlckmin_ diz q ações de a Petrobras “ viraram pó ” . http://t.co/vAdFc65rmh

(Interpretação: No último dia do governo FHC, a ação Petrobras (PETR3) **foi a** R$3,3 e Geraldo Alckmin diz que as ações da Petrobras viraram pó.).

Figura 60 – Atribuição de **root** para símbolo (“>”).

Deprel por inferência

Inferência de conjunção (conj)

No trecho em negrito do exemplo a seguir, infere-se uma coordenação (**conj**) entre “Petrobrás” (representada pela *hashtag* #PETR4) e “Pasadena”, mesmo sem conjunção presente. A seta que a representa **conj** parte do primeiro elemento da série (*head*) em direção a cada dependente.

Exemplo:

(60) Ôrrrra , finarmenti , hein ! ! | FHC muda discurso e diz que apoia CPI para investigar **#PETR4 #Pasadena** http://t.co/33WJ6NBvWW via @estadao

(Interpretação: “[...] FHC muda discurso e diz que apoia CPI para investigar Petrobrás **e** Pasadena.)

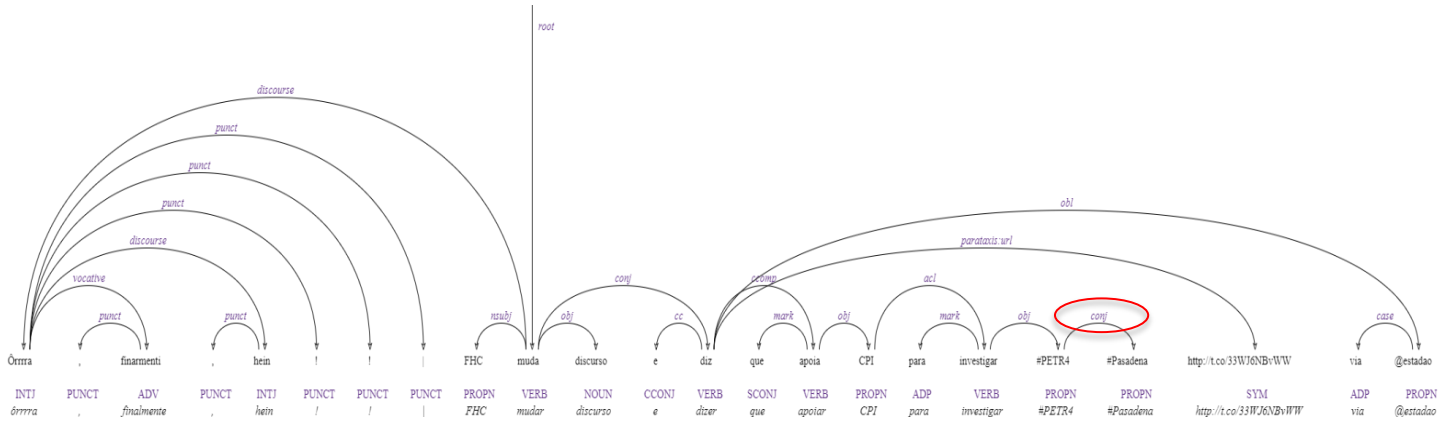


Figura 61 – Atribuição de **conj** sem ocorrência de conjunção (exemplo 60).

Exemplo:

(61) @KatiaAbreu como assim CPI prejudicará #PETR4 ? imagino q Sra entenda a corrente gravidade e saiba o q seja CPI . **@agenciapf @MP_PGR ctz fará**

(Interpretação: [...] . Agência PF (Polícia Federal) **e** o Ministério Público certamente farão).

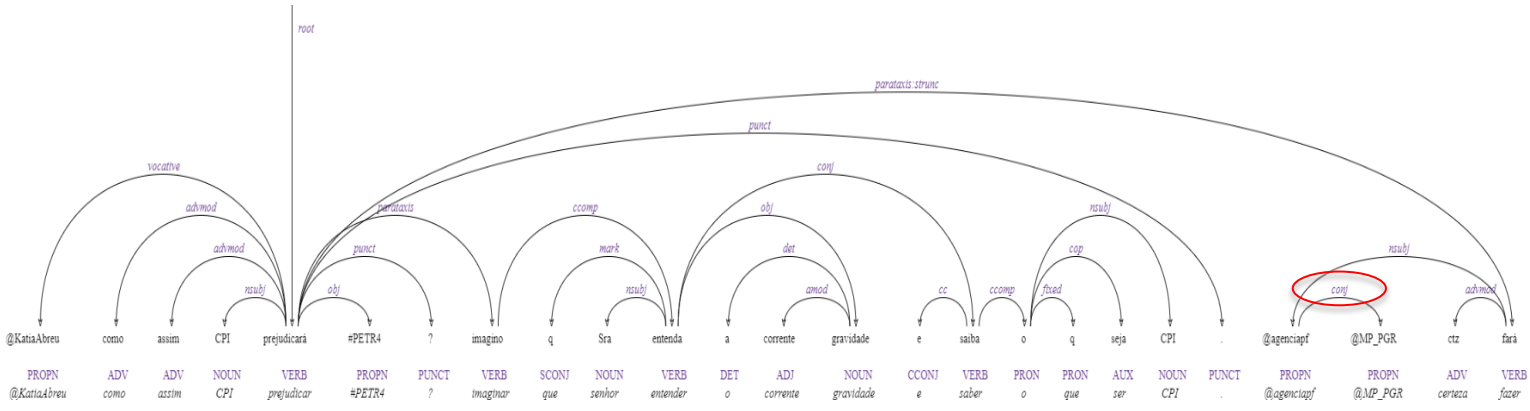


Figura 62 – Atribuição de **conj** sem ocorrência de conjunção (exemplo 61).

Inferência de conjunção e verbo

Exemplo:

(62) RT @reminiscences : Petr4 - CPI exclusiva de a Petrobras , Petr4 deve subir (pois afetará muito Dilma) CPI ampla , Petr4 deve cair , pois espalh …

(Interpretação: **Se for** uma CPI exclusiva da Petrobras, Petr4 deve subir, pois afetará muito Dilma; **se for** uma CPI ampla, Petr4 deve cair, pois espalh….).

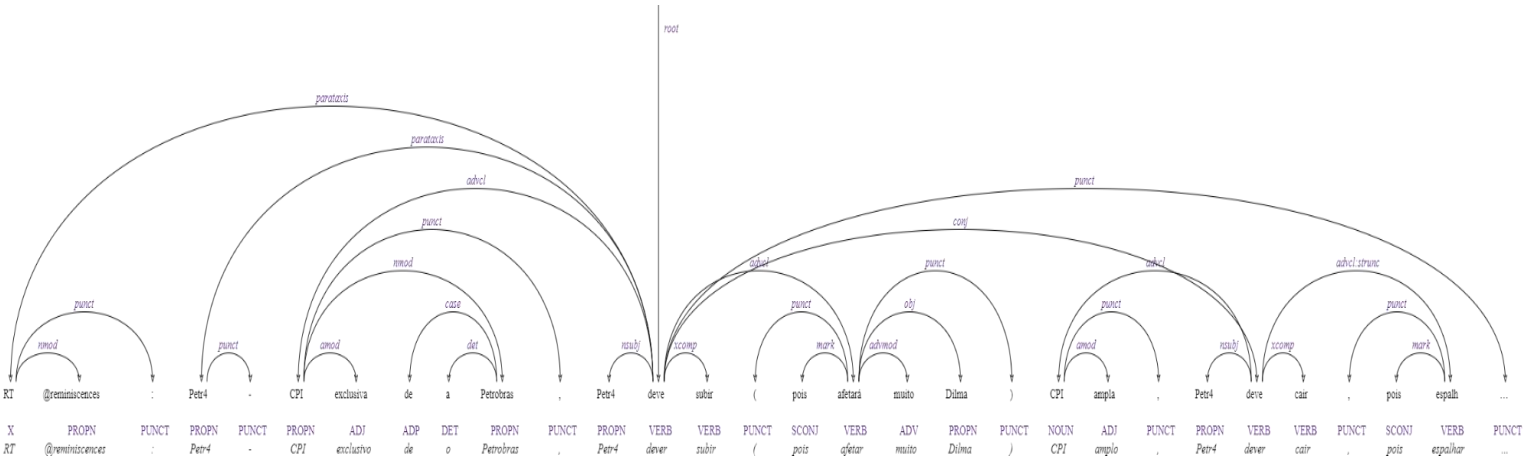


Figura 63 – Atribuição de **advcl** por inferência de conjunção e verbo (exemplo 62).

Exemplo:

(63) BVMF3 cancelado - bateu em o stop antes de acionar

(Interpretação: “BMF3 **foi** cancelado **pois** bateu no stop antes de acionar”) Essa interpretação explica a anotação de “cancelado” como **root**, o qual possui uma **advcl** causal).

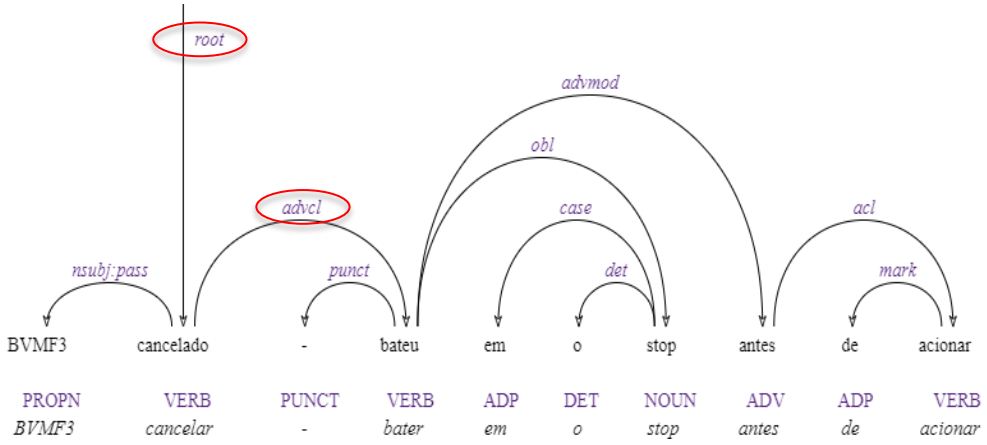


Figura 64 - Atribuição de **root** e **advcl** por inferência de conjunção e verbo (exemplo 63).

Inferência de verbo e preposição

Exemplo:

(64) #cyre3 postado hj antes de a abertura + 1,78

(Interpretação: "*#cyre3* **foi** *postado hj antes de a abertura* **a** + *1,78*") Essa interpretação justifica a anotação de "postado" como **root** e de "+ 1,78" como seu dependente por **obl**.

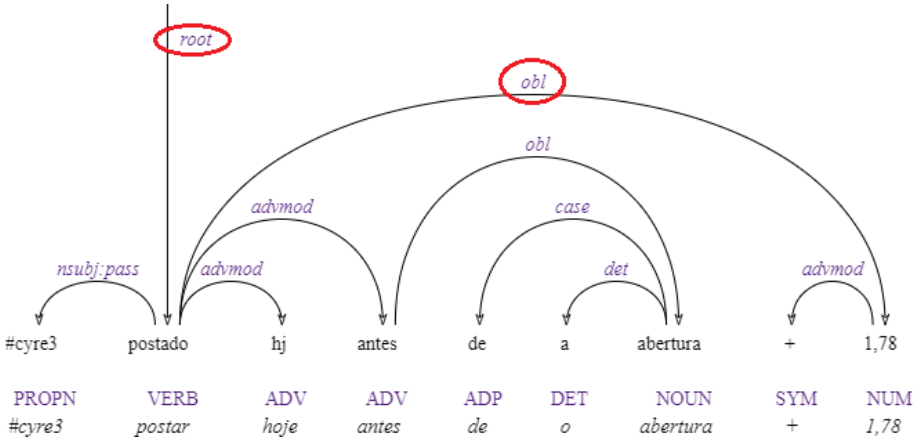


Figura 65 - Atribuição de **root** e **obl** por inferência de verbo e preposição (exemplo 64).

Exemplo:

(65) Day trade VALE5 Previsto e evitado zona de alto risco e falta de liquidez absurda 27-03-14 http://t.co/FZV3HPrEG3

(Interpretação: "Day trade VALE5 **foi** Previsto e **foram** evitado**(as)** zona de alto risco e falta de liquidez absurda **em** 27-03-14").

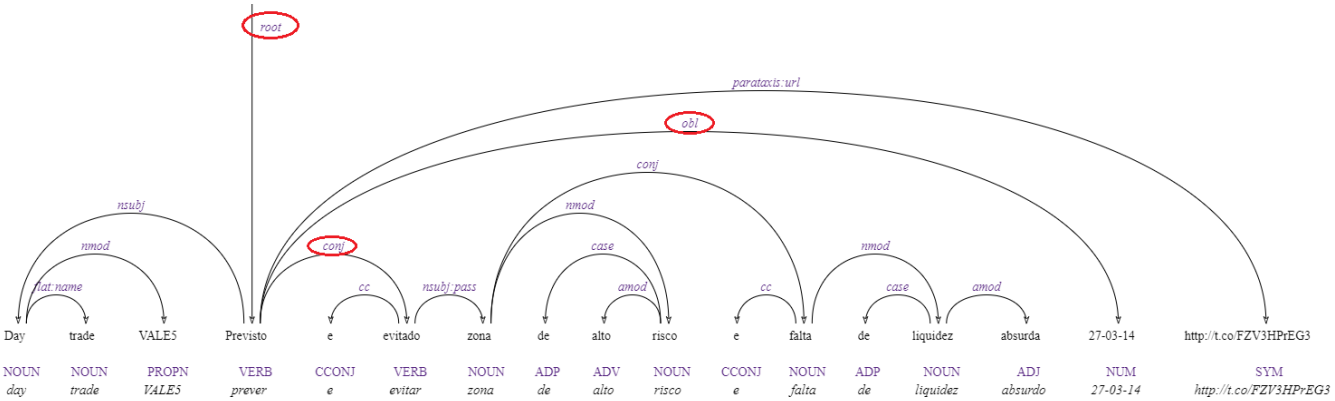


Figura 66 - Atribuição de ***deprel*** por inferência de verbos (exemplo 65).

Exemplo:

(66) @edmilsonpapo10 achei , olha : PETR4 , 2008 , por volta de R$ 45

(Interpretação: “@edmilsonpapo10, achei. Olha: PETR4, **em** 2008, **estava** por volta de R$45”). Essa interpretação explica a anotação de **vocative**, **nmod** (tempo) e **ccomp**.

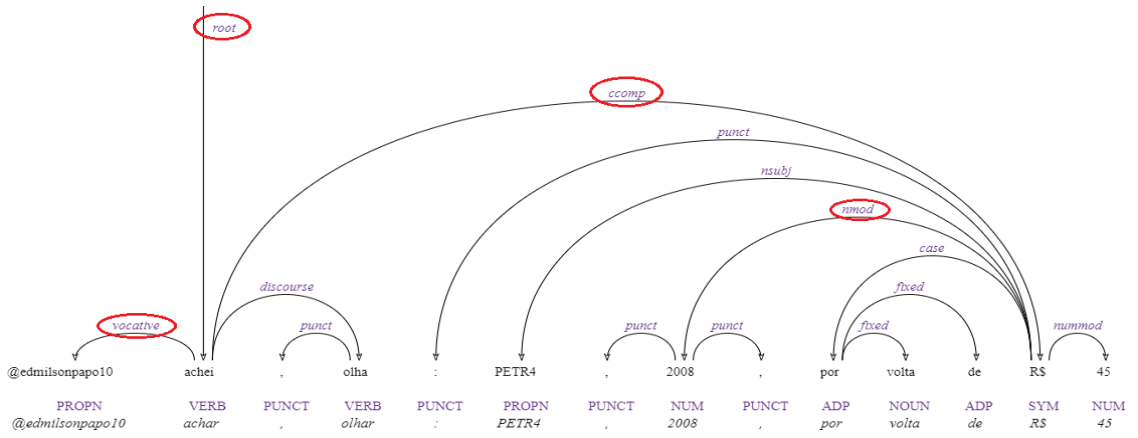


Figura 67 - Atribuição de **ccomp**, **obl** e **ccomp** por inferência de verbo e preposição (exemplo 66).

Inferências de símbolo como fórmula ou palavra de conteúdo

Os símbolos (PoS SYM) ocorrem em *tweets* do mercado financeiros que apresentam estruturas bastante distintas.

Muitos *tweets,* por veicularem conteúdo originalmente expresso na forma de tabela ou lista, apresentam estrutura bastante fragmentada. Neles, o símbolo de igual (“=”) parece iniciar uma sequência (trecho sublinhado no exemplo da Figura 68) de valores, *tickers* e outros símbolos (como “*”) na qual é extremamente complicado identificar uma hierarquia sintática entre os elementos. Assim, opta-se por anotar os elementos dessa sequência iniciada por “=” com **flat**.

Exemplo:

(67) Ouro : AEDU3 Fusão só depende de o CADE em jun/14 1 AEDU = 0,4548 KROT 43,7 * 0,4548 = 19,87 19,87 / 13,13 = 51 % 51 % de espaço pra altas ! !

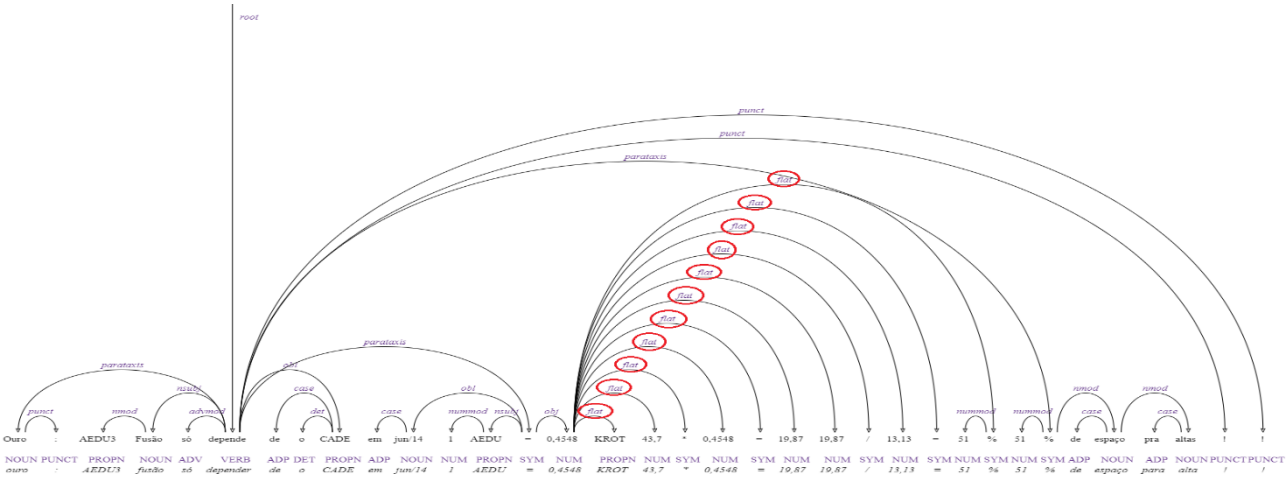


Figura 68 - Atribuição de **flat** a uma sequência de valores, *tickers* e outros símbolos iniciada por “=”.

O símbolo de igual também aparece em contexto não numérico, no qual é possível inferir que ele esteja funcionando como um verbo de cópula e, portanto, sendo anotado com **cop**. Dessa forma, o símbolo nesse contexto é responsável por ligar um sujeito de uma oração a um predicativo.

Exemplo:

(68) @CaciqueInvest não , aviãozinho **=** #goll4

(Interpretação: “[...] aviãozinho **é** #goll4”.)

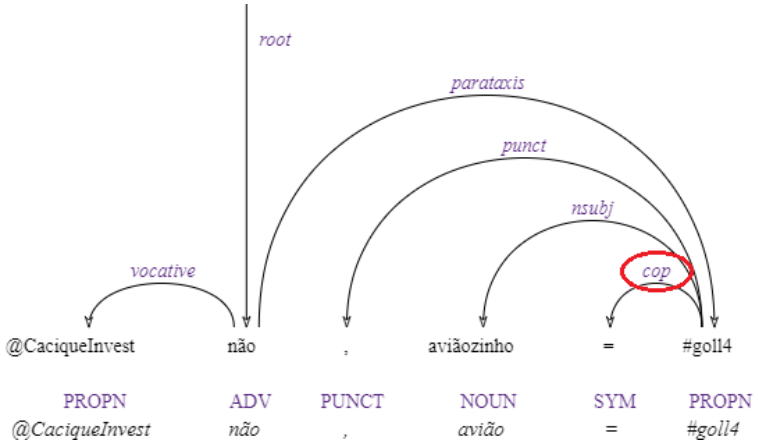


Figura 69 – Inferência de “=” como verbo de cópula, sendo anotado com **cop** (exemplo 68).

Exemplo:

(69) Ladeira abaixo ! #petr4 **=** caminhão velho com problemas em os freios descendo a rua CARREGADO de sardinhas atordoadas ! Vamos a os 10 pila msm ?

(Interpretação: “[...] #petr4 **é** caminhão velho com problemas nos freios descendo a rua [...]”)

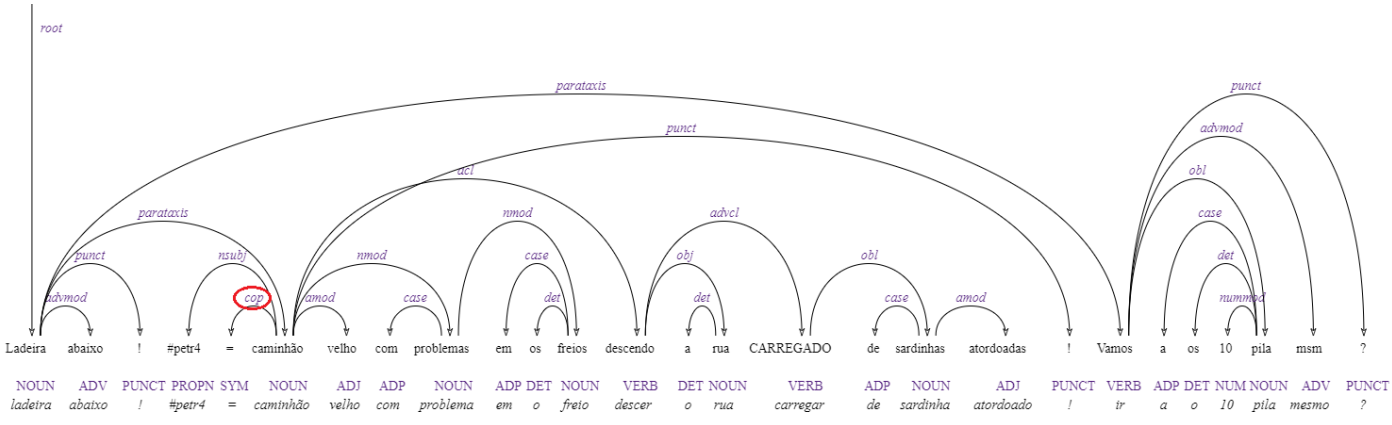


Figura 70 – Inferência de “=” como verbo de cópula, sendo anotado com **cop** (exemplo 69).

Quanto aos operadores matemáticos “+/-” que compõem os índices de (des-)valorização das ações (por exemplo, +1,4%), os quais estão anotados com a PoS SYM, infere-se que eles substituem as palavras de conteúdo “mais” e “menos”, funcionando como advérbios.
Dessa forma, opta-se por anotá-los com **advmod** (e conj se forem binários) (- + * /), e

Exemplo:

(70) Desempenho de as ações de a TIM em a semana passada : TIMP3 (Bovespa) : **+** 1,4 % , TSU (NYSE) : **-** 0.53 % , IBOV : **-** 1,8 % .

(Interpretação: “Desempenho das ações da TIM na semana passada: [...])

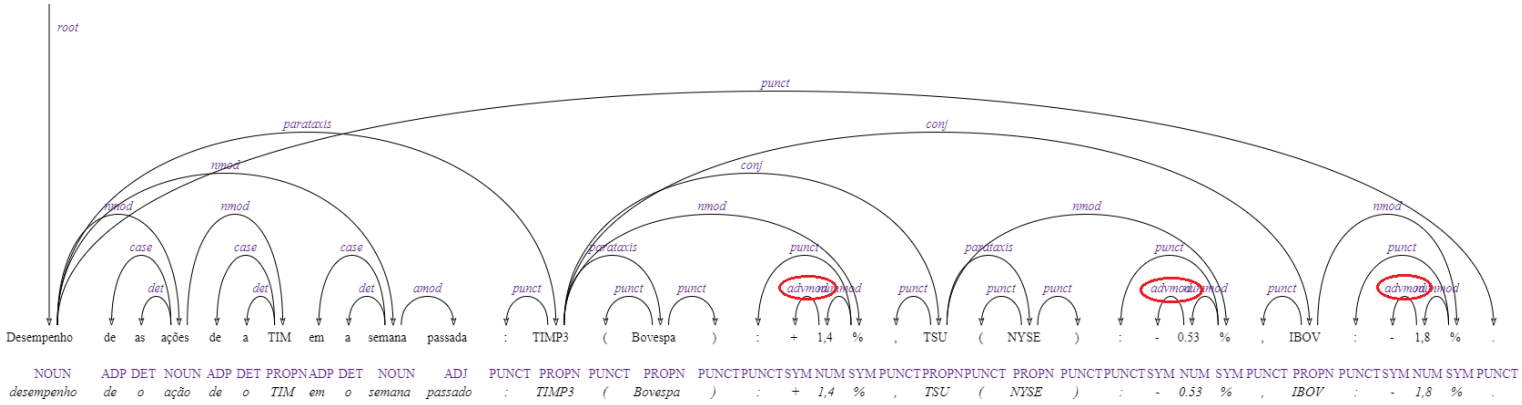


Figura 71 – Inferência de “+/-” como advérbios, sendo anotados com **advmod**.

SEGUNDA PARTE - PADRÕES ESTRUTURAIS

Como mencionado, as diretrizes descritas nesta segunda parte são específicas para certos padrões estruturais frequentes nos *tweets* do *corpus* DANTEStocks.

A maioria das diretrizes são descritas por um *template* composto por:
i. padrão estrutural
ii. lista de elementos constitutivos do padrão com direcionamentos para a anotação de cada um
iii. subpadrão (se houver)
iv. ao menos 1 exemplo anotado extraído do *corpus* DANTEStocks

Alguns *templates*, cujos padrões são particularmente fragmentados e/ou que apresentam fenômenos de domínios, também possuem uma interpretação (ou glosa) de seu conteúdo, a qual busca explicitar a interpretação que levou à anotação sintática (e escolha das *deprel*) do exemplo.

Template 1

Padrão: **notas gerais <sentença_truncada> ... <url>**, em que:

1. "notas gerais" é dependente por **parataxis** do **root**
2. <sentença_truncada> contém o **root**
3. ... (sinal de reticências que representa o truncamento) é dependente por **punct** do **root**
4. <url> é dependente por **parataxis:url** do **root**

5. Subpadrão
   5.1. **<sentença_truncada_com_mensagem_completa>**, em que:
      a. Embora haja um truncamento, a sentença possui estrutura sintática completa, com análise e anotação UD válidas

      Exemplo:

(71) Notas gerais A PETROBRAS (PETR4) concluiu a perfuração de o poço Pitu , localizado em águas profundas de a Bacia ... http://t.co/DJCj8f8xTH

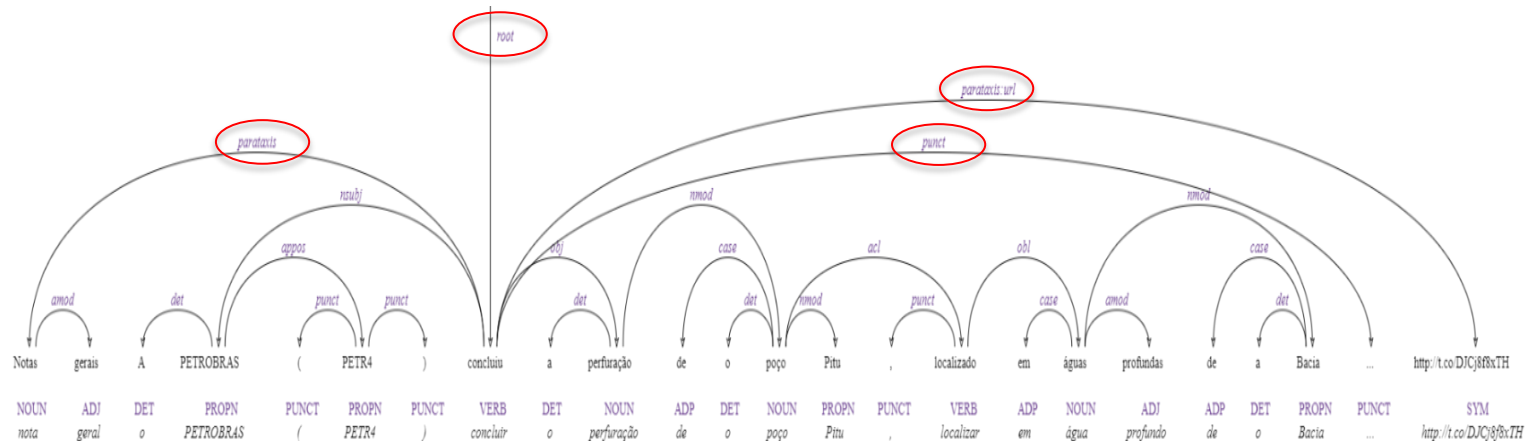


Figura 72 - Anotação do subpadrão 5.1 do *Template* 1.

5.2. **<sentença_truncada_incompleta>**, em que:

a. O truncamento impede a sentença de ter estrutura sintática completa e, por isso, haverá uma *deprel*, dependente de cada caso (pode ser **obj, obl, nmod, ccomp** e **root**), rotulada com a sub-relação ***:strunc*** (usada para indicar um truncamento no nível sintático, em oposição **:wtrunc**, que indica truncamento de palavras).

Exemplo [**obj:strunc**]:

(72) Notas gerais A BR PROPERTIES (BRPR3) vendeu a a LPP Empreendimentos e Participações , sociedade de o grupo GLP , a ... http://t.co/Ou2D3dYKDh

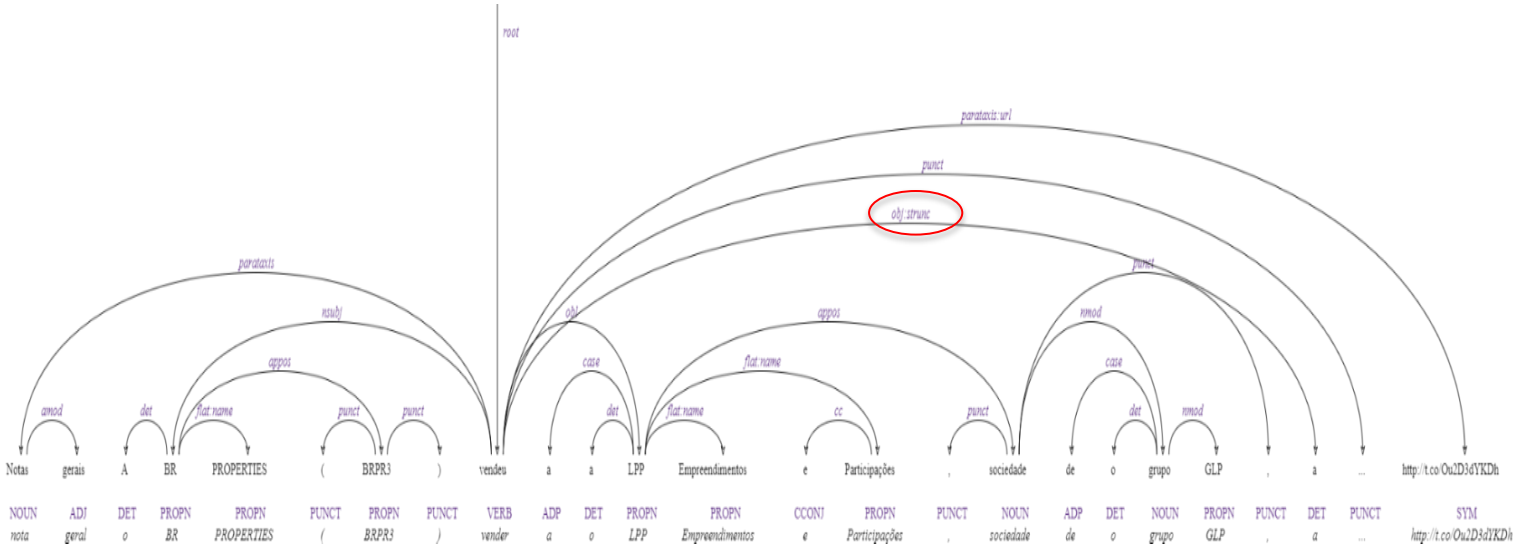


Figura 73 - Anotação do padrão 5.2 com **obj:strunc** do *Template* 1.

Exemplo [**obl:strunc**]:

(73) Notas gerais A LIGHT (LIGT3) pretender distribuir R$ 32 milhões referentes a o dividendo mínimo obrigatório a os ... http://t.co/p9jr05Re1l

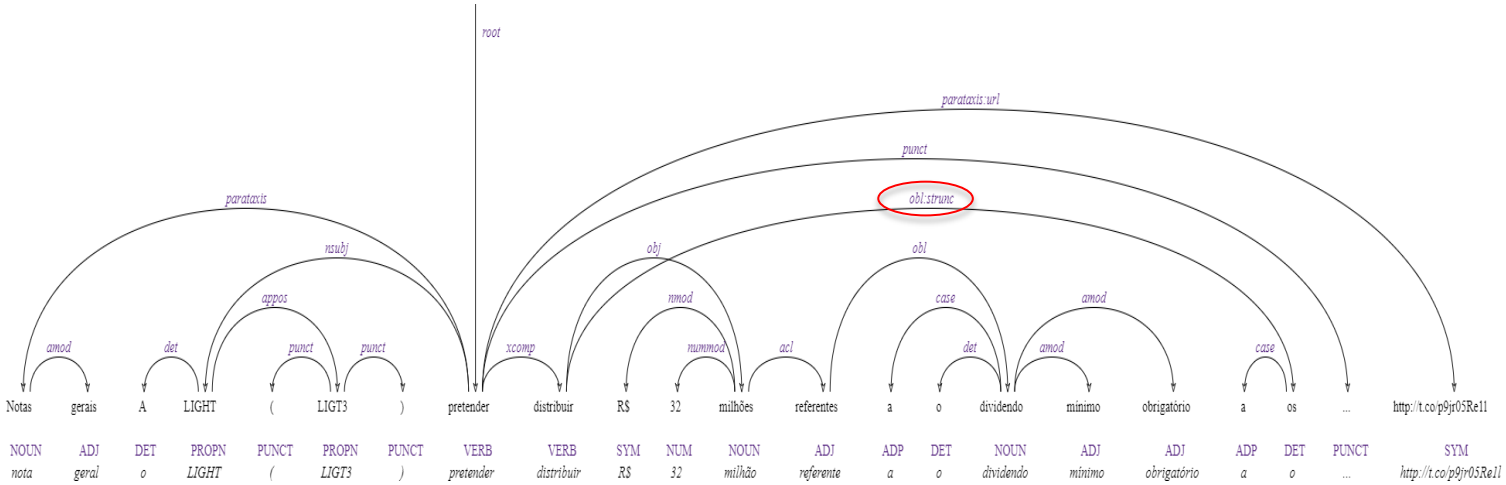


Figura 74 - Anotação do padrão 5.2. com **obl:strunc** do *Template* 1.[12]

[12] Em sentenças completas, não seria válido, pela UD, que um DET fosse dependente de uma relação **obl**, como na Figura 4. Isso só acontece porque DET é o último elemento antes do truncamento. Da mesma forma, um DET tampouco seria *head* de uma relação **case**, porém o truncamento faz com que essa seja uma alternativa para integrar a ADP "a" à árvore.

Exemplo [**nmod:strunc**]:

(74) Notas gerais A BROOKFIELD (BISA3) não conseguiu instalar a assembleia que definiria a saída de a companhia de a ... http://t.co/P0Nc4AO2et

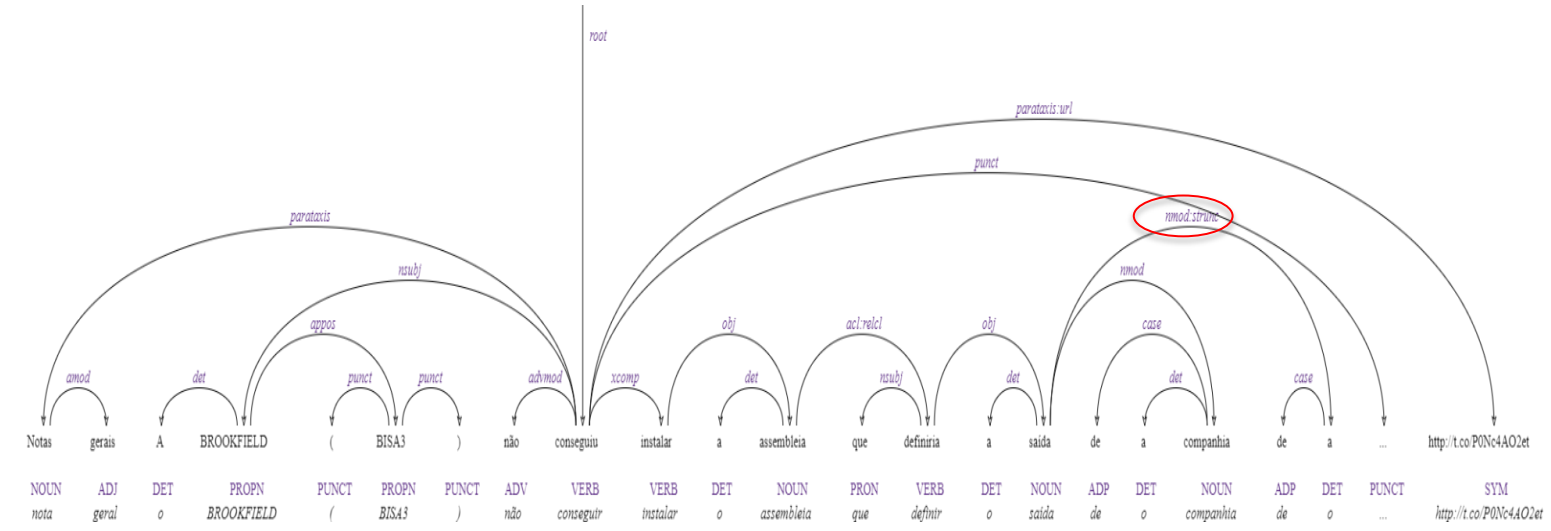


Figura 75 - Anotação do padrão 5.2. com **nmod:strunc** do *Template* 1.

Exemplo [**ccomp:strunc**]

(75) Notas gerais Por determinação de a justiça , a PETROBRAS (PETR4) precisará divulgar , em até cindo dias , qual é a ... http://t.co/uGFrxqliWy

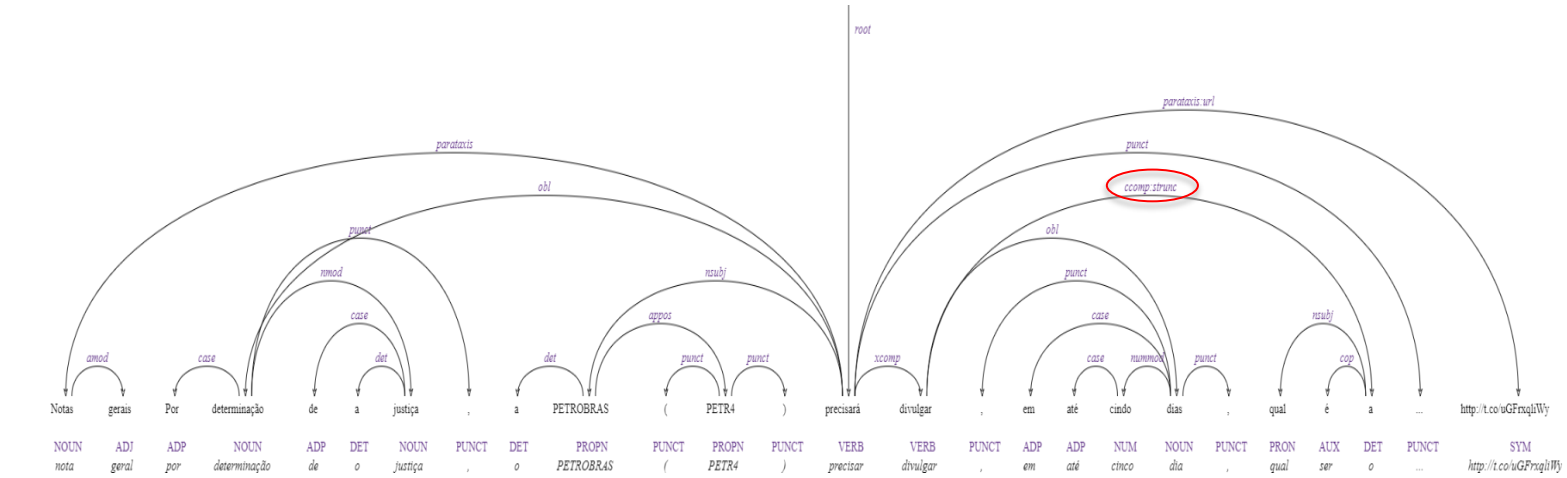


Figura 76 - Anotação do padrão 5.2 com **ccomp:struct** do *Template* 1.

Exemplo [**root:strunc**] (quando a última palavra antes do truncamento for o **root**):

(76) Notas gerais A produção total de petróleo e gás natural de a PETROBRAS (PETR4) em o Brasil em o mês de fevereiro foi ... http://t.co/qEEHshr7en

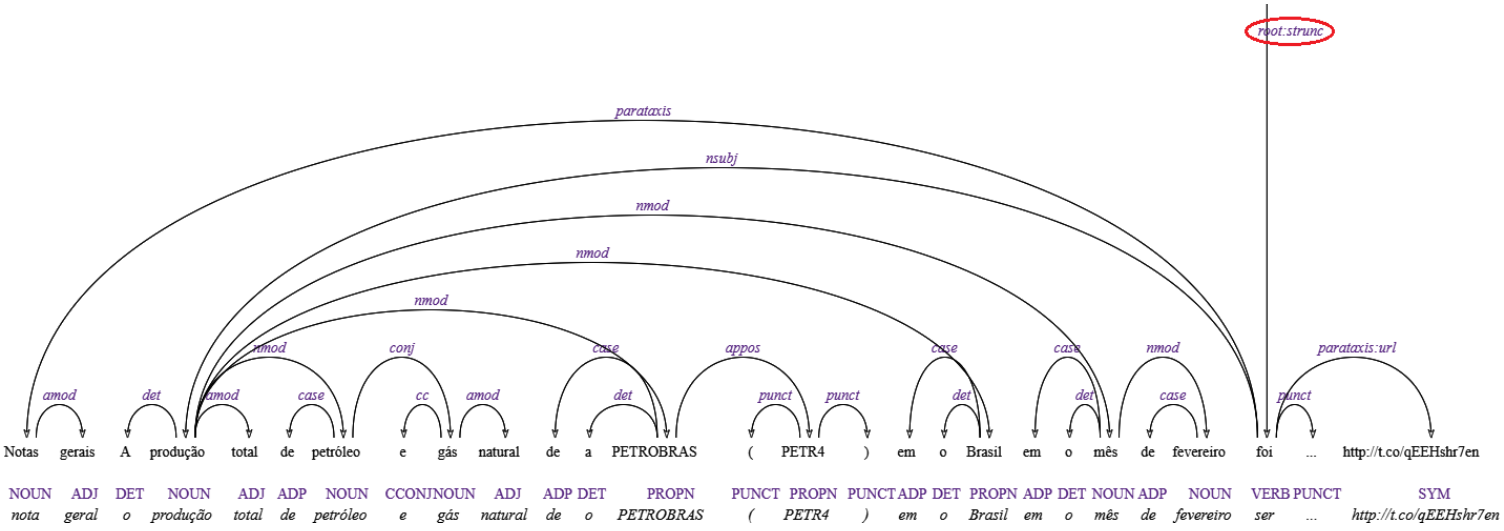


Figura 77 - Anotação do padrão 5.2. com **root:strunc** do *Template* 1.

Template 2

Padrão: ***<hashtag-ticker>* (mensagem:NNN) <url>**, em que:

1. <*hashtag-ticker*> será sempre **root**
2. (mensagem:NN) é dependente por **parataxis** de **root**
3. <url> é dependente por **parataxis:url** de *root*

Exemplo:

(77) #vale5 (mensagem : 950904) http://t.co/wfR8HEPu4k

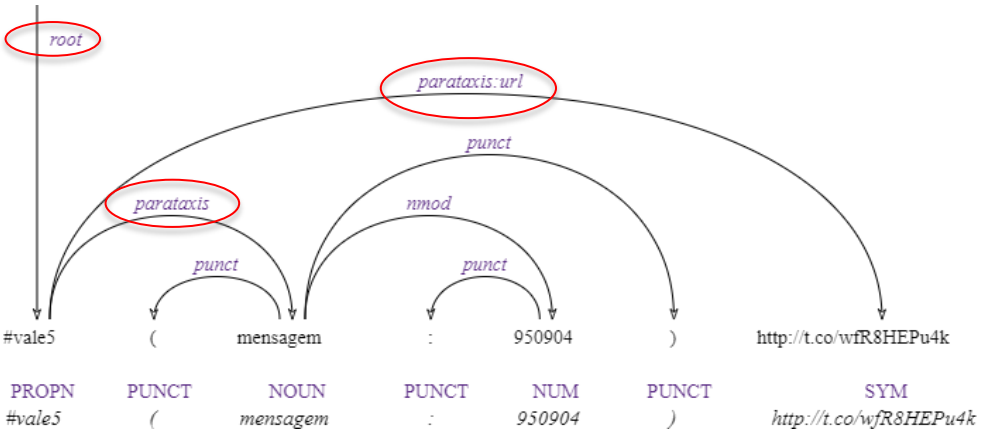


Figura 78 - Anotação do padrão do *Template* 2.

Template 3

Padrão: **<*hashtag-ticker*> <complemento> (mensagem:NNN) <url>**, em que:

1. <*hashtag-ticker*> é o **root**, a menos que ocupe um papel sintático em <complemento>
2. relação entre **root** e <complemento> depende da natureza deste último, podendo ser **nmod, appos, amod, advmod**, etc.
3. (mensagem:NN) é dependente por **parataxis** de **root**
4. <url> é dependente por **parataxis:url** de **root**

Exemplo: <complemento> do tipo [**nmod**]

(78) #PETR4 - 15 min (mensagem : 951348) http://t.co/7A5UINu9Mu

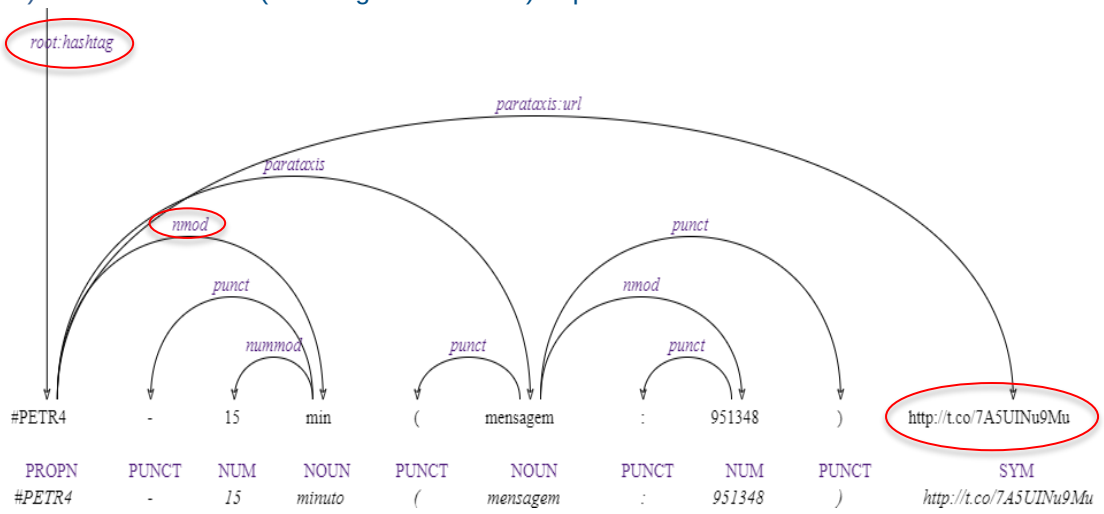


Figura 79 - Anotação do padrão do *Template* 3 com <complemento> do tipo **nmod**.

Exemplo: <complemento> do tipo [**appos**]

(79) #BBAS3 Banco de a Brasil (mensagem : 956467) http://t.co/75T8wtmEXw

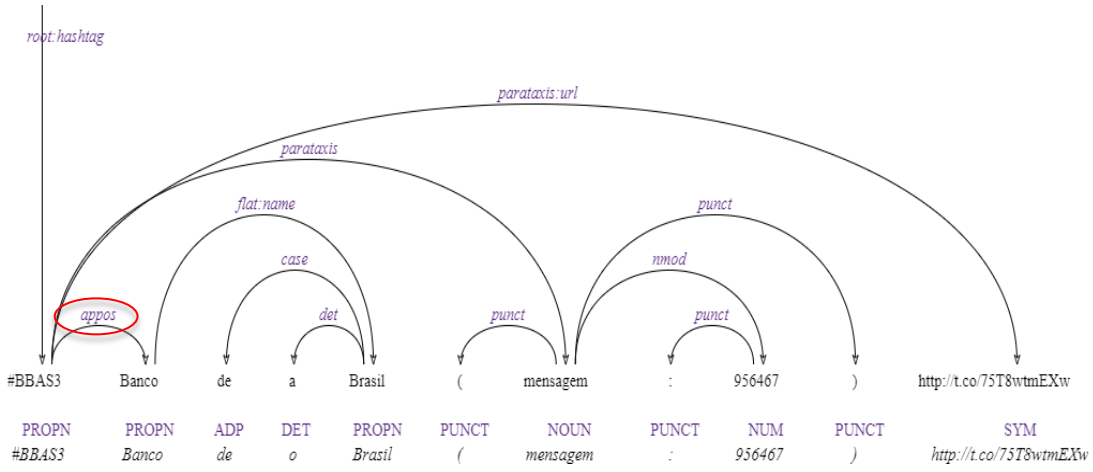


Figura 80 - Anotação do padrão do *Template* 3 com <complemento> do tipo **appos**.

Exemplo: <complemento> do tipo [**amod**]

(80) #csna3 semanal (mensagem : 950998) http://t.co/suRkLOSBUz

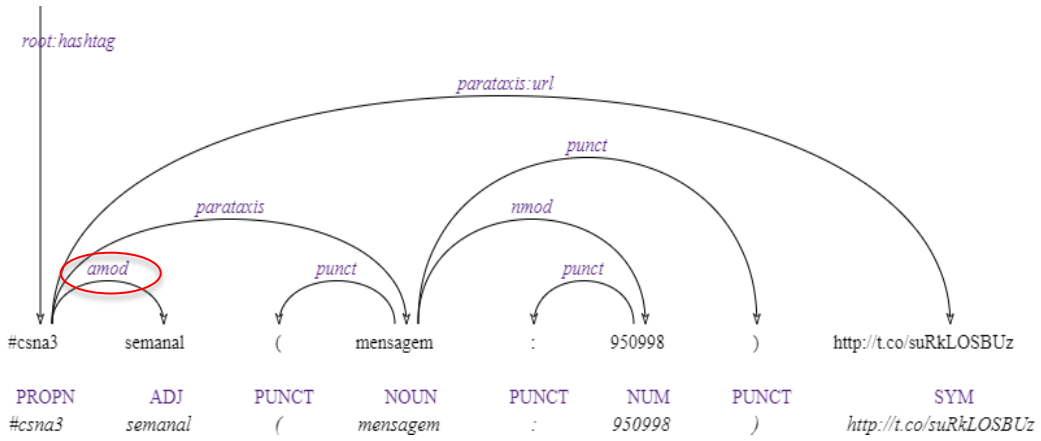


Figura 81 - Anotação do padrão do *Template* 3 com <complemento> do tipo **amod.**

Exemplo: <complemento> do tipo [**advmod**]

(81) #LLXL3 - acima de 1 (um) (mensagem : 952921) http://t.co/11sdL24xTr

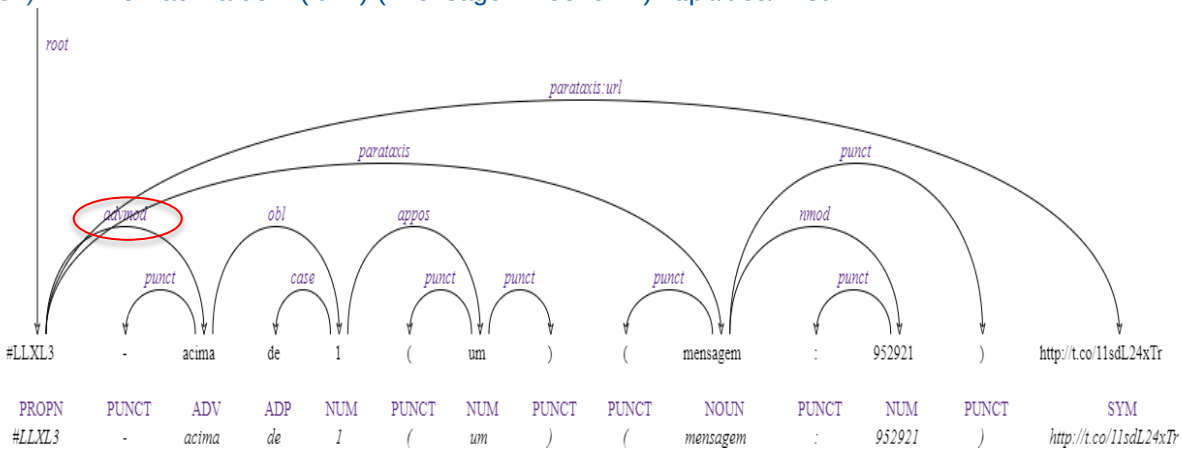


Figura 82 - Anotação do padrão do *Template* 3 com <complemento> do tipo **advmod.**

Exemplo: <complemento> do tipo [**nmod**] e **[parataxis**]

(82) #PETR4 15 min - acho que nao ! (mensagem : 952919) http://t.co/32XqwNSA6Y

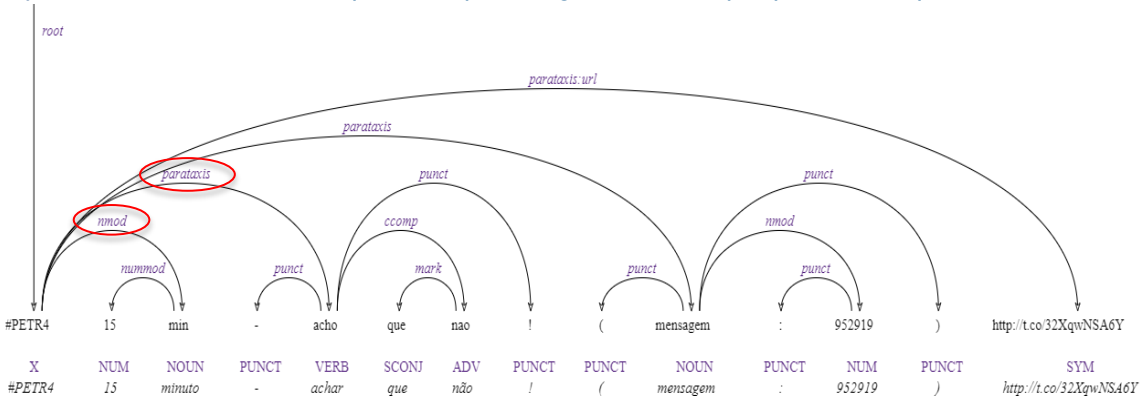


Figura 83 - Anotação do padrão do *Template* 3 com <complemento> **nmod** e **parataxis.**

Template 4

Padrão: **<sentença> (mensagem:NNN) <url>**, em que:

1. <sentença> contém o **root**
2. (mensagem:NN) é dependente por **parataxis** de **root**
3. <url> é dependente por **parataxis:url** de **root**

Exemplo:

(83) DANDO ZOOM EM A #OIBR4 (mensagem : 954226) http://t.co/Pk10JNN9fv

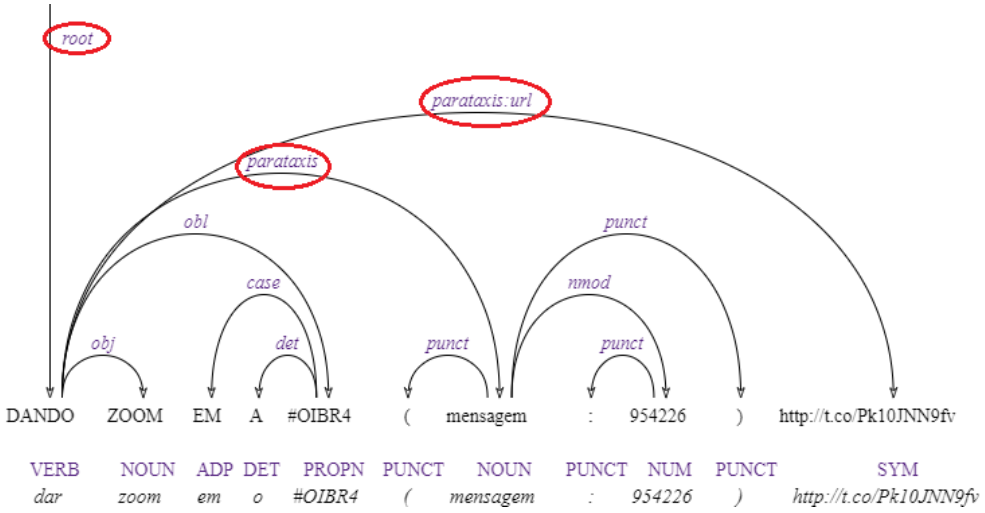


Figura 84 - Anotação do padrão do *Template* 4.

Template 5

Padrão: **<ação&cia> dividendos <lista info&data> <lista *ticker*&valor> || <url>,** em que:

1. **root** é o *token* “dividendos”
2. <ação&cia>: nome da ação e respectiva companhia (**appos**), ligados ao **root** por **nmod**
3. <lista info&data> são 3 elementos coordenados (**conj**) e ligados por **nmod** ao **root**:
   || x <data> || y <data> || z <data>, sendo
   x= aprov ou apro; y = dataex ou ex; z = pagto ou pg; data é da forma dd/mm/aa ou n/d (data não definida)
4. <lista *ticker*&valor> é uma lista de um ou mais itens coordenados (**conj**) e ligados por **appos** ao **root**:
   || <*ticker*> <valor-reais> || <*ticker*> <valor-reais> || <*ticker*> <valor-reais>....
5. <*ticker*> é dependente por **nmod** de <valor-reais>
6. <url> é dependente por **parataxis:url** de *root*

Exemplo:

(84) ELPL ELETROPAULO dividendos | | aprov 25/4/2014 | | dataex 28/4/2014 | | pagto n/d | | ELPL3 R$ 0,388977082 | | ELPL4 R$ 0,427874790 http://t.co/FhJ4SZQxMA

(Interpretação: **a respeito dos** dividendos da ELETROPAULO **:** aprovação em 25/05/2014; ex em 28/04/2014; pagamento n/d, **sendo:** R$ 0,388977082 para ELP3 e R$ 0,427874790 para ELP4; **fonte:** http://t.co/IjtHKlQlfr)

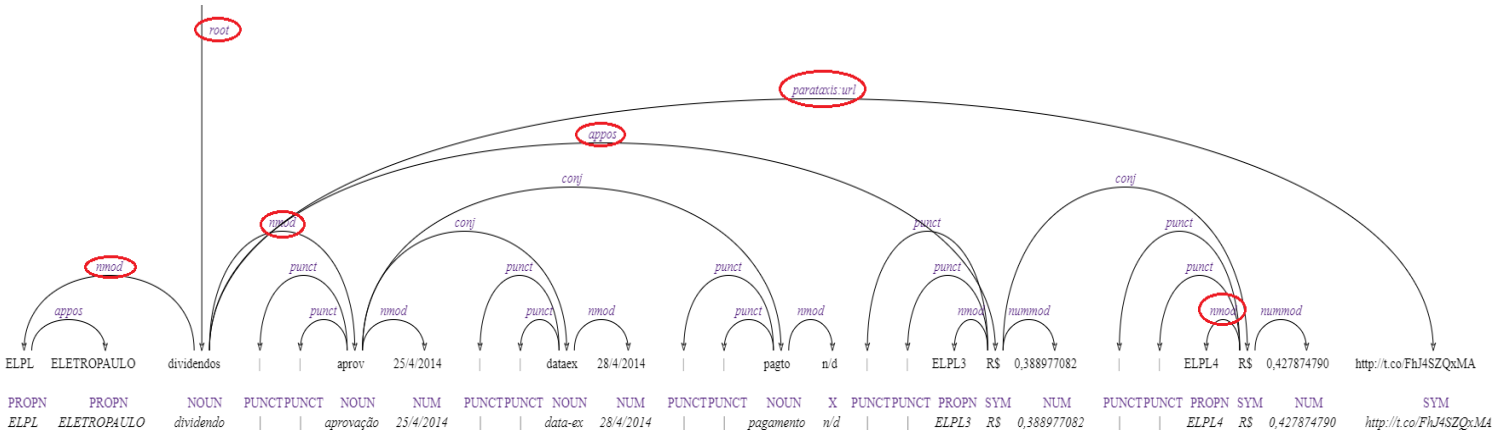


Figura 85 - Anotação do padrão do *Template* 5.

Template 6

Padrão: **<RT @xxxxx> : <ação e/ou cia> <dividendos/juros> <lista info&datas> <lista *ticker*&valor> || <url>,** em que:

1. <RT @xxxxx> é ligado por **parataxis** ao **root**
2. **root** é o *token* <dividendos/juros>
3. <ação cia>: nome da ação e/ou companhia ligado ao **root** por **nmod**
4. <lista info&data> são 3 elementos coordenados (**conj**) e ligados por **nmod** ao **root***:*
   || x <data> || y <data> || z <data>, sendo
   x= aprov ou apro; y = dataex ou ex; z = pagto ou pg; data é da forma dd/mm/aa ou n/d (não definida)
5. <lista *ticker*&valor> é uma lista de um ou mais itens coordenados (**conj**) e ligados por **appos** ao **root***:*
   || <*ticker*> <valor-reais> || <*ticker*> <valor-reais> || <*ticker*> <valor-reais>....
6. <*ticker*> é dependente por **nmod** de <valor-reais>
7. <url> é dependente por **parataxis:url** de **root**

Exemplo (com url trancada, por isso, **parataxis:wtrunc**):

(85) RT @dividendo_br : ELETROBRAS jscp | aprov 30/04/2014 | ex 02/05/2014 | | pg n/d | ELET3 R$ 0,399210837 | ELET5 R$ 2,178256587 | ELET6 R$ 1,63369244 htt …

(Interpretação: **a respeito dos** juros sobre o capital próprio (jscp) da ELETROBRAS **:** aprovação em 30/04/2014; ex em 02/05/2014; pagamento indefinido, **sendo:** R$ 0,399210837 para ELET3 **e** R$ 1,63369244 para ELET6; **fonte:** http://t.co/IjtHKlQlfr)

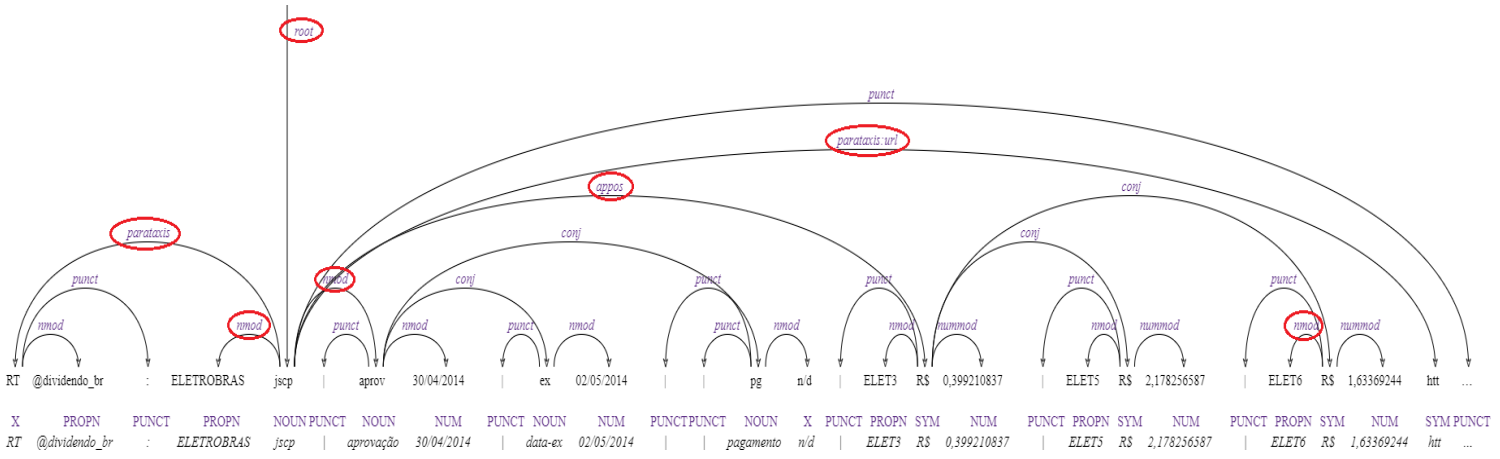


Figura 86 - Anotação do padrão do *Template* 6.

Template 7

Padrão: **<sentença> : <porção_truncada> ... <url> [hashtag opcional]**, em que:

1. <sentença> tem a mensagem principal e contém o **root**
2. <porção_truncada> é ligada ao *root* por **parataxis:strunc**
3. o ***head*** da <porção_truncada> dependerá de cada caso; <porção_truncada> pode terminar com uma palavra truncada (**wtrunc**)
4. <url> é dependente por **parataxis:url** de *root*
5. [hashtag opcional] lista de uma ou mais *hashtags* ao final do *tweet;* cada uma delas conectada por **parataxis:hashtag** ao **root**

Exemplo:

(86) ' Salvação ' de a OGX e OSX pode estar a a caminho , mais 5 empresas estão em o radar : Nível de o reser ... http://t.co/WYotABlDa5 #infomoney #vale5

Figura 87 - Anotação do padrão do *Template* 7 (exemplo 86).

Exemplo:

(87) O que há de melhor em a Bovespa : as ações mais indicadas por os analistas : A Vale ficou em prim ... http://t.co/tkdUiSqQUs #infomoney #vale5

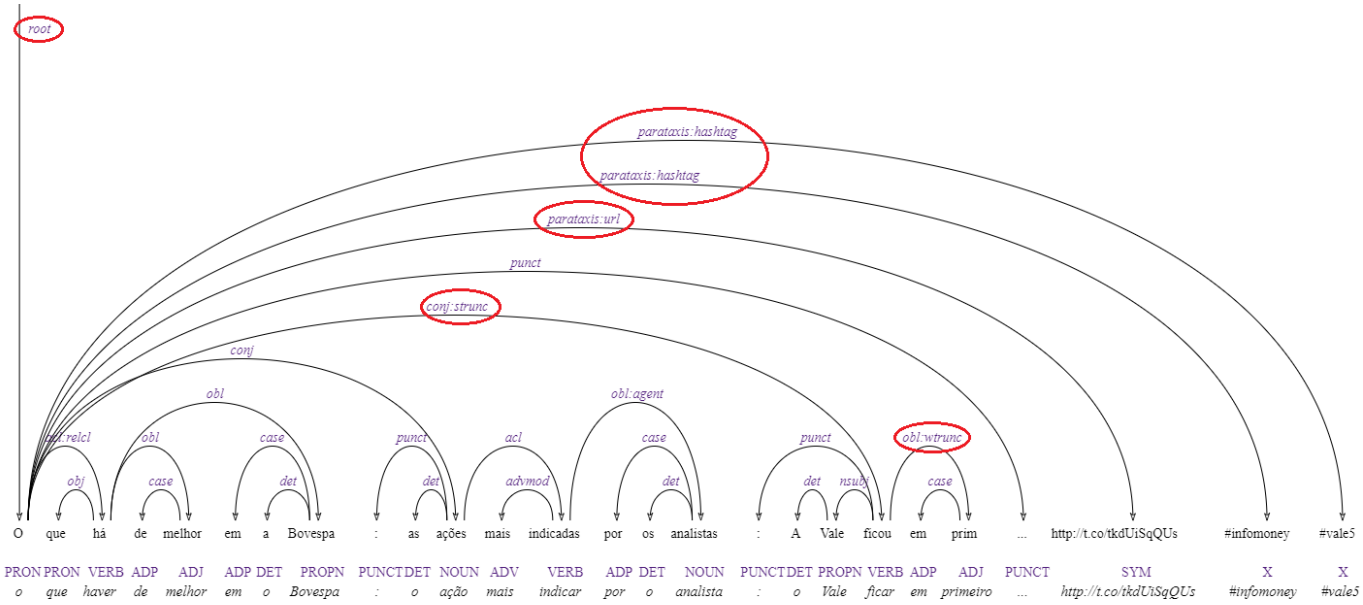


Figura 88 - Anotação do padrão do *Template* 7 (exemplo 87).

Exemplo 3:

(88) EM TEMPO REAL : Bancos caem , Vale mostra recuperação , Petrobras sobe 2 % : Mais informações em breve http://t.co/nQ3FofV6Jb #infomoney #vale5

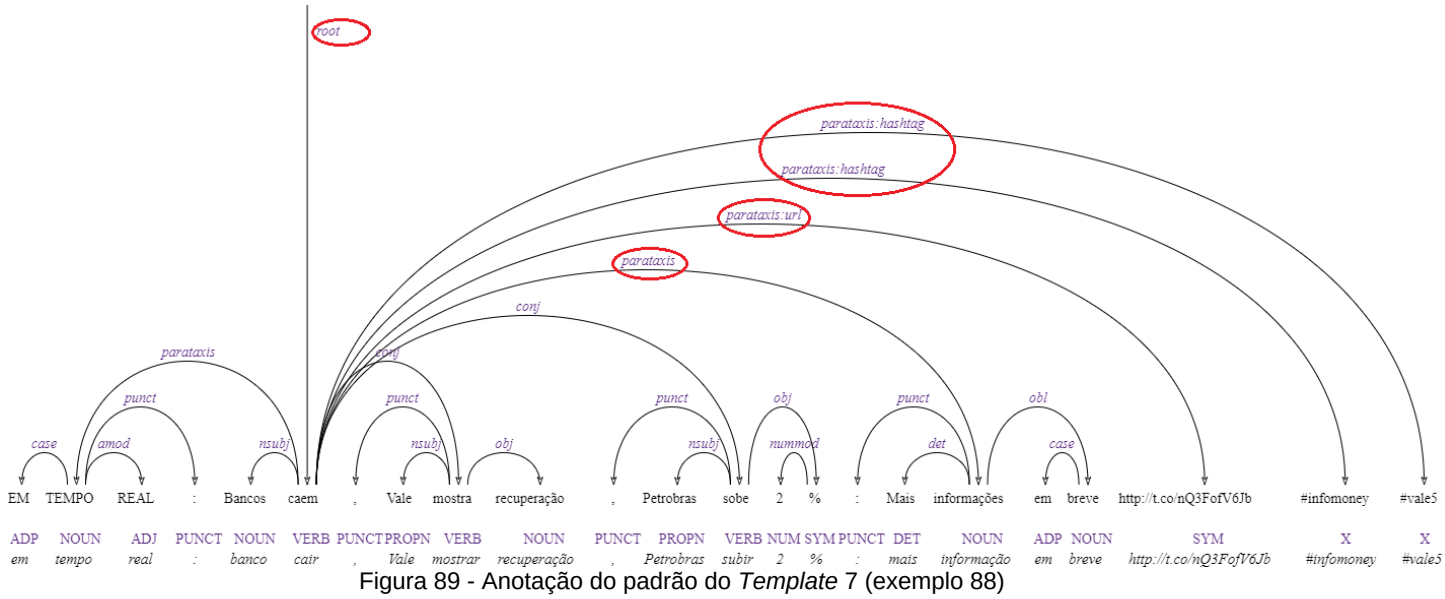


Figura 89 - Anotação do padrão do *Template* 7 (exemplo 88)

Exemplo:

(89) Usiminas e CSN caem 5 % , Petrobras e Vale recuam quase 3 % , uma small cap ganha destaque : Aind ... http://t.co/7vc5IkBJ30 #infomoney #vale5

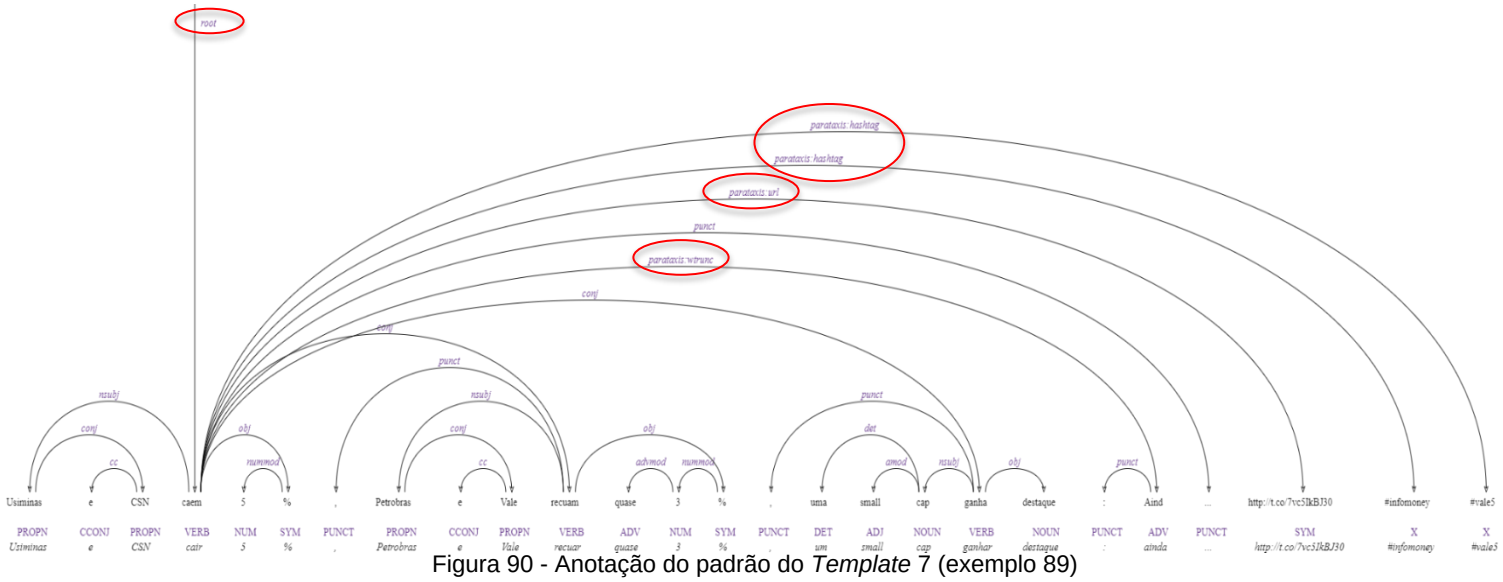


Figura 90 - Anotação do padrão do *Template* 7 (exemplo 89)

Template 8

Padrão : **Rastreamento ações – Gráfico diário – 14h. Analise se romper : <lista *ticker*&valor>**, em que:

1. **root** é o *token* “analise”
2. <lista *ticker*&valor> é dependente por **obj** de **root**,
3. <lista *ticker*&valor> é uma lista de um ou mais itens coordenados (**conj**), sendo: || <*ticker*> <valor> || <*ticker*> <valor> || <*ticker*> <valor>....
4. valor é dependente por **nmod** de ação
5. 3 primeiros elementos estão em lista (**list**) e como **parataxis** de **root**

Exemplo:

(90) Rastreamento ações - Gráfico diário - 14h . Analise se romper : ABEV3 16.97 BBTG11 27.62 BOVA11 48.15 BRML3 18,3 BVMF3 10.69 CPFE3 18.01 ICON 2301.35

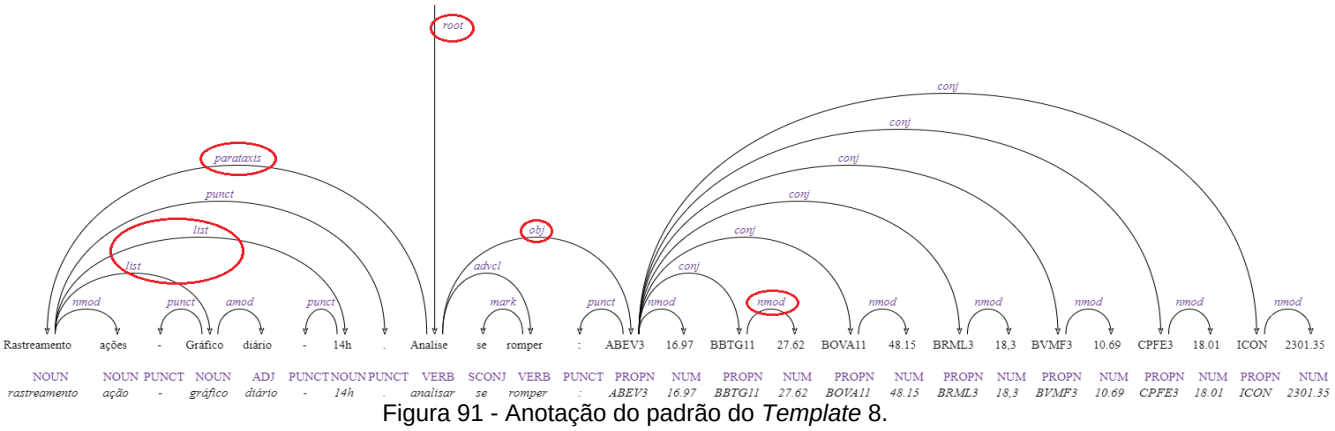


Figura 91 - Anotação do padrão do *Template* 8.

Template 9

Padrão: **Ativo c/ vol Financeiro Superior a a sua MM21 - <hora> : <lista *tickers*>**, em que:

1. *token* “Ativo” é **root**
2. <hora> é dependente por **parataxis** de **root**
3. As ações (*tickers*), conectadas por **conj**, estão como **appos** de **root**

Exemplo:

(91) Ativo c/ vol Financeiro Superior a a sua MM21 - 14h : BBSE3 PINE4

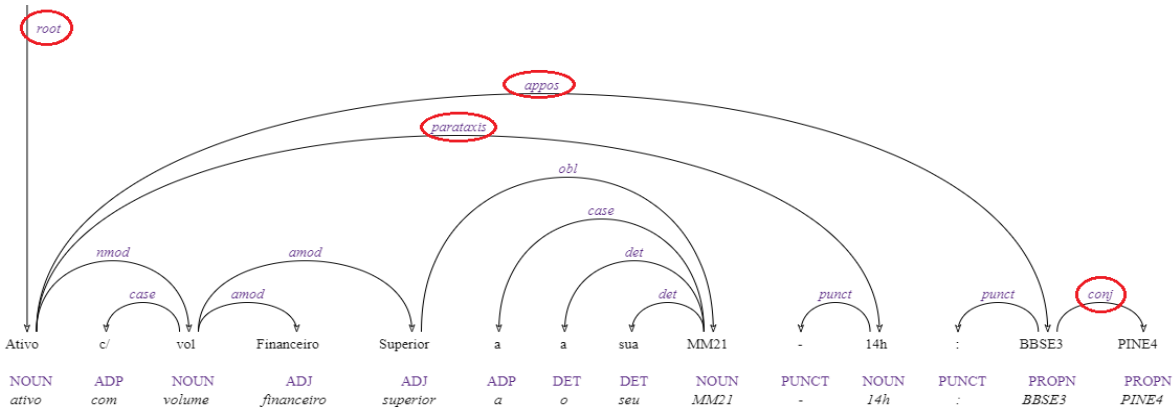


Figura 92 - Anotação do padrão do *Template* 9.

Template 10

Padrão: **<sentença>. Confira a nova indicação agora em url**, em que:

1. <sentença> é sintaticamente completa e contém o **root**
2. “Confira a nova indicação agora em url” é dependente por **parataxis** de <sentença>, e “Confira” é seu ***head***

Exemplo:

(92) A última indicação de a #MRVE3 resultou em - 1.25 % . Confira a nova indicação agora em http://t.co/kgt1YiTbF7

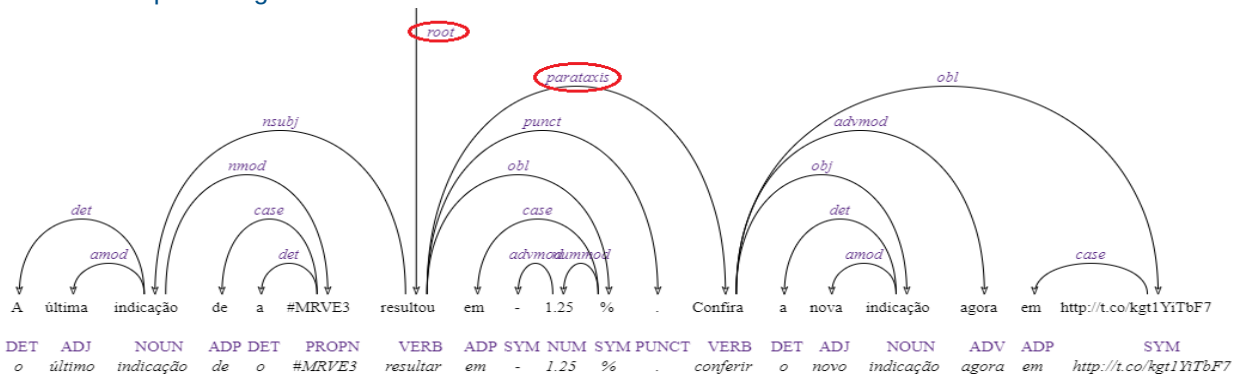


Figura 93 - Anotação do padrão do *Template* 10.

Template 11

Padrão: **<ticker> <tema> <url>**, em que:

1. <ticker> é dependente por **nmod** de **root**
2. <tema> contém “suportes e resistências”, sendo “suportes” o **root**
3. <url> é dependente por **parataxis:url** de **root**

Exemplo:

(93) #VALE5 suportes e resistências http://t.co/c8OrWXrECN

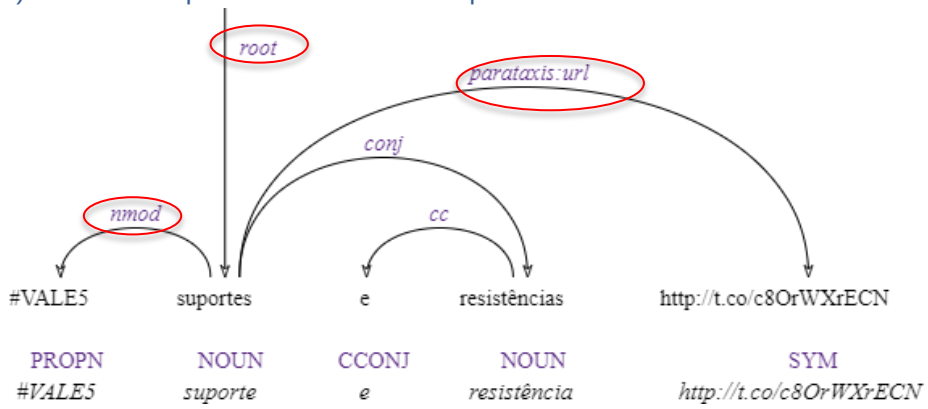


Figura 94 - Anotação do padrão do *Template* 11 (exemplo 93).

Exemplo:

(94) #VALE5 suportes e resistências, veja ainda notícia em o comentário. http://t.co/sJrLzoBlUT

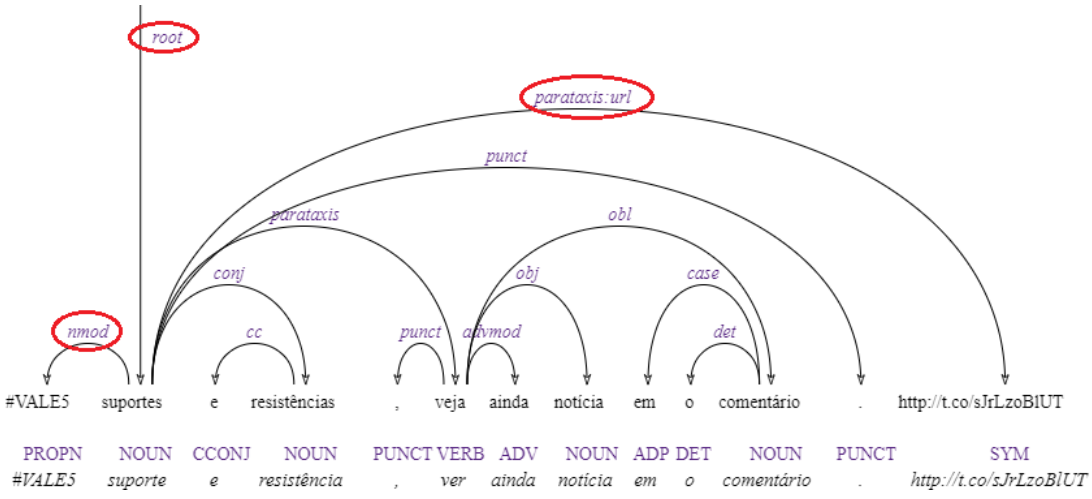


Figura 95 - Anotação do padrão do *Template* 11 (exemplo 94).

Template 12

Padrão: **<prefixo> Prepare-se para o próximo pregão! <sufixo> <lista *tickers*>. Assista! url**, em que:

1. <prefixo> e <sufixo> são opcionais; quando presentes, podem ser predicados ou sintagmas nominais, e se relacionam por **parataxis** ao **root**
2. **root** é o verbo “Prepare”
3. as ações em <lista *tickers*> estão em coordenação (**conj**) e podem estar em **parataxis** com **root**, se não houver <sufixo>, ou estão relacionadas apropriadamente ao <sufixo>

Exemplo:

(95) Prepare - se para o próximo pregão ! Análise : IBOV , PETR4 , VALE5 , TIMP3 , LPSB3 , MULT3 , GRND3 , LIGT3 e ABCB4 . Assista ! http://t.co/OVfdQj4UPe

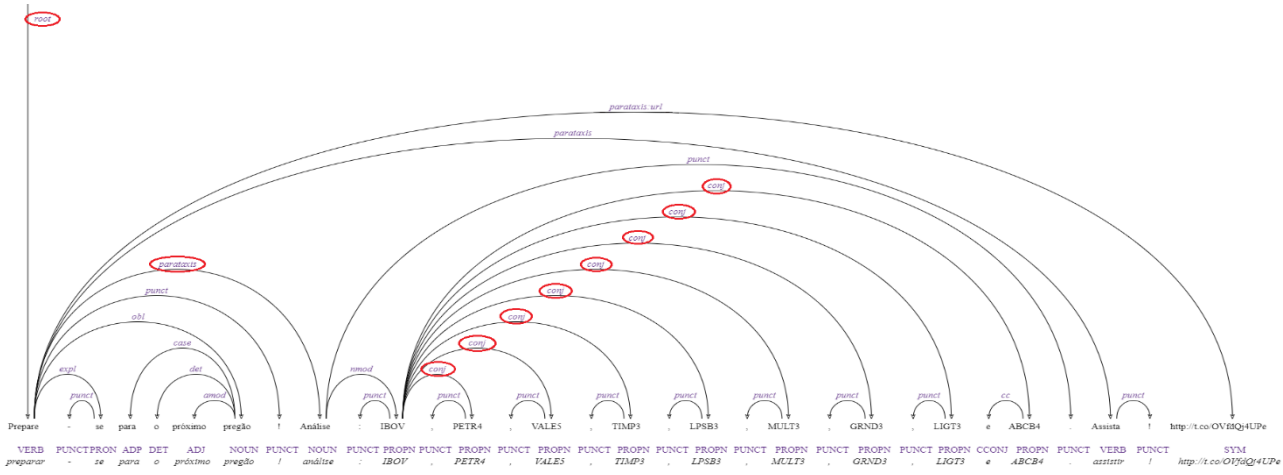


Figura 96 - Anotação do padrão do *Template* 12 (exemplo 95).

Exemplo:

(96) Bom dia ! Prepare - se para o pregão de esta 4ª . Assista : IBOV , PETR4 , VALE5 , EMBR3 , SMLE3 , HYPE3 , CRUZ3 , PMAM3 e NATU3 . http://t.co/SbHM6qlUrO

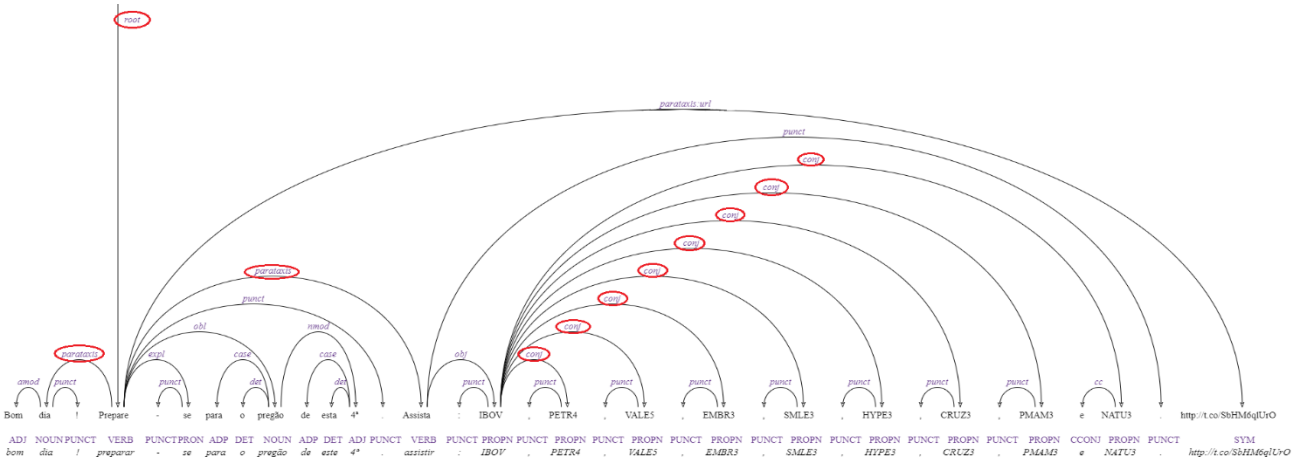


Figura 97 - Anotação do padrão do *Template* 12 (exemplo 96).

Exemplo:

(97) RT @daltonvieira : Prepare - se para o próximo pregão ! IBOV , PETR4 , VALE5 , SMLE3 , USIM5 , TCSA3 , LAME4 , MRVE3 , HYPE3 e MGLU3 . Assista ! http://t…

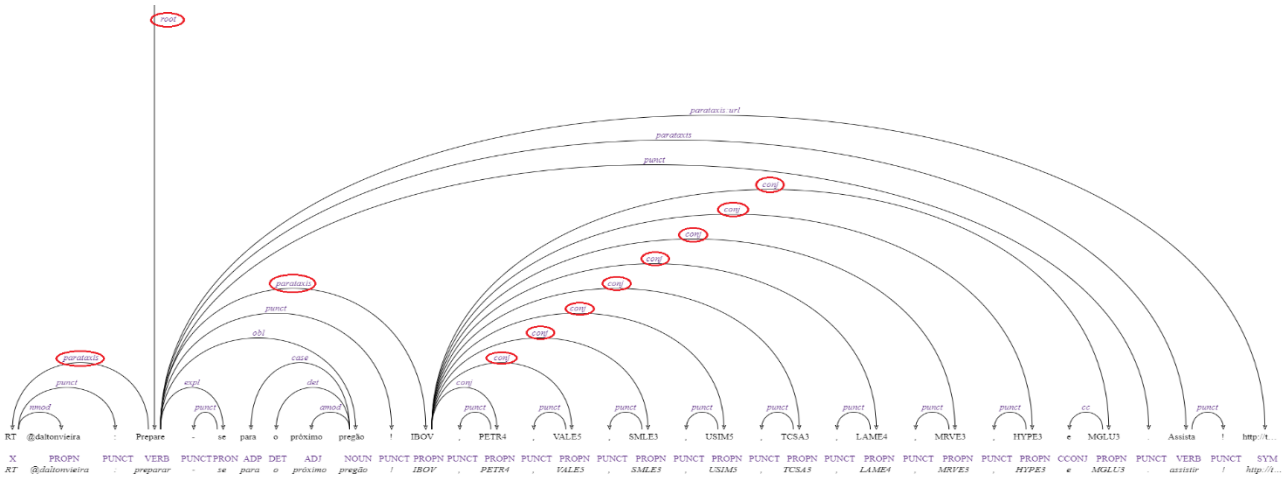


Figura 98 - Anotação do padrão do *Template* 12 (exemplo 97).

Template 13

Padrão: **<prefixo> : <sentença> url**, em que:

1. <prefixo> menciona empresa e ou ação
2. <sentença> é sintaticamente bem-formada e, por isso, contém o **root**
3. <prefixo> é dependente de **parataxis** de *root*, e seu ***head*** depende de sua forma
4. url é dependente por **parataxis:url** do *root*

Exemplo:

(98) Sabesp (SBSP3) : Responsável por o abastecimento de 47 % de a região metropolitana de São Paulo , o Sistema Cantareira ... http://t.co/xs8ji7To5z

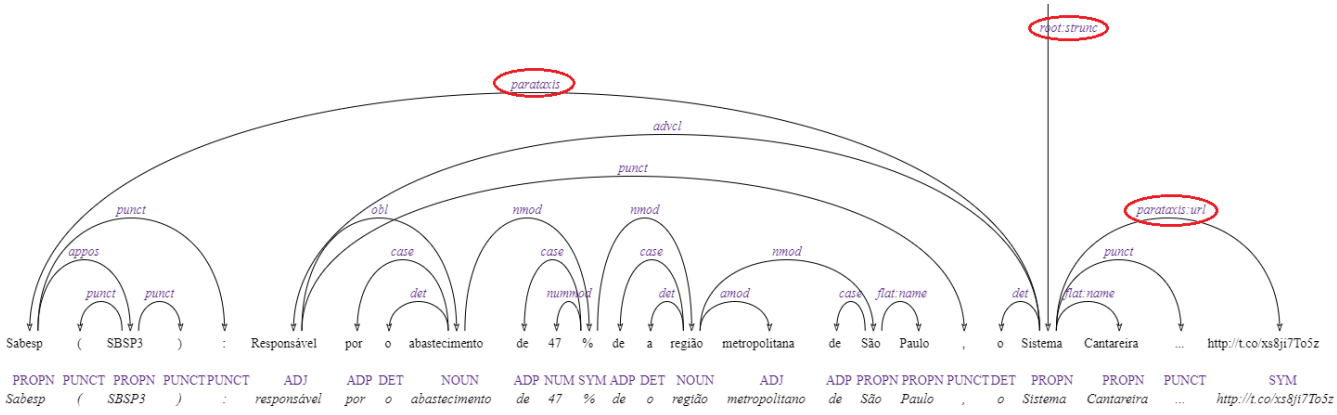


Figura 99 - Anotação do padrão do *Template* 13 (exemplo 98).

Exemplo:

(99) Notas gerais SABESP (SBSP3) : nível de o sistema cantareira caiu para 15 % , novo recorde de baixa segundo a ... http://t.co/cQZeQ7P09u

Figura 100 - Anotação do padrão do *Template* 13 (exemplo 99).

Exemplo:

(100) #SBSP3 E cadê o racionamento ! ! ! http://t.co/ApULfZt5Sr

(Interpretação: “E o racionamento está onde?”: onde é o predicativo, racionamento é o sujeito).

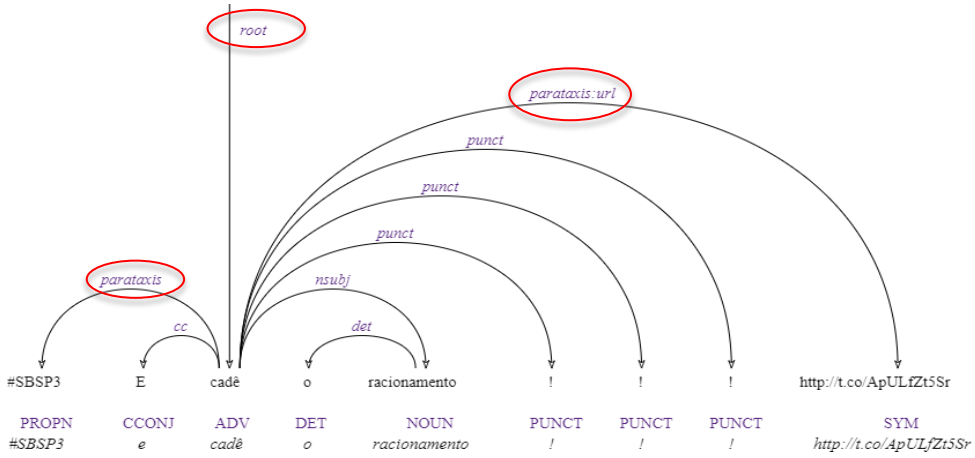


Figura 101 - Anotação do padrão do *Template* 13 (exemplo 100).

Template 14

Padrão: **<sentença> : <sentença truncada> url <lista *hashtags*>**, em que:

1. <sentença> é sintaticamente completa e contém o **root**
2. <sentença truncada> é uma sentença truncada e seu ***head*** é dependente de <sentença> por **parataxis:strunc**
3. url é dependente por **parataxis:url** do **root**
4. <lista *hashtags*> são dependentes por **parataxis:hashtag** do **root**

Exemplo:

(101) Confira o potencial de valorização de 66 ações , upside supera 120 % : Foram utilizadas as prev ... http://t.co/HbPlnX7Nxm #infomoney #vale5

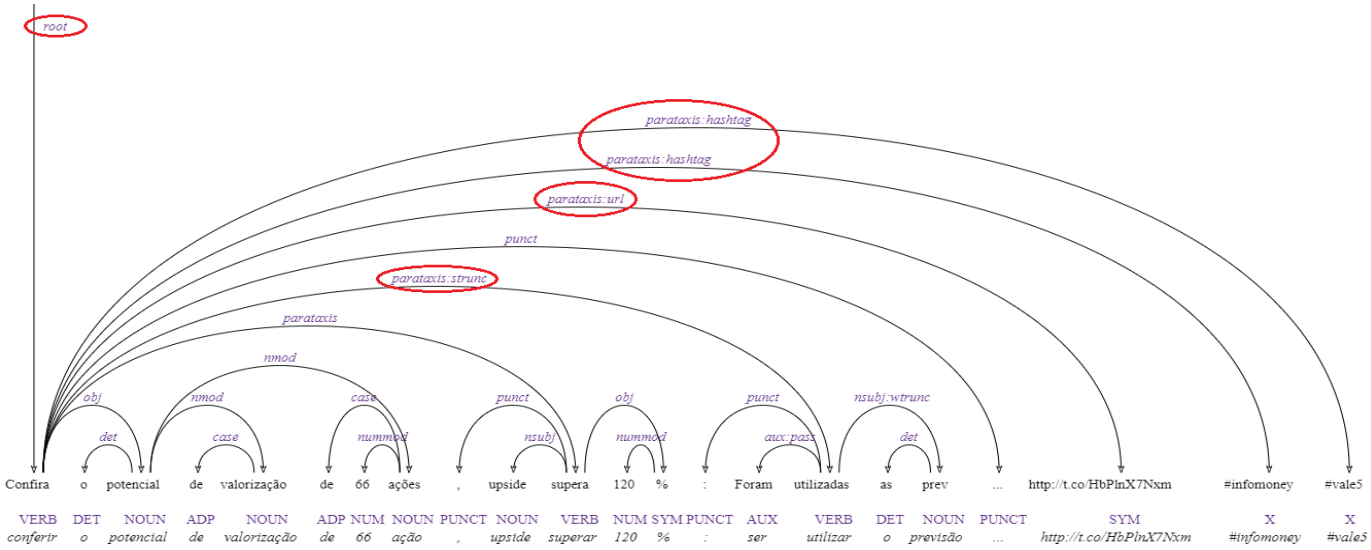


Figura 102 - Anotação do padrão do *Template* 14 (exemplo 101).

Exemplo:

(102) ' Quem montar uma carteira de ações agora vai morrer de rir em 10 anos ' , diz especialista : Se ... http://t.co/CrU7flmRmc #infomoney #vale5

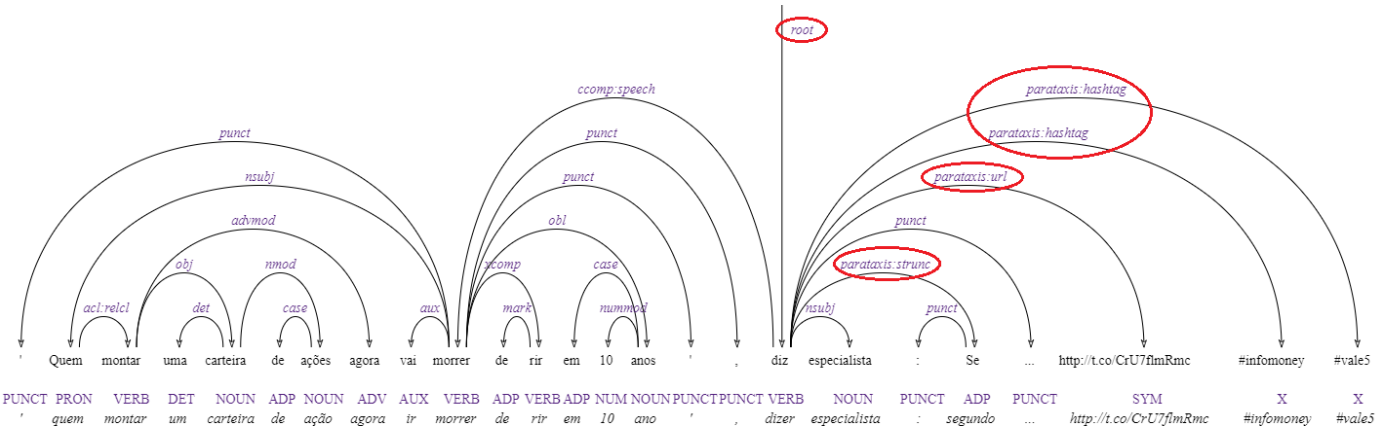


Figura 103 - Anotação do padrão do *Template* 14 (exemplo 102)

Exemplo:

(103) Wintrade recomenda uma nova ação para março , saiba qual : A nova escolhida por os analistas é ... http://t.co/fipOgc8hF4 #infomoney #vale5

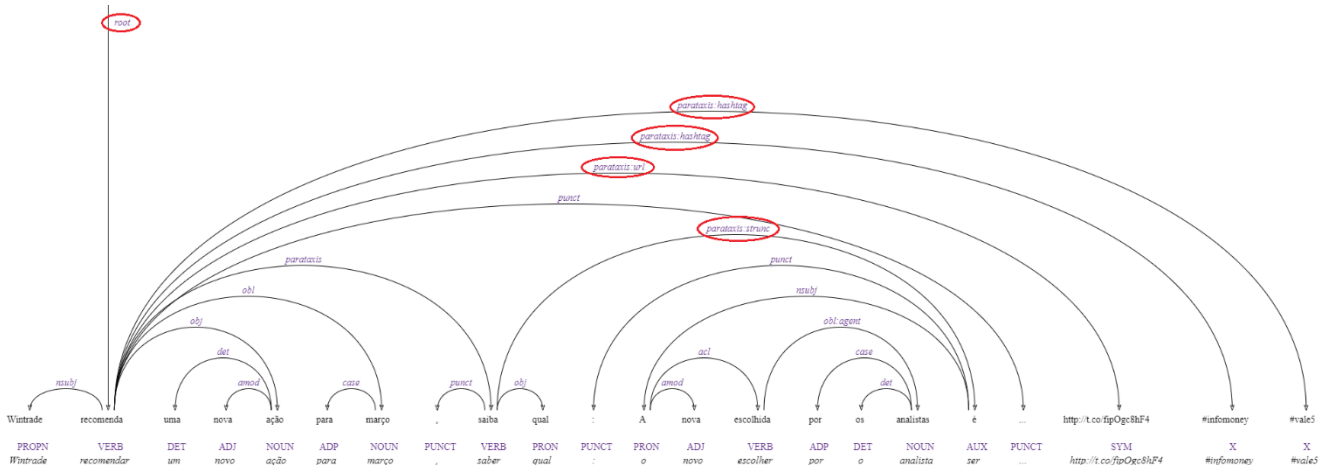


Figura 104 - Anotação do padrão do *Template* 14 (exemplo 103).

Template 15

Padrão: **<*cashtag*> <empresa> (tipo de ação) - Fato Relevante - <fato> url**

1. “Fato Relevante”, se presente, é o tema, e <fato> é **root**; se ausente, **root** está em <fato>
2. <fato> é dependente de **root** pela relação **parataxis**
3. <*cashtag*> é dependente por **nmod** de **root**
4. <empresa> é dependente por **nmod** de <*cashtag*>
5. (tipo de ação) é dependente por **appos** de <*cashtag*>
6. url está ligada por **parataxis:url** ao root

Exemplo:

(104) $MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo A o Acordo De Acionistas http://t.co/jBJzeY5U5V

(Intepretação: *Fato Relevante sobre ação $MRFG3 da Marfrig no novo mercado (nm): Primeiro aditivo ao acordo de acionistas*)

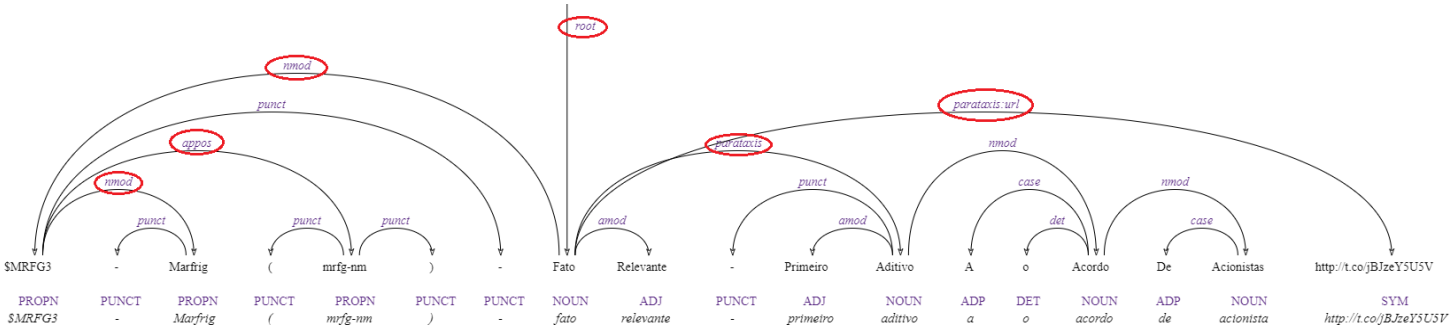


Figura 105 – Anotação do padrão do *Template* 15 (exemplo 104).

Exemplo:

(105) $RSID3 - Rossi Resid (rsid-nm) - Alteracao De a Politica De Divulgacao De Fato Relevante http://t.co/gJVTbOY0uA

(Intepretação: *Alteração da política de divulgação de fato relevante sobre $RSID3, da Rossi Resid, no novo mercado.*)

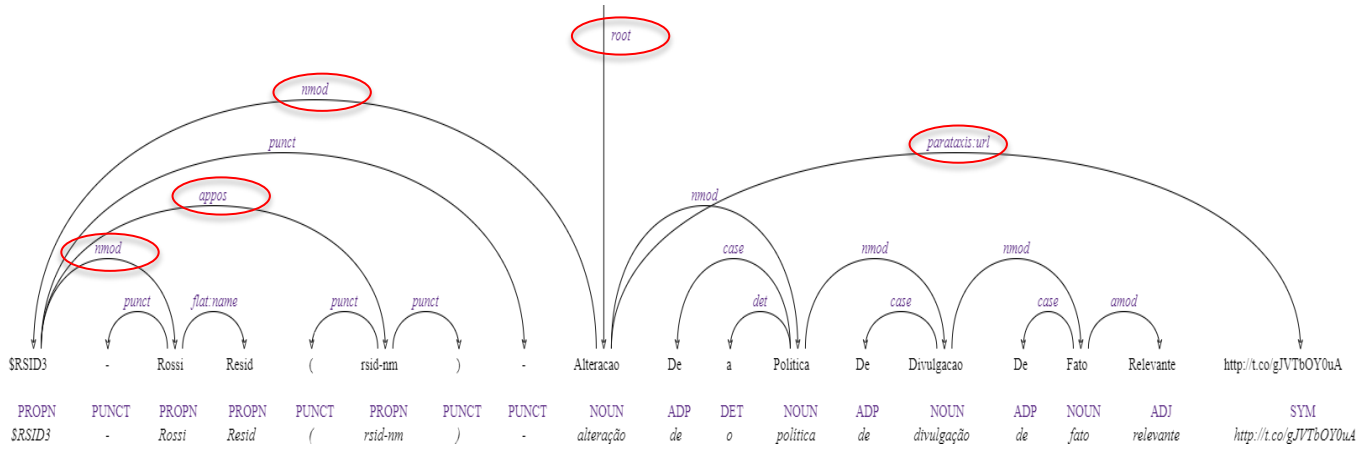


Figura 106 – Anotação do padrão do *Template* 15 (exemplo 105).

Exemplo:

(106) $BBAS3 - Banco De o Brasil (bbas-nm) - Esclarecimentos A Consulta De a Bm&fbovespa http://t.co/UaFXNfKifj

(Interpretação: *Esclarecimentos (sobre) a consulta da Bm&fbovespa (sobre) $BBAS3 (do) Banco do Brasil (bbas-nm*)

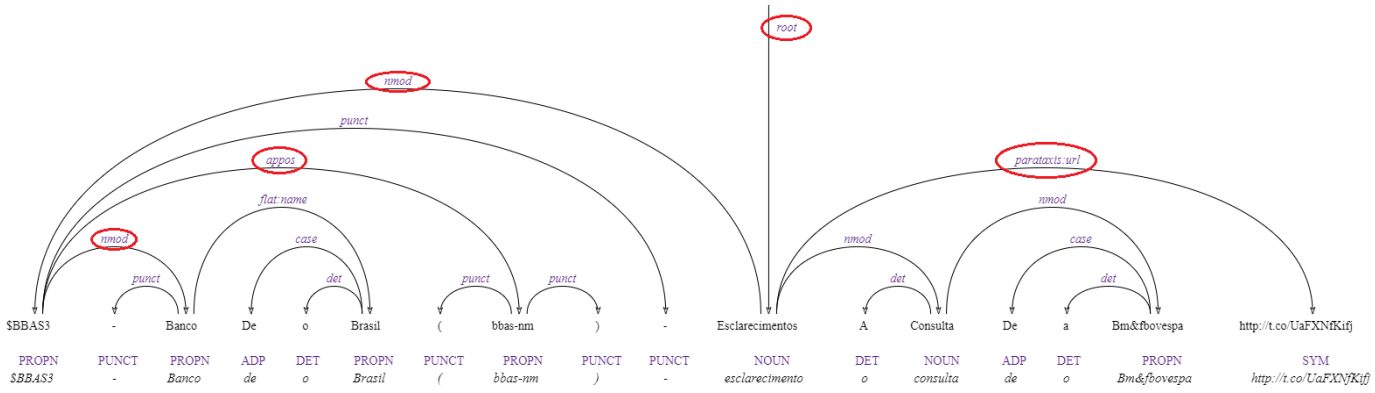


Figura 107 – Anotação do padrão do *Template* 15 (exemplo 106).

Template 16

Padrão: **[prefixo opcional] <sentença> url**, em que:

1. [prefixo] pode ter <lista *hashtags*>
2. <sentença> contém o *root*
3. cada elemento do [prefixo] é dependente por **parataxis** de **root**
4. url é dependente por **parataxis:url** de **root**

Exemplo :

(107) #BOVESPA #USIM5 S&P eleva perspectiva de nota de a Usiminas para estável : http://t.co/JdfGy10phl #BR

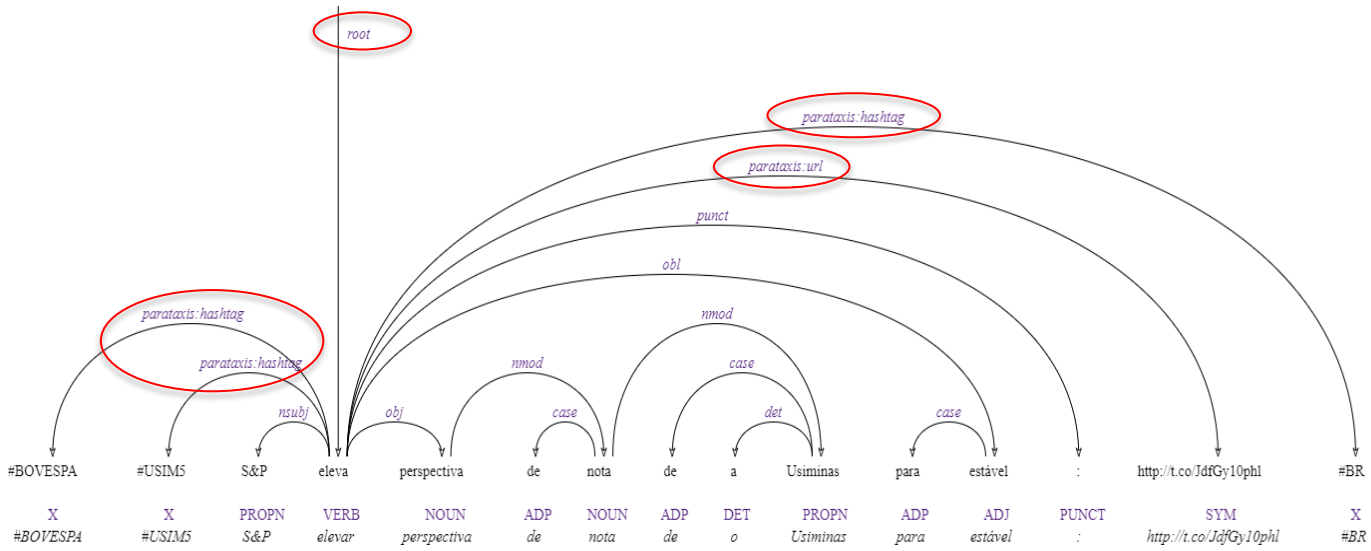


Figura 108 – Anotação do padrão com [prefixo] do *Template* 16 (exemplo 107).

Template 17

Padrão: **<lista *ticker*&[valor opcional]>**, em que:

1. <lista *ticker*&[valor opcional]> é uma lista de um ou mais itens coordenados (**conj**), sendo: || <*ticker*> [valor-reais] || <*ticker*> [valor-reais] || <*ticker*> [valor-reais]....
2. [valor opcional]: cada *ticker* pode ou não ocorrer seguido do valor da ação que representa

Exemplo:

(108) Prepare - se para o próximo pregão ! Análise : **IBOV , PETR4 , VALE5 , TIMP3 , LPSB3 , MULT3 , GRND3 , LIGT3 e ABCB4** . Assista ! http://t.co/OVfdQj4UPe

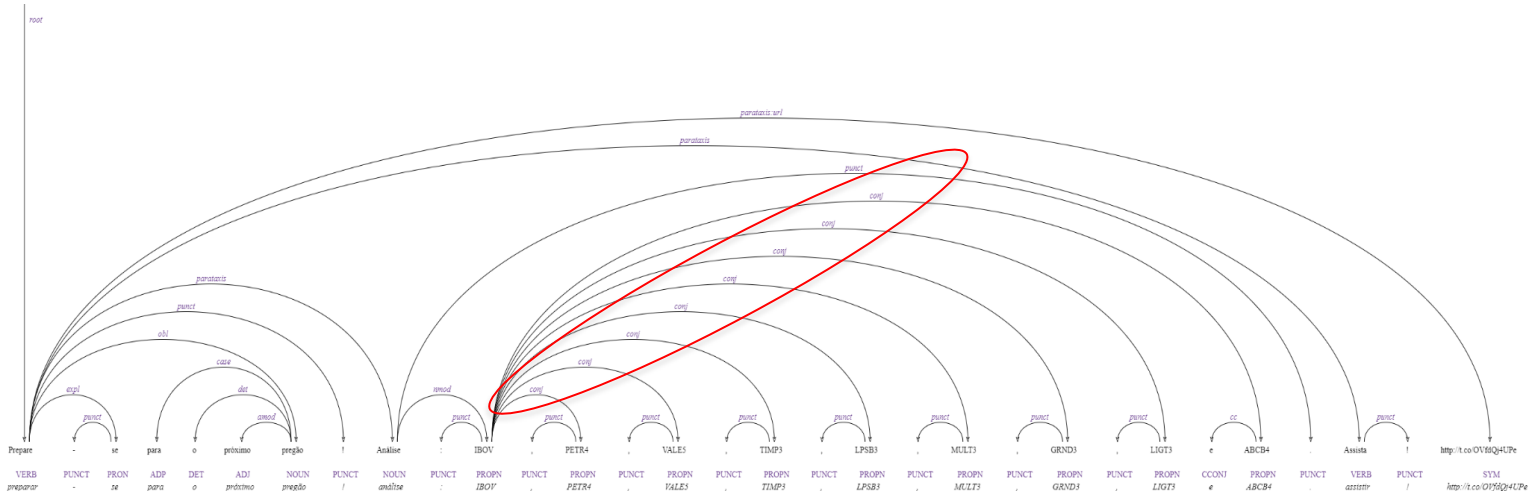


Figura 109 – Anotação do padrão do *Template* 17 (sem os valores expressos).

Exemplo:

(109) 14/03/2014 - 17:19 : Maiores Baixas : **MRVE3 - 12,5 % R$ 7,35 , DASA3 - 9,67 % R$ 15,13 , CMIG4 - 5,69 % R$ 12,94 , GFSA3 - 4,76 % R$ 3 , ELPL4 - 4,03 % R$ 7,62** .

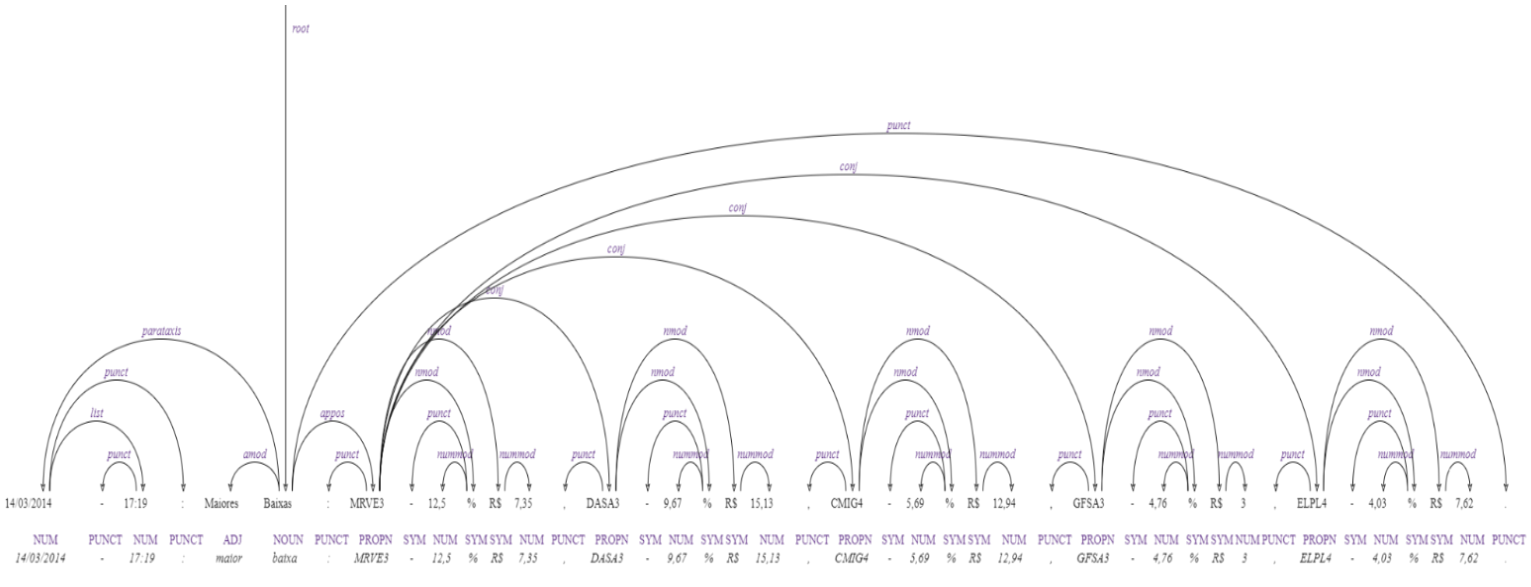


Figura 110 – Anotação do padrão do *Template* 17 (com os valores expressos).

Template 18

Padrão: **Gatilho position , semana que vem : Ativo cmg4 - Venda - Validade 14/mar - Start 12,89 - Stop 13,75 - RP 12,03 http://t.co/MPfzB6Vj0o**

(110) Gatilho position , semana que vem : Ativo cmg4 - Venda - Validade 14/mar - Start 12,89 - Stop 13,75 - RP 12,03 http://t.co/MPfzB6Vj0o.

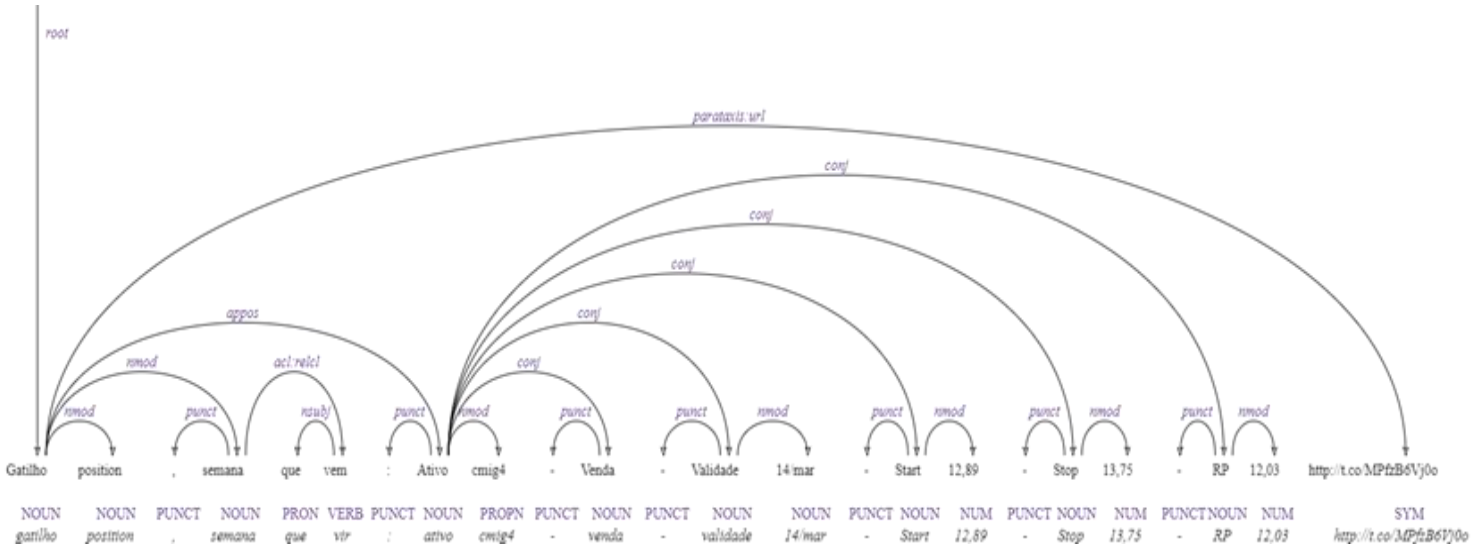


Figura 111 – Anotação do padrão do *Template* 18.

Template 19

Padrão: **INTRADAY PETR4 : Suportes 13,9 e 14,74 e resistências 15,99 e 16,4 INTRADAY VALE5 : Suportes 27,04 e 27,45 e resistências 28,12 e 28,38**

(111) INTRADAY PETR4 : Suportes 13,9 e 14,74 e resistências 15,99 e 16,4 INTRADAY VALE5 : Suportes 27,04 e 27,45 e resistências 28,12 e 28,38

(Interpretação: lista de informações de suportes e resistencias “intraday” de ações)

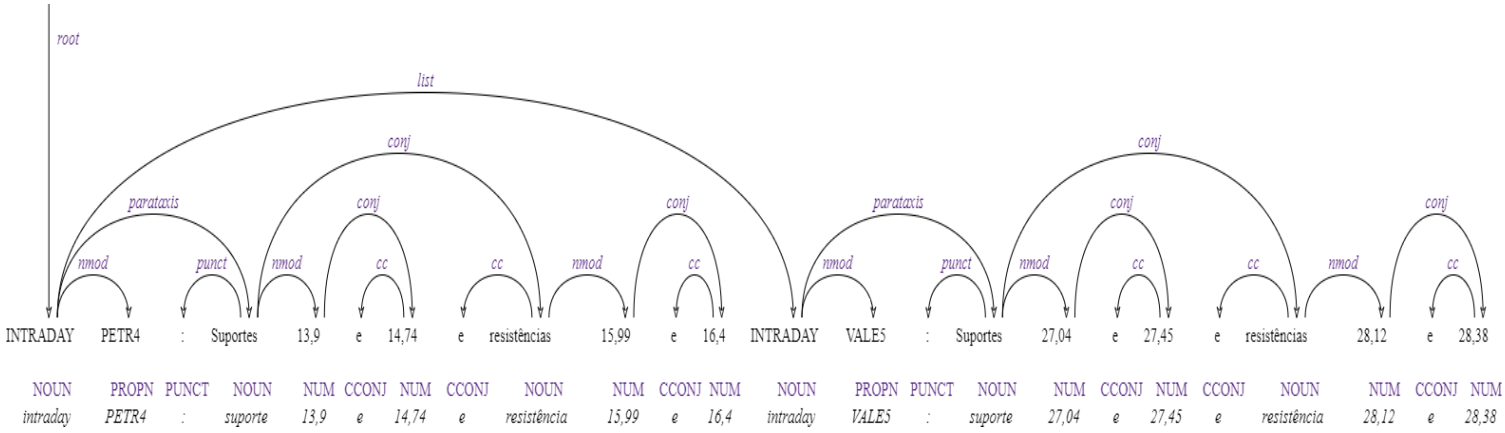


Figura 112 – Anotação do padrão do *Template* 19.

Template 20

Padrão: **Ações ex-dividendos hoje : GGBR3 , GGBR4 , GOAU3 , GOAU4 e MPLU3 . As cotações históricas foram ajustadas . Saiba mais ! http://t.co/C7k4DuDID2**

Exemplo:

(112) Ações ex-dividendos hoje : GGBR3 , GGBR4 , GOAU3 , GOAU4 e MPLU3 . As cotações históricas foram ajustadas . Saiba mais ! http://t.co/C7k4DuDID2

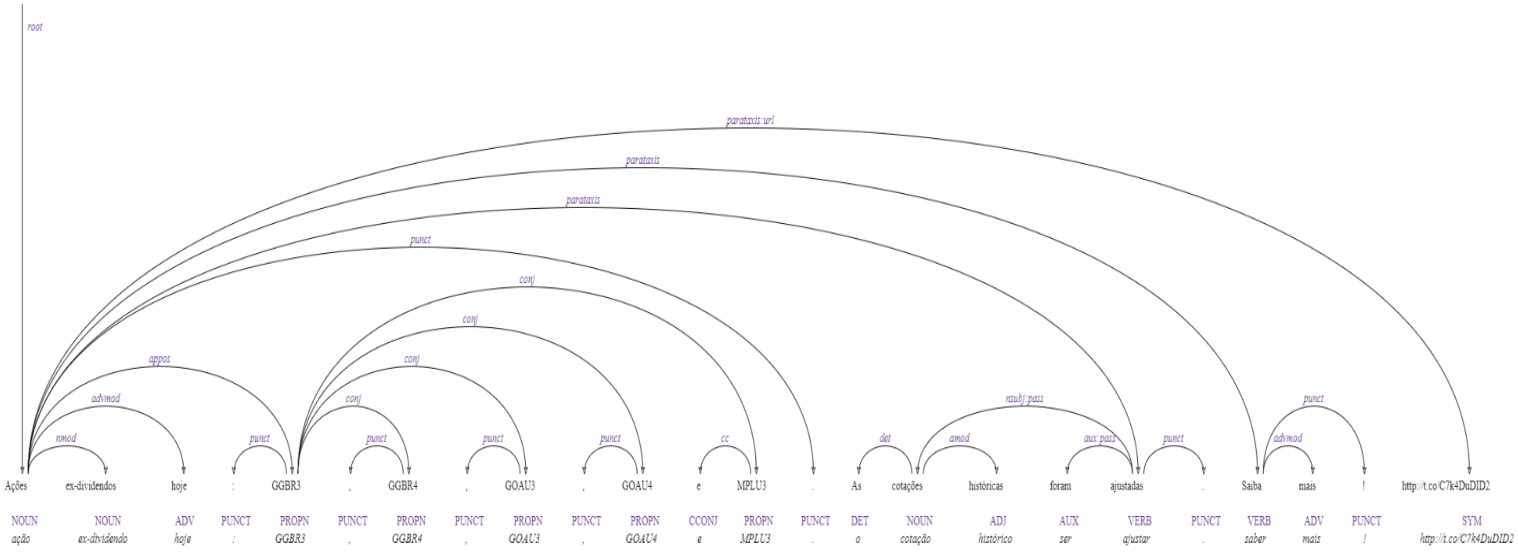


Figura 113 – Anotação do padrão do *Template* 20.

Template 21

Padrão: **Desempenho de as ações <empresa> em a semana passada : [bolsa] – <lista *ticker*&índice> , [bolsa – <lista *ticker*&índice>]**, em que:

1. *token* “desempenho” é **root**
2. <empresa> indica o nome da empresa, que é **nmod** de “ações”
3. [bolsa] indica bolsa de valores e é opcional; se presente, liga-se por **parataxis** a **root**
4. <lista *ticker*&índice> são 2 ou mais elementos coordenados (**conj**) e ligados por **parataxis** a [bolsa] (quando presente) (Figura 45) ou ao **root** (Figura 46), sendo:

|| <*ticker*> <índice> || <*ticker*> <índice> ...

5. [bolsa - <lista ticker&índice>] é opcional e, quando presente, liga-se ao primeiro [bolsa] por **conj**

Exemplo:

(113) Desempenho de as ações de o Itaú Unibanco em a semana passada : em a Bovespa - ITUB4 : - 0,9 % , ITUB3 : 0,58 % , em a NYSE – ITUB : - 1,13 % , Índice Bovespa : - 1,8 %.

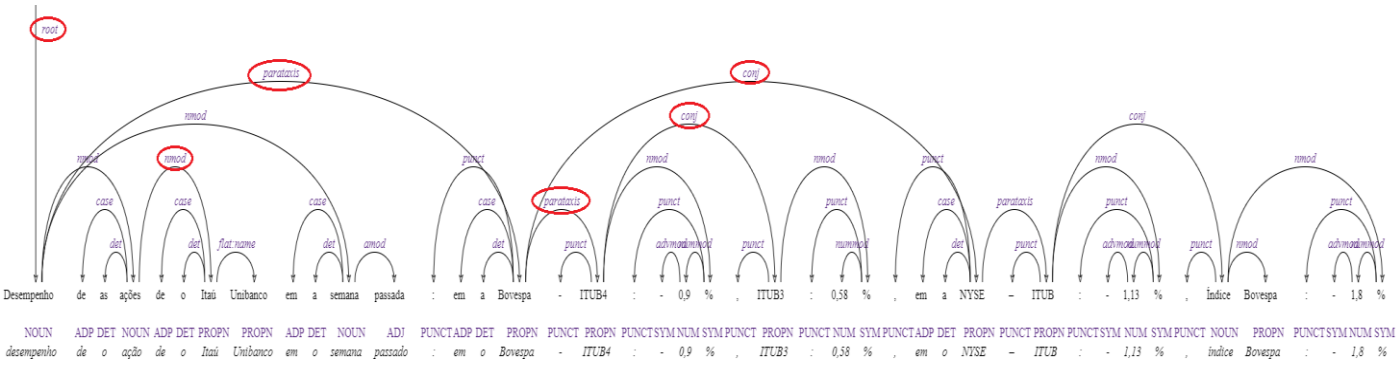


Figura 114 – Anotação do padrão (com os opcionais) do *Template* 21 (exemplo 114).

Exemplo:

(114) Desempenho de as ações de a TIM em a semana passada : TIMP3 (Bovespa) : - 0.94 % , TSU (NYSE) : + 0.58 % , IBOV : + 2.64 % .

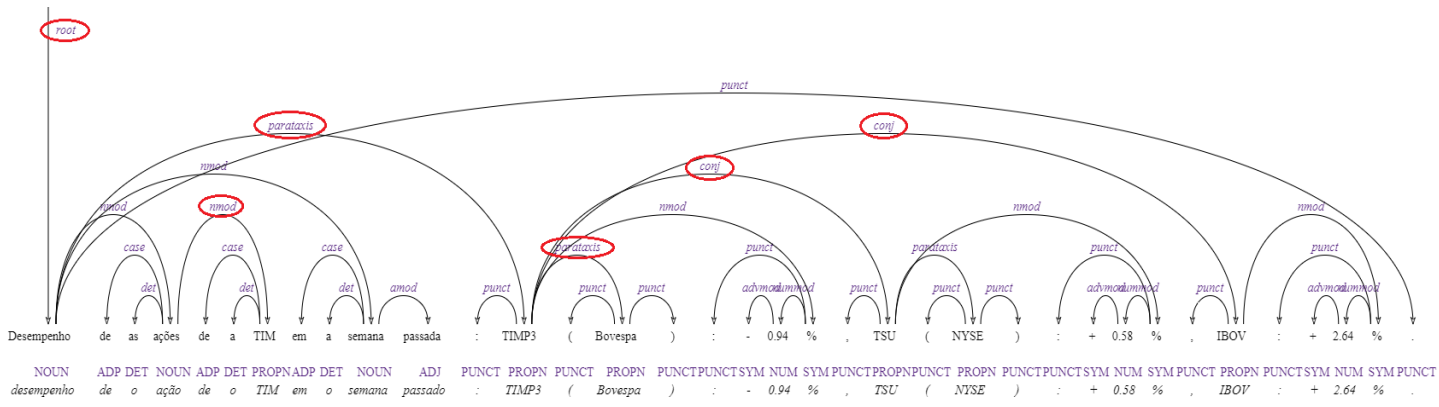


Figura 115 – Anotação do padrão (sem os opcionais) do *Template* 21 (exemplo 115).

Template 22

Padrão: **MRVE3 vender a R$ 7.99 indicado em 05/03/2014 13:29 e finalizou a compra com resultado de R$ - 0.11 ou - 1.36 % http://t.co/kgt1YiTbF7**, em que

1. MRVE3 é dependente por **obj** do **root**, já que é o objeto que complementa o verbo “vender” no infinitivo.

   Exemplo:

(115) **MRVE3 vender a R$ 7.99 indicado em 05/03/2014 13:29 e finalizou a compra com resultado de R$ - 0.11 ou - 1.36 % http://t.co/kgt1YiTbF7**,

(Interpretação: “Vender (ação) MRVE3 a R$7.99 (valor) indicado em 05/03/2014 (às) 13:29 e finalizou a compra com resultado de R$-0.11 ou -1.36%”)

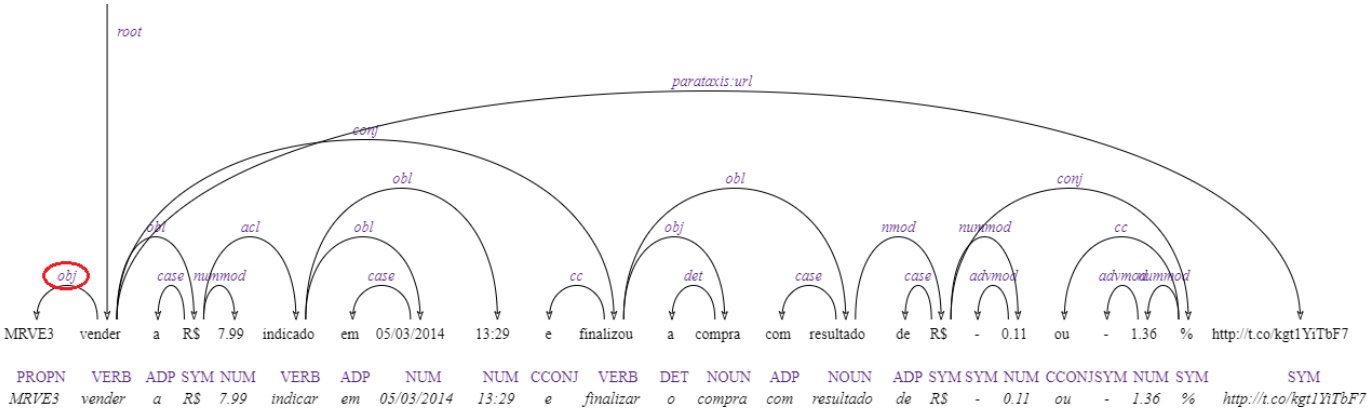


Figura 116 – Anotação do padrão do *Template* 22.

Bibliografia

Afonso, S.; Bick, E.; Haber, R.; Santos, D. (2002). Floresta sintá(c)tica: um treebank para o português. In Anais do XVII Encontro Nacional da Associação Portuguesa de Linguística, pp. 533-545.

Andrews, A. D. (2007). The major functions of the noun phrase. In Timothy Shopen, editor, Language Typology and Syntactic Description. Volume I: Clause Structure. Second edition, pp. 132-223. Cambridge University Press, Cambridge, UK.

Bouma, G.; Hajic, J.; Haug, D.; Nivre, J.; Solberg, E.; Øvrelid, L. (2018). Expletives in Universal Dependency Treebanks. In the Proceedings of the Second Workshop on Universal Dependencies (UDW), pp. 18-26.

Bresnan, J. (1982). Control and Complementation. Linguistic Inquiry, Vol. 13, N. 3, pp. 343-434. The MIT Press.

Di Felippo, A.; Postali, C.; Ceregatto, G.; Gazana, L.S.; Silva, E.H.; Roman, N.T.; Pardo, T.A.S. (2021). Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies. In the Proceedings of the VII Workshop on Portuguese Description (JDP), pp. 335-343. December, 1.

Duran, M.S. (2022). Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos-SP, Outubro, 166p.

Nivre, J. (2015). Towards a Universal Grammar for Natural Language Processing. In the Proceedings of the 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), pp. 3-16.

Nivre, J.; Marneffe, M-C.; Ginter, F.; Hajič, J.; Manning, C.D.; Pyysalo, S.; Schuster, S.; Tyers, F.; Zeman, D. (2020). Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In the Proceedings of the 12nd International Conference on Language Resources and Evaluation (LREC), pp. 4034-4043.

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 the Proceedings of the XIV Symposium in Information and Human Language (STIL), pp. 1-10.

Rademaker, A.; Chalub, F.; Real, L.; Freitas, C.; Bick, E.; Paiva, V. (2017). Universal Dependencies for Portuguese. In the Proceedings of the 4th International Conference on Dependency Linguistics (Depling), pp. 197-206.

Souza, E.; Cavalcanti, T.; Silveira, A.; Evelyn, W.; Freitas, C. (2020). Diretivas e documentação de anotação UD em português (e para língua portuguesa). Disponível em: https://nbviewer.jupyter.org/github/comcorhd/Documenta-o-UD-PT/raw/master/Documenta-o-UD-PT.pdf

Thompson, S. A. (1997). Discourse Motivations for the Core-Oblique Distinction as a Language Universal Directions in Functional Linguistics. In: Akio Kamio (ed.), Studies in Language Companion, Series 36, pages 59-82. John Benjamins, Amsterdam.

Zeman, D. (2017). Core Arguments in Universal Dependencies. In the Proceedings of the Fourth International Conference on Dependency Linguistics (Depling), pp. 287-296

Apêndice B – As classes e os tipos semânticos dos Npred

Npred	Frame	Classe	NOM-TYPE
abastecimento	supply	NOM	VERB-NOM
acesso	access	NOM	VERB-NOM
acordo	accord	NOM	VERB-NOM
acordo	accordance	NOM	VERB-NOM
acordo	agreement	NOM	VERB-NOM
aditivo	amendment	NOM	VERB-NOM
administração	administration	NOM	VERB-NOM
ajuste	adjustment	NOM	VERB-NOM
aliança	alliance	NOM	VERB-NOM
alavancagem	leverage	NOM	VERB-NOM
alienação	divestiture	NOM	VERB-NOM
alteração	change	NOM	VERB-NOM
amor	love	NOM	VERB-NOM
antro	hideaway	NOM	P-OBJ-PART
apreensão	seizure	NOM	VERB-NOM
aquisição	acquisition	NOM	VERB-NOM
assinatura	signing	NOMING	VERB-NOM
atestado	certificate	NOMLIKE	VERB-NOM
ataque	attack	NOM	VERB-NOM
avaliação	evaluation	NOM	VERB-NOM
briga	fight	NOM	VERB-NOM
caminho	path	ABILITY	NOM-REL
candidato	candidate	NOMLIKE	OBJECT
cara	appearance_02	NOM	VERB-NOM
carteira	portfolio_04	NOMLIKE	VERB-NOM
causa	reason_02	NOMLIKE	VERB-NOM
cenário	situation	NOM	VERB-NOM
chance	chance	NOM	VERB-NOM
comissão	commission_02	NOM	OBJECT
comparação	comparison	NOM	VERB-NOM
compartilhamento	sharing	NOMING	VERB-NOM
compra	buy	NOM	VERB-NOM
comprador	buyer	NOM	SUBJECT
confirmação	confirmation	NOM	VERB-NOM
confirmação	confirmation	NOM	VERB-NOM
construção	construction	NOM	VERB-NOM
contrato	contract	NOM	VERB-NOM
controle	control	NOM	VERB-NOM
conversa	conversation	NOM	VERB-NOM
conversão	conversion	NOM	VERB-NOM
convocação	convocation	NOM	VERB-NOM
coordenador	coordinator	NOM	SUBJECT
coragem	courage	NOMADJ	ADJ-NOM
corte	cut_02	NOM	VERB-NOM
cotação	quotation	NOM	VERB-NOM
curiosidade	interest	NOM	VERB-NOM
custódio	custody	NOMLIKE	VERB-NOM
daytrade	trade	NOM	VERB-NOM
decisão	decision	NOM	VERB-NOM
declaração	declaration	NOM	VERB-NOM

Npred	Frame	Classe	NOM-TYPE
deliberação	deliberation	NOM	VERB-NOM
demanda	demand	NOM	VERB-NOM
denúncia	complaint	NOM	VERB-NOM
descoberta	discovery	NOM	VERB-NOM
descolamento	gap	NOMLIKE	VERB-NOM
desova	sell-off	NOM	VERB-PART
diretor	director	NOM	SUBJECT
diretoria	board	GROUP	NOM-REL
discussão	discussion	NOM	VERB-NOM
distribuição	distribution	NOM	VERB-NOM
diverência	divergence	NOM	VERB-NOM
divisão	division	PARTITIVE	NOM-REL
divisão	division_02	NOM	VERB-NOM
divulgação	release	NOM	VERB-NOM
eleição	election	NOM	VERB-NOM
encerramento	closure	NOM	VERB-NOM
entendimento	understanding	NOMING	VERB-NOM
entendimento	understanding_02	NOMLIKE	VERB-NOM
entrada	entrance	NOM	VERB-NOM
esclarecimento	enlightenment	NOM	VERB-NOM
estimativo	estimate	NOM	VERB-NOM
exemplo	example	NOM	SUBJECT
expectativa	expectation	NOM	VERB-NOM
exploração	exploration	NOM	VERB-NOM
extensão	extension	NOM	VERB-NOM
falta	lack	NOM	VERB-NOM
fruto	reward	NOM	VERB-NOM
fusão	fusion	NOM	VERB-NOM
gestor	manager	NOM	SUBJECT
homem	associate	NOM	OBJECT
hora	hour	ENVIRONMENT	NOM-REL
importação	import	NOM	VERB-NOM
incorporação	fusion	NOM	VERB-NOM
indicador	indicator	NOM	SUBJECT
indicação	indication	NOM	VERB-NOM
inscrição	registration	NOM	VERB-NOM
instauração	establishment	NOM	VERB-NOM
investimento	investment	NOM	VERB-NOM
leilão	auction	NOM	VERB-NOM
locação	allocation	NOM	VERB-NOM
medo	fear	NOM	VERB-NOM
meio	middle	PARTITIVE	NOM-REL
membro	member	PARTITIVE	NOM-REL
mix (mistura)	mix	NOM	VERB-NOM
monte	lot	PARTITIVE	NOM-REL
mudança	change	NOM	VERB-NOM
necessidade	need	NOM	VERB-NOM
negociação	negotiation	NOM	VERB-NOM
notícia	news	WORK-OF-ART	NOM-REL

Npred	Frame	Classe	NOM-TYPE
número	number	PARTITIVE	NOM-REL
obrigação	obligation	NOM	VERB-NOM
oferta	offer	NOM	VERB-NOM
olhada	look	NOM	VERB-NOM
olho	monitoring	NOMING	VERB-NOM
outorga	grant	NOM	VERB-NOM
pagador	payer	NOM	SUBJECT
pagamento	payment	NOM	VERB-NOM
pedido	request	NOM	VERB-NOM
perspectiva	perspective	NOMLIKE	VERB-NOM
posição	position	NOMLIKE	NOM-REL
post	posting	NOMING	VERB-NOM
postagem	posting	NOMING	VERB-NOM
procura	search	NOM	VERB-NOM
projeto	project	ABILITY	NOM-REL
projeção	projection	NOM	VERB-NOM
proposta	proposal	NOM	VERB-NOM
racionamento	shortage	NOM	VERB-NOM
rateio	allocation	NOM	VERB-NOM
reapresentação	submission	NOM	VERB-NOM
recomendação	recommendation	NOM	VERB-NOM
recompra	buy	NOM	VERB-NOM
redução	reduction	NOM	VERB-NOM
reeleição	election	NOM	VERB-NOM
relatório	report	NOM	VERB-NOM
relação	relation	NOM	VERB-NOM
renúncia	resignation	NOM	VERB-NOM
resistência	resistance	NOM	VERB-NOM
resolução	resolution	NOM	VERB-NOM
responsabilidade	responsibility	NOMADJ	ADJ-NOM
retorno	return_04	NOMLIKE	VERB-NOM
reunião	meeting	NOM	VERB-NOM
risco	risk	NOM	VERB-NOM
solicitação	solicitation	NOM	VERB-NOM
spread (diferença)	spread	NOMLIKE	VERB-NOM
sugestão	suggestion	NOM	VERB-NOM
taxação	taxation	NOM	VERB-NOM
tendência	tendency	NOM	VERB-NOM
teste	test	NOM	VERB-NOM
trade	trade	NOM	VERB-NOM
transferência	transfer	NOM	VERB-NOM
transporte	transport	NOM	VERB-NOM
troca	change	NOM	VERB-NOM
venda	sell	NOM	VERB-NOM
vendedor	seller	NOM	SUBJECT
visão	vision	NOMLIKE	VERB-NOM
volta	return	NOMLIKE	VERB-NOM