Programa de Pós-Graduação em

LINGUÍSTICA

REPRESENTAÇÃO FORMAL DE SIGNIFICADO: O CASO DOS *TWEETS* DO MERCADO FINANCEIRO EM PORTUGUÊS.

Gabriel Ceregatto

SÃO CARLOS
2025

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

GABRIEL CEREGATTO
BOLSISTA FUSP/SOFTEX

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
Gabriel Ceregatto, 2025

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 Gabriel Ceregatto, realizada em 11/08/2025.

Comissão Julgadora:

Profa. Dra. Ariani Di Felippo (UFSCar)

Profa. Dra. Eloize Rossi Marques Seno (IFSP)

Prof. Dr. Jackson Wilke da Cruz Souza (UFBA)

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

Agradecimentos

Dirijo os meus mais profundos e sinceros agradecimentos à minha orientadora, Profa. Dra. Ariani Di-Felippo, por sua maravilhosa orientação repleta de paciência, empatia e apoio constante ao longo do desenvolvimento deste projeto. Sem a sua confiança, este trabalho não teria sido possível.

Agradeço também aos professores Norton Trevisan Roman e Thiago Alexandre Salgueiro Pardo, cujas sugestões, liderança e iniciativa foram fundamentais e indispensáveis ao longo da realização deste trabalho.

Agradeço à minha família, cujo apoio e confiança inquestionáveis foram, são e serão os pilares que me sustentam e me dão forças para seguir em frente diariamente.

Agradeço aos meus amigos, companheiros nos melhores e piores momentos, sempre dispostos, pacientes e bondosos. Em especial, agradeço ao Gabriel Kovacs Mellado, amigo de longa data e gentil até o fim, que permitiu que muitos dos aspectos computacionais deste trabalho fossem realizados. Espero que tenha encontrado a sua paz.

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.

Resumo

A anotação semântica de *corpora* desempenha papel essencial no desenvolvimento de ferramentas de Processamento de Línguas Naturais (PLN). No caso dos *corpora* de "conteúdo gerado por usuário" (CGU) compostos por *tweets*, cuja linguagem é predominantemente informal, as anotações semânticas pautam-se basicamente em aspectos lexicais, como entidade nomeada, emoção ou polaridade. Neste trabalho, investigou-se a representação semântica "sentencial" de *tweets* do mercado financeiro em português por meio do modelo *Abstract Meaning Representation* (AMR), sob a hipótese de que informações sintáticas podem auxiliar tal representação. Para tanto, anotou-se o *corpus* DANTEStocks, primeiro *tweebank* com anotação gramatical segundo o modelo *Universal Dependencies* (UD). Mais precisamente, anotou-se uma parcela de 1.128 dos 4.048 *tweets* (isto é, 30% do total) do *corpus*. A anotação AMR seguiu uma metodologia híbrida, com uma etapa manual para definição de diretrizes e modelos de referência, e outra semiautomática, que consistiu na revisão manual de grafos gerados por um *Large Language Model*. Tal anotação permitiu confirmar a hipótese de que as dependências sintáticas UD auxiliam na construção dos grafos AMR, sobretudo na identificação de subgrafos e relações entre conceitos. Outras contribuições são (i) desenvolvimento e validação de diretrizes de anotação AMR para fenômenos linguísticos específicos do português, CGU e domínio financeiro, (ii) proposição de rótulos para 3 tipos de entidades do domínio financeiro (URL, *ticker* e usuário) e (iii) proposição de um *frameset* para o repositório Verbo-Brasil referente ao verbo "repicar" do mercado financeiro. A principal dificuldade neste trabalho foi a interpretação dos *tweets* para a construção dos grafos AMR devido à sua fragmentação, truncamento, dependência contextual e vocabulário especializado, exigindo auxílio constante de *experts* e recursos externos sobre o domínio financeiro. Mesmo com base em um conjunot de validação pequeno, a medida-F de 89% para a concordância interanotador indica que a parcela do DANTEStocks com anotação AMR é um recurso confiável para subsidiar as primeiras pesquisas sobre *parsing* AMR para *tweets* no PLN e para a anotação do restante do *corpus*.

Palavras-chave: PLN. Conteúdo Gerado por Usuário. Semântica. Tweets. Corpus.

Abstract

Semantic annotation of corpora plays a crucial role in the development of Natural Language Processing (NLP) tools. In the case of usergenerated content (UGC) corpora composed of tweetswhose language is predominantly informalsemantic annotations typically focus on lexical aspects such as named entities, emotion, or polarity. In this study, we investigated the sentential semantic representation of Portugueselanguage financialmarket tweets via the Abstract Meaning Representation (AMR) framework, under the hypothesis that syntactic information can aid such representations. To this end, we annotated the DANTEStocks corpusthe first tweebank with grammatical annotation according to the Universal Dependencies (UD) formalismcovering a subset of 1,128 of the 4,048 tweets (i.e., 30% of the total). The AMR annotation followed a hybrid methodology, comprising a manual phase to establish guidelines and reference models, and a semiautomatic phase in which manually curated corrections were applied to graphs produced by a large language model. This process confirmed our hypothesis that UD syntactic dependencies facilitate the construction of AMR graphs, particularly in identifying subgraphs and relations among concepts. Additional contributions include: (i) the development and validation of AMR annotation guidelines for phenomena specific to Portuguese, UGC, and the financial domain; (ii) the proposal of labels for three types of financialdomain entities (URLs, tickers, and users); and (iii) the introduction of a frameset for the VerboBrasil repository covering the financial verb repicar. The primary challenge encountered was interpreting tweets for AMR graph construction, given their fragmentation, truncation, contextual dependence, and specialized vocabularynecessitating continuous expert consultation and external financial-domain resources. Nevertheless, even with a relatively small validation set, an inter-annotator F-score of 89% indicates that the AMR-annotated portion of DANTEStocks is a reliable resource for pioneering research on AMR parsing of tweets in NLP and for guiding the annotation of the remaining corpus.

Keywords: Natural Language Processing; Abstract Meaning Representation; User-Generated Content; Tweets; Financial Market; Annotated Corpus.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Sumário

1 Introdução

1.1 Contexto e Justificativa

No Processamento de Língua Natural (PLN), *corpus* é um conjunto de textos computacionalmente processável, tendo sido coletado com um propósito e produzido naturalmente (Sinclair, 2005). O PLN adotou os *corpora* como fonte de conhecimento e passou a utilizar métodos computacionais para aprender tarefas envolvendo a língua a partir deles. Esse aprendizado só foi possível porque se desenvolveu o que hoje se conhece por anotação de *corpus*. Tecnicamente, anotar significa delimitar um segmento do texto e atribuir-lhe uma etiqueta previamente definida, explicitando uma análise humana sobre esses segmentos (Freitas, 2022; Duran, Pardo, 2024).

Tais segmentos podem ser palavras, sintagmas, orações, sentenças ou parágrafos e as etiquetas a eles associadas podem indicar informações linguísticas de diferentes tipos ou níveis, como o morfossintático (p.ex.: *Part-of-Speech* (PoS) ou classes de palavras), sintático (p.ex.: relações sintáticas sintagmáticas ou de dependências), semântico (como estrutura argumental ou papéis semânticos, entidades nomeadas, polaridade e emoção), pragmático (p.ex.: atos de fala) e discursivo (como as relações discursivas), etc. (Jurafsky; Martin, 2025) As anotações podem ser *standoff* ou *in-line* (ou *embedded*). Se *standoff*, as informações linguísticas ficam armazenadas em um arquivo separado, porém conectadas ao texto original (ou a outras anotações) por meio de índices. Já nas *in-line*, as informações são acrescidas diretamente ao documento-base. (Jurafsky; Martin, 2025)

Como ressaltam Duran e Pardo (2024), é importante para o PLN que os métodos de Aprendizado de Máquina (AM) aplicados aos *corpora* sejam capazes de capturar a lógica expressa na anotação, gerando algoritmos que reproduzam automaticamente a ano-

tação humana. Tal anotação humana, aliás, pode ser feita de forma totalmente manual ou por máquinas com posterior revisão manual, isto é, semiautomaticamente. Em ambos os processos, o resultado são os chamados *corpora* padrão ouro (em inglês, *gold standard*).

A tarefa de anotação de *corpora* e os métodos de AM evoluíram bastante nas últimas décadas (Devlin et al., 2019). O aumento da capacidade de processamento dos computadores possibilitou, inclusive, o desenvolvimento de técnicas de AM sem o uso de *corpora* anotados, chegando-se aos grandes modelos de linguagem (*Large Language Models* LLM) gerativos da atualidade, como os da família GPT.

Embora as arquiteturas neurais, como as baseadas em *transformers* (como é o caso da família GPT), tenham atingido ou mesmo superado os resultados do estado-da-arte do processamento automático das línguas naturais, o conhecimento linguístico (em especial, o semântico, que é o foco deste trabalho) utilizado nessas arquiteturas está implícito em vetores densos (as chamadas *word embeddings*), o que torna as tarefas de análise e interpretação dos resultados mais difíceis (Senel et al., 2018).

Em outras palavras, isso quer dizer que, mesmo na era dos LLMs, em que os sistemas de ponta ou do estado-da-arte em PLN são baseados em redes neurais de caixa preta (do inglês, "*black-box neural networks*"), acredita-se que é importante continuar a desenvolver recursos linguísticos que possam ser usados para construir sistemas interpretáveis e controláveis para contextos em que a transparência é fundamental.

Assim, a anotação de *corpus* é bastante relevante, sobretudo para certas tarefas de PLN em que a explicitação do conhecimento pode ser necessária, como extração de informação, ferramentas de apoio à leitura e escrita e simplificação textual, etc. Aliás, outra utilidade dos *corpora* anotados é exatamente servir de referência para avaliar os resultados dos LLMs.

A crescente popularidade nas últimas décadas do conteúdo gerado pelos usuários (CGU) da *web* transformou principalmente as mídias sociais em fontes importantes de dados para muitas áreas. E isso pode ser comprovado pelo fato de que empresas, governos e comunidades cada vez mais precisam de dados em tempo real advindos dessas mídias dinâmicas e de larga escala (Derczynski; Bontcheva; Roberts, 2016). O termo CGU, aliás, designa todo conteúdo, seja no formato de imagem, vídeo, áudio e texto, que foi postado em plataformas online, como redes sociais, páginas de *review*, *blogs* e *wikis* (Krumm; Davies; Narayanaswami, 2008).

Diante da referida relevância, o PLN tem desenvolvido cada vez mais aplicações voltadas para CGU em formato texto, uma vez que a linguagem dos diferentes gêneros de CGU é repleta de fenômenos distintos da linguagem padrão, que até então era o foco das pesquisas nessa área. Entre essas aplicações, estão, por exemplo, a mineração de opinião e a análise de sentimento (Sanguinetti et al., 2023).

Para tanto, *corpora* anotados de CGU também têm sido construídos. Em um estudo recente, Sanguinetti et al. (2023) mostram que, desde 2011 até 2019, mais de 30 *corpora* anotados de CGU foram construídos para diversas línguas europeias, para o Inglês americano, Árabe, Chinês, Híndi, etc. A grande maioria desses recursos são parcial ou inteiramente compostos por conteúdo compilado da plataforma *Twitter*[1], que oferece um serviço que pode ser concebido como uma mistura de *microblog* e rede social (Freitas; Barth, 2015). O interesse pelos *tweets* tem como motivações principais o alcance das opiniões neles veiculadas em vários segmentos da sociedade e o fato de que apresentam uma série de características linguísticas que são comuns a outros gêneros CGU.

Enquanto gênero, o *tweet* 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, o *tweet* constitui um subgênero de CGU caracterizado pela informalidade e brevidade (Freitas; Barth, 2015). Assim, sua linguagem é repleta de fenômenos que dificultam qualquer tarefa de anotação, como truncamento, fragmentação, erros de digitação, abreviações informais, etc.

No que tange às anotações padrão-ouro, os *corpora* de *tweets* comumente possuem anotação sintática (Sanguinetti et al., 2023) , sendo chamados, por isso, de *tweebanks*. Tal denominação se deve ao fato de esse tipo de recurso linguístico ser um banco de *tweets* em que cada um deles está associado a uma representação formal de sua sintaxe em formato de árvore. No caso, o modelo gramatical *Universal Dependencies* (UD) (Nivre; Marneffe; Ginter; Goldberg et al., 2016; Nivre; Marneffe; Ginter; Haji et al., 2020) é um dos mais aplicados atualmente à anotação de *corpus*, incluindo dos *tweebanks*.

Para além da sintaxe, alguns *tweebanks* também possuem anotação de informação semântica, a qual diz respeito basicamente à explicitação das entidades nomeadas, emoções e polaridades, permitindo o desenvolvimento de métodos de análise semântica

[1]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

de *tweets* pautadas nas tarefas de Reconhecimento de Entidades Nomeadas (REN) (do inglês, *Named Entity Recognition*) (Derczynski; Bontcheva; Roberts, 2016), classificação de emoções (Cortis et al., 2017) e identificação de polaridade (Gomez et al., 2016).

Para as pesquisas sobre o processamento de *tweets* em português brasileiro, a comunidade do PLN conta com o DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024a), que possui aproximadamente 4 mil *tweets* distintos sobre o mercado financeiro, especificamente sobre as ações do Ibovespa. O fato de esse *corpus* ser do domínio do mercado financeiro tem como motivação a comprovada correlação entre o conteúdo do *Twitter* e a movimentação das ações nas bolsas, a qual tem fundamentado o desenvolvimento de formas mais eficazes de predição dos movimentos do mercado a partir dos *tweets* (Bollen; Mao; Zeng, 2011; Carosia; Coelho; Silva, 2019; Dhabe et al., 2023).

Esse *corpus* possui atualmente anotação gramatical (morfossintática e sintática) segundo o modelo UD, de emoções segundo os eixos de oposição (*joy-sadness*, *anger-fear*, *trust-disgust* e *surprise-anticipation*) de Plutchik e Kellerman (1986) (F. J. V. Silva; Carvalho, 2020) e de entidades nomeadas segundo a taxonomia do HAREM (Mota; Santos, 2008) como descrito em (Zerbinati; Roman; Di-Felippo, 2024).

Diante do exposto, pode-se dizer que as anotações semânticas que mais comumente estão presentes nos *tweebanks* se referem ao nível do léxico, adicionando informação ao segmento "palavra", como é o caso das entidades nomeadas. Mesmo as anotações que atribuem uma etiqueta ao *tweet* inteiro, tomado como a unidade básica de análise, pautam-se na ocorrência de certas pistas léxicas, como é o caso da anotação de emoção e polaridade.

O PLN, no entanto, dispõe de vários formalismos ou modelos semânticos sentenciais para explicitar o significado de um enunciado tomado como unidade básica de análise. Entre esses formalismos, citam-se, por exemplo, a tradicional *Lógica de Primeira Ordem* (LPO) (Pereira; Shieber, 2002; Jurafsky; Martin, 2025) e, mais recentemente, os modelos *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013), *Uniform Meaning Representation* (UMR) (Van Gysel et al., 2021) e *Lexical Meaning Representation* (LMR) (Baptista; Reis et al., 2024).

Entre eles, a AMR é um dos mais difundidos. Em linhas gerais, o modelo foi proposto com o objetivo de capturar o significado de um enunciado (no caso, sentenças), abstraindo elementos da estrutura sintática, como informação morfossintática e ordem

das palavras (Banarescu et al., 2013). Esse formalismo descarta, por exemplo, os artigos, que segundo os autores, pouco contribuem para a representação do significado essencial da sentença. Além disso, ele foca na estrutura predicado-argumento.

Com base nessas características, esse modelo ganhou muita atenção do PLN, principalmente pela "simplificação" da representação semântica. Assim, muitos *corpora* de linguagem formal ou padrão já foram anotados para diversas línguas, incluindo o português (Anchiêta, R.; Pardo, T., 2018; Inácio et al., 2023; Seno et al., 2022), os quais possibilitaram o desenvolvimento de analisadores semânticos automáticos (Anchiêta, R.; Pardo, T., 2018, 2022; Seno et al., 2022). Essas ferramentas são fundamentais para que as aplicações de PLN possam "compreender" o conteúdo dos enunciados.

Tendo em vista que o formalismo AMR ainda não foi aplicado na anotação de CGU, especificamente de *corpora* de *tweets*, os quais, por causa das características lingüísticas, impõem desafios para qualquer tipo de anotação, traçaram-se os objetivos e hipóteses descritos na próxima seção.

1.2 Objetivos e Hipóteses

Este trabalho teve como objetivo geral explorar a pertinência do modelo AMR para a representação semântica do *tweet* enquanto unidade básica de análise. Especificamente, objetivou-se anotar o DANTEStocks segundo esse formalismo sob a hipótese de que essa tarefa pode se beneficiar da anotação sintática dos *tweets*. Em outras palavras, a hipótese é a de que a anotação semântica-AMR dos *tweets* pode ser auxiliada por certas informações sintáticas codificadas na anotação-UD dos *tweets*.

Dessa forma, este trabalho contribui para a caracterização semântica dos *tweets* do domínio do mercado financeiro e para o enriquecimento do DANTEStocks com a inclusão de uma camada de anotação semântica. Com isso, o *corpus*, enquanto recurso linguístico, poderá fomentar o desenvolvimento de analisadores semânticos específicos para CGU. O DANTEStocks enriquecido com anotação segundo AMR é o primeiro *corpus* de *tweets* em língua portuguesa com esse nível de informação explicitada.

1.3 Metodologia

Com os objetivos em vista, este trabalho foi equacionado em 5 etapas, a saber:

- **Revisão da literatura:** estudo das referências centrais e demais pertinentes, publicadas no decorrer da pesquisa, sobre (i) o gênero *tweet* e suas características linguísticas, (ii) o modelo de representação semântica AMR e demais da literatura como LPO, UMR e LMR, (iii) metodologias de anotação de *corpus* e (iv) os recursos, ferramentas e analisadores semânticos para o português com base na AMR.

- **Estudo e descrição do modelo AMR**: investigação detalhada dos pressupostos de representação e as diretrizes/recursos de anotação do modelo AMR, com ênfase nas diretrizes já propostas para o português padrão, formal e normativo.

- **Estudo e descrição do *corpus* DANTEStocks**: investigação do *corpus* selecionado para esta pesquisa, o DANTEStocks, no que concerne às características estruturais e lexicais dos *tweets* com vistas à anotação semântica e as anotações prévias do recurso, com especial atenção à anotação sintática-UD.

- **Anotação semântica dos *tweets***: efetiva anotação do *corpus* com base no formalismo AMR. Essa tarefa foi realizada tanto de forma manual, auxiliada por um editor de anotação, quanto semi-automática (isto é, anotação automática com posterior revisão manual). Além disso, essa etapa também englobou a elaboração de um manual com as diretrizes para a representação semântica do DANTEStocks.

- **Descrição semântica do corpus** nesta etapa, realizou-se a análise da anotação realizada na etapa anterior com o objetivo de levantar as estruturas semânticas mais frequentes, problemas residuais de representação/anotação, correlatos entre a anotação sintática e a anotação semântica, entre outros.

2 Revisão de literatura

2.1 Conceitos fundamentais

Como mencionado, a revisão aqui apresentada engloba os seguintes tópicos fundamentais para a realização deste trabalho: (i) o gênero *tweet* e as suas características linguísticas, (ii) os modelos de representação semântica, como a LPO e os mais recentes AMR, UMR e LMR, (iii) metodologias de anotação de *corpus* e (iv) os recursos, ferramentas e analisadores semânticos com base na AMR, especialmente para o português.

2.1.1 O *Twitter* e o seu gênero textual

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

Em um estudo recente, (McGregor; Molyneux, 2020) observaram que jornalistas que utilizam o *Twitter* como fonte de informação consideram o conteúdo dessa mídia social tão importante quanto as manchetes de fontes oficiais, como a *Associated Press*. Diante disso, vê-se o quanto o *Twitter* se tornou especialmente central e influente ao se tratar da circulação de notícias na era digital. Por conseguinte, o conteúdo do *Twitter* tem sido alvo de aplicações desenvolvidas no âmbito do PLN que visam, por exemplo, (i) medir o interesse por certos tópicos ou assuntos, (ii) detectar eventos imprevisíveis em tempo real, (iii) predizer o comportamento dos ativos (ações) nas bolsas, etc.

Sobre esse último interesse, destaca-se que o trabalho de Bollen, Mao e Zeng

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

(2011) 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 base nessa constatação, inúmeros trabalhos têm sido desenvolvidos com o fim comum 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).

Do ponto de vista linguístico, o conteúdo veiculado no *Twitter* é um dos gêneros textuais cobertos pelo termo/sigla CGU. O termo CGU, na verdade, diz respeito a todo tipo de conteúdo na forma de imagem, vídeo, áudio ou texto, postado por usuários da *web* em redes sociais, fóruns de discussão, *wikis*, etc. (Krumm; Davies; Narayanaswami, 2008) Em geral, todo CGU é marcado pela acessibilidade e natureza colaborativa, contrapondo-se ao conteúdo produzido por meios de comunicação tradicionais.

Sanguinetti et al. (2023) dizem que o termo CGU recobre um contínuo de subgêneros textuais, os quais podem variar em função de alguns fatores, como (i) convenções e limitações do meio ou plataforma (isto é, *blog*, fórum, chat, etc.), (ii) grau de canonicidade em relação à língua padrão e (iii) mecanismos linguísticos.

Assim, o *tweet* está inserido nesse contínuo, podendo ser concebido como um subgênero textual que é uma mistura de outros, como notícia, propaganda e bilhete, os quais foram modificados para atender às necessidades comunicativas da plataforma. Dessa forma, o *tweet* constitui um subgênero caracterizado pela informalidade e brevidade, promovendo, assim, uma comunicação concisa e direta. Tal brevidade, aliás, advém da limitação de caracteres imposta pela plataforma, a qual, atualmente, é de 280 caracteres. Essa brevidade influencia a linguagem do gênero em questão, pois estruturalmente os *tweets* podem apresentar (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.

Sobre os aspectos lexicais, os *tweets* incluem fenômenos dependentes da plataforma (p.ex.: *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., 2023).

2.1.2 Os *tweebanks* e suas anotações

Como mencionado, a popularidade das mídias sociais motivou o desenvolvimento de várias aplicações para o processamento do conteúdo veiculado por elas, em especial pelo *Tweet*, e, com isso, a necessária construção de *tweebanks*. De 2011 até 2019, Sanguinetti et al. (2023) identificaram mais de 30 *corpora* anotados (os *tweebanks*) de CGU construídos para diversas línguas europeias, para o inglês americano, árabe, chinês, híndi, etc. Como se pode ver na Figura 2.1, a maioria dos recursos, especificamente 16 dos 30, são parcial ou inteiramente compostos por conteúdo extraído do *Twitter*.

Além dos já mencionados fatores que justificam essa proeminência, isto é, o alcance das opiniões veiculadas na plataforma em diferentes segmentos da sociedade e a representatividade dos fenômenos linguísticos típicos dos *tweets* (comuns a outros tipos de CGU), o fato de que apresentam, outras possíveis razões foram a então facilidade de obtenção dos dados por meio da API, bem como a política adotada pela plataforma em relação ao uso de dados para fins acadêmicos e não comerciais *Twitter*[2].

Com base na Figura 2.1, vê-se que apenas quatro recursos incluem dados de mídias sociais além do *Twitter*, especificamente *Facebook* (FSMB, Taiga), *Reddit* (GUM), *Sina Weibo* (CWT), *Instagram*, *YouTube* e *VK* (Taiga). A maioria dos recursos restantes compreende textos de fóruns de discussão de vários tipos. E apenas três *treebanks* englobam textos de diferentes subdomínios, ou seja, fóruns jornalísticos (NBZ), *blogs*, *reviews*, *e-mails*, grupos de notícias e perguntas e respostas (EWT), *Wikinews*, *Wikivoyage*, *wikiHow*, biografias da *Wikipedia*, entrevistas, escrita acadêmica, ficção sob licença *Creative Commons* ("*Creative Commons fiction*") (GUM). Dois recursos são compostos por dados genéricos rastreados automaticamente da *web* (EtWT, TDT).

A anotação sintática da maioria dos *tweebanks* segue o modelo gramatical *Universal Dependencies* (UD) (Nivre; Marneffe; Ginter; Goldberg et al., 2016; Nivre; Marneffe; Ginter; Haji et al., 2020), o que é indicado na quinta coluna ("UD-based) da Figura 2.1 com o índice "Yes". Alguns *corpora* em inglês também possuem anotação de informação semântica, sobretudo de entidades nomeadas, como *Broad Twitter Corpus* (Derczynski; Bontcheva; Roberts, 2016), *Temporal Twitter Corpus* (TTC) (Rijhwani; Preotiuc-Pietro, 2020) e *TweetNER7* (Ushio et al., 2022). Outros possuem anotação de

[2]Essa API, no entanto, tem passado por várias atualizações e mudanças e, com isso, o acesso gratuito é atualmente limitado.

emoções (Cortis et al., 2017) e polaridades (Gomez et al., 2016). O foco do levantamento de Sanguinetti et al. (2023) foi a anotação sintática e, por isso os *corpora* com informação semântica não estão na Figura 2.1.

Figura 2.1: Lista de *treebanks* contendo CGU.

Name	References	Source	Language	UD-based
ATDT	Albogamy and Ramsay (2017)	Twitter	AR	Yes
Hi-En-CS	Bhat et al. (2018)	Twitter	HI/EN	Yes
TwitterAAE (TAAE)	Blodgett et al. (2018)	Twitter	AAE, MAE	Yes
TWITTIRÒ-UD (TWRO)	Cignarella et al. (2019)	Twitter	IT	Yes
DWT	Daiber and Van Der Goot (2016)	Twitter	EN	No*
W2.0	Foster et al. (2011)	Twitter, sort fora	EN	No‡
Foreebank (Frb)	Kaljahi et al. (2015)	Technical fora	EN, FR	No‡
Tweebank (Twb)	Kong et al. (2014)	Twitter	EN	No*
Tweebank2 (Twb2)	Liu et al. (2018)	Twitter	EN	Yes
TDT	Luotolahti et al. (2015)	Various	FI	Yes
xUGC	Martínez Alonso et al. (2016)	Various	FR	Yes
Estonian Web Treebank (EtWT)	Martínez Alonso et al. (2016)	Various	ET	Yes
ITU	Pamay et al. (2015)	n.a.	TR	No*
WDC	Read et al. (2012b)	Various	EN	No‡
tweeDe	Rehbein et al. (2019)	Twitter	DE	Yes
PoSTWITA-UD (Pst)	Sanguinetti et al. (2018)	Twitter	IT	Yes
FSMB	Seddah et al. (2012)	Twitter, Facebook, discussions fora	FR	No‡
Narabizi (NBZ)	Seddah et al. (2020)	Newspaper fora	DZ/FR	Yes
EWT	Silveira et al. (2014)	Various	EN	Yes
LAS-DisFo (LDF)	Taulé et al. (2015)	Discussion fora	ES	No‡
MoNoise (MNo)	Van Der Goot and van Noord (2018)	Twitter	EN	Yes
STB	Wang et al. (2017)	Discussion fora	SgE	Yes
CWT	Wang et al. (2014)	Twitter, Sina Weibo	ZH	No*
GUM	Zeldes (2017)	Various	EN	Yes
HSE	n.a.	Various	BE	Yes
OOD	n.a.	Various	FI	Yes
TwittIrish (TwIr)	n.a. (Publication forthcoming)	Twitter	GA	Yes
Cadhan (Cdh)	n.a.	Various	GV	Yes
Taiga	n.a.	Various	RU	Yes
IU	n.a.	Various	UK	Yes

Fonte: Sanguinetti et al. (2023).

Para o português, tem-se alguns *corpora* de CGU: (i) *corpus* de 76.358 *tweets* que mencionam os candidatos à eleição presidencial de 2010 (Silva, I. S. et al., 2011), (ii) Corpus 7x1, que engloba 2.728 comentários postados no *Twitter* durante as semifinais da Copa do Mundo de 2014 (Moraes; Manssour; Silveira, 2015), (iii) *corpus* que contém 554.623 *tweets* com ocorrência de *emojis* positivos e 425.444 de *emojis* negativos (Junior et al., 2017), (iv) TweetSentBR, que engloba 15.000 *tweets* do domínio "show de TV" (Brum; Nunes, 2018), (v) *corpus* 4P (Silva; Pardo, 2019), que possui comentários sobre produtos eletrônicos e resumos comparativos entre eles, e (vi) o *corpus* de 4.517 *tweets* do domínio "bolsa de valores" (F. J. V. Silva; Carvalho, 2020). Alguns desses possuem anotação de emoção e outros de polaridade. O 4P é o único com anotação de aspectos.

O *corpus* de F. J. V. Silva e Carvalho (2020), em particular, deu origem ao

DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024a), que possui aproximadamente 4 mil *tweets* distintos sobre as ações do Ibovespa anotados com emoções, entidades nomeadas e dependências sintáticas segundo o modelo UD. Mais informações sobre esse *corpus* e suas anotações e o modelo UD serão fornecidas no capítulo 3.

2.1.3 Metodologias de anotação de *corpus*

Independentemente do tipo de informação linguística que se quer explicitar, a execução da anotação de *corpus* pode ser feita de três maneiras distintas, que variam em função do volume e complexidade do trabalho humano envolvido. Tais maneiras são: manual, semiautomática ou totalmente automática.

Se manual, os dados são anotados exclusivamente por humanos. Esse método é comumente aplicado no desenvolvimento de *esquemas de anotação*, o que envolve definir etiquetas e diretrizes de anotação, ou quando não há anotação automática confiável. Embora envolva um trabalho moroso, o método manual é, como aponta Stefanowitsch (2020), a única possibilidade a depender do fenômeno linguístico de interesse, uma vez que a automática ou não é possível ou resulta em uma qualidade tão baixa que simplesmente torna a revisão manual posterior inviável. Para ilustrar, cita-se o caso das metáforas, que são quase impossíveis de serem identificadas automaticamente, pois têm poucas ou nenhuma propriedade que as distingue sistematicamente da linguagem literal. Esse tipo de anotação, embora demorada, pode levantar questões que talvez não fossem identificadas em uma etapa de revisão, uma vez que os anotadores tendem a acatar a análise fornecida pela máquina (Freitas, 2022).

A anotação semiautomática é a mais frequente, sobretudo quando o objetivo é a construção de um recurso "padrão-ouro". Ela se caracteriza por englobar uma primeira etapa de anotação automática seguida por uma de revisão humana. Em outras palavras, esse tipo de anotação é correção manual da saída produzida por uma ferramenta computacional. Nesse caso, o tempo gasto no referido processo é consideravelmente menor, mas ainda assim se trata de uma tarefa custosa, demandando tempo e esforço por parte da equipe responsável pela execução da tarefa.

No contexto da anotação semiautomática, uma opção atualmente é realizar o primeiro passo automático por meio de um LLM, como o GPT-4o, disponível gratuitamente na *web*, explorando técnicas de *Engenharia de Prompt* (Liu et al., 2023). Em linhas gerais,

esse termo indica o processo de criar e ajustar *prompts* (comandos ou instruções) fornecidos a um LLM para obter respostas desejadas ou específicas. Isso envolve a elaboração de perguntas ou instruções de maneira precisa e clara para orientar o modelo a gerar respostas úteis e relevantes, que, no caso, seriam a anotação de um *corpus*.

As técnicas de *Engenharia de Prompt* são várias, como (i) *zero-shot*, (ii) *few-shot prompting*, (iii) (*few-shot*) *chain-of-thought prompting* e (iv) *zero-shot chain-of-thought prompting*, (v) *least-to-most prompting* e outras (Brown et al., 2020; Liu et al., 2023). A primeira, por exemplo, refere-se à capacidade de um LLM de entender e executar uma tarefa sem ter recebido exemplos específicos dessa tarefa anteriormente. A técnica *few-shot prompting*, por outro lado, envolve fornecer ao LLM um pequeno número de exemplos para ajudá-lo a entender o contexto e realizar uma tarefa específica. Já *chain-of-thought prompting* consiste em demonstrar ao modelo, no próprio *prompt*, a forma exata de raciocinar, dando uma resposta completa como exemplo, incluindo o passo a passo para chegar a ela. A *zero-shot chain-of-thought prompting*, por sua vez, considera que os LLMs não precisam necessariamente de toda essa explicação e exemplos (f*ew-shot*) para dar uma resposta correta, mas apenas da frase *"Let's think step by step."* (em português, Vamos pensar passo a passo.) no final do *prompt*.

Por fim, a anotação também pode ser totalmente automática. No entanto, como salienta Freitas (2022), a avaliação da qualidade desse tipo de anotação é feita pela comparação com o desempenho humano, evidenciando a necessidade de anotação humana.

Independentemente do esquema e da execução, vale ressaltar que não existe anotação neutra ou ateórica, sobretudo porque (i) as categorias e os esquemas de anotação são baseados em teorias linguísticas e escolhas metodológicas, sendo que diferentes teorias podem levar a diferentes interpretações e representações dos dados linguísticos e (ii) os anotadores humanos, mesmo com diretrizes claras, interpretam o texto com base em seu entendimento e conhecimento, o que pode introduzir viés e variação de interpretação entre eles. Ademais, vale lembrar que a anotação pode ser avaliada de duas maneiras (Artstein, 2017): (i) medida de concordância entre anotadores (em inglês, *inter-annotator agreeement*) humanos diferentes, ou (ii) comparação a um gabarito produzido por alguém. A primeira costuma ser a mais difundida, sendo aplicada com o objetivo de medir a consistência e não a corretude de uma anotação.

2.1.4 Visão geral sobre formalismos semânticos

Os esforços que envolvem a compreensão da estrutura semântica das línguas naturais, embora desafiadores, não são novos no campo dos estudos da linguagem, existindo, desde a Antiguidade, propostas de teorias e modelos representacionais do significado das línguas naturais. Entre os séculos 7 e 4 A.C., por exemplo, o gramático indiano Panini escreveu o famoso tratado gramatical do Sânscrito (Penn; Kiparsky, 2012), no qual descreve os *kãrakas*, isto é, relações semânticas entre predicados verbais e nominais e seus argumentos, como agente, paciente, instrumento (Jurafsky; Martin, 2025).

Mais recentemente na história, formulações teóricas com concepções similares foram desenvolvidas, como a *Gramática de Casos* de Fillmore (1968). O autor define que a ação ou o estado denotado pelo verbo seja representado por relações estabelecidas entre o verbo e seus argumentos. Fillmore propõe 9 relações, também chamadas "papéis semânticos", a saber: *agente, experienciador, instrumental, objeto, fonte, objetivo, localização, tempo* e *caminho*. Na sentença "João abriu a porta com uma chave", por exemplo, "João" exerce o papel de agente, "porta" exerce o papel de objetivo e "chave" o de instrumento.

O trabalho de Fillmore não apenas lançou luz sobre as estruturas semânticas subjacentes às línguas naturais, mas também embasou outros modelos de representação no âmbito do PLN (Specia; Rino, 2002). O projeto *PropBank* ("Banco de Proposições"[3]) (Palmer; Gildea; Kingsbury, 2005), por exemplo, utiliza o conceito de papéis semânticos de Fillmore ao adicionar uma camada de informação predicado-argumento à anotação sintática (sintagmática) do *subcorpus* financeiro do *Penn Treebank* (Marcus, 1993; Taylor; Marcus; Santorini, 2003).

O modelo de anotação de papéis semânticos proposto no *PropBank* foi amplamente difundido para a anotação de *corpora* em linguagem formal ou padrão (como a jornalística). O *PropBank* se destacou no PLN porque (i) os papéis semânticos relacionam os predicadores e seus argumentos à anotação sintática dos enunciados, pois, teoricamente, eles estão na interface sintaxe-semântica e (ii) a diversidade de rótulos é contornada pelo uso de argumentos numerados (ArgNs) (Arg0 a Arg5) e modificadores (ArgMs), que buscam facilitar a generalização para o AM.

Dada a complexidade da representação semântica, muitos têm focado no desen-

[3]Em linhas geral, proposição é uma estrutura (básica) que descreve as relações semânticas entre predicador e argumento(s), desconsiderando tempo, modo, aspecto, negação ou modificadores modais.

volvimento de modelos que buscam capturar o significado de um enunciado. Além dos já citados LPO, AMR, UMR e LMR, têm-se outros como as *Redes Semânticas* (Lehmann, 1992; Brachman, 2004), a *Universal Networking Language* (UNL) (Uchida; Zhu; Della Senta, 2005), etc. Neste trabalho, apresentam-se os fundamentos da LPO, um dos modelos mais clássicos do PLN, e sobretudo da AMR, que é o modelo recente mais difundido na área. Além desses, descrevem-se a UMR e a LMR, que são posteriores à AMR.

2.1.4.1 Lógica de Primeira Ordem

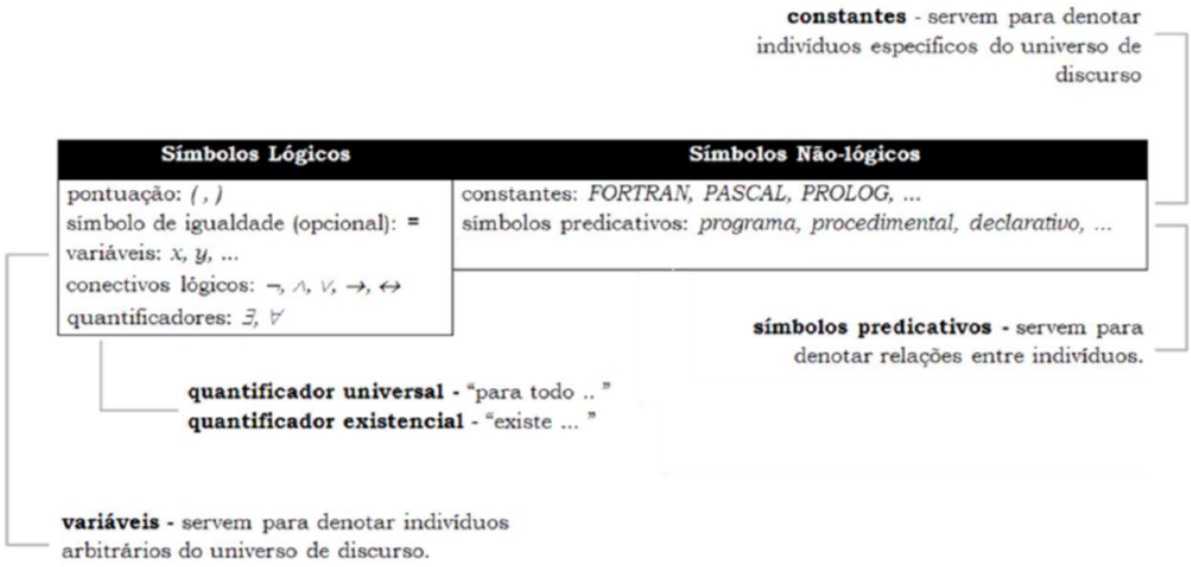
A *Lógica de Primeira Ordem* (Pereira; Shieber, 2002; Jurafsky; Martin, 2025) (LPO) é uma linguagem de representação do significado flexível, bem compreendida e tratável computacionalmente. Ela é uma extensão da lógica proposicional que permite o uso de quantificadores e predicados para expressar de forma mais rica e detalhada as relações entre objetos dentro de um domínio.

Basicamente, ela consiste de uma série de **símbolos lógicos** (variáveis, conectivos lógicos, quantificadores, pontuação e símbolo de igualdade) e **símbolos não-lógicos** (constantes, símbolos predicativos e símbolos funcionais), como é ilustrado na Figura 2.2.

Entre os símbolos lógicos, as **variáveis**, representadas como letras minúsculas únicas, denotam indivíduos arbitrários do universo de discurso. Em outras palavras, elas assumem valores dentro de um domínio e são utilizadas para expressar generalizações ou relações que envolvem múltiplos elementos. As variáveis podem ser usadas de duas maneiras: (i) referência a objetos anônimos específicos e (ii) referência genérica a todos os objetos em um conjunto. Esses dois usos são representados por meio dos operadores conhecidos como **quantificadores**.

Os dois operadores básicos da LPO são o **quantificador existencial** ($\exists$) (existe) e o **quantificador universal** ($\forall$) (para todo). O existencial indica que existe pelo menos um elemento no universo para o qual a proposição é verdadeira. Uma variável quantificada existencialmente é empregada, por exemplo, para representar um sintagma nominal indefinido, como em "*Uma lanchonete* que serve pão de queijo na UFSCar" (1). Informalmente, o quantificador existencial indica que, para (1) ser verdadeira, deve haver pelo menos um objeto tal que, ao substituir a variável x, resulte em uma sentença verdadeira. Assim, por exemplo, dado que PQ é uma lanchonete na UFSCar, substituir x por PQ resulta na fórmula lógica (2). Com base na semântica do operador $\wedge$, a sentença (2)

Figura 2.2: Construtos da LPO.



Fonte: https://acesse.dev/3Yf6T.

será verdadeira se todas as suas três fórmulas atômicas constitutivas forem verdadeiras. Estas, por sua vez, serão verdadeiras se estiverem presentes na base de conhecimento ou se puderem ser inferidas a partir de outros fatos da mesma base.

(1) $\exists x Lanchonete(x) \wedge Serve(x, PãodeQueijo) \wedge LocationOf(UFSCar)$

(2) $Lanchonete(PQ) \wedge Serve(PQ,\ PãodeQueijo) \wedge LocationOf(UFSCar)$

O quantificador universal $\forall$ afirma que, para a fórmula lógica (3) ser verdadeira ("Todas as lanchonetes servem pão de queijo"), a substituição da variável universalmente quantificada por qualquer objeto do universo que seja "lanchonete" deve gerar uma proposição verdadeira.

(3) $\forall x(Lanchonete(x) \Rightarrow Serve(x, PãodeQueijo))$

Entre os símbolos não-lógicos, as **constantes** referem-se a objetos específicos do universo e são convencionalmente representadas como letras maiúsculas. Os **predicados** representam propriedades ou relações entre os objetos, retornando, do processo de inferência, sempre um valor de verdade (verdadeiro ou falso). Em (1), por exemplo, *Serve*, que é um predicado de dois argumentos, estabele relação entre dois objetivos, isto é, *PãodeQueijo* e *UFSCar*, representados como constantes. Já na representação *Lanchonete(PQ)*, bastante comum para sentenças como "PQ é uma lanchonete", tem-se que "lanchonete" é um predicado de um lugar que afirma uma propriedade de um objeto. No caso, ele codifica a categoria a que pertence "PQ".

Além desses dois construtos, há também as **funções**, que não estão presentes na Figura 2.2. As funções representam relações entre objetivos. Diferentemente dos predicados, elas resultam em um objeto único do domínio. A função "localização de", por exemplo, representada como *LocationOf(x)*, mapeia um objeto a outro, sempre retornando um valor único, como *LocationOf(UFSCar)*[4].

Com a capacidade de se referir a objetos, afirmar fatos sobre objetos e relacionar objetos entre si, a LPO não se restringe a representações rudimentares (ou atômicas) na forma de *Predicate(Terms,...)*, como *Serve(PQ,PãodeQueijo)*. Representações compostas por mais de um predicado também podem ser formuladas por meio dos conectivos lógicos. Para ilustrar, considere a sentença Eu tenho apenas cinco dólares e não tenho muito tempo, para a qual uma possível representação lógica, extraída de Jurafsky e Martin (2025), é exibida em (4).

(4) $Tenho(Falante, CincoDólares) \wedge \neg Tenho(Falante, MuitoTempo)$

A representação semântica para a sentença-exemplo é composta diretamente a partir da semântica das cláusulas individuais (atômicas) e por meio dos operadores $\wedge$ (conjunção lógica) e $\neg$ (conectivo de negação). Assim, da mesma forma que ocorre com a sintaxe, pode-se usar um dispositivo finito (como uma gramática em LPO) para criar um número infinito de representações.

As representações de eventos e estados ilustradas até agora englobam um predicado simples e um conjunto de argumentos necessário para incorporar todos os papéis associados a determinada sentença-exemplo. A representação para "Eu como pão de queijo", isto é, *Come(Eu,PãodeQueijo)*, consiste em um único predicado cujos argumentos são a entidade que serve e o que é servido. Essa abordagem assume que o predicado usado para representar um verbo de evento tem o mesmo número de argumentos que está presente na estrutura de subcategorização sintática do verbo. No entanto, isso nem sempre é o caso, já que representação de eventos pode envolver uma série de participantes, objetos, tempos e locais (p.ex: "Eu como.", "Eu como pão de queijo na UFSCar.", "Eu como pão de queijo no almoço.", "Eu como pão de queijo de almoço na UFSCar.").

Assim, vê-se que escolher o número correto de argumentos para o predicado que representa o significado de "comer" é complicado. Esses exemplos introduzem vários argu-

[4]Em outras palavras, os predicados são usados para afirmar algo sobre objetos (retornando verdadeiro ou falso), enquanto as funções são usadas para produzir novos objetos a partir de outros (retornando um valor específico).

mentos distintos, ou papéis, em uma variedade de formas sintáticas, locais e combinações diferentes. Infelizmente, predicados em LPO têm aridade fixa, ou seja, eles aceitam um número fixo de argumentos.

Para resolver esse problema, a LPO, como salienta Jurafsky e Martin (2025), introduz a noção de uma variável de evento. Isso permite refatorar os predicados de evento para ter uma variável quantificada existencialmente como seu primeiro e único argumento. Usando essa variável de evento, é possível introduzir predicados adicionais para representar as outras informações sobre o evento. A fórmula em (5) ilustra esse esquema com a representação do significado de "Eu como pão de queijo de almoço na UFSCar.", anteriormente citado.

(5) $\exists e Comer(e) \wedge Eater(e, Speaker) \wedge Eaten(e, PodeQueijo) \wedge Meal(e, Almoo) \wedge Location(e, UFSCar)$

Representações de eventos como ilustrado em (5) são classificadas como neo-Davidsonianas (Parsons, 1990), em homenagem ao filósofo Donald Davidson Davidson (1967), que introduziu a noção de variável de evento.

Adicionalmente, uma característica atraente da LPO para o PLN é que ela faz poucos compromissos específicos sobre como as coisas devem ser representadas. Tais compromissos dizem respeito basicamente ao mundo representado consistir em objetos, propriedades de objetos e relações entre objetos. Esses poucos compromissos, no entanto, são compartilhados por muitos dos modelos ou formalismos mencionados anteriormente, incluindo os descritos na sequência.

Do ponto de vista histórico, a LPO desempenhou um papel fundamental no desenvolvimento do PLN, sendo um dos pilares para que os pesquisadores começassem a desenvolver sistemas que pudessem "compreender" e manipular as línguas naturais de forma lógica.

2.1.4.2 *Abstract Meaning Representation*

a) Pressupostos gerais e extensões

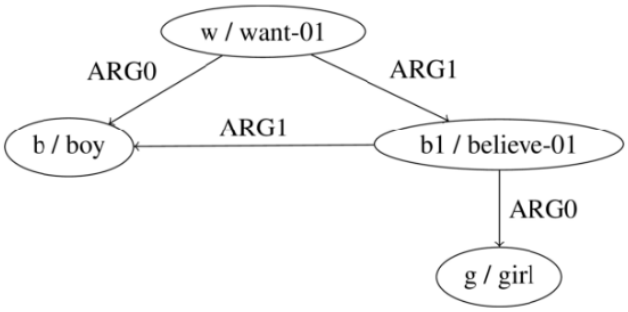
O modelo *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013) foi projetado com o objetivo de produzir grandes bancos de sentenças em inglês e suas respectivas representações semânticas, os chamados *sembanks*. Embora tenha sido originalmente

projetado para o inglês, a adaptação de suas diretrizes[5] para outras línguas é possível, definindo divergências e representações específicas para essas línguas (Li et al., 2016; Anchiêta, R.; Pardo, T., 2018).

No geral, a AMR tornou-se popular no PLN por empregar uma representação simplificada no formato de um grafo do tipo "enraizado, direcionado e com rótulos nas arestas e nós", abstraindo idiossincrasias sintáticas, e por usar recursos linguísticos externos (e abrangentes), como o *PropBank* (Palmer; Gildea; Kingsbury, 2005).

Por idiossincrasias sintáticas, os autores entendem todas as informações que contribuem pouco para o significado essencial da sentença, como determinantes e preposições, além da informação sobre a classe morfossintática e ordem das demais palavras. Assim, abstraindo-se da sintaxe, o modelo busca associar uma mesma representação AMR a diferentes sentenças que tenham o mesmo significado básico. A Figura 2.3, por exemplo, exibe o grafo que seria associado a sentenças como *"The boy wants the girl to believe him"* ("O menino quer que a menina acredite nele") e *"The boy wants to be believed by the girl"* ("O menino quer ser acreditado pela garota"). Nele, nota-se que o determinante *the* ("o/a") e as preposições *to* (formador do infinitivo) e *by* ("pela"), não aparecem. O pronome *him* ("nele") fica codificado pela relação ARG1 entre *believe-01* e *boy*.

Figura 2.3: Exemplo de grafo AMR.



Fonte: Banarescu et al. (2013)

Sobre o grafo, ressalta-se que ele é direcionado e tem raiz única. Esta, aliás, representa o foco da sentença associada ao grafo. Os nós representam conceitos como entidades, eventos, propriedades e estados. Mais precisamente, os conceitos são palavras da sentença (no caso, conceitos lexicalizados), estruturas predicado-argumento (isto é, os *frames/rolesets*, descritos a seguir) do *PropBank*, ou palavras-chave especiais provenientes do pró-

[5]https://github.com/amrisi/amr-guidelines

prio modelo AMR. As palavras-chave incluem tipos especiais de entidades (*data-entity*, *world-region*, etc.), quantidades (*monetary-quantity*, *distance-quantity*, etc.) e conjunções lógicas (*and*, etc.). Seguindo o esquema neo-Davidsoniano (Davidson, 1967), os nós/conceitos são codificados por variáveis de modo que *b / boy*, por exemplo, refere-se a uma instância (chamada *b*) do conceito *boy*.

As arestas representam as relações semânticas entre os conceitos. Assim, a aresta "w / want-01 :arg0 (b / boy))" da Figura 2.3 indica que é o menino (b) quem quer (w). Existem mais de 100 relações, sejam elas argumentos dos *frames/rolesets* (isto é, : arg0, : arg1, : arg2, : arg3, : arg4 e :arg5), relações semânticas gerais (como *:accompanier, :age, :beneficiary, :cause*, etc.) e relações para alguns tipos especiais de entidades, como quantidade (p.ex.: *:quant, :unit, :scale*), data (p.ex.: *:day, :month, :year, :weekday*, etc.) e listas (p.ex.: *:opN*, etc.).

O modelo também inclui as inversas de todas essas relações, por exemplo, *:arg0-of* e *:quant-of*, etc. Essas inversas têm várias funções, sendo uma delas a de permitir que o foco de uma sentença seja representado em um grafo de raiz única. Para ilustrar, considere a sentença *"The boy saw the girl who wanted him"* (O menino viu a menina que o quer). Nela, o foco é o predicado *see-01*, cujo :arg1 é a entidade (conceito) *girl*; essa mesma entidade é, por sua vez, :arg0-of do predicado *see-01*. Ademais, quando uma entidade desempenha múltiplos papéis, o grafo apresenta reentrância, isto é, nó com múltiplos pais. Na Figura 2.3, por exemplo, *"boy* é :arg0 de *want-01* e :arg1 de *believe-01*.

Como mencionado, a AMR utiliza o *PropBank* (Palmer; Gildea; Kingsbury, 2005; Bonial et al., 2014). O modelo faz uso do arquivo de *frames* do projeto *PropBank*. Um *frame* lista os sentidos de um predicador (no caso, verbos) e, para cada sentido, os seus respectivos argumentos nucleares, que são representados por etiquetas numeradas (Arg0 a Arg5). Um *roleset* também fornece exemplo de sentença em inglês anotada. O conjunto de ArgNs que expressa um sentido recebe o nome de *roleset*. Na Figura 2.4, tem-se o *frame/roleset* (ou simplesmente, *frameset*) do *PropBank* empregado na representação da Figura 2.3. No caso, trata-se do *frameset want.01*, que representa o sentido desejo de possuir ou fazer (algo). Esse *frameset* possui dois argumentos, Arg0 e Arg1, com os respectivos papeis semânticos *"wanter"* e *"thing wanted"*. Atualmente, o aquivo de *frames* já está em sua versão 3.4[6,7].

[6]https://propbank.github.io/v3.4.0/frames/

[7]Nessa versão, por exemplo, o *roleset* de *want.01* possui 4 ArgNs: ARG0-PAG (*wanter*), ARG1-PPT

Figura 2.4: *Roleset* fornecido pelo *PropBank*.

Frameset **want.01** "possession desiring"
Arg0: wanter
Arg1: thing wanted
Ex: [*Arg0* I] want [*Arg1* a flight from Ontario to Chicago].

Fonte: Banarescu et al. (2013)

Além de grafo, o significado sentencial segundo o modelo AMR também pode ser representado via LPO (Figura 2.5(a)) e notação PENMAN (Bateman et al., 1991) (Figura 2.5(b)), que pretende facilitar a leitura e a anotação feita por humanos.

Figura 2.5: Demais formatos de anotação AMR.

```
∃ w, b, b1, g:
instance(w, want-01) ∧
instance(b, boy) ∧
instance(b1, believe-01) ∧
instance(g, girl) ∧
ARG0(w, b) ∧
ARG1(w, b1) ∧
ARG0(b1, g) ∧
ARG1(b1, b)
```

(a) Lógica de Primeira Ordem

```
(w / want-01
    :ARG0 (b / boy)
    :ARG1 (b1 / believe-01
        :ARG0 (g / girl)
        :ARG1 b))
```

(b) Notação PENMAN

Fonte: Banarescu et al. (2013)

Como mencionado, a AMR foi empregada para a construção de *sembanks* em diferentes línguas. Para o processamento do inglês, a comunidade do PLN conta com alguns *corpora*, sendo que os do *Linguistic Data Consortium* (LDC) não são públicos. Quanto aos *corpora* públicos, destacam-se dois: o *corpus* produzido a partir do livro "O Pequeno Príncipe"[8] e o Bio AMR[9]. O primeiro contém o texto completo do livro de Antoine de Saint-Exupéry, publicado em 1943 e traduzido para 300 línguas. O segundo possui textos do domínio biomédico, extraídos da PubMed[10].

A respeito de *corpora* em outras línguas, existem algumas iniciativas como apontam Wein e Bonn (2023). Elas foram responsáveis basicamente por anotar versões do livro O Pequeno Príncipe de Saint-Exupéry (publicado em 1943) nas seguintes línguas:

(*thing wanted*), ARG2-GOL (*beneficiary*), ARG3-PPT (*in-exchange-for*) e ARG4-DIR (*from*)

[8] https://amr.isi.edu/download.html

[9] https://www.ncbi.nlm.nih.gov/pubmed/

[10] https://www.ncbi.nlm.nih.gov/pubmed/

chinês, espanhol, vietnamita, coreano, turco e persa. Isso proporciona uma oportunidade para comparar representações AMR do mesmo texto entre línguas diferentes. Além desses recursos monolíngues, existe um *corpus* multilíngue, o AMR 2.0*Four Translations*.[11] A iniciativa para o português será detalhada mais à frente.

A partir de corpora anotados via AMR, muitos analisadores semânticos foram desenvolvidos (Flanigan et al., 2014; Wang; Xue; Pradhan, 2015; Damonte; Cohen; Satta, 2017; Noord; Bos, 2017; Lyu, S.; Titov, 2018; Cai; Lam, 2020). Para avaliar as representações AMR produzidas automaticamente por um *parser* semântico, tem-se duas métricas disponíveis, *Smatch* (Cai; Knight, 2013) e SEMA (Anchieta; Cabezudo; Pardo, 2019).

Além disso, modificações/extensão do modelo AMR também vêm sendo propostas. Uma delas é a de (OGorman et al., 2018), que prevê uma estratégia para a representação de múltiplas sentenças, posto que a versão original da AMR não contemplava esse tópico. De acordo com essa proposição, a representação de várias sentenças em um único grafo AMR pode ser realizada por meio da inserção do construto *multi-sentencial* (m / multisentence) no topo do grafo, que se conecta às duas ou mais sentenças por meio das relações :snt1 e :snt2, como ilustrado na Figura 2.6, do grafo multi-sentencial das sentenças João correu. João não chegou.

Figura 2.6: Anotação de múltiplas sentenças em AMR segundo OGorman et al. (2018).

```
(m / multisentence
        :snt1 (c / correr-01
                :ARG0 (p / person :name (n / name :op1
"João"))
        :snt2 (c2 / chegar-01
                :ARG0 p
                :polarity -))
```

Fonte: O autor, 2025.

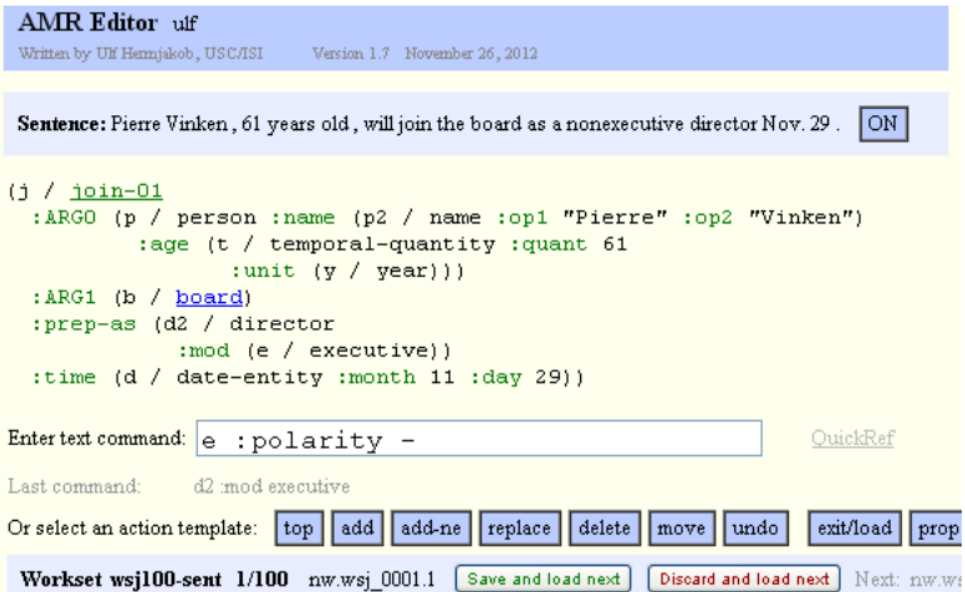
b) Editores ou ferramentas de anotação-AMR

Para a anotação em AMR, a literatura dispõe de alguns editores ou ferramentas de auxílio, em especial, o AMR Editor (Hermjakob, 2013) e o metAMorphosED (Heinecke, 2023). Essas ferramentas são importantes porque ajudam o anotador a garantir certa consistência na anotação, uma vez que o formato PENMAN, comumente empregado, torna a tarefa de anotação manual em um editor de texto comum praticamente impossível (Heinecke,

[11] https://catalog.ldc.upenn.edu/LDC2020T07

2023). A primeira ferramenta que a comunidade do PLN teve à disposição foi o AMR Editor (Hermjakob, 2013), cuja interface está ilustrada na Figura 2.7.

Figura 2.7: Interface do AMR Editor.



Fonte: Hermjakob (2013).

Trata-se de uma ferramenta *web-based*, isto é, disponível online, que possui as seguintes características principais: (i) interface gráfica interativa, que permite a edição visual de AMRs, facilitando a criação e modificação das representações semânticas sem precisar digitar diretamente em formato textual; (ii) visualização de grafos, isto é, exibe a estrutura AMR em um grafo direcionado, tornando a análise e a edição mais intuitivas; (iii) edição estruturada, pois suporta adição, remoção e modificação de nós e arestas no grafo, representando entidades, ações e suas relações, e (iv) conversão entre representações, uma vez que permite alternar entre a visualização gráfica e a forma textual AMR.

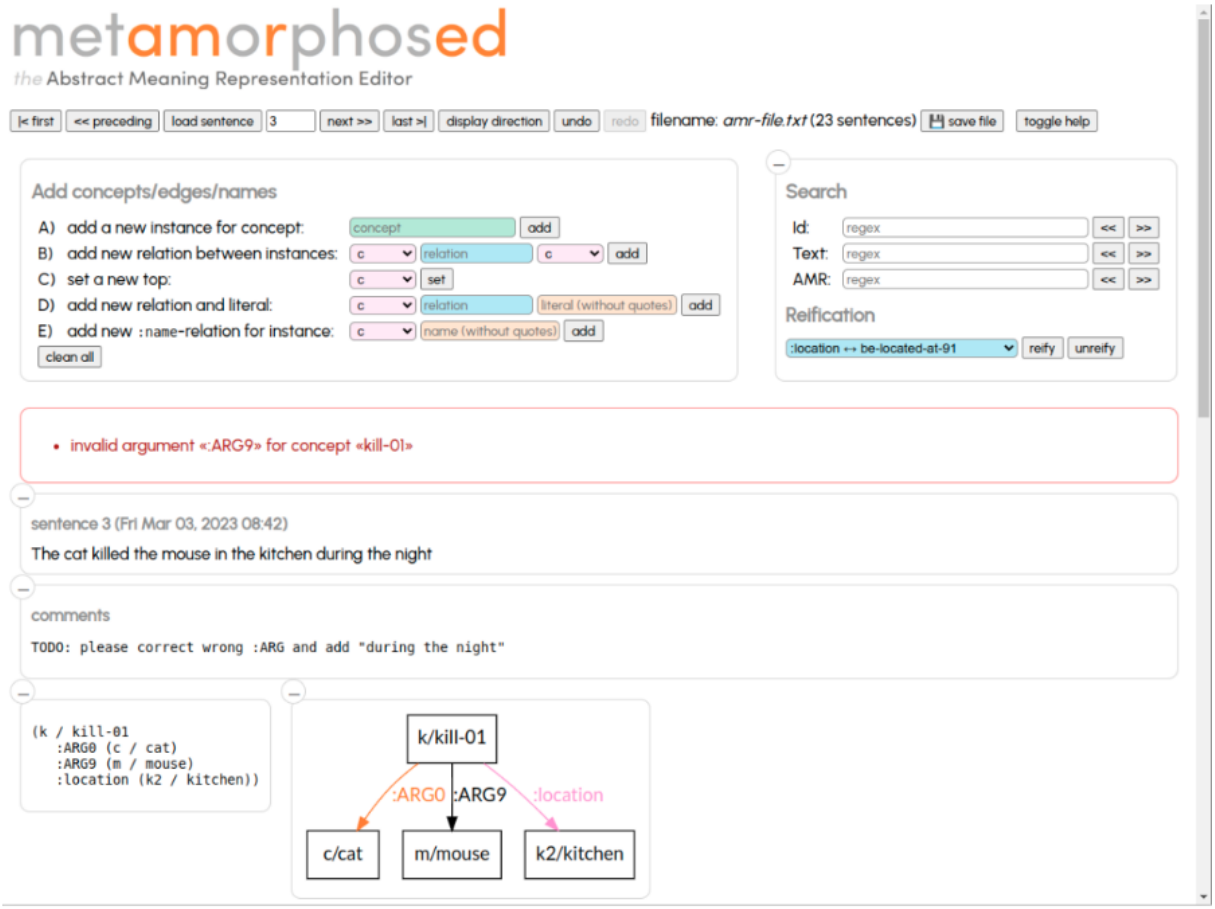
Para tanto, o editor dispõe ao anotador algumas funcionalidades, a saber: (i) diretrizes oficiais de anotação AMR, que podem ser invocadas digitando "*guidelines*" na caixa de comando de texto; (ii) *frames/rolests* do *PropBank* e uma lista de papéis semânticos com explicações e exemplos, que pode ser acessada de várias maneiras; (iii) lista de tipos de entidades nomeadas com explicações e exemplos, que também pode ser acessada de diferentes formas e (iv) função de busca que retorna exemplos prévios de um conjunto de mais de 400 representações AMR adjudicadas.

Essa ferramenta foi amplamente empregada na construção de *sembanks* em dife-

rentes línguas, incluindo o português (Inácio et al., 2023).

Outra ferramenta de edição é o metAMorphosED de Heinecke (2023). Trata-se de um *webserver* implementado em *python*, que conta com uma interface gráfica para a anotação acessível a partir de um navegador de internet. Ele foi desenvolvido a partir da necessidade do autor de anotar um *corpus* especializado de maneira facilitada. Por meio da interface gráfica, o metAMorphosED, assim como o AMR Editor, permite realizar a anotação de um grafo AMR e gerar o formato PENMAN, como ilustrado na Figura 2.8.

Figura 2.8: Interface gráfica do editor *metAMoRphosED*.

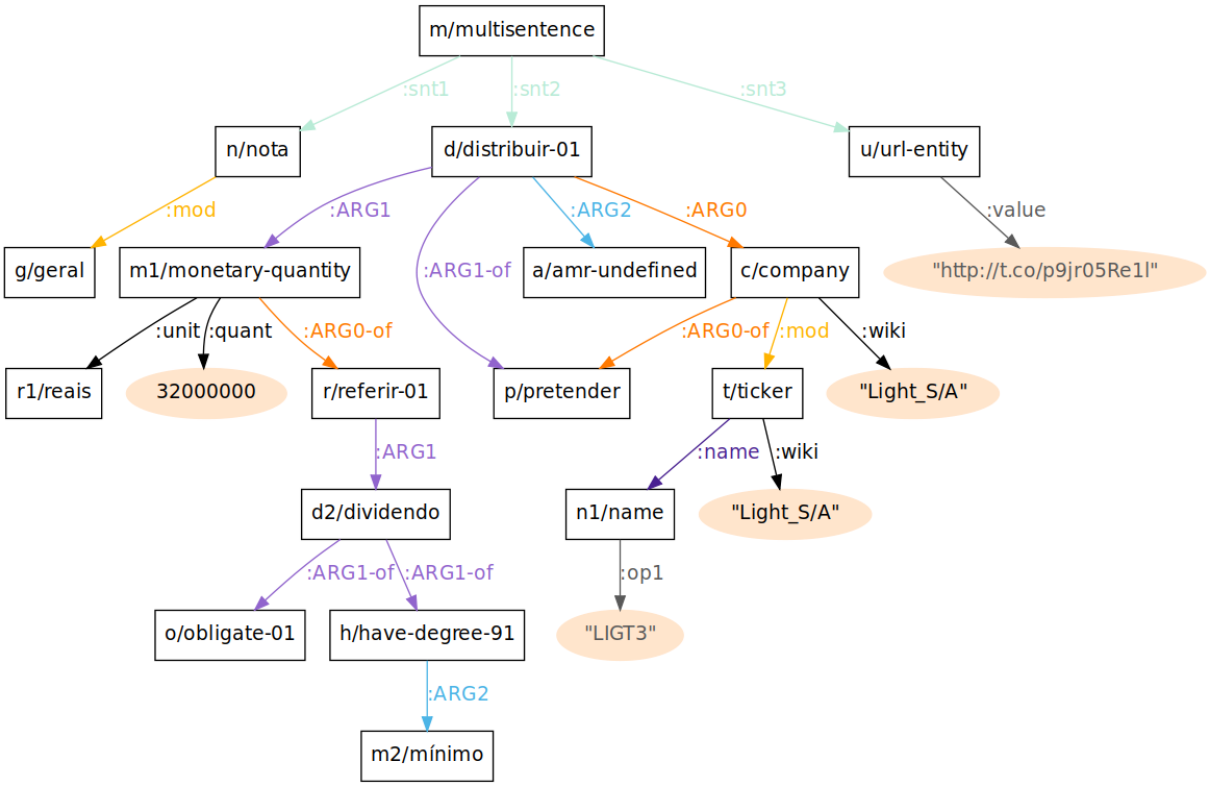


Fonte: Heinecke (2023)

Mais especificamente, o editor conta com diversas funções de validação do grafo, a fim de notificar os anotadores de possíveis erros durante a anotação referentes a: (i) conceitos do *PropBank*, (ii) relações semânticas, com base em uma lista que pode ser modificada, (iii) alcance das relações e (iv) relações reificáveis. A ferramenta dispõe também de (i) funções de pesquisa, seja pelo *id* da sentença, expressões regulares ou construções presentes no grafos e na notação PENMAN, e (ii) funcionalidades de edição em massa

dos grafos anotados, mudança do conceito principal e adição facilitada de conceitos e literais[12]. Os grafos gerados pelo editor também podem ser exportados em formato de imagem, como o grafo representado na Figura 2.9.

Figura 2.9: Grafo AMR gerado por meio do metAMoRphosED.



Fonte: O autor, 2025.

2.1.4.3 *Uniform Meaning Representation*

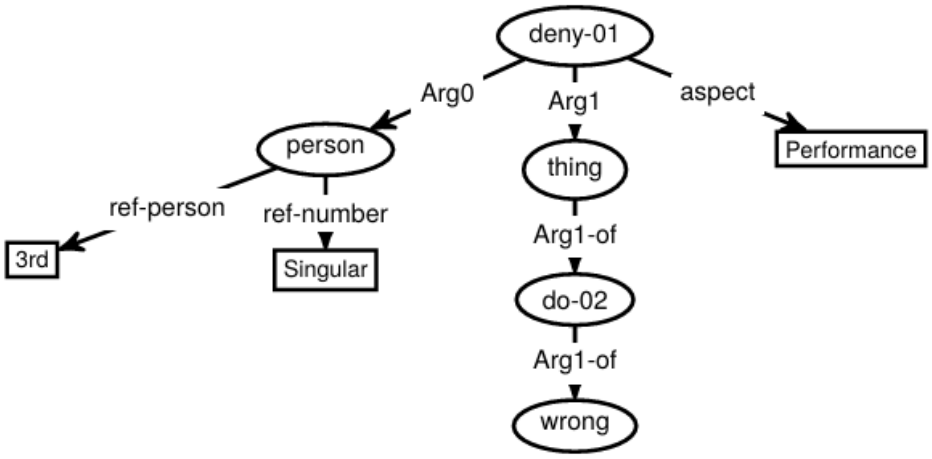
O modelo *Uniform Meaning Representation* (UMR) (Van Gysel et al., 2021) foi proposto com o objetivo de adaptar a AMR ao contexto *cross-language* ou multilíngue[13]. Isso quer dizer que os autores buscaram uniformizar o tratamento da diversidade linguística no processo de anotação AMR. Também por meio de grafos direcionados, a UMR tem dois níveis de representação: (i) nível de sentença, adaptada da AMR, e (i) nível de documento, que captura relações semânticas que ultrapassam os limites das sentenças.

[12]Os "literais" são as próprias palavras, isto é, as formas de superfície do texto.

[13]Outra proposta similar é a *BabelNet Meaning Representation* de Navigli, Blloshmi e Martinez Lorenzo (2022) e Martínez Lorenzo, Maru e Navigli (2022), que estende a AMR para o contexto multilíngue, afastando-se do *PropBank* e utilizando o VerbAtlas (Di-Fabio; Conia; Navigli, 2019) and BabelNet (Navigli; Ponzetto, 2010).

Assim como a AMR, o "nível da sentença" captura a representação das estruturas argumentais dos predicados, os sentidos das palavras, os tipos semânticos das entidades nomeadas, modo e polaridade. Nesse nível, a UMR acrescenta a representação dos atributos de aspecto, pessoa, número, grau e escopo de quantificadores. Uma representação em UMR no "nível da sentença" para *"He denied any wrong-doing."* ("Ele negou qualquer irregularidade.") é ilustrada na Figura 2.10.

Figura 2.10: Anotação UMR no "nível da sentença".



Fonte: Bonn, Buchholz et al. (2024).

Para comparação, destaca-se como os pronomes pessoais são representados na AMR e na UMR. Para tanto, considere o exemplo *"He saw rare birds today."* (Ele viu pássaros raros hoje.) (Figura 2.11). Na AMR, o pronome, no caso explícito, é sempre codificado como um conceito não-analisável, representado pela própria palavra do texto. A UMR, por sua vez, propõe representá-lo pelo conceito *(p / person)*, decompondo-o nos atributos *person* e *number*, buscando acomodar a diversidade das línguas. Assim, os pronomes são formalizados no "nível da sentença" como ilustrado na Figura (Figura 2.11). No caso, os atributos de *"he"* tem os respectivos valores *3rd* (terceira) e *SG* (singular). A subetiqueta ref- é usada para indicar que se trata de um índice de correferência.

Figura 2.11: Tratamento dado aos pronomes nos modelos AMR e UMR.

```
(s / see-01
   :ARG0 (h/ he)
   :ARG1 (b/ bird
           :mod (r/ rare)))
```

(a) AMR

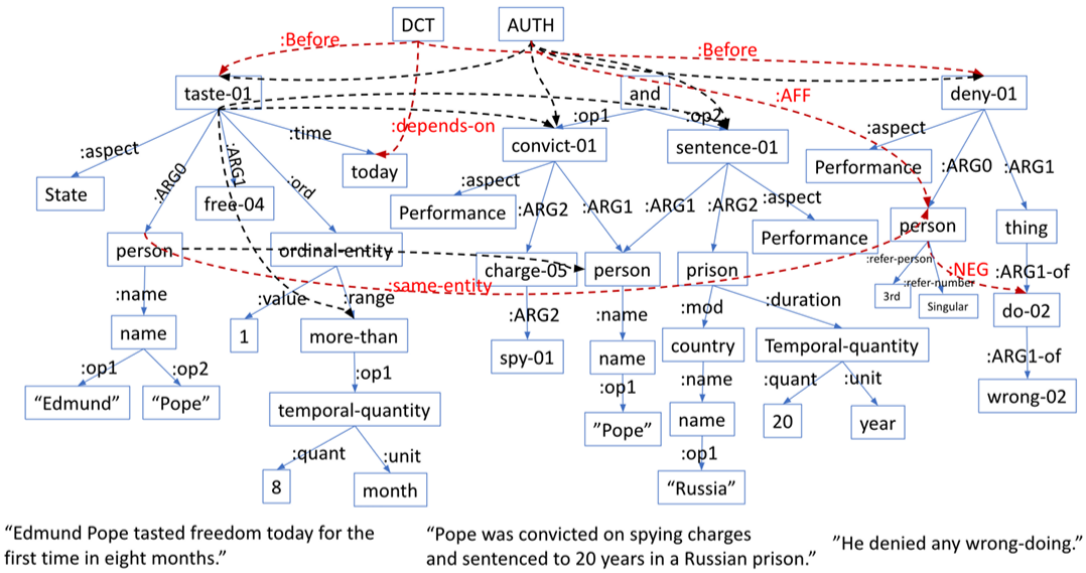
```
(s / see-01
   :ARG0 (p / person
            :ref-person 3rd
            :ref-number SG)
  :ARG1 (b / bird
           :mod (r/ rare)
           :ref-number Plural))
```

(b) UMR

Fonte: O autor, 2025.

No "nível do documento", as relações semânticas capturadas pelo modelo incluem correferência de entidades e eventos, relações temporais entre eventos e entre eventos e expressões de tempo, e dependências modais entre eventos e suas fontes chamadas *conceivers* (Van Gysel et al., 2021; Wein; Bonn, 2023). Isso significa, por exemplo, que, mesmo que se representa o conceito subjacente a um pronome pessoal por *(p / person)* e o decomponha nos atributos *person* e *number*, não seria possível saber a quem esse pronome se refere até que se identifique o seu antecedente, que pode ser encontrado em sentenças anteriores no documento. Isso é ilustrado na Figura 2.12, que é uma representação UMR no "nível do documento" para um curto texto de três sentenças. O conceito UMR *person* para o pronome "*he*" na última sentença se refere a outro conceito *person* da primeira sentença ("*Edmund Pope*"), o que é indicado pela linha tracejada em vermelho rotulada como :same-entity.

Figura 2.12: Anotação UMR no "nível do documento".

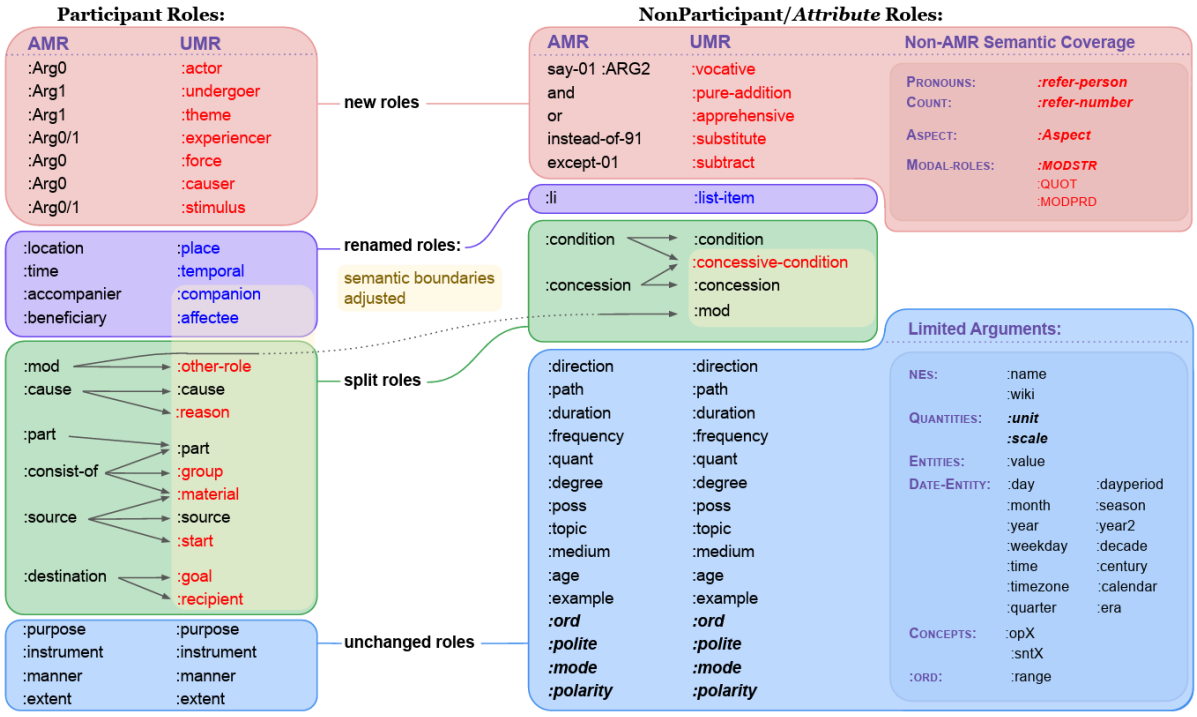


Fonte: Bonn, Myers et al. (2023).

Existem outras relações semânticas que vão além dos limites das sentenças. Na Figura 2.12, por exemplo, os eventos *convict-01* e *sentence-01* da segunda sentença ocorreram antes do evento *taste-01* na primeira sentença. A UMR captura isso por meio de dependências temporais entre um evento ou expressão de tempo e seu tempo de referência ou expressão temporal. Além disso, a UMR captura a factualidade dos eventos, relacionada ao nível de certeza que as fontes (*conceivers* na terminologia UMR) afirmam sobre os eventos. Isso é representado como dependências modais entre um evento e sua fonte no "nível do documento". Na Figura 2.12, por exemplo, tem-se que a fonte do evento *deny-01* é o autor AUTH, que tem força afirmativa (:AFF) sobre a ocorrência desse evento, enquanto a fonte do evento *do-02* é *person*, que nega (:NEG) que o evento *do-02* aconteceu.

Bonn, Myers et al. (2023), aliás, fornecem uma comparação bastante ilustrativa das diferenças e semelhanças entre o formalismo precursor AMR e o sucessor UMR na Figura 2.13.

Figura 2.13: Mapeamento entre AMR e UMR.



Fonte: Bonn, Myers et al. (2023)

Trabalhos recentes sobre UMR produziram um editor online (UMR-Writer) de auxílio à anotação UMR multilíngue (inclusive de línguas que não possuem um recurso ao estilo do *PropBank*), conjuntos de anotações para seis línguas. Além do inglês e chinês, esse conjunto inclui quatro indígenas com propriedades bastante distintas (kukama, arapaho,

sanapaná e navajo) e uma proposta de conversão determinística de AMR para UMR, especificamente os papéis semânticos e conceitos. Além disso, ressalta-se que, para usuários que já estão familiarizados com a AMR, as diretrizes UMR[14] incluem discussões sobre as diferenças entre os modelos, bem como mapeamentos úteis para conversão de um *corpus* AMR para UMR (Bonn; Buchholz et al., 2024).

2.1.4.4 *Lexicalized Meaning Representation*

O modelo *Lexicalized Meaning Representation* (LMR) (Baptista; Reis et al., 2024) é uma adaptação recente da AMR para o português europeu, motivada por desafios impostos pela língua e por questões linguísticas suscitadas pelas diretrizes da AMR na anotação de textos jurídicos. A anotação LMR é intimamente ligada ao texto, isto é, ancorada nas palavras das sentenças e, por isso, dita lexicalizada.

Com isso, o modelo não substitui palavras do texto por construções arbitrárias ou teóricas no grafo[15], ancorando a anotação semântica nas formas superficiais que são os únicos elementos visíveis que fornecem acesso ao significado da sentença, como enfatiza o autor. Assim, o modelo não adota os mecanismos da AMR de (i) reconstrução, como a inserção de pronomes em posições sintáticas lexicalmente não preenchidas[16], (ii) supressão de elementos textuais, como é o caso da maioria das preposições e (iii) substituição, como ocorre com algumas conjunções, representadas por relações semânticas por elas expressas.

Além disso, a LMR, ao contrário da AMR/UMR, adota o mesmo tratamento para os predicadores de diferentes categorias morfossintáticas. Isso quer dizer que, enquanto AMR e UMR tratam de forma distinta predicados verbais (representação padrão via estrutura de argumentos) e adjetivais (via relação :domain) e transformam os nominais em verbais, a LMR trata verbos, nomes e adjetivos da mesma maneira, isto é, por meio de estruturas predicado-argumento. A LMR também se diferencia da AMR e da UMR ao considerar em sua representação os verbos auxiliares, incluindo uma relação específica (:vaux) para a lexicalização desses casos, e ao optar por uma representação simplificada de entidades nomeadas, expressões temporais e quantificação.

[14] https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

[15] Em sentenças como "Eu tenho pouca água potável", por exemplo, o adjetivo *pouca* não constaria do grafo AMR/UMR, pois a representação estaria ancorada em uma conceitualização abstrata do predicado *have-degree-91*, que é empregado para representar construções adjetivais que indicam gradação.

[16] Esse é o caso, por exemplo, do sujeito nulo em português, para o qual a diretriz é a de representar esse argumento elíptico pelo pronome masculino singular *"he"* (Inácio et al., 2023).

As principais diferenças entre AMR e LMR estão sistematizadas no Quadro 2.1.

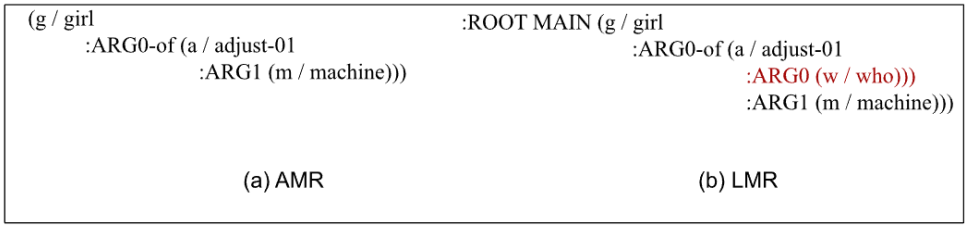
Quadro 2.1: Comparação resumida entre a AMR e a LMR.

AMR	LMR
Grafos acíclicos e direcionados carecem de um nó raiz, empregando em vez disso um arco rotulado como :TOP, que faz um *looping* sobre o nó do elemento predicativo principal da sentença.	Grafos acíclicos direcionados apresentam nó ROOT conectado ao predicador principal (:MAIN).
Reconstrução de elementos reduzidos.	Não ocorre a reconstrução de elementos reduzidos.
Não há relação direta entre as formas textuais e os nós do grafo; a representação gráfica é anexada à sentença.	Relação direta entre formas lexicais textuais e sua representação.
Elementos predicativos são representados por lemas verbais.	Elementos predicativos são preservados em sua forma lexical; sua associação a lemas e construções são feitas no pós-processamento.
Substituição de alguns elementos textuais (como conjunções e preposições) por relações semânticas.	Manutenção de todos os elementos textuais, juntamente de uma representação explícita das relações que expressam.
Verbos auxiliares, suporte e copulativos não são considerados.	Todos os tipos de verbos auxiliares são considerados.
Expressões multipalavra de complexidade variável são representadas, com uma representação sofisticada de entidades nomeadas, e particularmente expressões temporais e de quantificação.	Representação muito simplificada de expressões multipalavras, entidades nomeadas e expressões temporais e de quantificação.
Relações anafóricas intrafrásicas e interfrásicas são representadas; as interfrásicas são representadas pela extensão da notação de OGorman et al. (2018) através de cadeias de correferência no nível do texto.	Relações anafóricas intrafrasais são representadas apenas entre elementos explícitos no texto. A resolução de anáfora é considerada uma tarefa de pós-processamento; Não há diretrizes para a anotação multi-sentencial.
Predicados verbais (representação padrão via estrutura de argumentos) e adjetivais (via relação :domain) são representados de forma diferente, e os predicados nominais são representados pelas construções verbais	Predicados verbais, adjetivais e nominais são representados através da estrutura predicado-argumento.

Fonte: Adaptado de Baptista, Reis et al. (2024).

Para ilustrar uma diferença específica, cita-se o caso do pronome relativo "who" ("que") na sentença *"The girl who adjusted the machine."* ("A menina que ajustou a máquina."). A AMR omite o pronome relativo *"who"*, como ilustrado na Figura 2.14, enquanto a LMR opta por manter o pronome explícito na estrutura argumental do predicado verbal, representado pelo :arg0 do predicado verbal/*frame adjust01*.

Figura 2.14: Tratamento dado ao pronome *"who"* nos modelos AMR e LMR.



Fonte: Baptista, Reis et al. (2024).

Por fim, ressalta-se que o modelo utiliza os mesmos esquemas notacionais da AMR. Na representação PENMAN, como o do Quadro 2.14, fica evidente, aliás, outra particularidade da LMR, que é o fato do grafo direcionado ter o nó :ROOT, que se conecta ao predicado

principal da sentença (:MAIN), servindo como o nó ao qual os elementos com escopo sobre toda a sentença estão conectados. A LMR também faz uso de recursos léxicos externos para ancorar a representação semântica, que são o *Dicionário Gramatical de Verbos do Português* (Baptista; Mamede, 2020a) e ViPEr (Baptista, 2012, 2013) para os verbos e o SNIPER (Baptista; Mamede, 2020b) para os nomes. As diretrizes de anotação LMR estão publicamente disponíveis[17].

2.1.5 Medidas de avaliação

Quanto à avaliação automática de grafos AMR, existem três medidas amplamente utilizadas na literatura, a saber: Smatch (Cai; Knight, 2013), SEMA (Anchieta; Cabezudo; Pardo, 2019) e SemBleu (Song; Gildea, 2019).

A Smatch (CAI; KNIGHT, 2013) é a métrica de avaliação mais amplamente utilizada para comparar grafos AMR, sendo considerada padrão na literatura. O seu funcionamento baseia-se na comparação de triplas lógicas extraídas de dois grafos: um grafo de teste (gerado automaticamente por um parser ou outro anotador) e um grafo de referência (anotado manualmente por humanos). Cada grafo AMR pode ser representado como um conjunto de triplas na forma variável, relação, variável ou literal. A Smatch calcula três valores principais: precisão (P), *recall* (C) e medida-f, definidos pelas equações abaixo:

$$\text{Precisão (P)} = \frac{M}{N}1 \tag{2.1}$$

$$\textit{Recall}\ \text{(R)} = \frac{M}{T}2 \tag{2.2}$$

$$\text{F-score (F)} = \frac{2 \cdot P \cdot R}{P+R}3 \tag{2.3}$$

Na equação (1), M representa o número de triplas corretas, ou seja, aquelas que aparecem tanto no grafo de teste quanto no grafo de referência; N é o número total de triplas no grafo de teste. Já na equação (2), T é o número total de triplas no grafo de referência. Por fim, a equação (3) apresenta a medida-F, que é a média harmônica entre a precisão e *recall*.

[17] https://gitlab. hlt.inesc-id.pt/u000803/lmr4pt/.

Para realizar essa comparação, a Smatch implementa um algoritmo de mapeamento de variáveis que busca maximizar o valor de M, ou seja, o número de triplas coincidentes entre os dois grafos. Essa busca por correspondência ignora os nomes das variáveis (que podem diferir entre grafos) e foca na estrutura lógica representada pelas triplas.

A SEMA (Anchieta; Cabezudo; Pardo, 2019) constitui uma extensão da Smatch e introduz, como diferencial, a consideração da dependência entre os nós no momento da comparação. Enquanto a Smatch verifica apenas se duas triplas coincidem, a SEMA avalia se os nós das triplas possuem os mesmos antecedentes no grafo, o que torna a métrica mais sensível à estrutura interna dos grafos. Além disso, a SEMA exclui a relação TOP (usada para indicar a raiz do grafo) da contagem total de triplas, o que pode alterar significativamente os valores de precisão e cobertura.

A *SemBleu* (Song; Gildea, 2019) é uma métrica automática de avaliação de grafos AMR inspirada na métrica BLEU (em inglês, *Bilingual Evaluation Understudy*) (Papineni et al., 2002), tradicionalmente empregada na avaliação de sistemas de tradução automática. Assim como a BLEU, a SemBleu baseia-se na comparação entre n-gramas produzidos por um sistema automático e os n-gramas presentes em uma referência humana. No entanto, enquanto a BLEU opera sobre cadeias linguísticas (sentenças de texto), a SemBleu adapta esse princípio ao domínio de estruturas semânticas, utilizando os nós e arestas dos grafos AMR como unidades de comparação.

A métrica mensura o grau de sobreposição estrutural entre dois grafos AMR por meio da coocorrência de n-gramas extraídos a partir de seus componentes lineares. Para isso, os grafos são primeiramente transformados em listas lineares contendo sequências de até três unidades conectadas, ou seja:

- **unigramas**, correspondentes a nós isolados (conceitos);
- **bigramas**, compostos por pares do tipo `conceito relação conceito`;
- **trigramas**, compostos por sequências encadeadas com duas relações e três nós, como $\texttt{conceito}_1$ $\texttt{relação}_1$ $\texttt{conceito}_2$ $\texttt{relação}_2$ $\texttt{conceito}_3$.

A precisão de cada tipo de n-grama é então calculada pela razão entre a quantidade de n-gramas coincidentes entre o grafo de teste (a) e o grafo de referência (r), e o número total de n-gramas presentes no grafo de teste. Cada precisão parcial é ponderada de forma

igual (peso 1/3) e a média geométrica dessas precisões é combinada com um fator de penalidade por brevidade, denominado *Brevity Penalty (BP)*, de forma análoga à BLEU. O cálculo final da SemBleu é dado pela equação 2.4, em que:

$$\text{SemBleu} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) 4 \tag{2.4}$$

- p_n representa a precisão dos n-gramas de ordem n;
- w_n é o peso atribuído a cada ordem;
- N é o maior valor de n considerado;
- BP é o fator de penalidade por brevidade, que desestimula a produção de grafos muito menores que a referência.

  O valor de BP é definido da seguinte forma:

$$BP = \{\, 1 \,, \text{se } c > re^{(1-\frac{r}{c})}, \text{se } c \leq r$$

Em que c é o comprimento do grafo gerado e r o comprimento do grafo de referência (considerando o número total de nós + arestas).

2.2 Os formalismos semânticos e o português

Até o momento, somente recursos e ferramentas segundo a AMR já foram desenvolvidos para o português. Uma razão para isso talvez seja o fato de os modelos UMR e LMR serem os mais recentes da literatura, sobretudo LMR.

2.2.1 Recursos

a) Corpora

O primeiro recurso foi o *corpus* AMR-LittlePrince (Anchiêta, R.; Pardo, T., 2018). Trata-se da anotação da versão em português brasileiro do livro "O Pequeno Príncipe" de Saint-Exupéry. O *corpus* é o resultado de um alinhamento automático entre o *corpus* contendo a referida obra em inglês já anotada com AMR e a versão em português, com posterior revisão manual. Esse recurso contém 1.527 sentenças anotadas.

Além disso, existem outros três *corpora*. Dois deles são o AMRNews, composto por 870 sentenças de notícias do jornal *Folha de São Paulo*, e o OpiSums-PT-AMR que, contendo 404 sentenças opinativas de comentários sobre livros e produtos eletrônicos, é um *corpus* de CGU (Inácio et al., 2023). O OpiSums-PT-AMR também engloba resumos extrativos e abstrativos dos comentários. O AMRNews teve uma primeira versão com 299 sentenças (Cabezudo; Pardo, 2019), a qual foi estendida por Inácio et al. (2023) pelo acréscimo de mais 571 sentenças, resultando na atual versão. Inácio et al. (2023) também foram os responsáveis por definir as diretriz finais que fundamentaram a anotação do AMRNews. Em ambos os recursos, a anotação foi feita manualmente seguindo as diretrizes originais da AMR, com adaptações para alguns fenômenos da língua portuguesa.

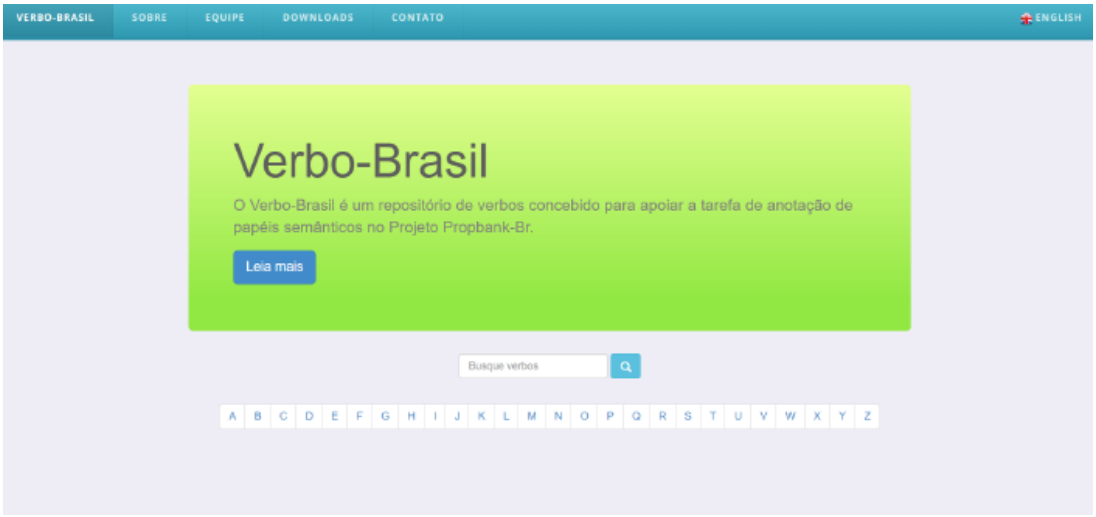
O AMRScien-Br-Corpus (Seno et al., 2022), por sua vez, possui 200 sentenças provenientes de um *corpus* paralelo português-inglês de notícias de divulgação científica da revista brasileira "Pesquisa FAPESP Online" (Aziz; Specia, 2011). A anotação semântica do AMRScien-Br-Corpus foi obtida por meio de um processo de mapeamento/-tradução do conhecimento semântico disponível na representação AMR da língua-fonte (inglês) para a língua-alvo (português). Em outras palavras, a partir dos grafos AMR produzidos para o inglês por meio de um *parser* de Chunchuan Lyu e Titov (2018), geraram-se os grafos paralelos no português, usando a configuração completa do *parser* XPTA (Seno et al., 2022). Posteriormente, os grafos das sentenças em português foram pós-editados por 4 anotadores nativos do português com experiência em PLN e anotação de *corpus*.

b) O repositório Verbo-Brasil

Com exceção do AMR-LittlePrince e AMRScien-Br-Corpus, os demais *corpora* foram anotados com base no Verbo-Brasil (Duran; Martins; Aluísio, 2013). Esse repositório é resultado do PropBank.Br (Duran; Aluísio, 2012), projeto que produziu dois *corpora* anotados nos moldes do *PropBank* do inglês. O Verbo-Brasil possui 1.060 verbos e foi concebido exatamente para apoiar a tarefa de anotação de papéis semânticos. Na Figura 2.15, 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.

Seguindo o *PropBank* original, o Verbo-Brasil apresenta um arquivo para cada verbo, em

Figura 2.15: Interface principal do repositório Verbo-Brasil.



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

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 de *corpus*, 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.

Na Figura 2.16, o verbo "comer" tem apenas um sentido (*comer.01*) ("ingerir, consumir"), o qual está mapeado para o *frame eat.01* do *PropBank*. O *frame/roleset* prevê 2 argumentos numerados e seus respectivos papéis semânticos: Arg0 (*comensal*) e Arg1 (*comida*). Na sequência, o repositório exibe três sentenças-exemplo advindas do *corpus* jornalístico PLN-Br (Bruckschen et al., 2008) e anotadas segundo o *roleset*.

c) Diretrizes e estratégias de anotação

Quanto às diretrizes de anotação AMR em português, não há um manual propriamente dito, uma vez que os trabalhos já realizados de acordo com esse formalismo seguiram, de maneira geral, as diretrizes originais para a língua inglesa. Entretanto, existem diretrizes específicas para alguns fenômenos do português (Inácio et al., 2023):

- **Diminutivo**: representar o conceito subjacente a um diminutivo semântico[18] pelo seu lema e relação :degree. No caso de *fininho* em "Lindo, *finhinho* e discreto", essa relação liga dois conceitos: *fino* e *muito*; representar o conceito subjacente a um diminutivo pragmático[19] ("Muito *engraçadinho*!") pela própria forma de superfície

[18]Indicam "tamanho/qualidade/intensidade reduzido" com base nas propriedades da entidade.
[19]Não expressam conceitos relacionados a :degree, mas conotam afeto ou agradabilidade.

Figura 2.16: Exemplo de arquivo do Verbo-Brasil para o verbo “comer”.

Predicate: *comer*

Roleset id: comer.01 , *ingerir; consumir (v?rios usos figurativos)*, vncls: 39.1, Mapeamento para o ingl?s: eat.01

Roles:

Arg0: *comensal* (vnrole: 39.1-agent)
Arg1: *comida* (vnrole: 39.1-patient)

Exemplo 1:

bosA. s2230: Os adultos que voltam a comer prote?nas podem ter piora da concentra??o e da mem?ria e uma maior agita??o.

Arg0: que
Rel: comer
Arg1: prote?nas

Exemplo 2:

bosC. s1637: Para chegar a suas conclus?es, Rosa entrevistou um grupo de funcion?rios p?blicos (t?cnicos administrativos) que trabalham e comem no centro velho da cidade de S?o Paulo (ruas L?bero Badar?, S?o Bento e adjac?ncias) .

Arg0: que
Rel: comem
Argm-loc: no centro velho da cidade de S?o Paulo (ruas L?bero Badar?, S?o Bento e adjac?ncias)

Exemplo 3:

bosE. s459: N?o sei se fui claro: seria mais f?cil um pinguim cruzar tr?s Saaras do que FHC comer carne de bode espremido a um magote de sertanejos suados.

Arg0: FHC
Rel: comer
Arg1: carne de bode espremido a um magote de sertanejos suados

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

da palavra (no caso, *engraçadinho*).

- **Sujeito oculto**: explicitar os sujeitos elípticos; especificamente, usar o pronome da terceira pessoa no masculino singular “ele para representar sujeito oculto que se refere a uma entidade animada, como em “[Eu/você/ele/ela] Não precisaria agir assim, e usar o pronome demonstrativo “isso para representar o conceito subjacente a um sujeito oculto inanimado, como em “[Isso/aquilo] Não é uma questão pessoal.

- **Sujeito indeterminado**: explicitar os sujeitos indeterminados; mais precisamente, usar o pronome de terceira pessoa no plural (isto é, “eles”) para representar um agente não identificado de uma ação que comumente é expressa por um verbo na terceira pessoal do plural, como em “*Dirão* até que é futebol raiz”.

- **Ambiguidade de sujeito indeterminado**: não explicitar o sujeito na anotação AMR quando não for possível determinar se a ação expressa por um verbo é atri-

buída a um sujeito animado ou inanimado, como em "[Ele/ela/isso/aquilo] Pode até criar problema emocional e retração social.

- **Ambiguidade de pronome possessivo**: representar o conceito subjacente a um pronome pronome possessivo ambíguo, como "sua" em "Ana encontrou Pedro no parque com *sua* [dele/dela] bicicleta, pelo pronome masculino "seu.

- **Expressões multipalavras**: representar o conceito de uma multipalavra através de uma unidade lexical sinônima (p.ex.: "ter direito e "merecer); representar o conceito de uma expressão (comumente idiomática) compostas por um verbo suporte e um nome predicador (p. ex.: "pagar mico) que não tem equivalente lexicalizado, considerando o verbo suporte como pleno; representar o conceito de uma expressão adverbial composta por palavra funcional, como "atrás de em "Ela fez uma besteira *atrás da outra*, por uma abstração resultante da hifenização das palavras da expressão e lematização ("atrás-de-outro).

- **Ambiguidade semântica**: utilizar o sentido mais frequente para a anotação em casos de ambiguidade.

- **Verbos modais**: utilização dos *framesets* da AMR para a anotação de verbos modais, uma vez que o Verbo-Brasil não contempla inteiramente esse tipo de verbo.

É importante ressaltar que, apesar dessas diretrizes terem sido utilizadas na anotação dos *corpora* AMRNews e OpiSums-PT-AMR, as decisões que levaram a elas ainda são questões em aberto e passíveis de discussão. Além disso, de acordo com Cabezudo e Pardo (2019), certas particularidades do português, como sentenças com negação dupla, levam muitas vezes os anotadores a omitirem uma das negações na anotação do grafo.

Além de diretrizes para a anotação de fenômenos linguísticos, Cabezudo e Pardo (2019) e Inácio et al. (2023) estabeleceram um procedimento geral para a anotação-AMR de *corpus*, especialmente em português. Esse procedimento de anotação consiste de duas etapas gerais: (i) análise linguística da estrutura da sentença e (ii) representação, com a ajuda de um editor AMR, do significado subjacente à sentença. Mais precisamente, esse procedimento envolve:

1. *Identificação do tipo de sentença*: identificar se a sentença é padrão, comparativa, superlativa, coordenada, subordinada e outras, pois isso determina se é necessário

construir dois ou mais subgrafos (no caso de sentenças coordenadas ou subordinadas) e, em seguida, uni-los usando uma conjunção (geralmente sentenças coordenadas) ou um conceito do subgrafo principal (no caso de sentenças subordinadas).

2. *Identificação de conceitos*: essa etapa consiste em determinar os conceitos subjacentes à sentença que precisam constar no grafo, sejam eles palavras da próprio sentença (conceitos lexicalizados), estruturas predicado-argumento (isto é, os *frames*), ou palavras-chave especiais da AMR. Para tanto, utilizam-se as diretrizes gerais do inglês, as específicas do português, assim como o Verbo-Brasil.

3. *Identificação do conceito principal e das relações entre os conceitos*: essa etapa é feita com base nas duas etapas anteriores.

2.2.2 *Parsers* ou analisadores semânticos

Para a análise semântica AMR do português, há dois *parsers* na literatura.

O RBAMR (Anchiêta, R. T.; Pardo, T. A. S., 2018) é um analisador baseado em regras. De acordo com Rafael Torres Anchiêta e Thiago Alexandre Salgueiro Pardo (2018), a ausência de um um grande *corpus* em português anotado em AMR levou os autores a desenvolver essa ferramenta com base em três etapas: (i) análise sintática, a fim de identificar relações de dependência, categorias morfossintáticas, entidades nomeadas e o verbo principal da sentença; (ii) anotação de papeis semânticos, com o objetivo de extrair a estrutura argumento-predicado e (iii) a aplicação de regras para gerar o grafo AMR. Para as etapas de análise sintática e anotação de papéis semânticos, foram utilizados, respectivamente, o *parser* sintático PALAVRAS (Bick, 2000) e o Brazilis SRL (Hartmann; Duran; Aluísio, 2016).

O RBAMR engloba especificamente seis regras para a geração dos grafos AMR, são elas: (i) anotação das relações :name e :opN para entidades nomeadas; (ii) geração da relação :mod para adjetivos seguidos de substantivos; (iii) aplicação da relação :manner quando ocorre a relação sintática :advmod; (iv) a anotação da relação :degree em casos de modificadores adverbiais; (v) anotação de :polarity negativa nas ocorrências de argumentos de negação e (vi) a aplicação da relação :time nas ocorrências de argumentos temporais.

O *parser* foi avaliado utilizando o *corpus* AMR-LittlePrince (Anchiêta, R.; Pardo,

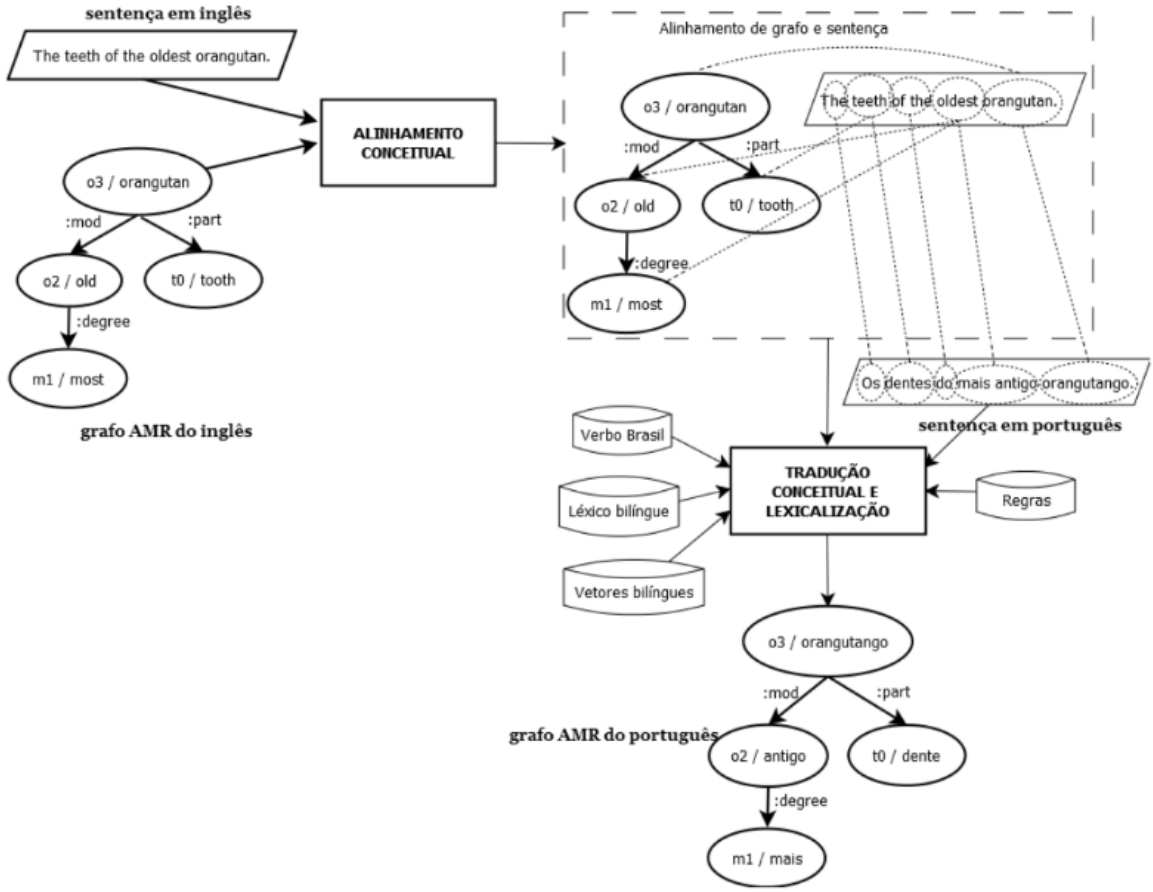
T., 2018) e sentenças curtas e longas. A ferramenta também foi comparada a *parsers cross*-linguísticos ou multilíngues, propostos por Damonte e Cohen (2018). Nessa comparação, o RBAMR obteve medida-F *Smatch* (Cai; Knight, 2013) de 61% em sentenças menores que a média dos testes, 46% em sentenças mais longas que a média e 53.5% em todas as sentenças de teste, atingindo resultados mais altos através da metodologia de aplicação de regras, em comparação aos parsers cross-linguísticos.

O RBAMR foi recentemente estendido (Anchiêta, R.; Pardo, T., 2022) pela a adição de duas novas regras para produzir (i) a relação :domain para os verbos "ser/estar" seguidos de adjetivos e substantivos que não são *framesets* do PropBank e (ii) os conceitos contrast-01 e and para as conjunções contrastivas e aditivas, respectivamente. A nova versão do *parser* também inclui um método de poda que remove os conceitos redundantes que possuem o mesmo nó-pai. A nova versão obteve resultados superiores a outros analisadores adaptados, como o CAMR (Wang; Xue; Pradhan, 2015) e o AMREager (Damonte; Cohen; Satta, 2017), alcançando, para sentenças curtas, medida-F de 0,66 ne métrica Smatch e de 0,48 na métrica SEMA (Anchieta; Cabezudo; Pardo, 2019) e, para sentenças longas, medida-F de 0,49 na Smatch e de 0,28 na SEMA.

O *parser* XPTA (Seno et al., 2022) se baseia em uma abordagem entre línguas (ou *cross-language*, indicado pelo X de XPTA). Mais precisamente, ele parte de um *parser* AMR existente para o inglês e de vários recursos linguístico-computacionais bilíngues inglêsportuguês e mapeia o conhecimento semântico disponível no inglês para a representação do significado equivalente em português, como ilustrado na Figura 2.17.
O mapeamento ocorre em duas etapas: (i) alinhamento conceitual entre grafos e sentenças em inglês, que consiste na detecção de correspondências das palavras do inglês com os nós do grafo, e (ii) tradução conceitual e lexicalização do grafo do inglês para o grafo em português, utilizando o alinhamento anterior e vários recursos para a escolha da unidade lexical em português que equivalente ao conceito AMR do inglês. Mais precisamente, a tradução conceitual e lexicalização ocorre em três etapas: (i) mapeamento de verbos, utilizando o Verbo-Brasil (com exceção dos modais), (ii) mapeamento de conceitos através da aplicação de regras e (iii) mapeamento de conceitos utilizando recursos bilíngues.

Os autores propuseram, a partir de uma análise de *corpus*, cinco regras para o mapeamento dos conceitos: (i) preservação da raiz e do conceito *:name* nas ocorrências de entidades nomeadas, com os operadores carregando o nome da entidade; (ii) mapeamento

Figura 2.17: Etapas de funcionamento do *parser* XPTA.



Fonte: Seno et al. (2022)

para a forma adjetiva de advérbios terminados em -mente, seguindo as diretrizes originais da AMR; (iii) preservação em inglês dos conceitos referentes a *frames* especiais (como have-org-role-91) ou reificação (p.ex.: be-located-at-91); (iv) manutenção em inglês dos conceitos-chaves da AMR, como monetary-quantity e temporal-quantity e (v) manutenção em inglês de conjunções como and, or, contrast-01.

Quanto à avaliação, destaca-se que o XPTA foi avaliado de forma automática com base no *corpus* paralelo AMRScien-Br-Corpus de 200 sentenças. A avaliação se deu pela comparação dos grafos AMR produzidos automaticamente pelo *parser* com grafos de referência feitos por anotadores humanos. Também foram avaliados diferentes configurações do modelo a fim de identificar a contribuição individual de cada um dos recursos usados no mapeamento dos grafos. Além disso, foi utilizado um normalizador AMR, que converte grafos não-reificados em grafos reificados, a fim de comparar estruturas equivalentes. Nessa avaliação, utilizaram-se três medidas: Smatch, SEMA e Sembleu (Song;

Gildea, 2019). Antes da normalização, o *parser* atingiu medidas-F de 58% na Smatch, 35% na SEMA e 26% na Sembleu. Após a normalização, os resultados alcançados foram de 66% na Smatch, 47% na SEMA e 34% na Sembleu.

3 Seleção e descrição do *corpus*

O *corpus* selecionado foi o DANTEStocks (Di-Felippo; Roman, 2025), originado dos 4.517 *tweets* do mercado financeiro compilados por F. J. V. Silva e Carvalho (2020) em 2014. A compilação foi automática a partir da ocorrência de pelo menos um *ticker* de uma das 73 ações do índice Ibovespa, que é o principal indicador de desempenho das ações negociadas na bolsa brasileira, a B3 ("Brasil, Bolsa, Balcão"). Essa compilação se justifica pelo fato de que os usuários ligados ao mercado financeiro geralmente usam esses para se referir às ações e também às empresas. Um *ticker* é uma sequência alfanumérica de cinco ou seis caracteres que representa um tipo específico de ação de uma empresa, como "PETR4" para a ação preferencial da Petrobras no exemplo (6).

3.1 Características estruturais e lexicais

Sobre as características estruturais dos *tweets* do DANTESTocks, destaca-se que os *posts* constitutivos do DANTEStocks possuem até 140 caracteres. Isso ocorre porque F. J. V. Silva e Carvalho (2020) compilaram as postagens que compõem o *corpus* em 2014, época em que o limite ainda não tinha sido expandido para 280 caracteres. Ariani Di-Felippo et al. (2021) apontam que os *tweets* do DANTEStocks apresentam uma combinação de linguagem padrão e não-padrão. Assim, ele possui *tweets* formados por uma ou mais sentenças bem delimitadas (6) e (7), mas também por *tweets* que apresentam ausência de pontuação (8) ou pontuação equivocada (9). *Tweets* com disfluências (10) e justaposição de fragmentados (11) também ocorrem. Em (10), tem-se um truncamento (ou quebra), indicado pelas reticências, que resulta em uma estrutura sintática incompleta. Todas essas características impõem desafios para qualquer anotação linguística.

(6) No momento PETR4 respeita o suporte de R$ 15,42.

(7) Um motivo a menos para a alta da PETR4. Que venha a correção!

(8) O #PT conseguiu fazer propaganda eleitoral antecipada O que a @dilmabr tem a dizer sobre isso?

(9) Bom dia Marcos, Alguma previsão para petr4?!

(10) Petrobrás PN (PETR4), Gráfico Semanal. Estudo das... http://t.co/5bHkUTy8AC

(11) OIBR4 (mensagem: 956643) http://t.co/VD2ApxqWqR

Scandarolli et al. (2023) fizeram um estudo descritivo sobre as idiossincrasias lexicais e ortográficas de DANTEStocks, sistematizando-as em uma tipologia com duas grandes classes: (i) "variação da norma padrão" e (ii) "norma inovadora" (Quadro 3.1).

Quadro 3.1: Taxonomia de fenômenos lexicais e ortográticos do DANTEStocks.

Phenomenon	Type	Subtype	Attested example	Standard form	Gloss
Standard Norm Variation	Substitution	*Diacritic* (cedilla)	**lancamento** das notas	*lançamento*	`notes issuing´
		Other	**segunda feira** **Neh?**	*segunda-feira* Né? (não é)	`Monday´ `Right?´
	Omission	*Diacritic*	capital **proprio**	*capital próprio*	`equity capital´
		Other	**valu** ferris	*valeu ferris*	`thanks ferry´
	Insertion	*Diacritic*	#PETR4 **fêz** uma Onda 2	*#PETR4 fez uma Onda 2*	`#PETR4 made a Wave 2´
		Other	montar um **Streaddle**	*montar um Straddle*	`to set a Straddle´
	Transposition	-	vc se manteve na **comrpa**?	*vc se manteve na compra?*	`did you stick with stocks?´
Innovative Norm	Abbreviation	*Initialism*	ação de **LP**	*ação de longo prazo*	`long-term stock´
		Shortening	(eles) falam **q** por **enqt**	(eles) *falam que* *por enquanto*	`(they) say that´ `for now´
		Contraction	**pq** será?	*por que será?*	`I wonder why´
	Neologism	*Agglutination*	44.6k no **Ibolixo**	*44.6 mil no Ibolixo*	`44.6 thousand in Ibotrash´
		Derivation	**diretassa** do morgan	*diretaça do morgan*	`straight from morgan´
		Foreign influence	#itub4 **estopou**	*#itub4 estopou*	`#itub4 stopped´
	Expressiveness	*Graphemic stretching*	**choooooram**!	*choram*	`Cry!´
		Punctuation repetition	linda!!!!	*linda!*	`beautiful!´
		Dialectal variation	De **zóio**!	*De olho!*	`(I am) keeping an eye!´
		Pictogram	:) 😊 muito $	- *muito dinheiro*	`smile´ `much money´
		Capitalization	**LINNDAA**	*linda*	`beautiful´
		Disguise	essa **p******a**	*essa puta*	`this bitch´
	Homophone Writing	*Phonetization*	é **d+**	*é demais*	`(it) is awesome´
		Graphemic substitution	**xatiado**	*chateado*	`upset´
		Onomatopoeia	**hahaha**	-	-
	Medium-dependent token	*Hashtag*	Presidente da **#PETR4**	-	`President of #PETR4´
		At-mention	né, **@user**?	não é, @user?	`isn´t it, @user?´
		URL	http://t.co/OQ3rDdWilf	-	-
		RT	**RT** @user...	-	-
		Truncation	ação sobe **fo**...	*ação sobe forte...*	`Stock rises sharply´
		Code-switching	E ponto final! **PERIOD**!	-	`Full stop! PERIOD´
	Domain-specific token	*Ticker*	**PETR4** subiu	-	`PETR4 went up´
		Cashtag	**$PBR** testando	-	`$PBR (is) testing´
		Decimal number	de **18,xx** a 21,00	-	`from 18,xx to 21,00´
		Valuation rate	ELET6 **+2,09%**	*ELET6 + 2,09 %*	-
		Temporal expression	**1T14, jun/14**	-	`first quarter of 2014`
		Monetary value	perdeu só **R$20,00**	*perdeu só R$ 20,00*	`(it) only lost R$ 20,00

Fonte: Di-Felippo e Roman (2025).

3.2 Anotações pré-existentes

Atualmente, o DANTEStocks tem três tipos de anotação: emoção (F. J. V. Silva; Carvalho, 2020), entidades nomeadas (Zerbinati; Roman; Di-Felippo, 2024) e gramatical (Di-Felippo; Nunes; Barbosa, 2024a).

3.2.1 Emoções

A anotação de emoções no DANTEStocks foi baseada na *Wheel of Emotions* (em português, "Roda de Emoções") de Plutchik e Kellerman (1980), que categoriza as emoções em quatro eixos emocionais fundamentais. Os eixos são compostos por pares de emoções opostas, que variam em intensidade, como: (i) *joy* v. *sadness*, (ii) *anger* v. *fear*, (iii) *trust* v. *disgust* e (iv) *surprise* v. *antecipation* (cf. Figura 3.1).

Figura 3.1: "Wheel of Emotions" de Plutchik.

Fonte: https://en.m.wikipedia.org/wiki/File:Plutchik-wheel.svg.

A tarefa de anotação se deu por meio de um método de contribuição colaborativa via *web* (*crowdsourcing*), no qual voluntários selecionaram uma emoção de cada par oposto para rotular os *tweets*. Esse método foi utilizado a fim de garantir a pluralidade dos anotadores, atenuando possíveis vieses na anotação devido aos perfis dos anotadores. Mais precisamente, todos os 4.517 *tweets* originalmente coletados por (F. J. V. Silva; Carvalho, 2020) foram submetidos à anotação por um conjunto de 442 voluntários. Para cada par de emoções, os anotadores tinham de escolher entre uma das emoções (do par), "neutro"

ou “Não sei”. Nesse processo, garantiu-se que todos os *tweets* fossem anotados por pelo menos 3 participantes diferentes.

Quando a maioria dos anotadores (ou seja, pelo menos dois) escolheu determinada emoção, o *tweet* foi então rotulado como tal. Para ilustrar, a *tweet* (7) Um motivo a menos para a alta da PETR4. Que venha a correção!, por exemplo, tem as seguintes etiquetas de emoções finais: *joy*, *anger*, *trust* e *anticipation*. Dos 4.517 *tweets* analisados, ressalta-se que 240 foram descartados por terem sido rotulados com “Não sei” em todos os pares de emoções pela maioria dos anotadores, resultando em 4.277 *tweets* anotados.

Como avaliação da confiabilidade das anotações, cada emoção etiquetada foi avaliada com base na proporção de anotadores que escolheram aquela emoção. Dos 4.277 *tweets* anotados, 2.340 receberam uma etiqueta majoritária em, no mínimo, um par de emoções, enquanto os restantes foram classificados como neutros. Tendo em vista o tamanho do conjunto original de 4.517, o montante de 2.340 significa que 52% dos *tweets* do *corpus* foram rotulados com pelo menos uma emoção, o que torna a anotação útil como padrão-ouro.

3.2.2 Entidades nomeadas

Além da anotação de emoções, o DANTEStocks conta com uma anotação preliminar de Entidades Nomeadas (EN) (Zerbinati; Roman; Di-Felippo, 2024). Diz-se “preliminar” porque a anotação de EN está sob revisão. A anotação preliminar foi manual e conduzida por apenas 1 anotador, utilizando a taxonomia de EN do Segundo HAREM[1] (Mota; Santos, 2008). Nessa versão, os autores seguiram majoritariamente as diretrizes do HAREM, delimitando e classificando as entidades em contexto. Entretanto, a anotação diverge das diretrizes HAREM ao não anotar entidades com múltiplas categorias.

Mais precisamente, os autores utilizam as 10 classes da taxonomia do HAREM, que inclui as etiquetas: ABSTRAÇÃO, ACONTECIMENTO, COISA, LOCAL, OBRA, ORGANIZAÇÃO, PESSOA, TEMPO, VALOR e OUTRO (Figura 3.2). E eles empregaram o padrão de anotação BIOES (Jurafsky; Martin, 2025), no qual o *token* equivalente ao início da EN é etiquetado com B, os *tokens* internos são rotulados com I e o *token* final é marcado com a etiqueta S (cf. Figura 3.3).

[1]HAREM é uma avaliação de reconhecedores de entidades mencionadas organizado pela Linguateca, que é um um centro de recursos - distribuído - para o processamento computacional da língua portuguesa. Mais informações em: https://www.linguateca.pt/.

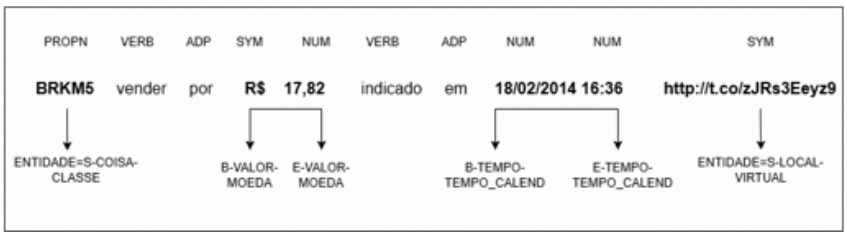
Figura 3.2: Taxonomia de EN do Segundo HAREM.

Categorias	Tipos	Subtipos
ABSTRACCAO (5)	DISCIPLINA	
	ESTADO	
	IDEIA	
	NOME	
	OUTRO	
ACONTECIMENTO (4)	EFEMERIDE	
	EVENTO	
	ORGANIZADO	
	OUTRO	
COISA (5)	CLASSE	
	MEMBROCLASSE	
	OBJECTO	
	SUBSTANCIA	
	OUTRO	
LOCAL (4)	FISICO (7)	ILHA, AGUACURSO, PLANETA, REGIAO, RELEVO, AGUAMASSA, OUTRO
	HUMANO (6)	RUA, PAIS, DIVISAO, REGIAO, CONSTRUCAO, OUTRO
	VIRTUAL (4)	COMSOCIAL, SITIO, OBRA, OUTRO
	OUTRO	
OBRA (4)	ARTE	
	PLANO	
	REPRODUZIDA	
	OUTRO	
ORGANIZACAO (4)	ADMINISTRACAO	
	EMPRESA	
	INSTITUICAO	
	OUTRO	
PESSOA (8)	CARGO	
	GRUPOCARGO	
	GRUPOIND	
	GRUPOMEMBRO	
	INDIVIDUAL	
	MEMBRO	
	POVO	
	OUTRO	
TEMPO (5)	DURACAO	
	FREQUENCIA	
	GENERICO	
	TEMPO CALEND (4)	HORA, INTERVALO, DATA, OUTRO
VALOR (4)	CLASSIFICACAO	
	MOEDA	
	QUANTIDADE	
	OUTRO	
OUTRO (1)		

Fonte: Baseado em Mota e Santos (2008).

A anotação resultou na identificação de 23,453 *tokens* como EN, o que equivale a praticamente 28% do total de 84.397 *tokens* do DANTEStocks. Isso se deu principalmente por conta da metodologia de compilação do *corpus*, baseada na ocorrência de ao menos um *ticker* do Ibovespa, o que gera uma quantidade considerável desse tipo de EN.

Figura 3.3: Exemplo de anotação de EN no formato BIOES.



Fonte: O autor, 2025.

Tendo em vista que a referida anotação incluiu apenas as classes mais genéricas do HAREM, ela passou por um refinamento que consiste na anotação das classes do HAREM, inclusão dos tipos previstos na taxonomia do Segundo HAREM (cf. Figura 3.2) e proposição de novos tipos em função das particularidades do *corpus*. Os novos tipos, que

buscam melhor categorizar as entidades do DANTEStocks, são: *ticker*, (ii) *indicador* e *usuário*, todos da categoria COISA (Piai, 2025). Esses novos tipos foram propostos pra refinar a categorização mais genérica inicial, especificando etiquetas que explicitam EN que são típicas e relevantes no domínio do DANTEStocks, como os ativos/ações, perfis de usuários e indicadores financeiros de forma geral.

3.2.3 Anotação gramatical

Como mencionado, o DANTEStocks possui anotação segundo o modelo *Universal Dependencies* (Nivre; Marneffe; Ginter; Goldberg et al., 2016; Nivre; Marneffe; Ginter; Haji et al., 2020). Ele resulta de uma iniciativa colaborativa internacional que busca criar padrões consistentes de anotação para representar a estrutura gramatical de diferentes línguas/gêneros. Ao oferecer um padrão unificado para representar a sintaxe em diferentes línguas, o modelo permite a análise comparativa e o compartilhamento de recursos entre elas, ampliando a inclusão linguística em aplicações tecnológicas.

Desde 2021, o projeto POeTiSa[2] tem se dedicado à construção de um *corpus* multigênero significativo para fomentar o desenvolvimento de ferramentas e sistemas de análise sintático-semântica, contribuindo para que o português do Brasil não seja mais uma língua pobre de recursos e ferramentas. Nomeado *Porttinari*, esse grande *corpus* atualmente engloba 2 porções de gêneros distintos (jornalístico e CGU) com anotação-UD.

A porção jornalística contém 167,048 notícias do jornal Folha de São Paulo, totalizando 3.964,292 sentenças. Ela é composta por três subcorpora (Duran; Lopes et al., 2023): (i) *Porttinari-base* (revisado para servir de padrão-ouro), (ii) *Porttinari-check* (pequeno e estruturalmente similar ao *Porttinari-base* para servir de *testbed* e ilustrar o contraste entre anotação manual e automática), e (iii) *Porttinari-automatic* (grande e automaticamente anotado). O DANTEStocks corresponde à porção de CGU do *corpus* multigênero *Porttinari*, anotado com UD.

a) Pressupostos gerais da UD

Trata-se de um modelo de dependência, que prevê 2 níveis de representação. No nível

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

morfológico, especificam-se lema, categoria morfossintática (ou *tag* PoS) e traços gramaticais (*features*). No nível sintático, a anotação se dá por relações de dependência (*deprels*) binárias e assimétricas.

A anotação-UD é codificada no formato CoNLL-U, como ilustrado no Quadro 3.2. Nele, tem-se a anotação-UD da sentença (11), extraída do *Porttinari-base*.

(11) O jornalista viajou a convite do Festival do Rio.

Quadro 3.2: Formato CoNll-U típico do modelo UD.

Id	Form	Lemma	Upos Tag	Xpos Tag	Feats	Head	DepRel	Deps	Misc
1	O	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	2	det	_	_
2	jornalista	jornalista	NOUN	_	Number=Sing	3	nsubj	_	_
3	viajou	viajar	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin	0	root	_	_
4	a	a	ADP	_	_	5	case	_	_
5	convite	convite	NOUN	_	Gender=Masc|Number=Sing	3	obl	_	_
6-7	do	_	_	_	_	_	_	_	_
6	de	de	ADP	_	_	8	case	_	_
7	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	8	det	_	_
8	Festival	Festival	PROPN	_	_	5	nmod	_	_
9-10	do	_	_	_	_	_	_	_	_
9	de	de	ADP	_	_	11	case	_	_
10	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	11	det	_	_
11	Rio	Rio	PROPN	_	_	8	nmod	_	_
12	.	.	PUNCT	_	_	3	punct	_	_

Fonte: O autor, 2025.

Cada uma das 10 colunas do CoNLL-U é destinada a uma informação específica:

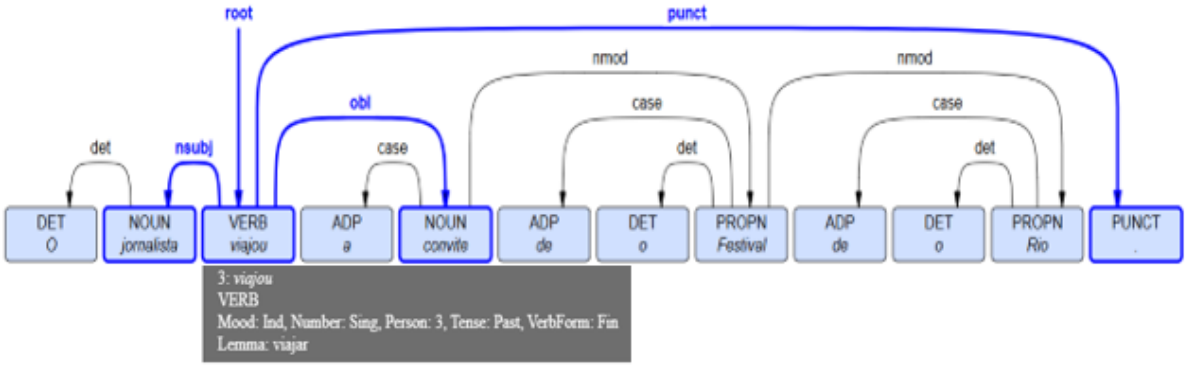
1. ID: o identificador da posição do *token* na sentença (índice numérico a partir de 1).
2. FORM: *token* na forma como ocorre na sentença.
3. LEMMA: lema ou forma canônica da palavra.
4. POS: etiqueta de classe de palavra (ou *part-of-speech* (PoS) *tag*).
5. XPOS: etiqueta PoS específica da língua.
6. FEAT: atributos morfológicos do *token*.
7. HEAD: ID do *head* da *deprel* cujo token (dependente) que está sendo descrito.
8. DEPREL: relação de dependência que conecta o *token* ao seu *head*.
9. DEPS: relação de *enhanced dependency* do *token*.
10. MISC: informações adicionais sobre o *token*.

A partir de um arquivo CoNLL-U, ferramentas de visualização geram representações em grafo. O grafo da Figura 3.4 foi gerado por uma dessas ferramentas[3] a partir do CoNLL-U do Quadro 3.2, que contém a anotação da sentença (11). Nela, vê-se que apenas um

[3]https://urd2.let.rug.nl/ kleiweg/conllu/

token é o **root**[4] da árvore e que as *deprels* estão indicadas por setas rotuladas que se originam no *head* e se destinam ao dependente. O *token* destacado "viajou" é o *root* e suas informações morfológicas estão no retângulo cinza; ele é o *head* das *deprels* **advcl** (oração adverbial), **ccomp** (complemento oracional fechado) e **punct** (pontuação).

Figura 3.4: Exemplo de representação arbórea da anotação-UD.



Fonte: O autor, 2025.

A UD dispõe de 17 *tags* ou etiquetas PoS (Figura 3.5) e um conjunto extenso de traços, além de37 (*deprels*) (Figura 3.6).

Seguindo a decisão do projeto POeTiSA, a anotação-UD do DANTEStocks foi fatorada nos níveis morfológico e sintático, pois, segundo Pardo et al. (2021), a separação dos níveis produz melhores resultados dado que a tarefa é sofisticada. Antes, porém, da anotação em si, o *corpus* passou um pré-processamento.

b) Pré-processamento do DANTEStocks

O conjunto inicial de 4.517 *tweets* de (F. J. V. Silva; Carvalho, 2020) passou por um refinamento, consistindo na exclusão de 469 *tweets* repetidos e/ou não pertencentes ao domínio. O refinamento resultou em 4.048 *posts* que foram efetivamente submetidos à anotação-UD. Ademais, o *tweet* foi tomado como unidade de análise e, com isso, eles não foram segmentados em unidades estruturais menores como sentenças ou sintagmas, e não se aplicou nenhum processo de normalização da linguagem.

Na sequência, o *corpus* foi *tokenizado* segundo os pressupostos da UD em um processo semiautomático (Silva, E. H. et al., 2021). Como o modelo se baseia em uma visão lexicalista da sintaxe, as unidades básicas da anotação são as palavras sintáticas[5]. Assim,

[4]Toda sentença tem uma raiz, normalmente o predicado da oração principal, como dependente da *deprel* **root**.

[5]Tradução do termo em inglês *syntactic word*, que é definido como a unidade mínima a que corresponde

Figura 3.5: 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: Baseada em Nivre, Marneffe, Ginter, Goldberg et al. (2016).

Figura 3.6: As 37 relações de dependência (*deprels*) 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: Baseada em Nivre, Marneffe, Ginter, Goldberg et al. (2016).

as palavras sintáticas (*tokens*) foram segmentadas automaticamente por uma versão do NLTK TweetTokenizer, enriquecida com regras específicas para o DANTEStocks (Silva, E. H. et al., 2021) baseadas na taxonomia de fenômenos de (Scandarolli et al., 2023). A ferramenta preservaria a maioria dos *tokens* delimitados por espaços em branco, incluindo fonetização (por exemplo, d+ (demais), *hashtag*, *cashtag*, *at-mention*, *emoticon* e URL, e separou *tokens* ortográficos únicos que correspondem a várias palavras (sintáticas), como clíticos, contrações (canônicas e não-canônicas), sinais de pontuação (exceto abreviações), taxas de avaliação das ações na bolsa e valores monetários com ortografia não-convencional. Após a revisão manual da saída da ferramenta, o corpus totalizou 81.037 *tokens*.

c) Anotação das informações morfológicas

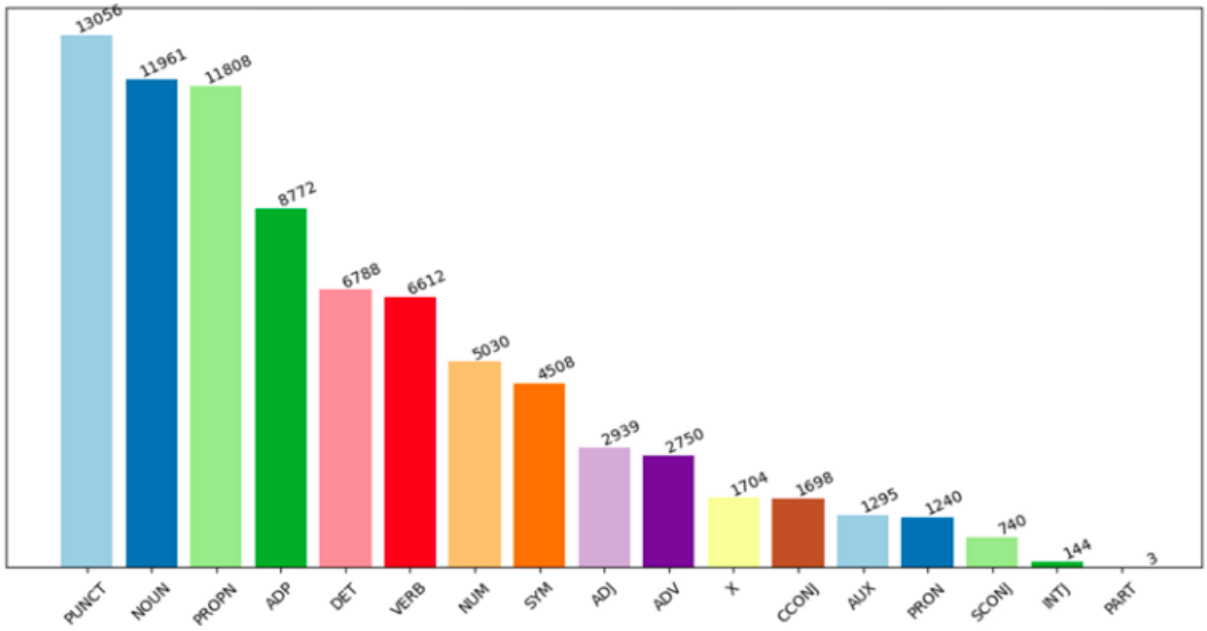
Como mencionado, as informações morfológicas no modelo UD englobam: lema, etiqueta PoS e traços gramaticais. As *tags* PoS foram as primeiras a serem anotadas. Um um processo semiautomático, o *corpus* foi submetido ao *parser* UDPipe 2 (Straka, 2018), treinado incrementalmente sobre o UD-Portuguese Bosque (Rademaker et al., 2017) e *tweets* (Silva, E. H. et al., 2021). Os resultados do *parser* foram analisados manualmente por três anotadores, auxiliados por diretrizes de anotação de PoS definidas para os *tweets* do DANTEStocks (Di-Felippo, A. et al., 2022) e para língua portuguesa (Duran, 2021). Os casos de discordância entre os anotadores foram adjudicados por um linguista sênior com base nos mesmos manuais.

Na Figura 3.7, exibe-se a distribuição das PoS no DANTEStocks. Nela, vê-se que todas as 17 etiquetas UD ocorrem no DANTEStocks. PUNCT é a mais frequente, com cerca de 16% de todos os tokens recebendo essa tag, seguida por NOUN, com cerca de 15%, e PROPN, correspondendo a aproximadamente 14% de todas as tags. Juntas, essas três etiquetas somam quase metade de todas as *tags* PoS (cerca de 45%). A anotação de PoS padrão-ouro do DANTEStocks possibilitou o desenvolvimento do primeiro *tagger* para CGU em português, o Porttagger (Silva, E. et al., 2023), que obteve resultados do esta-da-arte para a tarefa.

De acordo com Di-Felippo e Roman (2025), os lemas e traços (*features*) também foram anotados de forma semiautomática, no caso, a partir do PortiLexicon-UD (Lopes et al.,

uma função sintática. Na anotação-UD, palavras sintáticas são sinônimas de *tokens*.

Figura 3.7: Frequência das etiquetas PoS no DANTEStocks.



Fonte: Barbosa (2024).

2022). Os dados gerados por esse dicionário ou léxico precisaram de uma revisão manual relativamente grande devido à alta taxa de palavras/*tokens out-of-vocabulary* (isto é, não previstas em dicionário). Com relação aos traços, o cenário foi bastante diferente. A extração dos traços do PortLexicon foi guiada pelas *tags* PoS e lemas já validados manualmente, o que diminuiu o esforço de revisão manual dos traços. A maioria das correções foi referente a erros decorrentes da ambiguidade dos traços da classe VERB (*VerbFom, Mood, Tense, Genre, Number* e *Person*).

d) Anotação sintática

A anotação dos deprels no DANTEStocks foi feita em duas etapas semiautomáticas (Di-Felippo; Nunes; Barbosa, 2024a; Di-Felippo; Roman, 2025). A primeira criou um *subcorpus* de referência e a segunda etapa ajustou um *parser* pré-treinado para *tweets*, usando o *subcorpus* de referência como parte de seu conjunto de treinamento inicial, e anotou o restante do *corpus*. Para tanto, os 4.048 *tweets* foram agrupados em três grandes conjuntos em função do tipo de linguagem/estrutura: linguagem relativamente padrão, padrões estruturais recorrentes e outros (*tweets* que não pertencem aos outros dois conjuntos). Os *tweets* foram agrupados por meio do algoritimo *k-means* e *tf-idf* (*term frequencyinverse document frequency*.

A organização dos *tweets* nos referidos conjuntos permitiu selecionar instâncias de

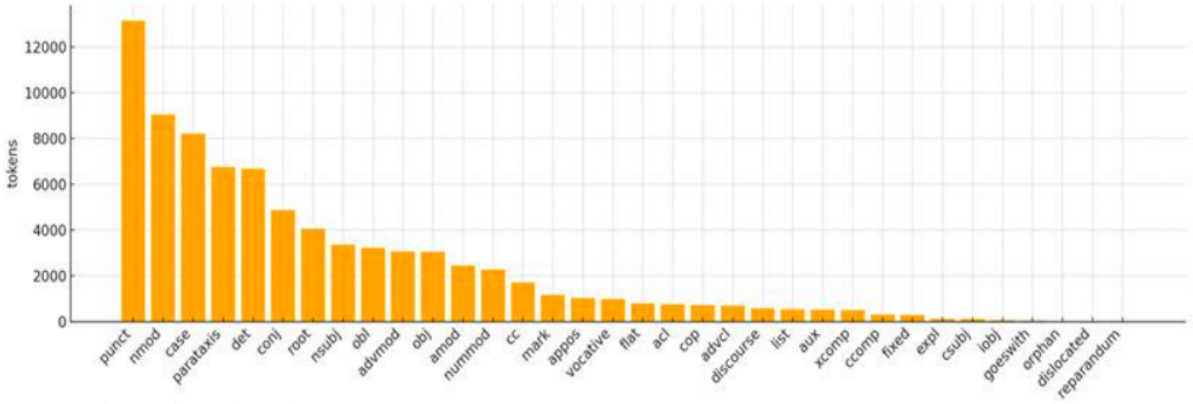
cada um deles para compor um *subcorpus* de referência de 1.000 tweets, cobrindo, assim, a diversidade estrutural do DANTEStocks. Além disso, a anotação semiautomática do *subcorpus* também foi baseada nessa classificação. Visando consistência na anotação, cada conjunto foi anotado e revisado manualmente em separado, começando pelos *tweets* com linguagem padrão, os quais foram seguidos pelos *tweets* com padrões estruturais recorrentes e, por fim, pelos outros.

O 1.000 *tweets* foram submetidos ao UDPipe 2, treinado sobre o UD-Portuguese Bosque. A anotação gerada pelo *parser* foi posteriormente revisada de forma manual por um único especialista. Ao final, obteve-se um *subcorpus* de referência com anotação padrão-ouro. O processo de revisão deu origem a diretrizes específicas para a anotação das *deprels* do modelo UD em *tweets* do domínio do mercado financeiro. Elas foram usadas na revisão manual do restante do *corpus*, que foi feita ao treinar um *parser* de última geração nos *tweets* do DANTEStocks.

O restante do *corpus* foi anotado personalizando o Stanza (Qi et al., 2020) para o DANTEStocks. O Stanza é um modelo pré-treinado para o português, tendo a vantagem de ser um *pipeline* amigável para análise de texto. O processo começou com a arquitetura base do Stanza, ajustada no *Porttinari-base* acrescido pelo *subcorpus* de referência de *tweets*. O modelo de *parser* resultante desse treinamento inicial foi usado para anotar um novo pacote de dados (proveniente dos 3.048 *tweets*), que foi revisado manualmente e incorporado ao conjunto de dados inicial, sendo então usado para iniciar uma nova execução de treinamento do Stanza. Esse ciclo continuou de forma incremental até que o último pacote de *tweets* tivesse sido anotado/revisado. Os pacotes de *tweets* foram adicionados na mesma ordem aplicada na anotação do *subcorpus* de referência: *tweets* de linguagem padrão, *tweets* de padrões estruturais e *tweets* com propriedades lexicais/estruturais variadas.

O desempenho do Stanza foi medido pelas medidas *Unlabeled Attachment Score* (UAS) e *Labeled Attachment Score* (LAS), obtendo UAS 95,78% e LAS de 94,62%. Esses resultados são considerados bons, dada a complexidade da tarefa. Figura 3.8 ilustra a distribuição geral das relações de dependência (sem subrelações) no DANTEStocks. Ao final, esse treinamento gerou o UGC Parser (Barbosa, 2024), que é o primeiro do tipo voltado para CGU em português.

Figura 3.8: Frequência das etiquetas PoS no DANTEStocks.



Fonte: Barbosa (2024).

A respeito da frequência das relações de dependência do modelo UD no DANTEStocks exibidas na Figura 3.8, observa-se 34 das 37 previstas pelo modelo foram empregadas na anotação do corpus. As relações não empregadas foram **clf**, **compound** e **dep**. Das 34, ressalta-se que **parataxis** é a quarta mais frequente, com 6.733 ocorrências. Isso comprova aquilo que Ariani Di-Felippo et al. (2021) já haviam observado sobre a alta frequência de *tweets* fragmentados, uma vez que a **parataxis** se estabelece entre dois elementos que poderiam ter relação sintática entre si, porém essa relação não está explicitada.

Como todo o processo de revisão foi feito por apenas 1 anotador, um segundo especialista em PLN revisou manualmente a anotação automática de 100 *tweets* aleatórios com base nas mesmas diretrizes do primeiro anotador. As árvores de dependência analisadas pelo anotador adicional poderiam ser do *subcorpus* de referência ou geradas pelo Stanza em uma de suas interações. A pontuação do "*Inter-annotator Agreement*" (IAA) foi calculada usando o coeficiente Kappa (Carletta, 1996) em dois cenários diferentes (Di-Felippo; Roman, 2025). No primeiro, o foco foi avaliar a anotação de *head* e *deprel* separadamente. Os resultados do Kappa para *head* e *deprel* foram 0,96 e 0,97, respectivamente. No segundo cenário, a avaliação visou a combinação de *head* e *deprel*, obtendo a pontuação Kappa de 0,95. A IAA por *deprel* foi medida usando a pontuação de concordância total, uma vez que o Kappa não é apropriado devido à distribuição desequilibrada das relações. Nesse caso, obteve-se concordância total de 100% para mais da metade dos 46 diferentes *deprels* (incluindo sub-relações) que ocorrem na amostra de 100 *tweets*.

4 Anotação AMR do DANTEStocks

4.1 Seleção do formalismo AMR

Para a realização deste trabalho, o formalismo semântico selecionado foi a AMR. Essa escolha se deu por motivos práticos/pragmáticos/metodológicos do que propriamente linguísticos. Diz-se isso porque a LMR seria a opção mais linguisticamente motivada, uma vez que ela é lexicalizada, o que dá maior visibilidade ao léxico da língua e pressupõe menos abstrações arbitrárias, e foi proposta tendo o português como uma das línguas de referência cujos fenômenos particulares estão previstos pelo modelo. E, do ponto de vista da robustez, isto é, da capacidade de representar maior número de fenômenos, a UMR seria a opção nesse caso, posto que recobre relações intra- e inter-sentenciais.

No entanto, a publicação da LMR ocorreu apenas em maio de 2024, época em que este trabalho já havia avançado muito na anotação do DANTEStocks a ponto de mudar de formalismo. Quanto ao modelo UMR, destaca-se que sua referida robustez já faz do processo de anotação uma tarefa mais complexa e essa complexidade poderia ser potencializada pelos fenômenos linguísticos dos *tweets*, os quais impõem desafios a qualquer tipo de anotação.

Dito isso, a AMR foi selecionada pelas seguintes razões: (i) consolidação na área do PLN, pois tem sido amplamente empregada na anotação de *corpora* em diversas línguas, inclusive o português; (ii) diretrizes de anotação para certos fenômenos em português, propostas com base na anotação de textos de linguagem padrão (Anchiêta, R.; Pardo, T., 2018; Cabezudo; Pardo, 2019; Inácio et al., 2023); (iii) *corpora* em português com anotação AMR de diferentes gêneros, isto é, jornalístico (AMRNews), literário (AMR-LittlePrince) e CGU (OpiSums-PT-AMR), que podem servir para estudo, consulta e comparação; (iv)

recurso lexical de apoio à anotação AMR em português (Verbo-Brasil), (v) equivalência dos grafos (enraizados, direcionados e com rótulos nas arestas e nós) a outros formatos de representação como conjunções de triplas lógicas (cf. Figura 2.5) e (vii) proposta simplificada de representação de múltiplas sentenças (OGorman et al., 2018), que parece pertinente, já que muitos *tweets* apresentam multiplicidade de segmentos.

4.1.1 Metodologia de anotação

4.1.1.1 Estratégia geral a partir da sintaxe

Tendo em vista que os *tweets* do DANTEStocks apresentam linguagem não-padrão, marcada por fenômenos estruturais e lexicais que dificultam qualquer tipo de anotação, optou-se por não realizar a anotação AMR do zero, mas sim tomando como ponto de partida os grafos de dependência-UD. Essa decisão foi tomada sob a hipótese de que a sintaxe pode contribuir para o processo de geração de representações semânticas do tipo AMR, mesmo que alguns elementos sintáticos sejam reconstruídos, suprimidos ou substituídos na AMR.

A anotação sintática de dependência da UD pode auxiliar principalmente na identificação do *token* que corresponde ao conceito principal do *tweet*, pois se acredita que deve haver em muitos uma correlação entre o *root* da UD e o nó-raiz do grafo AMR. Além disso, ela pode contribuir para a identificação dos ArgN dos *frames/rolesets* provenientes do Verbo-Brasil (ou PropBank). Isso porque há relação entre certas *deprels* e os papéis semânticos dos ArgN dos *frames/rolesets*. Com base em Freitas e Pardo (2024), sabe-se que há uma correlação bastante frequente em textos jornalísticos entre, por exemplo, **obj** e Arg1 (com 92,03%) e **nsubj** e Arg0 (63,49%). Assim, essas correlações podem contribuir para que a tarefa de identificação de ArgN na anotação AMR seja menos ambígua e mais consistente.

Para tanto, tomou-se como ponto de partida a classificação feita por Barbosa (2024), na qual os *tweets* foram organizados em três blocos. Esses blocos foram usados no processo de anotação sintática semi-automática do *corpus* (Di-Felippo; Nunes; Barbosa, 2024a) e guiaram a confecção do manual de diretrizes para anotação de *deprels* no DANTEStocks (Di-Felippo; Nunes; Barbosa, 2024b).

Com base na similaridade estrutural e lexical, os blocos compreendem *tweets* que possuem (i) linguagem relativamente padrão (isto é, compostos ou uma ou mais

sentenças relativamente bem estruturadas) (12), (ii) padrões estruturais recorrentes, como os apresentados pelos exemplos (13a,b), e (iii) estrutura variada (14), a qual levou à classificação dos mesmos como "miscelânea".

(12) **(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

(13) **(a)** #OIBR4 (mensagem : 956643) http://t.co/VD2ApxqWqR
(b) #vale5 (mensagem : 950904) http://t.co/wfR8HEPu4k

(14) **(a)** Tô de olho no HB esperando o MOMENTO HISTÓRICO de PETR4 na era PT . Falta $0,01 pra 13 . E 13 é ... PT!
(c) R$ 13 ... que ironia hein ? ,) #PETR4

A anotação AMR iniciou com os *tweets* que possuem padrões estruturais recorrentes. Até o momento, apenas esse bloco foi analisado. Barbosa (2024) identificou 22 padrões distintos que possuem 1.143 ocorrências em 1.128 *tweets* distintos do total de 4.048 do *corpus* (Figura 4.1). A diferença entre a quantidade de instânciaa e de *tweeta* diz repeito ao fato de que 15 *tweets* distintos possuem mais de um padrão. As ocorrências em (13), por exemplo, correspondem ao Padrão 2 do manual de Di-Felippo, Nunes e Barbosa (2024b), formalizado como: [**<hashtag-ticker> (mensagem:NNN) <url>**].

A opção por iniciar o processo por esse bloco de *tweets* foi motivada por: (i) a proposição de anotação AMR para cada um dos 22 padrões permite anotar suas instâncias de forma consistente, totalizando a quase 1/4 dos *tweets*; (ii) a anotação do bloco composto por *tweets* com linguagem relativamente formal pode se constituir em uma tarefa mais simples, pois poderá contar com a experiência, diretrizes e ferramentas de anotação já existentes para o português padrão, e (iii) a anotação do bloco composto por *tweets* "miscelânia" pode se beneficiar da experiência adquirida com a proposição de diretrizes para os *tweets* com padrões estruturais e os bem-formados.

4.1.1.2 Execução e procedimentos

Das diferentes maneiras para a execução de uma anotação de *corpus*, empregaram-se duas: manual e semiautomática.

Figura 4.1: Padrões estruturais recorrentes e sua frequência no *corpus*.

#	Padrão	Qt.
1	notas gerais <sentença_truncada> ... <url>	25
2	<hashtag-ticker> (mensagem:NNN) <url>	13
3	<hashtag-ticker> <complemento> (mensagem:NNN) <url>	14
4	[<hashtag-ticker><complemento>] <sentença> (mensagem : número) <url>	5
5	<ação&cia> dividendos <lista info&data> <lista ticker&valor> || <url>	10
6	<RT @user> : <ação e/ou cia> <dividendos/juros> <lista info&datas> <lista ticker&valor> || <url>	7
7	<sentença> : <segmento_truncada> ... <url> [hashtag opcional]	174
8	Rastreamento ações – Gráfico diário – <hora>. Analise se romper : <lista ticker&valor>	200
9	Ativo c/ vol Financeiro Superior a sua MM21 - <hora> : <lista ticker>	115
10	<sentença>. Confira a nova indicação agora em url	155
11	<ticker> <tema> <url>	20
12	<prefixo> Prepare-se para o próximo pregão! <sufixo> <lista ticker>. Assista! <url>	11
13	<prefixo> : <sentença> <url>	19
14	<sentença> : <sentença truncada> <url> <lista hashtags>	176
15	<cashtag> <cia> (tipo de ação) - Fato Relevante - <fato> url	29
16	[prefixo opcional] <sentença> <url>[lista ticker]	32
17	<lista ticker&[valor opcional]>	38
18	Gatilho position , semana que vem : Ativo <ticker> - Venda - Validade <data> - Start <valor> - Stop <valor> - RP <valor> < url>	16
19	INTRADAY <ticker>: Suportes <valor> e <valor> e resistências <valor> e <valor> INTRADAY <ticker> : Suportes <valor> e <valor> e resistências <valor> e <valor>	40
20	Ações ex-dividendos hoje : <lista ticker>. As cotações históricas foram ajustadas . Saiba mais ! <url>	19
21	Desempenho de as ações <empresa> em a semana passada : [bolsa] – <lista ticker&índice> , [bolsa – <lista ticker&índice>]	15
22	<ticker> vender a <valor> indicado em <data> e finalizou a compra com resultado de <moeda> <valor> ou <porcentagem> <url>	10
	TOTAL	**1143**

Fonte: Baseada em Di-Felippo, Nunes e Barbosa (2024b) e Barbosa (2024).

4.1.1.2.1 Anotação manual

A anotação manual foi empregada para a definição de uma proposta de anotação AMR para cada um dos 22 padrões a partir das árvores UD de suas instâncias. Ela foi conduzida de forma linear (*tweet* a *tweet*) e auxiliada por uma ferramenta de edição. Nesse processo, ela permitiu a familiarização com a AMR e a linguagem do *corpus* e a criação de diretrizes para fenômenos do português ainda não contemplados e dos *tweets*.

Com base em Cabezudo e Pardo (2019) e Inácio et al. (2023), a proposição de uma anotação AMR para cada um dos 22 padrões foi feita nas seguintes etapas: (i) identificação do tipo estrutural do *tweet*, isto é, se composto ou não por relações **parataxis**, pois isso pode determinar se a representação semântica será multissentencial[1], sendo necessário assim construir dois ou mais subgrafos e os unir usando o construto **m/multisentence** de (OGorman et al., 2018), (ii) identificação dos conceitos que compõem o *tweet*, (iii) identificação das relações semânticas, e (iv) construção do grafo via metAMorphosED.

A etapa (i) tem início com a verificação da ocorrência de **parataxis** na anotação-UD do *tweet*, responsável por conectar segmentos justapostos que não tem relação sintática clara entre eles. A correspondência entre **parataxis** e multiplicidade de segmentos (via m/multisentence), no entanto, nem sempre ocorre, como se verá a seguir. Caso a corres-

[1] O termo "multissentencial" se refere ao fato de um grafo AMR ter dois ou mais segmentos conectados ao nó-raiz por **m/multisentence**; esses segmentos não necessariamente correspondem a sentenças.

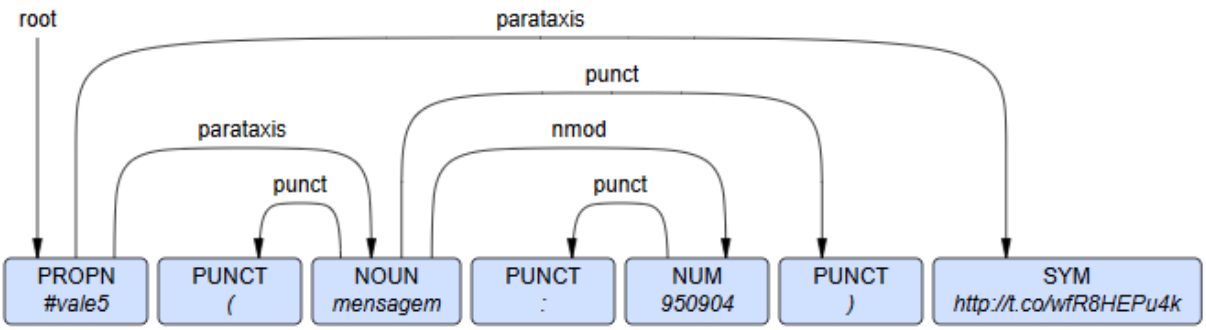
pondência seja pertinente e o *tweet* possua, por exemplo, apenas uma **parataxis**, isso indica que é necessário construir dois subgrafos e, em seguida, uní-los em um só grafo AMR. Caso contrário, a representação AMR não tem subgrafos. A etapa (ii) consiste no mapeamento entre palavras e conceitos e na definição de como representá-los, seja por palavras do próprio *tweet* (conceitos lexicalizados), estruturas predicado-argumento (*frames*) ou palavras-chave especiais da AMR. A etapa (iii) visa identificar as relações semânticas, sejam elas *core*, *non-core* ou específicas.

As etapas (ii) e (iii) pautam-se nas diretrizes gerais AMR do inglês e nas específicas do português e no repositório Verbo-Brasil. Caso um verbo x não estivesse contemplado no Verbo-Brasil ou nenhum de seus sentidos descritos fosse adequado, a diretriz foi consultar o PropBank em busca de um equivalente de x e propor um *frame* para x nos moldes do Verbo-Brasil. Nesse caso, a estrutura argumental que representa o sentido de x é acompanhada de uma na ocorrência do *corpus* para exemplificação.

Quanto ao editor utilizado na etapa (iv), ressalta-se que, inicialmente, o escolhido tinha sido o AMR Editor (Hermjakob, 2013). Logo após o início do trabalho, no entanto, esse editor deixou de estar disponível na *web* e, por isso, adotou-se o metAMorphosED (Heinecke, 2023), que foi empregado para a definição de uma proposta de anotação para a maioria dos 22 padrões estruturais.

Para propor uma anotação do Padrão 2, por exemplo, selecionou-se aleatoriamente uma instância do padrão no *corpus*, que podia ser equivalente a um *tweet* completo ou parte de um *tweet*, pois alguns *posts* são compostos por um ou mais padrões. No caso, a instância selecionada (13a) equivale ao *tweet* completo. A árvore de dependências UD da instância é exibida na Figura 4.2.

Figura 4.2: Anotação UD de uma das instâncias do Padrão 2.



Fonte: Di-Felippo, Nunes e Barbosa (2024b).

Tomando a árvore-UD como ponto de partida, a etapa (i) teve início com a observação de que o *tweet* possui *ticker* como *root* e 2 **parataxis**, as quais indicam 3 segmentos justapostos. Sobre o **root**, vale uma ressalva importante. Como o *tweet* não possui propriamente uma estrutura sintática, a escolha do *ticker* como **root** (predicado principal) parece ter tido motivação semântica. Diz-se isso porque a interpretação do exemplo considerada neste trabalho levou a estabelecer o conceito subjacente ao *ticker* como o principal da representação AMR, coincidindo com a classificado do mesmo como **root** da UD. Mais precisamente, a ocorrência do *ticker* indica o assunto, tópico ou tema, pois, sem ele, não se sabe a que o *post* se refere. Nesse sentido, as demais informações foram interpretadas como relacionadas ao assunto principal e, por isso, conectadas (diretamente ou indiretamente) a ele por relações semânticas específicas. Isso permitiu compor um grafo AMR único de nó-topo lexicalizado e, com isso, a mutliplicidade de segmentos capturada na árvore de dependência sintática não está expressa em AMR via m/multisentence.

Na etapa de mapeamento entre palavras e conceitos, identificou-se que: (i) o PROPN *ticker* expressa entidade nomeada (t/ticker), (ii) o NOUN "mensagem" expressa entidade codificada pelo próprio *token* do *tweet* (m/mensagem), e (iii) a URL e o número da mensagem expressam entidades representadas por construtos específicos da AMR, isto é, u/url-entity e o/ordinal-quantity[2], respectivamente. Por se tratar de uma entidade nomeada, o conceito t/ticker tem uma relação (*non-core*) :name, que conecta o conceito expresso pela entidade (t/ticker) à conceitualização abstrata do nome da entidade (n/-name), e uma relação :wiki, responsável por conectar o conceito t/ticker ao nome da empresa utilizado na página correspondente na Wikipedia (no caso, "Vale_S.A")[3]. O construto n/name da AMR tem relações :opN a depender da quantidade de *tokens* que compõem o nome da entidade. No caso, tem-se apenas :op1, que relaciona n/name a "#vale5". Todo conceito o/ordinal-quantity, por sua vez, tem uma relação :value, assim como u/url-entity.

Após a etapa de mapeamento entre as palavras e os conceitos, que engloba inclusive a definição de como efetivamente cada conceito é representado na AMR, fez-se a cone-

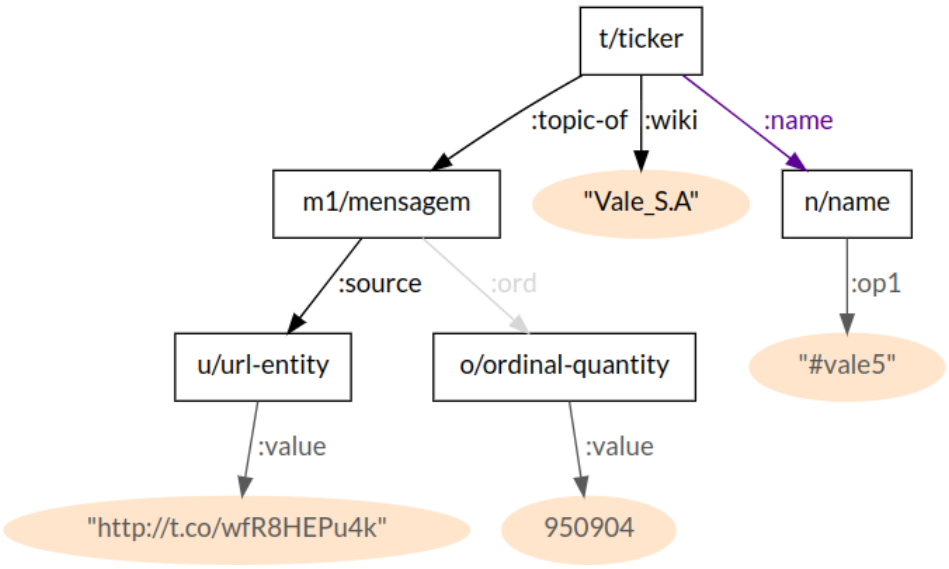
[2]Mesmo que "950904" não seja um numeral ordinal, a opção por representar o conceito a ele subjacente por meio de o/ordinal-quantity pautou-se na interpretação de que parece haver certa ordem na divulgação das mensagens feita pelo site indicado na URL.

[3]Na AMR, a "wikificação" é importante quando um conceito tem diferentes expressões linguísticas, pois a relação :wiki serve para anotar essas expressões com a mesma forma canônica. A "wikificação" é igualmente importante para casos de ambiguidade, como "Santos Dumont" (pessoa/cidade/aeroporto/-rua). Caso não haja página da Wikipedia correspodente à entidade, anota-se :wiki-.

xão dos 4 conceitos anteriormente identificados (m/mensagem, u/url-entity, o/ordinary-quantity, t/ticker) de forma a compor um grafo AMR único. O conceito m/mensagem foi conectado a o/ordinary-quantity via relação :ord e o conceito u/url-entity foi conectado a m/mensagem por meio de :source. Essas duas conexões criaram um subgrafo parcial. Na sequência, seguindo a interpretação na qual t/ticker é o conceito principal, isto é, o nó-topo do grafo, este foi conectado a m/mensagem via :topic-of.

Ao final, a identificação das relações semânticas descritas permitiu construir o grafo AMR para o exemplo (13a) com o auxílio do metAMorphosED, definindo assim uma proposta de anotação para o Padrão 2 (Figura 4.3). Sobre o Padrão 2, vale dizer que a representação AMR se caracteriza por (i) ter um conceito nominal como topo, (ii) não englobar nenhum conceito do tipo *frame* e (iii) apresentar certa correspondência com a anotação-UD no que tange à informação principal.

Figura 4.3: Proposta de anotação AMR para o Padrão 2.



Fonte: O autor, 2025.

A partir da anotação da instância do padrão (como o *tweet* (13a) para o Padrão 2), concebida como proposta de referência, outra instância aleatória do mesmo padrão foi manualmente anotada via metAMorphosED, produzindo assim 2 instâncias de referência por padrão para serem utilizadas na abordagem semiautomática, descrita na sequência. A representação AMR definida para um dos padrões compõe o **Apêndice 1** deste documento.

4.1.1.2.2 Anotação semiautomática

Posteriormente, empregou-se a abordagem semiautomática, baseada em técnicas de *Engenharia de Prompt*. Para tanto, utilizaram-se a versão *web* do ChatGPT o1[4] da OpenAI e a técnica *few-shot prompting*. Tal técnica, como mencionado, consiste em fornecer ao LLM um pequeno número de exemplos para ajudá-lo a entender o contexto e realizar a tarefa específica (Brown et al., 2020). Neste trabalho, a técnica foi empregada para anotar todas as demais instâncias de cada um dos 22 padrões.

A aplicação da técnica *few-shot* foi conduzida da seguinte forma. A partir do padrão (estrutura recorrente) selecionado, forneceram-se como exemplos as 2 instâncias anotadas na etapa manual anteriormente descrita. Na sequência, uma nova instância aleatória do padrão foi submetida ao LLM para a geração da anotação AMR. O resultado do modelo foi então revisado de forma manual com posterior devolutiva se necessário, ou seja, retorno do exemplo corrigido como *feedback* ao LLM. Na devolutiva, explicitaram-se as diretrizes que não foram seguidas corretamente na anotação automática. Essas devolutivas ajudam a garantir que os exemplos estejam corretos e sejam úteis para treinar o modelo de maneira eficaz, facilitando a generalização do aprendizado para novos exemplos. Caso o LLM ainda estivesse cometendo muitos equívocos, uma nova rodada de aprendizado com devolutivas foi feita. Na sequência, as demais instâncias do padrão em questão foram submetidas ao LLM, sob a hipótese de que ele tenha aprendido a tarefa de representação AMR a partir dos exemplos e devolutivas.

Na Figura 4.4, ilustra-se a instância do Padrão 2 anotada pelo LLM após a rodada de aprendizado *few-shot prompt*. Nela, vê-se que a anotação gerada pelo modelo está praticamente correta. A única exceção diz respeito à relação :wiki que caracteriza a representação semântica das entidades nomeadas (como t/ticker), a qual não consta no grafo automático. Diante disso, definiu-se a devolutiva ilustrada na Figura 4.5, fornecida como *feedback* ao LLM juntamente com a instância por ele gerada corrigida.

[4]https://openai.com/o1/

Figura 4.4: Exemplo de anotação AMR problemática gerada pelo LLM para o Padrão 2.

```
(t / ticker
   :topic-of (m1 / mensagem
                 :source (u / url-entity
                            :value "http://t.co/qJauDAJSRq")
                 :ord (o / ordinal-quantity
                         :value 951417))
   :name (n / name
            :op1 "#USIM5"))
```

Fonte: O autor, 2025.

Figura 4.5: Exemplo de devolutiva para a anotação AMR gerado por LLM.

A anotação segundo o modelo AMR está bem boa! Entretanto, toda entidade nomeada, como ticker, possui uma relação :wiki, responsável por conectar o conceito referente ao ticker no grafo AMR (t/ticker) ao nome da empresa correspondente a ele na página da Wikipedia. Observe a versão corrigida da anotação AMR, com a relação :wiki explícita:

```
(t / ticker
  :topic-of (m1 / mensagem
            :source (u / url-entity
                    :value "http://t.co/qJauDAJSRq")
            :ord (o / ordinal-quantity
                 :value 951417))
  :name (n / name
        :op1 "#USIM5")
  :wiki "Usiminas")
```

Fonte: O autor, 2025.

No geral, o modelo não apresentou dificuldades ao lidar com os padrões estruturais mais fragmentados e fixos (como o Padrão 2), "aprendendo" as diretrizes e as aplicando corretamente. A técnica de *few-shot prompting*, combinada com as devolutivas de grafos corrigidos, permitiu que o modelo aprendesse a tarefa de anotação AMR para os referidos padrões mesmo diante de poucos exemplos. Notadamente, os resultados gerados pelo LLM para os padrões que englobam construções efetivamente sintáticas (como sintagmas longos ou mesmo orações) precisaram de mais revisão manual. Em outras palavras, isso quer dizer que, nessas situações, a anotação automática gerada pelo modelo foi menos precisa, indicando que ele não "aprendeu" tão bem, diante de poucos exemplos, a realizar a tarefa quando se trata de *tweets* com maior variabilidade linguística.

4.1.1.3 Fenômenos linguísticos e propostas de diretrizes AMR

Como mencionado, a anotação manual visou, entre outras coisas, à proposição de diretrizes de anotação AMR para fenômenos do português ainda não contemplados na literatura e para os específicos dos *tweets*. Nesta seção, discorre-se sobre os fenômenos observados e analisados durante a anotação manual até então sistematizados, acompanhados pelas diretrizes adotadas para a representação AMR dos mesmos.

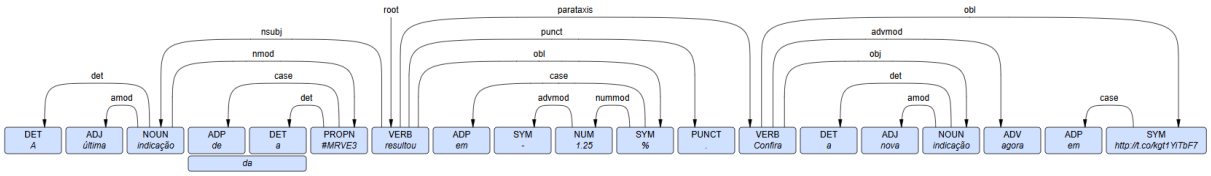
4.1.1.3.1 Fenômenos gerais do português

1. Multiplicidade sentencial

Como mencionado por Di-Felippo, Nunes e Barbosa (2024b), os *posts* que compõem o *corpus* de estudo apresentam mistura de linguagem padrão e não-canônica. Nesse sentido, o Padrão 10 do manual de Di-Felippo, Nunes e Barbosa (2024b) é um dos que apresentam estrutura bem formada, considerando a linguagem padrão como referência. O Padrão 10 é: [**<sentença>. Confira a nova indicação agora em <url>**], exemplificado pelo *tweet* (15), cuja anotação-UD é ilustrada na Figura 4.6.

(15) A última indicação da #MRVE3 resultou em -1.25 %. Confira a nova indicação agora em http://t.co/kgt1YiTbF7

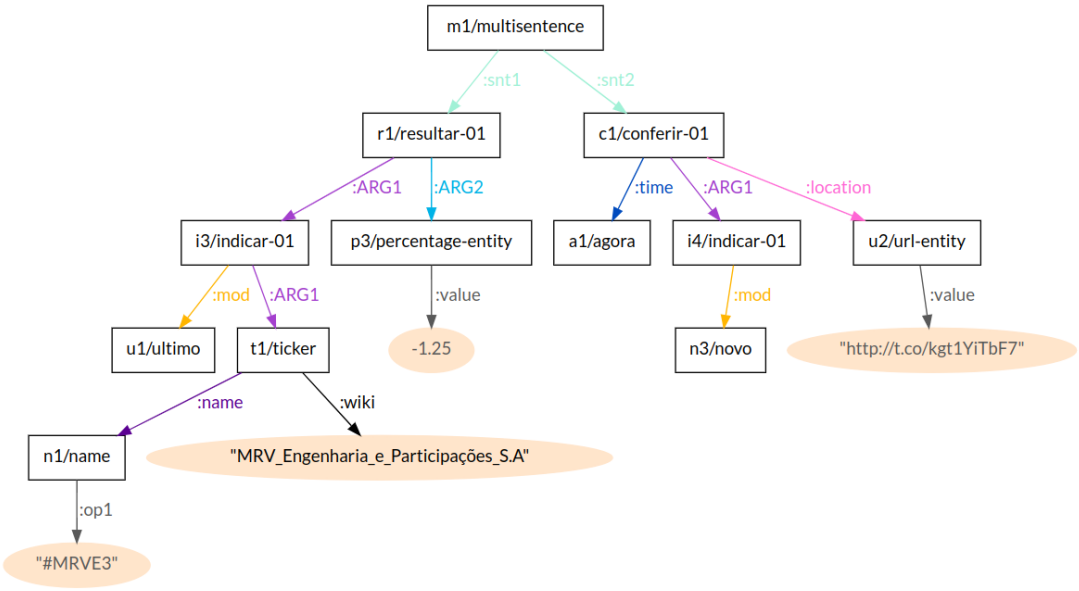
Figura 4.6: Anotação-UD do *tweet* (15).



Fonte: O autor, 2025.

Trata-se de um *tweet* composto por duas sentenças completas e relativamente bem pontuadas. A única exceção é a ausência de pontuação ao final da segunda sentença. Na anotação-UD do DANTEStocks, a multiplicidade de sentenças é capturada pela **parataxis**. No *tweet* em questão, a *parataxis* se estabelece entre os predicadores principais de cada sentença, que, no caso, são verbos. Para representar esse fenômeno na AMR, adotou-se-se a estratégia de OGorman et al. (2018), como ilustrado na Figura 4.7.

Figura 4.7: Proposta de representação AMR para multissentenças (*tweet* (15)).



Fonte: O autor, 2025.

O construto m/multisentence foi originalmente definido para tratar a correferência na anotação AMR em *corpus* multissentencial. Aliás, a esse respeito, salienta-se que há um caso de correferência por repetição no *tweet* (15), pois em ambas as sentenças o nome "indicação" se refere ao mesmo conceito. Ao adotar essa estratégia, a representação AMR de um *tweet* como o da Figura 4.7 é composta por 2 subgrafos ligados ao nó-topo m/multisentence por :sntN (no caso, :snt1 e snt:2). Como as sentenças que compõem o *tweet*-exemplo têm predicados verbais, a raiz de cada subgrafo é representada por um *frameset*, isto é, resultar-01 e conferir-01, ambos herdados do Verbo-Brasil.

O mapeamento específico dos elementos textuais para eventos e seus argumentos pode contar com algumas pistas sintáticas. Quanto ao NOUN "indicação", por exemplo, destaca-se que, sendo um nome predicador, a diretriz da AMR é a representá-lo como evento, no caso, indicar-01, segundo o Verbo-Brasil. Os nomes predicadores comumente projetam um complemento sintático na forma de um sintagma preposicionado. Na anotação-UD, o complemento é dependente do nome por **nmod**. Expressando "aquilo que é indicado", o **nmod** "#MRVE3" da primeira sentença é o :Arg1 do evento correspondente i3/indicar.01, sendo codificado na anotação AMR pela entidade t/ticker. A ocorrência de i4/indicar-01 na segunda sentença não possui nenhum ArgN expresso. O ADJ "nova", mapeado para n/novo, conecta-se ao evento por :mod.

Para a representação do evento r1/resultar-01, as *deprels* **nsubj** e **obl** foram usadas como pistas para mapear o evento i3/indicar-01 como :Arg0 (“origem”) e o conceito p3/porcentagem como :Arg2 (“resultado”), respectivamente.

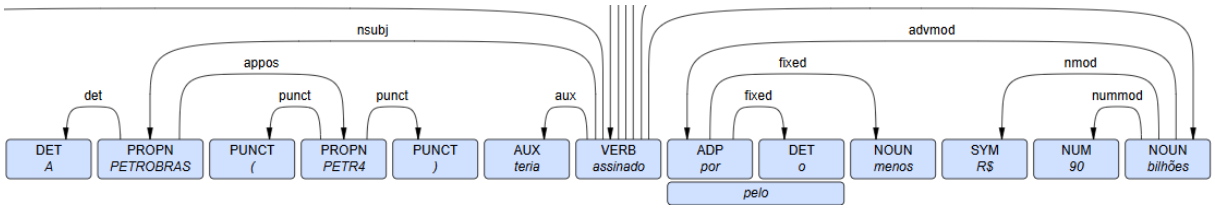
Quanto ao evento conferir-01, destaca-se que, com exceção de **obl** entre o predicador e “indicação”, que foi mapeada para um ArgN (no caso, Arg1), as demais pistas sintáticas levaram à identificação de relações *non-core*. Mais precisamente, **advmod** entre “conferir” e ADV “agora” resultou em :time e **obl** entre “conferir” e a URL permitiu representar :location.

2. Locução adverbial pelo menos

A Figura 4.8 exibe o trecho com anotação-UD do *tweet* (16) em que há a ocorrência de “pelo menos”. Nela, os elementos constitutivos da locução “pelo menos” foram relacionados entre si por *fixed*, indicando que não há relação sintática entre eles, mas funcionam como um conjunto. A função sintática da locução em (16) é de **advmod** (modificador adverbial), especificamente do nome “bilhões”.

(16) Notas gerais A PETROBRAS (PETR4) teria assinado **pelo menos** R$ 90 bilhões em contratos sem licitação desde 2011... http://t.co/u9Bft3wAw6

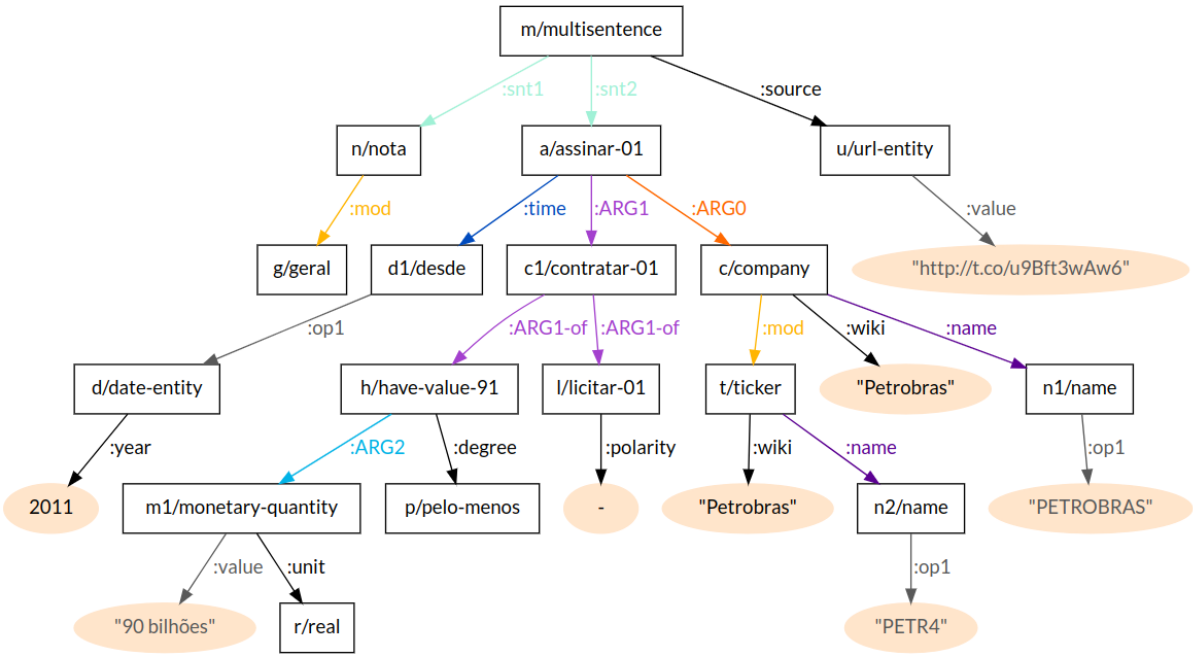
Figura 4.8: Anotação-UD da locução “pelo menos” no *tweet* (16).



Fonte: O autor, 2025.

Dito isso, a proposta de representação AMR para a locução está ilustrada na Figura 4.9. Ela consiste em representar esse conceito de quantidade por meio de uma abstração resultante da hifenização das palavras da expressão (“pelo-menos”).

Figura 4.9: Proposta de representação AMR para a locução “pelo menos” (*tweet* (16)).



Fonte: O autor, 2025.

3. Expressão “por (NOUN) ADJ nisto”

As expressões multipalavra “longo (prazo)” e “curto prazo” no *tweet* (17) (Padrão 3) se referem a gráficos. A ênfase do usuário por meio da expressão “poe longo nisto”, pode significar, por exemplo, que O gráfico mostra (i) uma tendência clara de alta ou baixa no longo prazo ou (ii) uma projeção baseada em um período muito grande, como gráficos semanais ou mensais. Embora seja a única ocorrência do *corpus*, expressões do tipo “por (NOUN) ADJ nisto/nisso” (como “poe (prazo) longo nisto”, “poe (anel) caro nisso”, etc.) são comuns na língua geral, sobretudo em contexto de informalidade.

(17) #PETR4 - longo (**poe longo nisto**) e curto prazo (mensagem: 955011) http://t.co/cIh5ZtcFko

A interpretação dada à expressão “poe longo nisto” foi a de “muito longo” ou “longo demais”, cujo equivalente em inglês seria “*too long*”. Essa interpretação levou ao emprego da mesma estratégia representacional dada pela AMR aos superlativos, que consiste em empregar o *frame* have-degree-91[5] (Figura 4.10) como o conceito raiz e não o adjetivo intensificado (“longo”), ou a entidade caracterizada por esse adjetivo (“prazo”).

[5]Por se tratar de um conceito específico da AMR, empregou-se o rótulo original em inglês.

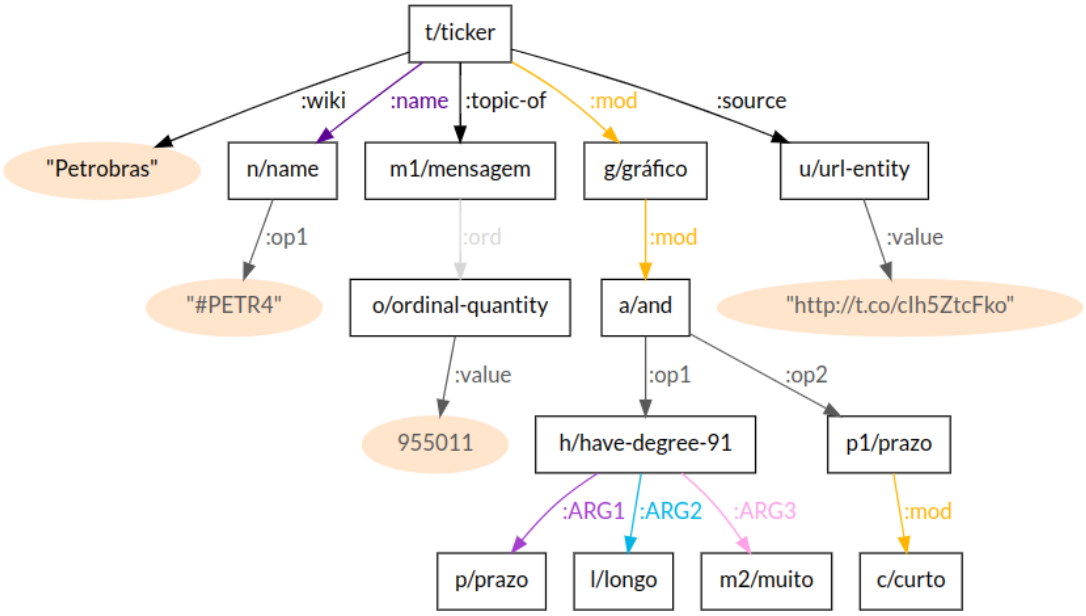
Figura 4.10: *Frameset* do conceito *have-degree-91*.

```
Have-degree-91
Arg1: domain, entity characterized by attribute (e.g. girl)
Arg2: attribute (e.g. tall)
Arg3: degree itself (e.g. more, less, equal, most, least, enough, too, so, to-the-point, at-least, times)
Arg4: compared-to (e.g. (than the) BOY)
Arg5: superlative: reference to superset
Arg6: reference, threshold of sufficiency (e.g. (tall enough) TO RIDE THE ROLLERCOASTER)
```

Fonte: https://github.com/amrisi/amr-guidelines/blob/master/amr.md.

A diretriz para "poe longo nisto" envolveu duas decisões: (i) "alucinar o conceito "muito, pois a interpretação foi a de "(prazo) muito longo, e (ii) utilizar o conceito reificado *have-degree-91*, relacionando-o aos conceitos "prazo, "longo e "muito pelas relações de ARG1 (entidade caracterizada pelo atributo), ARG2 (atributo) e ARG3 ("grau), respectivamente. A proposta de anotação AMR para o *tweet* (17) está na Figura 4.11.

Figura 4.11: Proposta de anotação da locução "poe longo nisto" em AMR.



Fonte: O autor, 2025.

Na representação AMR do *tweet* inteiro, o ramo do grafo formado pela coordenação da expressão em questão e "curto prazo" (isto é, o nó a/and) se relaciona ao conceito alucinado g/gráfico por :mod e este se conecta a t/ticker também por :mod.Isso busca traduzir em AMR que as duas perceptivas temporais dizem respeito ao gráfico de análise da ação Petr4.

4.1.1.3.2 Fenômenos CGU No DANTEStocks

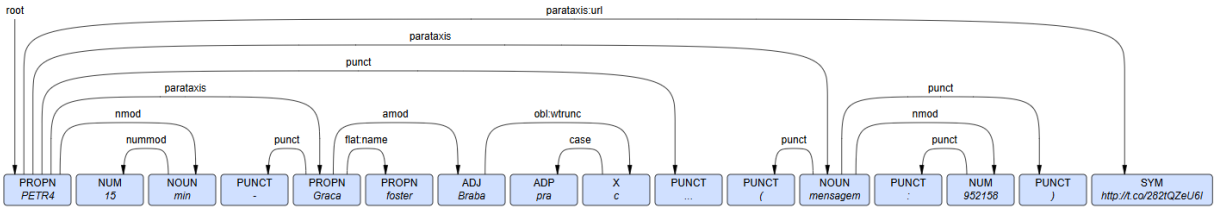
1. Multiplicidade de segmentos/fragmentos

Se, por um lado, o DANTEStocks apresenta *tweets* compostos por múltiplas sentenças bem delimitadas (Padrão 10, exemplo (15)), por outro lado, ele apresenta postagens que se caracterizam pela justaposição de segmentos (não necessariamente sentenças) sem relação de conteúdo aparente entre eles. O exemplo (18) ilustrada esse fenômeno. Trata-se em particular de uma das instância do Padrão 4, formalizado como: **[{<hashtag-ticker><complemento>} <sentença> (mensagem : número) <url>]**. As chaves aqui indicam que [<hashtag-ticker><complemento>] é opcional. Em (18), esse segmento está presente no *tweet*.

(18) PETR4 15 min - Graca foster Braba pra c ... (mensagem : 952158) http://t.co/282tQZeU6I

Com base na árvore de dependência-UD (Figura 4.12), observa-se que há 3 *deprels* **parataxis**, indicando a existência de 4 segmentos justapostos na anotação sintática.

Figura 4.12: Anotação-UD do *tweet* (18).

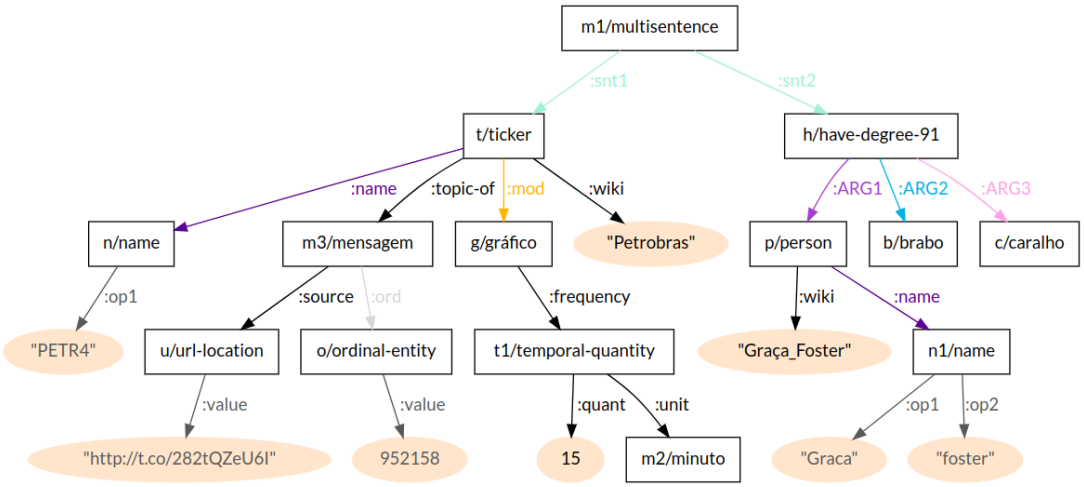


Fonte: O autor, 2025.

O Padrão 4 é composto exatamente pelo Padrão 3 (17), acrescido de <sentença>. O grafo AMR referente ao exemplo (17) (Figura 4.11 do Padrão 3 é monossentencial, tendo topo representado por t/ticker. Isso indica que, na representação semântica AMR, nem todas as **parataxis** indicaram a necessidade de criação de subgrafos. No exemplo (18), a sentença "Graca Foster Braba c..." parece não ter relação direta com a mensagem veiculada pelo *tweet*, sendo um comentário do usuário sobre um elemento externo ao conteúdo da mensagem. Assim, interpreta-se que o grafo AMR de (18) tem dois subgrafos, um deles encabeçado por t/ticker, da mesma forma que no Padrão 3, e outro referente à <sentença>. Nesse caso, emprega-se também o construto m/multisentence como nó-raiz

ao qual os subgrafos se ligam via :snt1 e :snt2, mesmo que o subgrafo que tem t/ticker como topo não represente de fato uma sentença. A Figura 4.13 apresenta a proposta de anotação AMR para *tweets* como (18), com múltiplos segmentos.

Figura 4.13: Proposta de representação AMR para múltiplos segmentos (*tweet* (18)).



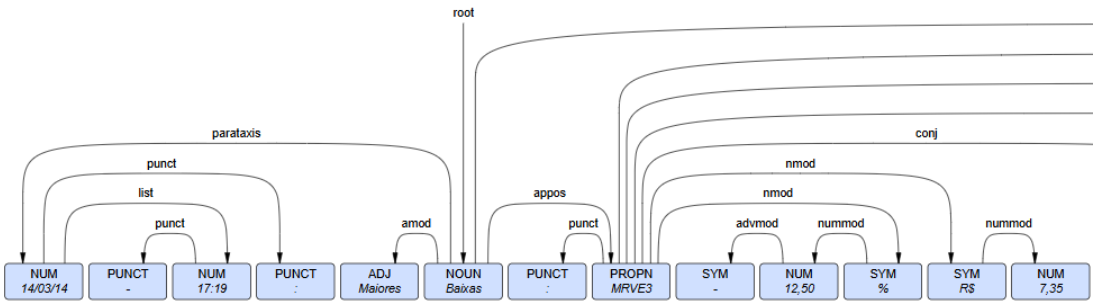
Fonte: O autor, 2025.

2. Segmentos complexos

O *tweet* (19) possui uma lista de segmentos complexos justapostos. Diz-se "complexo" porque cada segmento da lista é composto pela sequência "*ticker* + porcentagem + valor-monetário". Esse tipo de lista é frequente no *corpus* e caracteriza o Padrão 17: **<lista ticker&porcentagem&valor-monetário>**. O exemplo (19) é uma instância do padrão, cuja anotação-UD referente ao primeiro segmento complexo da lista está na Figura 4.14.

(19) 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 4.14: Anotação-UD parcial do *tweet* (19).



Fonte: O autor, 2025.

Antes, porém, de demonstrar a representação semântica do segmento, destaca-se que a proposta de anotação AMR para o *tweet*-exemplo inteiro inclui apenas 1 subgrafo, embora haja uma **parataxis**. O segmento ligado por **parataxis** ao *root* relaciona-se semanticamente ao nó-topo b/baixa por :time, uma vez que todo o trecho "14/03/14 17:19" expressa um conceito de tempo. Isso pode ser observado na Figura 4.14, que exibe a notação PENMAN da proposta de representação AMR para o exemplo. O formato PENMAN foi escolhido em detrimento do grafo por questão de limitação de espaço, uma vez a representação em grafo de (19) é grande.

Figura 4.15: Proposta de anotação AMR para "lista de segmentos complexos" (*tweet* (19)).

```
(b / baixa
   :time (d / date-entity
            :year 2014
            :month 03
            :day 14
            :time "17:19")
   :degree (m1 / maior)
   :mod (a1 / and
            :op1 (t1 / ticker
                     :name (n1 / name
                               :op1 "MRVE3")
                     :wiki "MRV_Engenharia_e_Participações_S.A"
                     :value (p1 / percentage-entity
                                :value -12.5)
                     :ARG1-of (h1 / have-quant-91
                                  :ARG2 (m2 / monetary-quantity
                                            :value "7,35"
                                            :unit (r / real))))
            :op2 (t2 / ticker
                     :name (n2 / name
                               :op1 "DASA3")
                     :wiki "Diagnósticos_da_America_S.A."
                     :value (p2 / percentage-entity
                                :value -9.67)
                     :ARG1-of (h2 / have-quant-91
                                  :ARG2 (m3 / monetary-quantity
                                            :value "15,13"
                                            :unit r)))
            :op3 (t3 / ticker
                     :name (n3 / name
                               :op1 "CMIG4")
                     :wiki "Companhia_Energética_de_Minas_Gerais"
                     :value (p3 / percentage-entity
                                :value -5.69)
                     :ARG1-of (h3 / have-quant-91
                                  :ARG2 (m4 / monetary-quantity
                                            :value "12,94"
                                            :unit r)))
            :op4 (t4 / ticker
                     :name (n4 / name
                               :op1 "GFSA3")
                     :wiki "Gafisa_S.A."
                     :value (p4 / percentage-entity
                                :value -4.76)
                     :ARG1-of (h4 / have-quant-91
                                  :ARG2 (m5 / monetary-quantity
                                            :value "3,00"
                                            :unit r)))
            :op5 (t5 / ticker
                     :name (n5 / name
                               :op1 "ELPL4")
                     :wiki "Eletropaulo_Metropolitana_Eletricidade_de_São_Paulo_S.A."
                     :value (p5 / percentage-entity
                                :value -4.03)
                     :ARG1-of (h5 / have-quant-91
                                  :ARG2 (m6 / monetary-quantity
                                            :value "7,62"
                                            :unit r)))))
```

Fonte: O autor, 2025.

Ressalta-se também que, na proposta, os segmentos em lista estão coordenados (somente por vírgulas), sendo a lista, portanto, representada pelo conceito a1/and e as relações :opN. O conceito a1/and, por sua vez, está relacionado ao conceito b/baixa "(maiores) baixas" pela relação :mod, indicando que os vários :opN são pertencentes às maiores baixas (concebidas como "domínio").

Dito isso, a proposta de representação AMR (Figura 4.15) do segmento compleco em questão tem o *ticker* (PROPN) como conceito principal (t1/ticker) da sequência e, sendo uma entidade nomeada, possui as relações :name e :wiki. A porcentagem é representada como um tipo especial de entidade e se liga ao conceito t1/ticker pela relação :value. Especificamente, o conceito p1/percentagem-entity possui a relação :value, que o liga ao conceito "-12,5". O valor monetário, por sua vez, foi representado pela reificação have-quant-91, que se liga ao conceito t1/ticker por ARG1-of. O valor monetário especificamente é tratado como uma entidade especial do tipo monetary-quantity, que possui as relações :value e :unit, responsáveis por ligá-lo aos conceitos "7,35" e "real", respectivamente.

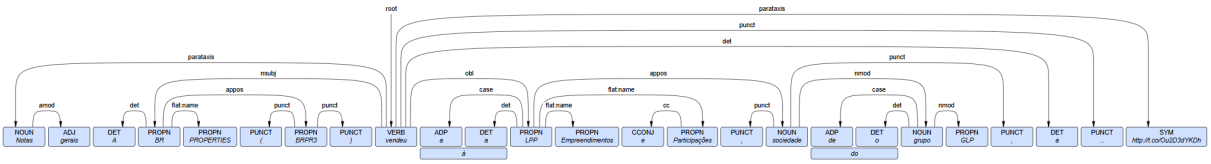
3. Truncamento estrutural e lexical

Os truncamentos são, segundo a revisão da literatura, um fenômeno linguístico típico dos gêneros CGU, que ocorrem nos *tweets* pela limitação de caracteres imposta pela plataforma. Eles ocorrem no final dos *posts*, sendo indicados pelas reticências. O truncamento pode ocorrer no nível lexical ou estrutural, ou seja, pode ser uma de quebra na última palavra ou *token*, gerando ou não problema estrutural que interfere na interpretação da postagem, ou quebra na estrutura sintática. O exemplo (20) contém um caso de truncamento estrutural. Trata-se uma das instâncias do Padrão 1: **notas gerais <sentença-truncada> ... <url>**. A anotação-UD de (20) está na Figura 4.16.

(20) Notas gerais A BR PROPERTIES (BRPR3) vendeu a a LPP Empreendimentos e Participações , sociedade de o grupo GLP , **a ...** http://t.co/Ou2D3dYKDh .

Com base na árvore de dependência, vê-se que o *tweet* possui 3 segmentos, o que é evidenciado pela ocorrência de 2 **parataxis**, uma à esquerda e outra à direito do **root**. O segmento que contém o **root** está claramente quebrado devido às reticências ao seu final.

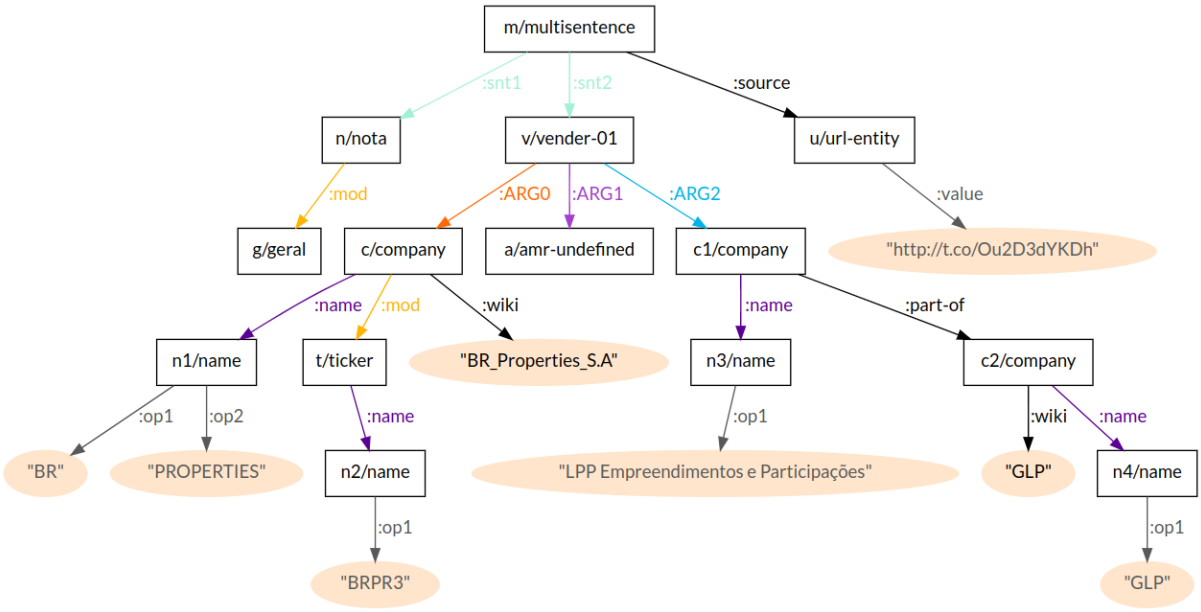
Figura 4.16: Anotação-UD do *tweet* em (20).



Fonte: O autor, 2025.

Na representação AMR proposta na Figura 4.17, há apenas 2 segmentos conectados ao nó-raiz (m/multisentence) por :sntN. O primeiro tem n/nota como conceito central, e o segundo, v/vender-01. O conceito subjacente à URL está conectado ao topo da representação AMR (m/multisentence) pela relação semântica *non-core* :source.

Figura 4.17: Proposta de representação AMR para truncamento estrutural (*tweet* (21)).



Fonte: O autor, 2025.

O predicador principal do segundo segmento (sentencial) é v/vender-01, que é o nó-raiz do subgrafo AMR desse segmento. O *frame* vender-01 no Verbo-Brasil prevê 4 ArgNs: Arg0 (vendedor), Arg1 (coisa vendida), Arg2 (comprador) e Arg3 (preço). Desses 4 argumentos, 2 estão explícitos: Arg0 (“BR PROPERTIES”) e Arg2 (“LPP Empreendimentos e Participações (sociedade de o grupo GLP)”). O Arg1, por sua vez. não está completo devido ao truncamento. Mais precisamente, o núcleo do sintagma nominal que realiza o Arg1 “coisa vendida” não está efetivamente explícito, mas apenas o determinante (DET) que o constitui (“a”) (cf. Figura 4.16). Em outras palavras, tem-se no *tweet*-exemplo um truncamento que incide sobre a estrutura argumental do verbo “vendeu”.

Para representar isso na anotação AMR, empregou-se o conceito *amr-undefined*, que foi originalmente proposto para ser atribuído a casos em que a representação semântica do conteúdo de uma sentença não é totalmente definida ou não pode ser expressa diretamente através das relações e conceitos conhecidos da AMR. Assim concebido, a impossibilidade de mapear uma expressão (isto é, o núcleo do sintagma) para a estrutura semântica da AMR (mais precisamente, para o :Arg1) devido ao truncamento parece ser um caso tratável via conceito amr-undefined.

Por conseguinte, o conceito v/vender-01 está relacionado via :Arg1 à abstração amr-undefined. Como a pista sintática sobre a realização do Arg1, o DET, é suprimida na AMR, amr-undefined indica que o argumento não está efetivamente expresso por causa do truncamento, mas há rastro na superfície do *tweet* que dica seu preenchimento pelo verbo. A proposta de anotação AMR para esse fenômeno está ilustrada na Figura 4.17.

Os conceitos subjacentes a truncamentos lexicais são expandidos e representados pela sua forma estendida. As formas completas de palavras truncadas são recuperadas da coluna MISC no arquivo CoNLL-U. Nessa coluna, a *feature* FullForm de um *token* registra a forma canônica do *token* como seu valor. Esse é o caso, por exemplo, de “reser” do *tweet* (21), cuja forma estendida é “reservatório” no CoNLL-U (cf. Figura 4.18).

(21) ‘ 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

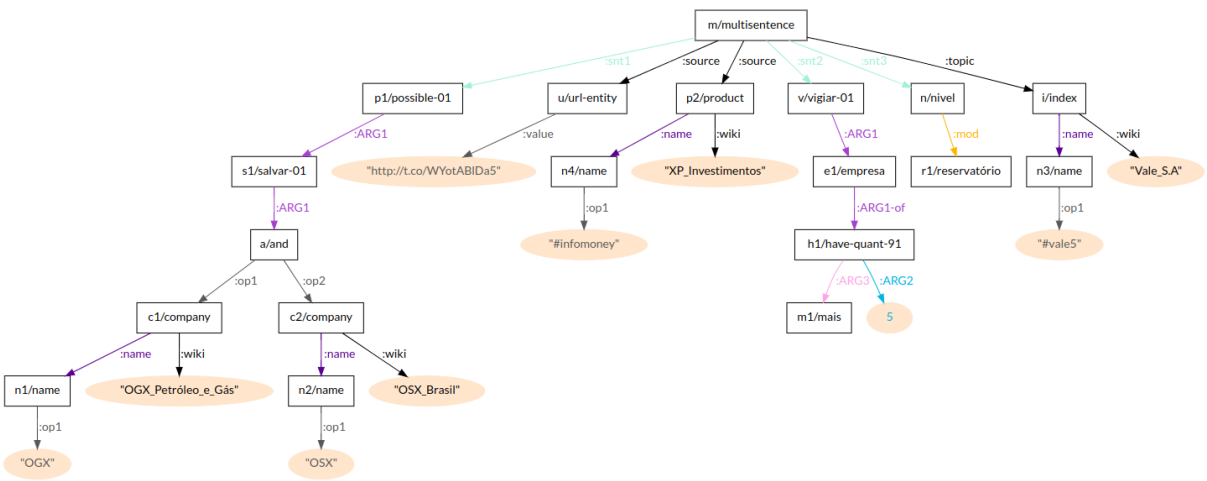
Até o momento, a recuperação da forma estendida com base no CoNLL-U foi possível. Essa decisão, aliás, vai ao encontro da diretriz geral da AMR que é a de expandir as abreviações de nomes comuns. Os truncamentos lexicais, ao geram uma espécie de forma reduzida da palavra, encaixam-se nessa regra. A representação em AMR de um *tweet* com truncamento lexical é ilustrada pela Figura 4.19.

Figura 4.18: *Feature* FullForm preenchida na coluna MISC do ConLL-U.

# sent_id = dante_01_441233076748681216l									
# text = 'Salvação' da OGX e OSX pode estar à caminho, mais 5 empresas estão no radar: Nível do reser... http://t.co/WYotABlDa5 #infomoney #vale5									
1	'	'	PUNCT	_	_	2	punct	_	SpaceAfter=No
2	Salvação	salvação	NOUN	_	Gender=Fem|Number=Sing	9	nsubj	_	SpaceAfter=No
3	'	'	PUNCT	_	_	2	punct	_	_
45781	da	_	_	_	_	_	_	_	_
4	de	de	ADP	_	_	6	case	_	_
5	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	6	det	_	_
6	OGX	OGX	PROPN	_	_	2	nmod	_	_
7	e	e	CCONJ	_	_	8	cc	_	_
8	OSX	OSX	PROPN	_	_	6	conj	_	_
9	pode	poder	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	0	root	_	_
10	estar	estar	AUX	_	VerbForm=Inf	13	cop	_	_
11-12	à	_	_	_	_	_	_	_	_
11	a	a	ADP	_	_	13	case	_	_
12	a	o	DET	_	Definite=Def|Gender=Fem|Number=Sing|PronType=Art	13	det	_	_
13	caminho	caminho	NOUN	_	Gender=Masc|Number=Sing	9	xcomp	_	SpaceAfter=No
14	,	,	PUNCT	_	_	21	punct	_	_
15	mais	mais	ADV	_	_	16	advmod	_	_
16	5	5	NUM	_	NumType=Card	17	nummod	_	_
17	empresas	empresa	NOUN	_	Gender=Fem|Number=Plur	21	nsubj	_	_
18	estão	estar	AUX	_	Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin	21	cop	_	_
19-20	no	_	_	_	_	_	_	_	_
19	em	em	ADP	_	_	21	case	_	_
20	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	21	det	_	_
21	radar	radar	NOUN	_	Gender=Masc|Number=Sing	9	conj	_	SpaceAfter=No
22	:	:	PUNCT	_	_	23	punct	_	_
23	Nível	nível	NOUN	_	Gender=Masc|Number=Sing	9	parataxis:strunc	_	_
24-25	do	_	_	_	_	_	_	_	_
24	de	de	ADP	_	_	26	case	_	_
25	o	o	DET	_	Definite=Def|Gender=Masc|Number=Sing|PronType=Art	26	det	_	_
26	reser	reservatório	NOUN	_	Gender=Masc|Number=Sing	23	nmod:wtrunc	_	FullForm=reservatório|Trunc=Yes|SpaceAfter=No
27	...	...	PUNCT	_	_	9	punct	_	_
28	http://t.co/WYotABlDa5	http://t.co/WYotABlDa5	SYM	_	_	9	parataxis:url	_	_
29	#infomoney	#infomoney	X	_	_	9	parataxis:hashtag	_	CorrectSpaceAfter=Yes
30	#vale5	#vale5	X	_	_	9	parataxis:hashtag	_	SpaceAfter=No

Fonte: O autor, 2025.

Figura 4.19: Proposta de representação AMR para truncamento lexical (*tweet* (21)).



Fonte: O autor, 2025.

4. Variações (léxico-ortográficas) da norma padrão

Como a taxonomia de Scandarolli et al. (2023) evidencia, os *tweets* são marcados por vários fenômenos que se caracterizam por serem variações lexicais e gráficas da norma padrão de escrita. Segundo a AMR, os desvios ortográficos (ou *typos*) que ocorrem em palavras das classes abertas devem ser normalizados para o inglês americano, mas variações dialetais não são. Neste trabalho, seguiu-se a mesma diretriz. Por conseguinte, a representação dos conceitos subjacentes a formas de superfície que apresentavam um ou mais fenômenos (substituição/omissão/inserção/transposição de caractere) da variação da norma padrão de Scandarolli et al. (2023) demandou a normalização do *token* para o português padrão no grafo AMR. Esse é o caso de "Gráfi" e "diári" em (22), os quais têm variação por omissão de vogal final. Para representar o conceito, a forma de superfície "Gráfi" e foi normalizada para o português padrão ("gráfico") enquanto a "diári" foi representado por meio da relação unit ligada ao conceito temporal-quantity.

(22) Rastreamento ações-**Gráfic diári**-12h. Analise se romper: BBPO11 106,01 CSAN3 37,64 IGTA3 22,85 LREN3 69,02 MPLU3 33,42 RENT3 36,79 VLID3 36,65

Figura 4.20: Tratamento de desvios (*types*) da norma padrão na AMR (*tweet* (22)).

```
:snt1 (r / rastrear-01
        :ARG1 (a / ação)
        :mod (g / gráfico
                :frequency (t1 / temporal-quantity
                                :unit (d1 / dia)
                                :quant 1)
                :time (t2 / temporal-quantity
                          :unit (h1 / hora)
                          :quant 12)))
```

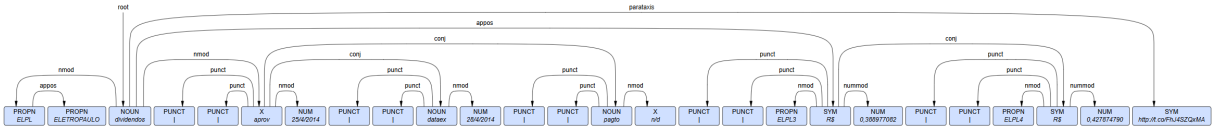
Fonte: O autor, 2025.

5. Fenômenos lexicais da norma inovadora

Scandarolli et al. (2023) também mostraram que o DANTEStocks apresenta vários fenômenos que expressam o caráter inovador/criativo da linguagem dos gêneros CGU. Entre ele estão, por exemplo, abreviações de diferentes tipos. Por se tratar do domínio do mercado financeiro, em que os nomes predicadores ocupam lugar de destaque (Voskaki et al., Barbosa, 2024), muitas dessas abreviações são de nomes predicadores verbais ou reduções de expressões nominais. Algums exemplos de abreviações são "aprov/apro (19 ocorrências), "dataex (15 ocorrências), "pagto/pg (13 ocorrências) e n/d (13 ocorrências). O *tweet* (23) apresenta ocorrências de todas. A anotação-UD correspondente a ele está na Figura 4.21.

(23) 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

Figura 4.21: Anotação-UD do *tweet* (20).



Fonte: O autor, 2025.

Tomando como base apenas as que ocorrem em (23), destaca-se que “aprov e “pagto são formas reduzidas por encurtamento de uma única palavra. Sobre “aprov, não se sabe ao certo qual é a sua forma estendida ou canônica, uma vez que não foi possível recuperar nenhuma ocorrência da plataforma Twitter ou da web que esclarecesse essa questão. A opção mais provável, segundo especialistas de domínio, é que seja “aprovação (ou “aprovado). “Pagto, por sua vez, é claramente abreviação de “pagamento.

“Dataex é a redução ou abreviação da expressão multipalavra “data ex-dividendo. O processo linguístico que formou “dataex a partir de “data ex-dividendo é um tipo de abreviação chamada concatenação ou combinação de palavras. Mais precisamente, o termo complexo “data ex-dividendo foi encurtado pela eliminação de “dividendo e combinação de “data e “ex, sem a adição de hífen. Já “n/d é uma redução de “não divulgado.

Tendo em visa que no processo de mapeamento entre palavras e conceitos as quatro abreviações do tweet foram mapeadas a conceitos que precisam constar no grafo, foi necessário estabelecer diretrizes para isso.

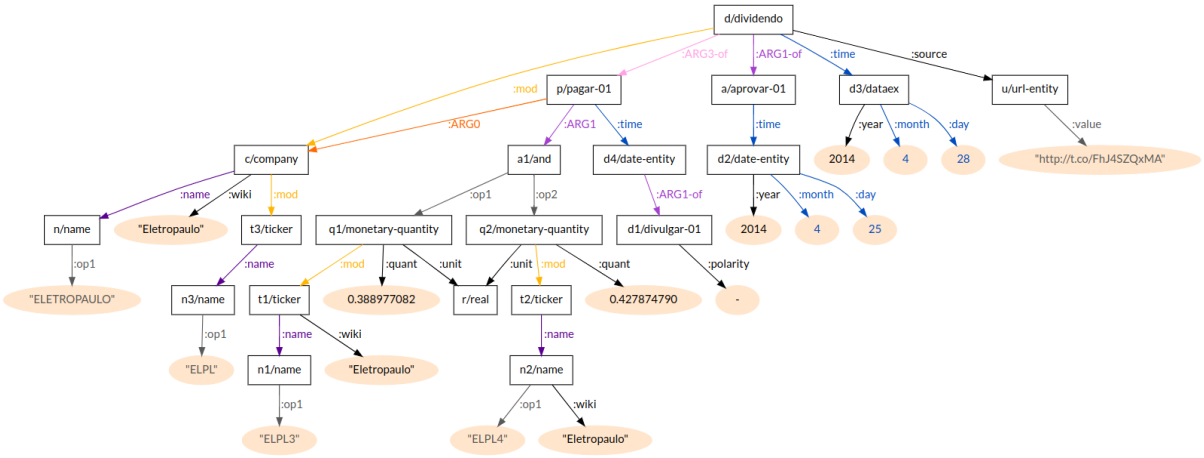
A diretriz geral da AMR sobre abreviações de nomes comuns é a de sempre empregar a forma expandida, buscando mapear diferentes expressões linguísticas ao mesmo conceito. Assim, a forma “aprov (e “aprot́) (seja redução de “aprovação ou mesmo de “aprovado) foi expandida e representada como evento e, por isso, por meio de um *frame*, isto é, “aprovar-01 do Verbo-Brasil.

As abreviações, como dito, nem sempre são versões curtas de uma úni ca palavra. Embora seja uma abreviação, “dataex, por exemplo, é uma redução de “data ex-dividendot́t́, que significa “data em que as ações começam a ser negociadas sem direito ao dividendo. Segundo especialistas de domínio, o termo “dataex é extremamente popular

e, assim, acredita-se que ele possa ser concebido como uma “lexicalização desse conceito complexo. Dessa forma, definiu-se que a própria abreviação seria usada como rótulo para a representação do conceito. Na Figura 4.22, esse conceito é d3/dataex. Para expressar que se trata de um conceito de tempo, aliás, a proposta de anotação AMR do exemplo (23) inclui a relação :time, conectando d3/dataex ao conceito d/dividendo.

As formas abreviadas “pagto e “n/d, em particular, coocorrem formando uma espécie de expressão fixa composicional (“pagamento não divulgado), cujo significado, por isso, pode ser decomposto segundo a diretrizes da AMR. Para tanto, o nome “pagto foi expandido e representado pelo *frame* “pagar-01, assim como a forma “n/d foi codificada pelo evento/*frame* divulgar-01; ambos provenientes do Verbo-Brasil. A forma “n/d, por envolver uma negação, foi representada da seguinte forma: (i) emprego do *frame* “pagar-01 relacionado via :time a uma entidade do tipo data (d4/date-entity), que é ARG1-of do *frame* d1/divulgar-01, (ii) representação da negação que incide sobre esse ARG1-of com o construto (:polarity -).

Figura 4.22: Proposta de tratamento AMR para abreviações no geral (*tweet* (23)).



Fonte: O autor, 2025.

As decisões sobre o mapeamento entre as abreviações e os conceitos buscam também garantir uma macro estruturação coerente do grafo AMR. Para ilustrar, ressalta-se que o *tweet* (23) é uma das instâncias do Padrão 5, definido como: **<ação&cia> dividendos <lista info&data> <lista ticker&valor> || <url>**. Nele, a <lista info&data> indica três informações sobre os dividendos veiculadas no *tweet* (aprov, dataex e pagto), sendo cada uma delas especificada de certa forma por uma informação de data.

No grafo, essas informações, codificadas nos conceitos a/aprovar-01, d3/dataex e p/pagar-01, ligam-se ao conceito d/dividendo pelas relações :ARG1-of, :ARG3-of e :*time*, respectivamente. E cada um dos conceitos, a depender do seu tipo semântico (*frame* ou conceito lexicalizado), possui a informação de data relacionada, o que parece garantir uma representação coerente do conteúdo do *tweet*/padrão. Por fim, destaca-se que a url, dependente por **parataxis** do **root** "dividendo na anotação UD, é :source na representação AMR. Os conceitos subjacentes ao *ticker* ELPL e à companhia ELETROPAULO são representados por dois tipos de entidades nomeadas e conectados a d/dividendo por :mod. Assim, tem-se a proposta de tratamento AMR para o Padrão 5 (Figura 4.22).

Além das abreviações por concatenação, os *tweets* até então anotados também apresentam o tipo de abreviação chamado inicialismo, ou seja, forma normalmente nominal composta pelas letras iniciais das palavras de uma expressão. Esse é o caso, por exemplo de "jscp em (24), que reduz o sintagma "juro sobre capital próprio.

(24) ELETROBRAS **jscp** | aprov 30/4/2014 | ex 2/5/2014 | | pg n/d | ELET3 R$0,39921083663 | ELET5 R$2,17825658673 | ELET6 R$1,63369244005 http://t.co/IjtHKlQlfr

Embora a diretriz geral da AMR seja a de compor o significado sempre que possível, não há uma diretriz específica para inicialismos. Com base na prática geral de anotação semântica e na flexibilidade do modelo AMR, por um lado, e no amplo uso desses inicialismos (representando claramente o conceito da expressão ou sintagma completo), do outro lado, optou-se por não o expandir, mantendo o inicialismo como nó semântico (cf. Figura 4.23).

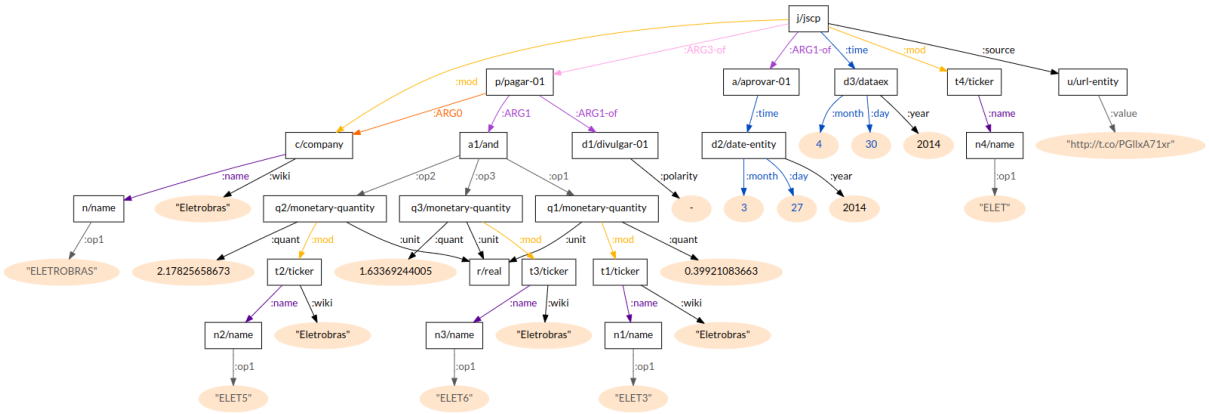
6. Fenômenos da plataforma: RT, menção e URL

Entre os fenômenos típicos da plataforma *Twitter*, destacam-se aqui três deles: marca de *retweet* (RT), menção a perfis (*at-mention*) e url. O *tweet* em (25) possui os três, sendo a URL um caso de truncamento lexical que não afeta a representação AMR.

(25) **RT @dividendo_br** : ELETROBRAS jscp | aprov 30/4/2014 | ex 2/5/2014 | | pg n/d | ELET3 R$0,39921083663 | ELET5 R$2,17825658673 | ELET6 R$1,63369244005 **htt**

A marca de *retweet* (RT) e a menção sempre ocorrem de forma justaposta, com RT precedendo a menção, em uma estrutura fixa que da plataforma para registrar que se trata de

Figura 4.23: Proposta de tratamento AMR para inicialismos (*tweet* (24)).



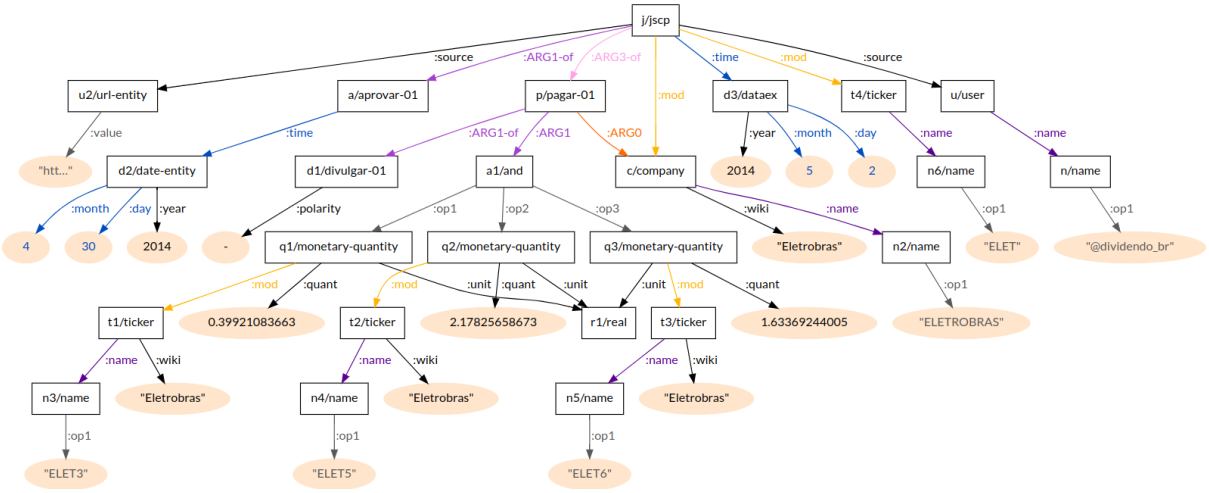
Fonte: O autor, 2025.

uma repostagem de uma mensagem originalmente veiculada por outro perfil. A marca de RT em si é apenas uma função da plataforma para a republicação, sendo a fonte da mensagem original indicada de fato pela menção. Essa marca, aliás, deixou de ser necessária em 2015, quando a plataforma introduziu a funcionalidade de *retweet* sem necessidade de adicionar a marca RT manualmente. Não compondo de fato a mensagem/conteúdo veiculado pelos *tweets*, optou-se por suprimir o RT da representação AMR.

A menção, em particular, foi concebida como um tipo específico de entidade nomeada (u/user), relacionada ao nó-topo por meio de :source. A entidade u/user ser :source do nó-topo justifica-se porque a menção se refere ao conteúdo completo do *tweet* inteiro. Quando se trata de uma entidade com página da Wikipedia, como @JornalOGlobo, o conceito u/user tem :wiki preenchido. A massiva maioria dos usuários da plataforma, no entanto, não têm página correspondente da Wikipédia e, por isso, anota-se u/user com :wiki -. As URLs foram caracterizadas como entidades (u/url-entity), cujo papel semântico :value é desempenhado pelo próprio endereço eletrônico; elas são conectadas ao conceito-topo por :source.

Observa-se que nesse caso os conceitos u/url-entity e u/user são ambos relacionados ao nó-topo do grafo por :source. Isso parece ser coerente, pois a URL e a menção são a fonte primária e secundária, respectivamente, do conteúdo da mensagem/postagem. A Figura 4.24 exemplifica a proposta aqui sugerida para os três fenômenos da plataforma citados.

Figura 4.24: Proposta de representação AMR para menções e URL (*tweet* em (25)).



Fonte: O autor, 2025.

7. Fenômeno do domínio: ticker e indicador financeiro

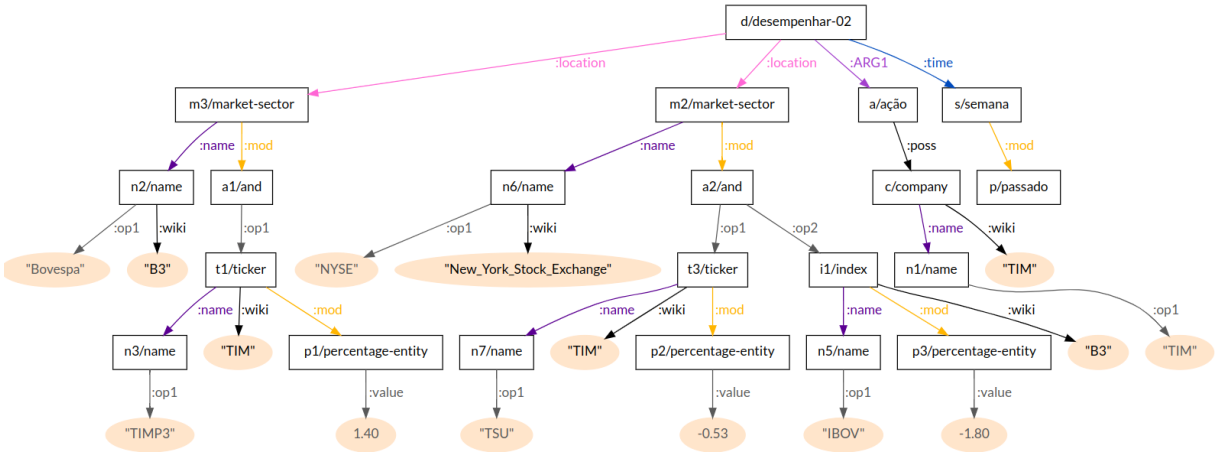
Sobre os conceitos do mercado financeiro propriamente ditos, o DANTEStocks possui muitas ocorrências de (i) *tickers*, sejam precedidos ou não dos sinais de *hashtag* ou *cashtag* (como $Petr4), e (ii) indicadores financeiros usados para avaliar o desempenho das ações no mercado, tais como IMOB (índice imobiliário), IFNC (índice financeiro), SMLL (índice das small caps), IBOV (índice Bovespa), e muitos outros. No exemplo (26), ocorrem os *tickers* TIMP3 e o indicador IBOV; ambos PROPN na anotação-UD.

(26) Desempenho das ações da TIM na semana passada: **TIMP3** (Bovespa) : + 1,40 % , **TSU** (NYSE) : - 0,53 % , **IBOV** : - 1,80 %.

Para a representação AMR, eles foram considerados entidades nomeadas, pois um *ticker* é identificar de forma única uma ação, título ou outro ativo financeiro na bolsa e um indicador é o identificador único para um grupo de ações. Sendo uma entidade desse tipo, a AMR sugere um conjunto de rótulos que engloba as categorias mais abrangentes e difundidas no PLN, como pessoa, organização, local, etc. Caso nenhum dos rótulos se aplique, o modelo sugere o uso de *thing*. Ao mesmo tempo que a AMR sugere a utilização de rótulo tão genérico como *thing* para as entidades não cobertas pelas classes da lista, o modelo aconselha utilizar rótulos que indicam com maior especificidade o tipo da entidade. Isso tanto é verdade que a própria AMR fornece rótulos específicos para entidades do domínio biomédico.

Diante disso, optou-se por empregar dois rótulos distintos dos previstos no modelo, "*ticker* e "*index*, para mais adequadamente representar as ações ou ativos da bolsa e indicadores. Assim, ambos terão das relações :name e :wiki. A decisão por concebê-los como entidade e propor rótulos específicos para a sua representação também foi motivada pelo trabalho em andamento junto ao PPGL de Piai, Di-Felippo e Roman (2025) , no qual esses rótulos, juntamente com "*user*, são propostos para a anotação de entidades nomeadas no DANTEStocks. Muitos dos indicadores possuem página da Wikipédia e, por isso, o conceito u/index terá :wiki preenchido. Na proposta de representação AMR exibida na Figura 4.25 para o *tweet* (26), vê-se o emprego de t/ticker e i/index.

Figura 4.25: Proposta de representação AMR para *tickers* e indicadores financeiros.



Fonte: O autor, 2025.

8. Fenômeno do domínio: hashtag/cashtag como indexadores de conteúdo

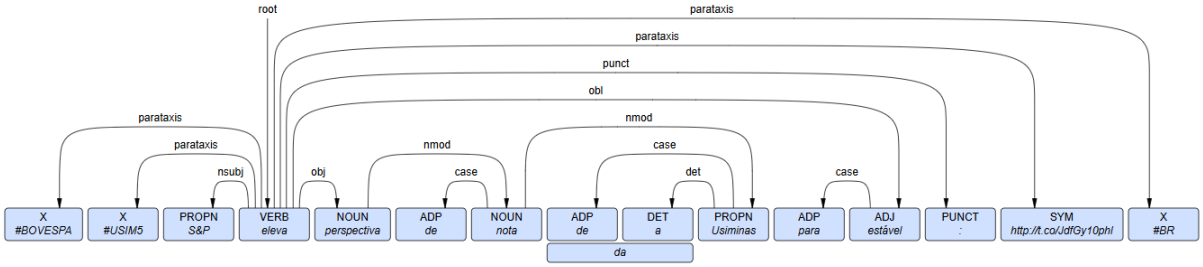
O *tweet* (27) é uma instância do Padrão 16, cuja formalização é: **[prefixo opcional] <sentença> url [<lista ticker>].**

As *hashtags*, como #BOVESPA, #USIM5 e #BR em (27), estão sintaticamente em contexto *standalone*, isto é, não possuem relação sintática clara com o predicador principal do *tweet*, que é o verbo "elevou". Diz-se que as *hashtags* em tal contexto (27) funcionam como indexadores, servindo para organizar e categorizar o conteúdo e, com isso, facilitar a localização de tópicos específicos.

(27) **#BOVESPA #USIM5** S&P eleva perspectiva de nota da Usiminas para estável: http://t.co/JdfGy10phl **#BR**

Por essa razão, aliás, elas receberam a etiqueta PoS X[6] na anotação-UD (Di-Felippo, A. et al., 2022) e foram conectadas ao **root** por **parataxis** (Di-Felippo; Nunes; Barbosa, 2024b), como evidenciado na Figura 4.26.

Figura 4.26: Anotação-UD do *tweet* (26), com *hashtag-ticker* como indexador.



Fonte: O autor, 2025.

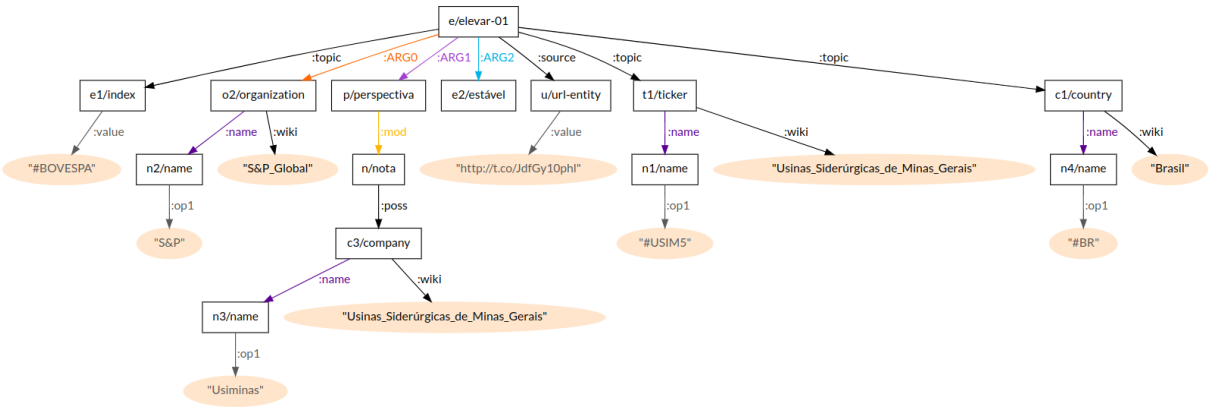
Esse caso é diferente, por exemplo, da ocorrência de "#PETR4" em (17), cuja *tag* PoS PROPN evidencia que a *hashtag-ticker* está integrada à mensagem, na qual, aliás, desempenha a função de **root**. No nível semântico, a *hashtag-ticker* inicial em (17) é o assunto central, pois, como argumentado, sua exclusão impede que se saiba a respeito do quê se trata a postagem, e, por isso, representado como o nó-raiz do grafo AMR.

Em (27), o conteúdo principal está veiculado no componente <sentença> do Padrão 17. Trata-se do enunciado "S&P eleva perspectiva de nota da Usiminas para estável. Ainda que ocorra o sinal de dois-pontos após "estável", indicando que o enunciado tem continuação, este veicula uma mensagem completa, sendo o verbo "eleva" o predicador principal na anotação-UD. Para a anotação AMR, entende-se que o *frame* elevar-01 é o nó-topo do grafo AMR e que as *hashtags* iniciais e final parecem indicar informações secundárias, que seriam: (i) ação específica da Usiminas que teve a nota elevada (#USIM5), (ii) Usiminas/ações fazem parte do índice IBOVESPA (#BOVESPA) e (iii) a notícia é relacionada à economia do Brasil (#BR).

Dessa forma, uma alternativa para a representação semântica de uma *hashtag-ticker* como indexador é relacionar essa entidade nomeada pela relação :*topic* ao nó-topo, como ilustrado na Figura 4.27.

[6]Na modelo gramatical UD, a etiqueta X é usada para palavras que, por algum motivo, não podem ser anotadas com uma categoria gramatical.

Figura 4.27: Proposta de grafo AMR para *hashtag-ticker* como indexador (*tweet* (27)).



Fonte: O autor, 2025.

5 Descrição semântica do DANTEStocks

5.1 Estatísticas da anotação AMR

Como mencionado, os 22 padrões estruturais recorrentes de Di-Felippo, Nunes e Barbosa (2024b) ocorrem em 1.128 *tweets* distintos, os quais foram inteiramente anotados, seja de forma manual ou semiautomática. Com base na anotação desse *subcorpus* do DANTEStocks, fez-se o levantamento estatístico dos diferentes *framesets* e relações AMR, por meio de uma extração utilizando expressões regulares (Aho, 1991) através de um script em *Python*.

5.1.1 Dos *framesets*

Na Tabela 5.1, tem-se os *framesets* mais frequentes no *corpus*. Vê-se na tabela que entre eles não estão, por exemplo, *framesets* relativos a verbos modais (*possible-01, likely-01, obligate-01, permit-01, recommend-01, prefer-01*, etc.). Vale ressaltar que os verbos modais foram anotados no DANTEStocks diretamente com base no repositório de *frames* do PropBank, posto que o Verbo-Brasil não os possui. Com isso, manteve-se a mesma estratégia aplicada aos demais recursos AMR em português.

Sobre os modais não estarem entre os mais frequentes, a hipótese é a de que isso se deve ao estilo conciso, direto e objetivo do mercado financeiro. Em outras palavras, as mensagens financeiras tendem a se concentrar em fatos e dados (numéricos, percentuais) diretos sem expressar incertezas ou possibilidades. Assim, usar verbos modais, como "poderia" e "deveria", poderia criar incerteza ou subjetividade, o que não é desejável em um ambiente que preza pela objetividade e pela precisão. É possível que o limite de caracteres, atualmente em 280, também influencie o uso de verbos modais.

No entanto, entre os mais frequentes está um dos *framesets* exclusivos da AMR, o *have-degree-91*. A hipótese mais provável para isso é que esse *frame* permite capturar a intensidade do desempenho das ações, os graus de valorização/desvalorização e as expectativas de mercado, que são aspectos frequentemente discutidos nas mensagens rápidas e concisas do *Twitter* sobre o mercado financeiro.

Tabela 5.1: Os 10 *framesets* mais frequentes no DANTEStocks.

Frameset	**Freq.**	**%**
indicar-01	367	18.80
analisar-01	207	10.60
rastrear-01	202	10.35
romper-03	201	10.30
resultar-01	186	9.53
conferir-01	156	7.99
have-degree-91	117	5.99
vender-01	63	3.23
comprar-01	51	2.61
concluir-02	32	1.64

Fonte: O autor, 2025.

Os conceitos codificados pelos demais *framesets* (indicar-01, analisar-01, rastrear-01, romper-03, resultar-01, conferir-01, vender-01, comprar-01, concluir-02) são característicos do domínio do mercado financeiro. Os usuários do *Twitter* interessados em finanças (especificamente em ações da bolsa) empregam tais conceitos porque precisam se comunicar a respeito de processos de análise, decisões de compra e venda e movimentações do mercado de ações no geral. Pode-se dizer que os verbos que os expressam permitem que os investidores e analistas compartilhem informações e alertas de forma eficiente e objetiva, essencial para o ambiente dinâmico e ágil do mercado de ações, especialmente no *Twitter*.

Ainda com relação à anotação AMR dos 1.125 *tweets*, que correspondem à totalidade daqueles que apresentam os 22 padrões estruturais identificados, destacam-se os seguintes pontos: (i) o evento mais frequente é indicar-01, cuja alta ocorrência se deve, principalmente, à sua presença nos Padrões 10 e 22, sobretudo no Padrão 10, que conta com 155 instâncias, (ii) em seguida, os *framesets* mais frequentes são analisar-01, rastrear-01 e romper-03, cuja distribuição está fortemente concentrada no Padrão 8, que possui 200 instâncias, (iii) a frequência de conferir-01 também se relaciona diretamente à sua recorrência no Padrão 10 e (iv) os demais *framesets* apresentam frequências mais distribuídas entre os diferentes padrões e suas respectivas instâncias.

5.1.2 Das *relações semânticas*

A Tabela 5.2 exibe a frequência de ocorrência das principais relações semânticas no *corpus*.

Tabela 5.2: As 10 relações semânticas mais frequentes no DANTEStocks.

Relação	Freq.	%
:op1	3430	17.33
:name	2739	13.84
:wiki	2534	12.80
:mod	2429	12.27
:value	1686	8.52
:ARG1	1331	6.73
:unit	938	4.74
:quant	937	4.73
:op2	684	3.46
:time	664	3.36

Fonte: O autor, 2025.

As relações *:opN* (isto é, *:op1* e *:op2*), *:name* e *:wiki* estão entre as mais frequentes. A relação *:opN*, em particular, constitui a estratégia empregada para representar os elementos ordenados que compõem nomes compostos de entidades nomeadas e conjunções.

O fato de as relações *:op1*, *:name* e *:wiki* ocuparem as três primeiras posições na classificação pode ser explicado pela alta frequência de entidades nomeadas no *corpus*, as quais são representadas justamente pela combinação dessas relações. O *ticker*, aliás, é um dos tipos mais frequentes de entidade, uma vez que a compilação do *corpus* foi realizada com base na ocorrência desses códigos, o que faz com que praticamente todo *tweet* do DANTEStocks contenha ao menos um *ticker*. O fato desses códigos serem entidades unitárias contribui para a alta frequência da relação *:op1*. Além disso, a frequente ocorrência de sequências de *tickers*, como no exemplo (19), em que ocorrem 5 códigos em um único *post*, é outro fator que pode explicar a frequência elevada de *:opN*.

Quanto à alta frequência de *:mod*, esta relação, além de representar os casos típicos de adjetivação, foi empregada no DANTEStocks para capturar explicitamente a relação entre um ativo (seja na forma de *ticker* ou não) e a empresa que ele representa, como em “ações da Petrobras” ou “A PETROBRAS (PETR4) concluiu...”. Embora ocorrências como essas possam ser interpretadas inicialmente como envolvendo uma relação de “propriedade” (*:poss*), isto é, as ações pertencem à Petrobras, as ações no geral não pertencem diretamente às empresas no sentido de “posse”. As empresas emitem ações, que

são então compradas e vendidas pelos investidores no mercado financeiro. Assim, seria mais pertinente do ponto de vista técnico dizer que as "ações" representam a empresa, ou melhor, representam parte do capital da empresa. Por essa razão, descartou-se o emprego de *:poss* para esses casos. A alternativa encontrada foi a de utilizar *:mod* para representar a associação entre ação/ticker e o conceito mais amplo de empresa/organização, no sentido de que a empresa é o *:domain-of* (ou apenas *:mod*) das ações. Essa decisão de projeto justifica a frequência da relação AMR *:mod* no *corpus*.

A presença de *:time* entre as mais frequentes também é justificável pelo domínio a que se refere os *tweets*, pois o tempo é um fator essencial para a análise e interpretação das flutuações do mercado de ações relativas ao desempenho das ações/ativos.

A frequência das relações *:value*, *:unit* e *:quant* também refletem o tipo de conteúdo expresso nos *tweets* do mercado financeiro. As URLs, os ordinais e as porcentagens são representados por *:value*, enquanto as quantidades são representadas pelo seu tipo (monetário, temporal, frequência, etc.) e os argumentos *:unit* e *:quant*. Por fim, vale ressaltar que, com o avanço da anotação, acredita-se que relações ARGN, como :ARG1 e :ARG0, etc., serão mais frequentes, já que compõem estruturas de predicado-argumento completas, mais frequentes em *tweets* que possuem linguagem mais similar à padrão.

5.2 Avaliação da anotação

5.2.1 Concordância interanotador

Com o intuito de avaliar a consistência da camada de anotação AMR proposta para o *corpus*, conduziu-se uma etapa de verificação da consistência ou concordância interanotador (em inglês, *Inter-Annotator Agreement - IAA*). Mais precisamente, a consistência foi verificada por meio da dupla anotação de um subconjunto de 20 *tweets* aleatoriamente selecionados entre os 1.125 *posts* anotados neste trabalho.

A quantidade reduzida de *tweets* no conjunto de validação justifica-se, principalmente, por restrições de tempo impostas pelo cronograma deste trabalho, aliadas à alta complexidade envolvida na tarefa de anotação em AMR. Trata-se de um processo manual, detalhado e intensivo, que exige não apenas conhecimento linguístico, mas também domínio das diretrizes específicas da representação semântica, o que limita a viabilidade de uma amostragem mais ampla dentro dos recursos disponíveis.

O subconjunto de validação foi anotado por um segundo anotador previamente treinado, que aplicou a mesma metodologia adotada na anotação manual, realizando a tarefa de forma independente e sem acesso à anotação original, a fim de assegurar a imparcialidade na verificação e permitir a avaliação da consistência interanotador.

O segundo anotador utilizou o editor *metAMoRphosed* (Heinecke, 2023) e o mesmo conjunto de diretrizes da anotação original, o qual contempla fenômenos linguísticos do português, bem como específicos do gênero "*tweet* do mercado financeiro".

Para o cálculo de concordância, utilizou-se a tradicional métrica *Smatch* (Cai; Knight, 2013). Trata-se de uma métrica que calcula o grau de sobreposição ou similaridade entre dois grafos por meio do melhor alinhamento possível entre os conceitos e relações de cada par de grafos. A métrica realiza buscas por todas as possíveis correspondências entre variáveis dos dois grafos, de modo a maximizar o número de triplas coincidentes, considerando a estrutura predicado-argumento. A métrica calcula três valores: (i) precisão (isto é, número de triplas corretas recuperadas em relação ao total de triplas produzidas), (ii) revocação (ou seja, número de triplas corretas recuperadas em relação ao total de triplas do grafo de referência) e (iii) medida-F (no caso, média harmônica entre precisão e *recall*).

A avaliação foi realizada com a biblioteca Python *smatch*, utilizando o parâmetro padrão de *restart*. Na execução, as duas anotações manuais foram comparadas entre si, sendo a primeira delas atribuída como referência (**Anotador 1**) e a segunda como sistema (**Anotador 2**) apenas para fins operacionais, conforme exigido pela ordenação do parâmetro *-f*. Com isso, obtiveram-se os seguintes valores: 88% de precisão, 90% de *recall* e medida-F de 89%.

Os valores obtidos indicam um alto nível de concordância entre os anotadores, o que pode ser atribuído à clareza e consistência das diretrizes de anotação adotadas. Além disso, outro fator que pode contribuir para isso é a relativa regularidade estrutural dos *tweets* presentes no subconjunto de validação. Embora esse conjunto represente os 22 padrões distintos mapeados no *corpus*, os *tweets* selecionados apresentam construções recorrentes e menos ambíguas, o que tende a facilitar decisões anotativas mais alinhadas entre diferentes anotadores. No entanto, é importante relativizar a interpretação desses resultados, pois o subconjunto de validação, embora representativo dos diferentes padrões estruturais, é quantitativamente reduzido.

5.2.2 Análise dos casos de discordância

5.2.2.1 Dos conceitos

Quanto ao mapeamento das palavras para conceitos, destaca-se que, no geral, a sobreposição entre as duas anotações foi significativa. Do total de 357 conceitos anotados pelo Anotador 1, apenas 34 apresentaram algum tipo de discordância com o Anotador 2, o que representa 9,52% do total. Esse índice pode ser considerado relativamente baixo, sobretudo diante da complexidade inerente à representação semântica AMR dos *tweets* do domínio financeiro.

O Quadro 5.1 sistematiza os tipos de discordância, indicando o fenômeno e o elemento linguístico envolvidos nos casos. Vale ressaltar que a quantidade de casos (14) difere da quantidade de conceitos divergentes (34), pois algumas decisões de representação feitas pelos anotadores envolviam mais de um conceito por caso. As divergências ocorreram em 9 dos 20 grafos. Com base no Quadro 5.1, observa-se que os fenômenos que deram origem às discordâncias são bem variados, podendo ser atribuídos a ambiguidade, variação lexical, informalidade ou estruturas semanticamente complexas.

Quadro 5.1: Tipos de discordância entre os anotadores.

Tipo de discordância	Palavra/Expressão linguística	Qt. de casos
Substantivo agentivo	“analista”	3
Truncamento lexical	“este...”, “c...”	2
Seleção de sentido/*frame*	“indicar”	1
Locução prepositiva	“referente a”	1
Expressão quantificadora	“mais NUM NOUN”	1
Expressão de domínio	“15 min”	1
Adjetivo qualificador deverbal (modal)	“obrigatório”	1
Adjetivo qualificador denominal	“milionário”	1
Adjetivo de grau superlativo	“mínimo”	1
Adjetivo de grau comparativo	“melhor”	1
Expressão informal	“braba pra c...”	1
TOTAL		14

Fonte: O autor, 2025.

Um dos fenômenos diz respeito a substantivos agentivos. A anotação do nome “analista”, que ocorre três vezes, variou entre uma representação baseada na forma verbal correspondente e outra mais genérica. O Anotador 1, seguindo as *guidelines* originais da AMR, mapeou o nome agentivo para uma instância do conceito *person* e o relacionou ao predicado verbal correspondente (*a/analisar-01*) por meio da relação inversa *:ARG0-of*. O

Anotador 2 mapeou o nome diretamente para o conceito (*a/analista*).

Outro tipo de desafio está relacionado ao truncamento lexical no final dos *tweets*, que é um fenômeno típico dos *tweets*. Nos casos envolvendo "(OSX) este..." ("estender") e "(braba pra) c..." ("caralho"), os anotadores produziram representações AMR diferentes, mesmo que a diretriz contida no Apêndice A seja a de recuperar a forma completa dos truncamentos sempre que possível (cf. exemplo (21)). Mais especificamente, o Anotador 1 não a representou no grafo, enquanto o Anotador 2 recuperou a forma completa "estender" a partir do valor da *feature* CorrectForm(=estender) descrita na coluna MISC dos arquivos CoNLL-U, e a mapeou para o *frame estender-01* que tem "OSX" como :ARG0. No caso de "caralho", que é parte da expressão "braba pra caralho", a decisão foi inversa. O Anotador 1 recuperou a forma completa e o Anotador 2, não. Isso, aliás, demosntra uma incosistência entre as próprias decisões de um mesmo anotador.

A seleção do sentido e do *frameset* correspondente diante de verbos polissêmicos também foi um fator de discordância. Esse foi o caso do verbo "indicar" (em "Concórdia indica 15 ações"), para o qual o Anotador 1 selecionou o *roleset indicar-01* ("dar indícios, apontar, mostrar") e o Anotador 2 selecionou *indicar-02* ("fazer indicação, recomendar").

Outro fenômeno que gerou divergência foi a locução prepositiva "referente(s) a" em (28). O Anotador 1 representou essa expressão explicitamente como uma instância do evento *referir-01*. O Anotador 2, por sua vez, interpretou a expressão como o propósito de "pagar dividendos". Por isso, a representação AMR inclui o *frameset pagar-01* ligado por *:purpose* à proposição principal *distribuir-01*. Em outras palavras, o Anotador 2 não modelou explicitamente o sentido relacional de "referentes a", mas substituiu essa relação por uma inferência pragmática: que o dinheiro seria distribuído com a finalidade de pagar dividendos obrigatórios mínimos.

(28) Notas gerais A LIGHT (LIGT3) pretender distribuir R$ 32 milhões **referentes ao dividendo mínimo obrigatório** aos... http://t.co/p9jr05Re1l

Quanto à expressão quantificadora do tipo mais NUM NOUN (p.ex.: "mais 9 empresas"), a divergência foi de detalhamento ou granularidade representacional. O Anotador 1 atribuiu diretamente o valor "9" à quantidade (*:quant* 9), ignorando o advérbio "mais". O Anotador 2, por sua vez, empregou o conceito reificado *have-quant-91*. Assim, o conceito *e/empresa* é ARG1-of de *have-quant-91*, cujo :ARG3 (m1/mais) indica acréscimo/adicionalidade e o :ARG2 (9) é o valor do acréscimo.

O par de grafos referente ao *tweet* (29) (Figura 5.1) concentra os casos referentes às expressões de domínio e informal. Nele, tem-se ao todo 4 conceitos divergentes.

(29) PETR4 15 min - Graca foster **Braba pra c...** (mensagem: 952158) http://t.co/282tQZeU6I

Figura 5.1: Anotações discordantes sobre expressões de domínio e informal.

Anotador 1

```
(m1 / multisentence
  :snt1 (t / ticker
        :name (n / name
              :op1 "PETR4")
        :topic-of (m3 / mensagem
                  :source (u / url-entity
                          :value "http://t.co/282tQZeU6I")
                  :ord (o / ordinal-entity
                       :value 952158))
        :mod (g / gráfico
             :frequency (t1 / temporal-quantity
                        :quant 15
                        :unit (m2 / minuto)))
        :wiki "Petrobras")
  :snt2 (h / have-degree-91
        :ARG1 (p / person
              :wiki "Graça_Foster"
              :name (n1 / name
                    :op1 "Graca"
                    :op2 "foster"))
        :ARG2 (b / brabo)
        :ARG3 (c / caralho)))
```

Anotador 2

```
:snt1 (t / ticker
      :name (n / name
            :op1 "PETR4")
      :frequency (t1 / temporal-quantity
                 :quant 15
                 :unit (m / minuto)))
:snt2 (i / irritar-01
      :ARG0 (p / person
            :name (n1 / name
                  :op1 "Graca"
                  :op2 "foster")))
:snt3 (m1 / mensagem
      :ord (o / ordinal-entity
           :value 952158)
      :source (u / url-entity
              :value "http://t.co/282tQZeU6I")))
```

Fonte: O autor, 2025.

A expressão de domínio em questão é “15 min” precedida de um *ticker*. No mercado financeiro, essa expressão se refere a um gráfico com velas (*candlesticks*[1]) de 15 minutos. Esse tipo de visualização é comum entre *traders* que realizam operações de curto prazo, pois permite observar com maior detalhe os movimentos intradiários do mercado.

A representação construída pelo Anotador 1 parece indicar exatamente a interpretação de que se trata do “gráfico de 15 min (da Petr4)”, uma vez que ele pode ter inferido e representado o conceito *g/gráfico* como *:mod* de *t/ticker*, o qual se relaciona a *t1/temporal-quantity* por meio de *:frequency*. O Anotador 2 fez uma representação simplificada, ao não representar o conceito g/gráfico, mas apenas *t1/temporal-quantity*

[1]Um candlestick é uma representação gráfica usada em finanças para mostrar a variação de preço de um ativo em um intervalo de tempo específico. O termo *candlestick* (“vela”) é usado porque cada unidade no gráfico se assemelha visualmente a uma vela: o corpo representa a variação entre os preços de abertura e fechamento, enquanto os pavios (ou sombras) indicam os preços máximo e mínimo do período.

diretamente relacionado ao *ticker*. Ao que parece, essa variação reflete graus diferentes de familiaridade dos anotadores com o domínio discursivo.

O outro caso observado no par da Figura 5.1 envolve a expressão informal "braba pra c...". Embora a última palavra (ou *token*) esteja truncada, infere-se que se trata de caralho. Nesse contexto, o truncamento, que em muitos casos decorre do limite de caracteres imposto pela plataforma, parece funcionar como uma estratégia de atenuação ou disfarce, uma vez que "caralho" constitui um tabu lexical.

O Anotador 1 optou por empregar o conceito reificado *have-degree-91*, o qual é justificado nesse caso porque a construção "brava pra caralho" expressa uma intensificação explícita de um adjetivo. A estrutura semântica da expressão não envolve apenas a atribuição de uma propriedade ("brava"), mas também a indicação de seu grau elevado, veiculado pelo intensificador informal "pra caralho". Assim, o conceito *have-degree-91* permite representar essa relação de forma sistemática, decompondo a expressão em: (i) o sujeito que possui a propriedade (:ARG1), a propriedade em si (:ARG2, no caso, "brava") e o grau ou intensidade dessa propriedade (:ARG3, isto é, "pra caralho"). Essa representação preserva nuances semânticas da linguagem natural e é compatível com práticas formais de anotação em AMR.

Já o Anotador 2 optou por uma representação mais literal da expressão, recorrendo ao conceito verbal *irritar-01* do Verbo-Brasil para capturar o conteúdo proposicional subjacente. Essa escolha sugere uma interpretação pragmática do enunciado, em que "brava pra caralho" é compreendido como equivalente a "estar muito irritada". Assim, em vez de decompor a estrutura semântica da intensificação adjetival, o anotador interpreta a expressão como a manifestação de um estado emocional com possível impacto causativo.

Pode-se dizer que, no caso dessa expressão informal, a escolha entre representar literalmente a emoção evocada (*irritar-01*) ou decompor a estrutura em graus de propriedade (isto é, *have-degree-91* com bravo + intensificador) depende da leitura pragmática, do estilo do anotador e de sua experiência com a linguagem coloquial.

Os casos restantes do Quadro 5.1 dizem respeito a adjetivos de diferentes tipos.

Seguindo as diretrizes gerais da AMR, o Anotador 1 capturou a modalidade expressa pelo adjetivo (modal) "obrigatório" (28) por meio do emprego do *roleset* do verbo correspondente, no caso, *obligate-01*. O Anotador 2, por sua vez, anotou-o como sendo um modificador (*:mod*) de "dividendo".

Quanto aos adjetivos que indicam grau, o Anotador 1 mostrou consistência na representação do superlativo (“mínimo”) e comparativo (“melhor”). Por meio de uma estrutura reificada para expressar grau, ambos foram tratados lexicalmente. No contexto do *tweet* (28), “mínimo” é o :ARG2 (isto é, o “atributo”) de *have-degree-91*. No contexto “melhores ações”, “melhor” é o :ARG3 (isto é, o “grau em si”) de *have-degree-91*.

O Anotator 2 adotou estratégias diferentes para esses dois adjetivos. O adjetivo “mínimo” foi tratado como modificador (*:mod*) direto de “dividendo”. Já o adjetivo “melhor” recebeu um tratamento mais detalhado. Por meio da estrutura reificada, esse adjetivo foi interpretado como uma construção analítica, na qual “bom” é o :ARG2 e “máximo” é o :ARG3 de *have-degree-91*. Como “melhor” não equivale semanticamente a “máximo” e sim a mais bom, talvez uma possibilidade mais adequada às *guidelines* da AMR seja empregar “mais” como :ARG3 para indicar o grau superioridade.

Por fim, o adjetivo denominal “milionário” (em “perda milionária”) foi representado como modificador (*:mod*) direto do evento *perder-01* pelo Anotador 1. Na representação do Anotador 2, o adjetivo é o :ARG1 de *perder-01*, sendo especificamente codificado pelo conceito *amount-quantity* (usado para quantidades genéricas sem unidade específica) e a relação *:quant*, cujo valor é 1000000. Ao se analisar as *guidelines* da AMR, a interpretação do Anotador 1 parece ser mais adequada, posto que *:quant 1000000* sugere que a quantia seja de fato “1 milhão”.

5.2.2.2 Das relações semânticas

Sobre as relações semânticas, a sobreposição entre as duas anotações também foi significativa. Do total de 575 relações anotadas pelo Anotador 1, apenas 28 apresentaram algum tipo de discordância com o Anotador 2, representando 4,87% do total. Assim como ocorre com os conceitos, as discordâncias na representação das relações AMR envolvem fenômenos variados, as quais estão sistematizadas no Quadro 5.2.

Entre eles, as entidades nomeadas do tipo *ticker* e *pessoa* geraram as discordâncias mais recorrentes. Segundo as diretrizes da AMR, as entidades nomeadas devem ser representadas por um conceito que indica sua categoria, como *t/ticker* e *p/person*, sendo ambos especificados pelas relações *:name* (nome literal da entidade) e *:wiki* (associa a entidade à Wikipedia). A discordância, no caso, foi produzida pela não inclusão de *:wiki* pelo Anotador 2.

Quadro 5.2: Tipos de discordância entre os anotadores.

Tipo de discordância	Palavra/Expressão linguística	Qt. de casos
Entidade nomeada (*ticker* e pessoa)	LIGT3, PETR4, Graça Foster	4
URL	–	2
Valor monetário	"32 milhões"	1
Locução prepositiva	"referente a"	1
Hashtag como índice	–	1
TOTAL		9

Fonte: O autor, 2025.

Algo parecido ocorreu com o valor monetário "32 milhões", que, segundo a AMR, deve ser representado pelo conceito *:monetary-quantity*, especificado pelas relações *:quant* e *:unit*. A discordância ocorreu devido à troca de *:quant* por *:value* feita pelo Anotador 2.

No que tange às URLs, os anotadores discordaram em duas ocasiões, nas quais o Anotador 1 as representou como sentenças relacionadas ao conceito de topo por *:sntN*, enquanto o Anotador 2 empregou a relação *:source* em ambos os casos.

Como mencionado, a Anotador 2 interpretou a expressão "referente a" como o propósito de pagar dividendos e, por isso, a representação AMR do *tweet*-exemplo (28) inclui o conceito *pagar-01* e a relação *:purpose*, que o liga à proposição principal *distribuir-01*. Como essa relação não está presente no grafo do Anotador 1, daí a discordância.

Por fim, a ocorrência de *hashtag* ao final do *tweet*, funcionando meramente como índice de assunto na plataforma (sem função sintática), gerou apenas um caso de discordância. Especificamente, o Anotador 1 não seguiu a diretriz de anotá-las como *:topic* relacionado ao nó-topo (*m/multisentence*), mas sim como *:sntN*.

De um modo geral, pode-se dizer que as discordâncias entre as anotações resultam de múltiplos fatores interligados, como (i) a complexidade dos itens linguísticos (como verbos polissêmicos ou adjetivos subjetivos) exige escolhas interpretativas finas, (ii) a informalidade dos *tweets* (com truncamentos e estruturas elípticas) dificulta a recuperação do sentido completo, (iii) a ambiguidade entre leitura literal e pragmática (p.ex.: braba pra caralho) leva a decisões distintas, (iv) lacunas ou ambiguidades nas próprias guidelines da AMR deixam margem a interpretações variadas, (v) o domínio temático específico (p.ex.: finanças) pode favorecer quem o domina, e (vi) o perfil interpretativo dos anotadores (mais ou menos inferencial) também contribui para as divergências.

Apesar do conjunto de validação ser reduzido, os resultados do cálculo da concordância são promissores, indicando que a metodologia adotada e o manual desenvolvido

oferecem um caminho pertinente para futuras expansões da anotação semântica AMR de *tweets* do mercado financeiro em português.

6 Considerações finais

A anotação semântica realizada neste trabalho é pioneira no que tange a *tweet*, um dos gêneros CGU. Diz-se isso porque a maioria dos *tweebanks* possui anotação semântica em nível lexical, restringindo-se a entidades nomeadas e polaridade (lexicalmente marcada). Ao empregar um formalismo semântico como a AMR, que busca explicitar o significado do enunciado (no caso, do *tweet* completo), este trabalho é, além de pioneiro, desafiador, pois as características lexicais e estruturais dos *tweets* do *corpus* selecionado, o DANTEStocks, impõem desafios adicionais à originalmente complexa tarefa de anotação AMR.

Enquanto contribuições desta pesquisa, citam-se:

- Proposta de representação AMR para cada um dos 22 padrões recorrentes identificados por Di-Felippo, Nunes e Barbosa (2024b) (Apêndice 1);
- Anotação AMR de 1.128 *tweets* distintos do *corpus* DANTEStocks em um processo manual e semiautomático, o que representa quase 30% do total de 4.048 postagens que compõem o *corpus*;
- Disponibilização da parcela do DANTEStocks com anotação AMR *gold-standard*, cujo IAA, embora baseado em um conjunto de validação pequeno, indica confiabilidade do *subcorpus* anotado, o qual será pioneiro na literatura com esse tipo de anotação;
- Diretrizes de anotação AMR para aspectos do português que ocorrem nos *tweets* (multiplicidade de sentenças, locução "pelo menos" e expressão "põe ADJ nisto");
- Diretrizes de anotação AMR para fenômenos gerais do gênero CGU (multiplicidade de segmentos, segmentos complexos, truncamento estrutural e lexical, variações ortográficas da norma padrão e inovações lexicais), específicos da plataforma *Twitter* (marca de *retweet* (RT), menção e URL) e típicos do domínio do mercado financeiro

(*ticker*, indicador financeiro e *hashtag/cashtag* como indexador);

- Emprego de 3 rótulos específicos para representar mais adequadamente o conhecimento de domínio via AMR, a saber: *u/url-entity, t/ticker* e *u/user*;
- Proposição de 1 *frameset* em português para o sentido do verbo "repicar" (e sua nominalização "repique") no mercado financeiro, que é "recuperação temporária ou pico de valorização de um ativo/ação após um período de queda ou estabilização"[1]; o *frameset* foi proposto nos moldes do Verbo-Brasil, incluindo numeração do sentido em "01", definição dos ArgN e anotação de instâncias do *corpus* com as diferentes realizações dos argumentos, como descrito a seguir:

  Roleset id: repicar.01

  Arg1: coisa repicada

  Arg2: valor/preço do repique

  Exemplo 1: petr4 vai *repicar* quando bater no 12,46

  Rel: repicar

  Arg1: petr4

  Arg2: 12,46[2]

  Exemplo 2: Na #petr4 nos 15 min na máx de ontem uma Onda 2 e na mín uma sub onda 3 com alvos de repique 17,80 obj de fibo 16,90 http://t.co/De60uvFQbc

  Rel: repique

  Arg1: petr4

  Arg2: 17,80

  Exemplo 3: #VALE5 será q hj vc irá *repicar* Valedita??

  Rel: repicar

  Arg1: #VALE5

  Arg2: -

  Exemplo 4: Hoje é o dia da verdade para a PETR4. Ao longo do dia informarei se ela vai *repicar* ou despencar de vez.

  Rel: repicar

  Arg1: #PETR4

[1] http://www.igf.com.br/aprende/glossario/glo_Resp.aspx?id=2613

[2] Nesse caso, "12,46" indica o valor/preço que a ação precisa atingir para o movimento de *repique*.

Arg2: -

A principal dificuldade enfrentada residiu na interpretação dos *tweets* para a representação semântica AMR. Essa dificuldade advém de vários fatores: (i) necessidade de conhecimento especializado, o que demanda consultas constantes a especialistas de domínio para a compreensão da postagem, (ii) estrutura fragmentada dos *tweets*, (iii) falta de contexto e (iv) todos os fenômenos típicos de CGU.

Diante disso, pode-se dizer que não há uma representação semântica "correta" dos *tweets*, mas que os grafos AMR produzidos neste trabalho são resultados de uma interpretação possível das postagens. Nesse sentido, uma preocupação constante do trabalho foi a de tentar identificar padrões de conteúdo e representá-los de forma regular ou uniforme segundo o modelo AMR.

Nesse cenário, a identificação do conceito a ser representado como nó-raiz do grafo AMR é a tarefa mais desafiadora. Por essa razão, essa tarefa foi pautada, sempre que possível, na sintaxe, principalmente no **root** da anotação-UD.

Como trabalho futuro, propõe-se a expansão da anotação AMR para os 70% restantes do *corpus*, de modo a consolidar um recurso mais amplo e robusto para o treinamento e a avaliação de *parsers* AMR para *tweets* em português. Com a disponibilização da parcela anotada e revisada do DANTEStocks, reforça-se a possibilidade de aplicar abordagens baseadas em LLMs com técnicas de *few-shot learning* para a anotação do restante do *corpus*.

Além disso, seria relevante explorar possíveis mapeamentos entre as relações sintáticas (de dependência) superficiais da UD e a representação AMR, focada em relações semânticas profundas. Do ponto de vista descritivo, essa exploração pode, por exemplo, indicar o mapeamento sistemático entre funções sintáticas e papéis semânticos. Do ponto de vista do PLN, tal correspondência pode, por exemplo, favorecer o *transfer learning*[3] entre tarefas, possibilitando o uso de *parsers* sintáticos UD como ponto de partida para a geração de grafos AMR.

Outro caminho promissor consiste em verificar a possibilidade de generalização das diretrizes de anotação para *tweets* de outros domínios, como é o caso do *corpus* de 6.420 *tweets* denominado COVID-19 (Barberia; Schmalz; Roman, 2023). Compilados

[3]Refere-se à técnica em que um modelo treinado em uma tarefa (por exemplo, análise sintática-UD) é usado como ponto de partida, ou fornece representações intermediárias, para treinar outro modelo em uma tarefa relacionada, mas mais complexa (como geração de grafos AMR) (Jurafsky; Martin, 2025).

entre 2020 e 2021, os *tweets* desse *corpus* expressam o posicionamento de candidatos às eleições locais brasileiras sobre as vacinas e a vacinação contra a COVID-19.

Referências

Aho, Alfred V. **Algorithms for finding patterns in strings, Handbook of theoretical computer science (vol. A): algorithms and complexity**. [S.l.]: MIT Press, Cambridge, MA, 1991.

Anchieta, Rafael T.; Cabezudo, Marco A. S.; Pardo, Thiago A. S. SEMA: an Extended Semantic Evaluation Metric for AMR, 2019. arXiv: `1905.12069 [cs.CL]`. Disponível em: <`https://arxiv.org/abs/1905.12069`>.

Anchiêta, Rafael; Pardo, Thiago. Análise Semântica com base em AMR para o Português. **Linguamática**, v. 14, n. 1, p. 33–48, 2022.

Anchiêta, Rafael; Pardo, Thiago. Towards AMR-BR: A SemBank for Brazilian Portuguese Language. In: PROCEEDINGS of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), mai. 2018. Disponível em: <`https://aclanthology.org/L18-1157/`>.

Anchiêta, Rafael Torres; Pardo, Thiago Alexandre Salgueiro. A rule-based AMR parser for Portuguese. In: Simari, Guillermo R. et al. (Ed.). **Advances in Artificial Intelligence - IBERAMIA 2018**. [S.l.: s.n.], 2018. P. 341–353.

Artstein, Ron. Inter-annotator Agreement. In: Ide, Nancy; Pustejovsky, James (Ed.). **Handbook of Linguistic Annotation**. [S.l.]: Springer, 2017. P. 297–313.

Aziz, W.; Specia, L. Fully Automatic Compilation of Portuguese-English and Portuguese-Spanish Parallel Corpora. In: PROCEEDINGS of the 8th Brazilian Symposium in Information and Human Language Technology. Cuiabá, MT, Brazil: Sociedade Brasileira de Computação, 2011. P. 234–238.

Banarescu, Laura et al. Abstract Meaning Representation for sembanking. In: PROCEEDINGS of the 7th linguistic annotation workshop and interoperability with discourse. [S.l.: s.n.], 2013. P. 178–186.

Baptista, Jorge. ViPEr: A Lexicon-Grammar of European Portuguese Verbs. In: 31E Colloque International sur le Lexique et la Grammaire. [S.l.: s.n.], 2012. P. 10–16.

Baptista, Jorge. Viper: uma base de dados de construções léxico-sintáticas de verbos do português europeu. In: ACTAS do XXVIII Encontro da APL-Textos Selecionados. [S.l.: s.n.], 2013. P. 111–129.

Baptista, Jorge; Mamede, Nuno. **Dicionário gramatical de verbos do português**. [S.l.: s.n.], 2020.

Baptista, Jorge; Mamede, Nuno. Syntactic Transformations in Rule-Based Parsing of Support Verb Constructions: Examples from European Portuguese. In: SCHLOSS DAGSTUHL-LEIBNIZ-ZENTRUM FÜR INFORMATIK. 9TH Symposium on Languages, Applications and Technologies (SLATE 2020). [S.l.: s.n.], 2020.

Baptista, Jorge; Reis, Sónia et al. Lexicalized Meaning Representation (LMR). In: Bonial, Claire; Bonn, Julia; Hwang, Jena D. (Ed.). **Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024**. Torino, Italia: ELRA e ICCL, mai. 2024. P. 101–111. Disponível em: <https://aclanthology.org/2024.dmr-1.11/>.

Barberia, Lorena; Schmalz, Pedro Henrique; Roman, Norton. When Tweets Get Viral - A Deep Learning Approach for Stance Analysis of Covid-19 Vaccines Tweets by Brazilian Political Elites. In: ANAIS do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. Belo Horizonte/MG: SBC, 2023. P. 104–114. DOI: 10.5753/stil.2023.233961. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25442>.

Barbosa, B.K.S. **Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português**. 2024. F. 208. MSc Dissertation – Universidade Federal de São Carlos, São Carlos, SP.

Bateman, John et al. The re-use of linguistic resources across languages in multilingual generation components. In: PROCEEDINGS of the 12th international joint conference on Artificial intelligence-Volume 2. [S.l.: s.n.], 1991. P. 966–971.

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

Bollen, Johan; Mao, Huina; Zeng, Xiaojun. Twitter mood predicts the stock market. **Journal of Computational Science**, Elsevier BV, v. 2, n. 1, p. 1–8, mar. 2011. ISSN 1877-7503. DOI: 10.1016/j.jocs.2010.12.007. Disponível em: <http://dx.doi.org/10.1016/j.jocs.2010.12.007>.

Bonial, Claire et al. PropBank: Semantics of New Predicate Types. In: PROCEEDINGS of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA), mai. 2014. P. 3013–3019. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2014/pdf/1012_Paper.pdf>.

Bonn, Julia; Buchholz, Matthew J. et al. Building a Broad Infrastructure for Uniform Meaning Representations. In: PROCEEDINGS of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Torino, Italia: ELRA e ICCL, 2024. P. 2537–2547.

Bonn, Julia; Myers, Skatje et al. Mapping AMR to UMR: Resources for Adapting Existing Corpora for Cross-Lingual Compatibility. In: Dakota, Daniel et al. (Ed.). **Proceedings of the 21st International Workshop on Treebanks and Linguistic Theories (TLT, GURT/SyntaxFest 2023)**. Washington, D.C.: Association for Computational Linguistics, mar. 2023. P. 74–95. Disponível em: <https://aclanthology.org/2023.tlt-1.8/>.

Brachman, Ronald J. **Knowledge Representation and Reasoning**. [S.l.]: Morgan Kaufman/Elsevier, 2004.

Brown, Tom et al. Language models are few-shot learners. **Advances in neural information processing systems**, v. 33, p. 1877–1901, 2020.

Bruckschen, M. et al. **Anotação Linguística em XML do Corpus PLN-BR**. [S.l.: s.n.], 2008. [S.l.]

Brum, H. B.; Nunes, M. G. V. N. Building a sentiment corpus of tweets in Brazilian Portuguese. In: PROCEEDINGS of the 11th International Conference on Language Resources and Evaluation (LREC). Miyazaki, Japan: [s.n.], 2018. P. 4167–4172.

Cabezudo, M.; Pardo, T. Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese. In: PROCEEDINGS of the 13th Linguistic Annotation Workshop. Florence, Itália: Association for Computational Linguistics, 2019. P. 236–244.

Cai, Sheng; Lam, Wai. AMR Parsing with Biaffine Attention. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2020 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2020. P. 2020–2025.

Cai, Shu; Knight, Kevin. Smatch: an evaluation metric for semantic feature structures. In: PROCEEDINGS of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). [S.l.: s.n.], 2013. P. 748–752.

Carletta, Jean. Assessing Agreement on Classification Tasks: The Kappa Statistic. Edição: Julia Hirschberg. **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. DOI: 10.1080/08839514.2019.1673037. Disponível em: <http://dx.doi.org/10.1080/08839514.2019.1673037>.

Cortis, Keith et al. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In: Bethard, Steven et al. (Ed.). **Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)**. Vancouver, Canada: Association for Computational Linguistics, ago. 2017. P. 519–535. DOI: 10.18653/v1/S17-2089. Disponível em: <https://aclanthology.org/S17-2089/>.

Damonte, M.; Cohen, S. B. Cross-lingual Abstract Meaning Representation Parsing. In: 16TH Annual Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies. [S.l.: s.n.], 2018. P. 1146–1155.

Damonte, Marco; Cohen, Shay B.; Satta, Giorgio. An Incremental Parser for Abstract Meaning Representation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. [S.l.: s.n.], 2017. P. 536–546.

Davidson, Donald. The Logical Form of Action Sentences. Edição: Nicholas Rescher. **The Logic of Decision and Action**, University of Pittsburgh Press, p. 81–95, 1967.

Derczynski, Leon; Bontcheva, Kalina; Roberts, Ian. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In: Matsumoto, Yuji; Prasad, Rashmi (Ed.). **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**. Osaka, Japan: The COLING 2016 Organizing Committee, dez. 2016. P. 1169–1179. Disponível em: <https://aclanthology.org/C16-1111/>.

Devlin, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, Jill; Doran, Christy; Solorio, Thamar (Ed.). **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)**. Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423. Disponível em: <https://aclanthology.org/N19-1423/>.

Dhabe, Priyadarshan et al. Stock Market Trend Prediction Along with Twitter Sentiment Analysis. In: Balas, Valentina Emilia; Semwal, Vijay Bhaskar; Khandare, Anand (Ed.). **Intelligent Computing and Networking**. Singapore: Springer Nature Singapore, 2023. P. 45–59. ISBN 978-981-99-0071-8.

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)**. São Carlos, 2021. P. 79.

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, Turkey: ELRA, 2012. P. 1862–1867.

Duran, Magali; Lopes, Lucelene et al. 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. Belo Horizonte/MG: SBC, 2023. P. 115–124. DOI: 10.5753/stil.2023.233975. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/25443>.

Duran, Magali Sanchez; Martins, J. P.; Aluísio, Sandra Maria. Um repositório de verbos para a anotação de papéis semânticos, disponível na Web. In: PROCEEDINGS of the 9th Brazilian Symposium in Information and Human Language Technology. [S.l.: s.n.], 2013. P. 168–172.

F. J. V. Silva, N. T. Roman; Carvalho, A. M. Stock Market Tweets Annotated with Emotions. **Corpora**, v. 15, n. 3, p. 343–354, 2020.

Di-Fabio, Andrea; Conia, Simone; Navigli, Roberto. VerbAtlas: a Novel Large-Scale Verbal Semantic Resource and Its Application to Semantic Role Labeling. In: p. 627–637. DOI: 10.18653/v1/D19-1058.

Di-Felippo, A. et al. **Diretrizes de Anotação de PoS Tags em Tweets do Mercado Financeiro: Orientações para Anotação em Língua Portuguesa segundo a Abordagem Universal Dependencies**. São Carlos-SP, 2022. P. 24. Relatório Técnico n. 438 – ICMC, USP.

Di-Felippo, Ariani; Nunes, Maria das Graças Volpe; Barbosa, Bryan Khelven da Silva. A Dependency Treebank of Tweets in Brazilian Portuguese: Syntactic Annotation Issues and Approach. In: ANAIS do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. Belém/PA: SBC, 2024. P. 192–201. DOI: 10.5753/stil.2024.245383. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31131>.

Di-Felippo, Ariani; Nunes, Maria das Graças Volpe; Barbosa, Bryan Khelven da Silva. **Diretrizes de anotação de relações de dependência em tweets do mercado financeiro**. São Carlos, 2024. P. 70. Relatório Técnico n. 446 – ICMC, USP.

Di-Felippo, Ariani; Roman, Norton Trevisan. DANTEStocks: a multi-layered annotated corpus of stock market tweets for Brazilian Portuguese. **Brazilian Journal of Applied Linguistics**, Corpus Linguistics: Studies and Applications, p. 1–23, 2025. To appear.

Di-Felippo, Ariani et al. 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: [s.n.], 2021. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17813>.

Fillmore, C. J. The Case for Case. In: UNIVERSALS in linguistic theory. [S.l.]: Holt, Rinehart e Winston, Inc., 1968. P. 1–88.

Flanigan, Jeffrey et al. A Discriminative Graph-Based Parser for the Abstract Meaning Representation. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). [S.l.: s.n.], 2014. P. 1426–1436.

Freitas, Cláudia. **Linguística Computacional**. [S.l.]: Parábola Editorial, 2022. ISBN 978-85-7934-278-3.

Freitas, Cláudia; Pardo, Thiago. PropBank e anotação de papéis semânticos para a língua portuguesa: O que há de novo? In: ANAIS do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. Belém/PA: SBC, 2024. P. 118–128. DOI: 10.5753/stil.2024.245377. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31123>.

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.

Gomez, Helena et al. Cicbuapnlp at semeval-2016 task 4-a: Discovering twitter polarity using enhanced embeddings. In: PROCEEDINGS of the 10th International Workshop on Semantic Evaluation (SemEval-2016). [S.l.: s.n.], 2016. P. 145–148.

Hartmann, N. S.; Duran, Magali Sanchez; Aluísio, Sandra Maria. Automatic Semantic Role Labeling on Non-Revised Syntactic Trees of Journalistic Texts. In: INTERNATIONAL Conference on Computational Processing of the Portuguese Language. [S.l.: s.n.], 2016. P. 202–212.

Heinecke, J. metAMoRphosED, a Graphical Editor for Abstract Meaning Representation. In: PROCEEDINGS of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19). Nancy, France: Association for Computational Linguistics, 2023. P. 27–32.

Hermjakob, U. **AMR Editor: A Tool to Build Abstract Meaning Representations**. [S.l.: s.n.], 2013. Acesso em: 25 ago. 2024. Disponível em: <https://amr.isi.edu/papers/amr-editor-ulf2013a.pdf>.

Inácio, Marcio et al. The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. **DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada**, v. 39, set. 2023. DOI: 10.1590/1678-460x202339355159.

Junior, E. Correa et al. PELESent: Cross-domain polarity classification using distant supervision. In: PROCEEDINGS of the 6th Brazilian Conference on Intelligent Systems (BRACIS). Uberlândia, Brazil: [s.n.], 2017. P. 49–54.

Jurafsky, D.; Martin, J. H. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition**. 3rd (draft). [S.l.: s.n.], 2025. Acesso em: 20 jan. 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>.

Krumm, John; Davies, Nigel; Narayanaswami, Chandra. User-Generated Content. **IEEE Pervasive Computing**, v. 7, n. 4, p. 10–11, 2008. DOI: 10.1109/MPRV.2008.85.

Lehmann, Fritz. **Semantic networks in artificial intelligence**. [S.l.]: Elsevier Science Inc., 1992.

Li, Bin et al. Annotating the Little Prince with Chinese AMRs. In: PROCEEDINGS of the 10th Linguistic Annotation Workshop (LAW-X 2016) held in conjunction with ACL 2016. [S.l.]: Association for Computational Linguistics, 2016. P. 7–15. Disponível em: <https://aclanthology.org/W16-1702/>.

Liu, Yuxuan et al. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. **arXiv**, v. 2305, n. 13860, 2023. Disponível em: <https://arxiv.org/abs/2305.13860>.

Lopes, Lucelene et al. PortiLexicon-UD: a Portuguese lexical resource according to Universal Dependencies model. In: PROCEEDINGS. [S.l.: s.n.], 2022.

Lyu, Chunchuan; Titov, Ivan. AMR parsing as graph prediction with latent alignment. In: PROCEEDINGS of the 56th Annual Meeting of the Association for Computational Linguistics. [S.l.]: Association for Computational Linguistics, 2018. P. 397–407.

Lyu, Shiqi; Titov, Ivan. A Transition-Based AMR Parser with Stack-LSTM. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2018. P. 1101–1106.

Marcus, M. P. et al. Building a Large Annotated Corpus of English: The Penn Treebank. **Computational Linguistics**, v. 19, n. 2, p. 313–330, 1993.

Martínez Lorenzo, Abelardo Carlos; Maru, Marco; Navigli, Roberto. Fully-Semantic Parsing and Generation: the BabelNet Meaning Representation. In: Muresan, Smaranda; Nakov, Preslav; Villavicencio, Aline (Ed.). **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**. Dublin, Ireland: Association for Computational Linguistics, mai. 2022. P. 1727–1741. DOI: 10.18653/v1/2022.acl-long.121. Disponível em: <https://aclanthology.org/2022.acl-long.121/>.

McGregor, S. C.; Molyneux, L. Twitters influence on news judgment: An experiment among journalists. **Journalism**, v. 21, n. 5, p. 597–613, 2020.

Moraes, S. M. W.; Manssour, I. H.; Silveira, M. S. 7x1-PT: um corpus extraído do Twitter para análise de sentimentos em língua portuguesa. In: PROCEEDINGS of the 10th Symposium in Information and Human Language Technology (STIL). Natal, Brazil: Sociedade Brasileira de Computação, 2015. P. 21–25.

Mota, Cristina; Santos, Diana. **Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM**. [S.l.]: Linguateca, 2008.

Navigli, Roberto; Blloshmi, Rexhina; Martinez Lorenzo, Abelardo Carlos. Babelnet Meaning Representation: A Fully Semantic Formalism to Overcome Language Barriers. In: PROCEEDINGS of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI, 2022. v. 36, p. 12274–12279.

Navigli, Roberto; Ponzetto, Simone Paolo. BabelNet: Building a Very Large Multilingual Semantic Network. **Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics**, Association for Computational Linguistics, p. 216–225, 2010.

Nivre, Joakim; Marneffe, Marie-Catherine de; Ginter, Filip; Goldberg, Yoav et al. Universal Dependencies v1: A Multilingual Treebank Collection. In: Calzolari, Nicoletta et al. (Ed.). **Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)**. Portoro, Slovenia: European Language Resources Association (ELRA), mai. 2016. P. 1659–1666. Disponível em: <https://aclanthology.org/L16-1262/>.

Nivre, Joakim; Marneffe, Marie-Catherine de; Ginter, Filip; Haji, Jan et al. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In: Calzolari, Nicoletta et al. (Ed.). **Proceedings of the Twelfth Language Resources and Evaluation Conference**. Marseille, France: European Language Resources Association, mai. 2020. P. 4034–4043. ISBN 979-10-95546-34-4. Disponível em: <https://aclanthology.org/2020.lrec-1.497>.

Noord, Rob van; Bos, Marie. A Transition-Based Parser for AMR. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2nd Workshop on Representation Learning for NLP. [S.l.: s.n.], 2017. P. 21–30.

OGorman, Tim et al. AMR beyond the sentence: the multi-sentence AMR corpus. In: PROCEEDINGS of the 27th international conference on computational linguistics. [S.l.: s.n.], 2018. P. 3693–3702.

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.

Papineni, Kishore et al. BLEU: a method for automatic evaluation of machine translation. In: PROCEEDINGS of the 40th Annual Meeting on Association for Computational Linguistics. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002. (ACL '02), p. 311–318. DOI: 10.3115/1073083.1073135. Disponível em: <https://doi.org/10.3115/1073083.1073135>.

Pardo, T. A. S. et al. 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.

Parsons, Terence. **Events in the Semantics of English: A Study in Subatomic Semantics**. [S.l.]: MIT Press, 1990.

Penn, Gerald; Kiparsky, Paul. On Panini and the Generative Capacity of Contextualized Replacement Systems. In: Kay, Martin; Boitet, Christian (Ed.). **Proceedings of COLING 2012: Posters**. Mumbai, India: The COLING 2012 Organizing Committee, dez. 2012. P. 943–950. Disponível em: <https://aclanthology.org/C12-2092/>.

Pereira, F.C.N.; Shieber, S.M. **Prolog and Natural-language Analysis**. [S.l.]: Microtome Publishing, 2002. ISBN 9780971977709. Disponível em: <https://books.google.com.br/books?id=1ahu2QEuYaIC>.

Piai, Laís. **Anotação de corpus: caracterização de Entidades Nomeadas em tweets do mercado financeiro**. 2025. Dissertação de Mestrado em Linguística – Universidade Federal de São Carlos, São Carlos. Disponível em: <https://repositorio.ufscar.br/handle/20.500.14289/22651>.

Piai, Laís; Di-Felippo, Ariani; Roman, Norton Trevisan. **Guia de anotação de entidades nomeada sem tweets do mercado financeiro: adaptação da taxonomia hierárquica do segundo HAREM**. [S.l.: s.n.], 2025.

Plutchik, R.; Kellerman, H. **Emotion: Theory, Research and Experience**. Nova Iorque: Academic Press, 1986.

Qi, Peng et al. Stanza: A Python natural language processing toolkit for many human languages. **arXiv preprint arXiv:2003.07082**, 2020.

Rademaker, Alexandre et al. Universal dependencies for Portuguese. In: PROCEEDINGS of the fourth international conference on dependency linguistics (Depling 2017). [S.l.: s.n.], 2017. P. 197–206.

Rijhwani, Shruti; Preotiuc-Pietro, Daniel. Temporally-informed analysis of named entity recognition. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. P. 7605–7617. Disponível em: <https://www.aclweb.org/anthology/2020.acl-main.676>.

Sanguinetti, M. et al. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. **Language Resources Evaluation**, v. 57, p. 493–544, 2023. DOI: https://doi.org/10.1007/s10579-022-09581-9.

Scandarolli, Clarissa Lenina et al. 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). [S.l.]: Sociedade Brasileira de Computação, 2023. (STIL 2023). DOI: 10.5753/stil.2023.233948. Disponível em: <http://dx.doi.org/10.5753/stil.2023.233948>.

Senel, Lutfi Kerem et al. Semantic Structure and Interpretability of Word Embeddings. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, Institute of Electrical e Electronics Engineers (IEEE), v. 26, n. 10, p. 1769–1779, out.

2018. ISSN 2329-9304. DOI: 10.1109/taslp.2018.2837384. Disponível em: <http://dx.doi.org/10.1109/TASLP.2018.2837384>.

Seno, Eloize et al. XPTA: um parser AMR para o Português baseado em uma abordagem entre línguas. **Linguamática**, v. 14, p. 49–68, jul. 2022. DOI: 10.21814/lm.14.1.359.

Silva, E. H. et al. 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, E.H. et al. Etiquetagem morfossintática multigênero para o português do Brasil segundo o modelo Universal Dependencies. In: PROCEEDINGS of the 14th Symposium in Information and Human Language Technology. Belo Horizonte, Brazil: SBC, 2023. P. 63–73.

Silva, I. S. et al. Effective sentiment stream analysis with self-augmenting training and demand-driven projection. In: PROCEEDINGS of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. Beijing, China: [s.n.], 2011. P. 475–484.

Silva, R. R.; Pardo, T. A. S. Córpus 4P: um córpus anotado de opiniões em português sobre produtos eletrônicos para fins de sumarização contrastiva de opinião. In: ANAIS da 6a Jornada de Descrição do Português (JDP). Salvador, Bahia, Brazil: [s.n.], out. 2019. P. 330–338.

Sinclair, John McH. Corpus and Text: Basic Principles. In: Wynne, Martin (Ed.). **Developing Linguistic Corpora: A Guide to Good Practice**. Oxford: Oxbow Books, 2005. P. 1–16.

Song, Linfeng; Gildea, Daniel. SemBleu: A robust metric for AMR parsing evaluation. **arXiv preprint arXiv:1905.10726**, 2019.

Specia, L.; Rino, L. **Representação Semântica: Alguns Modelos Ilustrativos**. [S.l.], 2002.

Stefanowitsch, Anatol. **Corpus Linguistics: A Guide to the Methodology**. [S.l.: s.n.], mai. 2020. ISBN 978-3-96110-224-2. DOI: 10.5281/zenodo.3735822.

Straka, Milan. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In: PROCEEDINGS of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. [S.l.: s.n.], 2018. P. 197–207.

Taylor, A.; Marcus, M.; Santorini, B. The Penn Treebank: An Overview. In: TREEBANKS. [S.l.]: Springer, Dordrecht, 2003. v. 20. (Text, Speech and Language Technology). DOI: 10.1007/978-94-010-0201-1_1.

Uchida, Hiroshi; Zhu, Meiying; Della Senta, Tarcisio. Universal networking language. **UNDL foundation**, v. 2, 2005.

Ushio, Asahi et al. Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts. In: PROCEEDINGS of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Online only: Association for Computational Linguistics, 2022. P. 309–319. Disponível em: <https://www.aclweb.org/anthology/2022.aacl-main.32>.

Van Gysel, Jens EL et al. Designing a uniform meaning representation for natural language processing. **KI-Künstliche Intelligenz**, Springer, v. 35, n. 3, p. 343–360, 2021.

Wang, Chuan; Xue, Nianwen; Pradhan, Sameer. A Transition-based Algorithm for AMR Parsing. In: ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S.l.: s.n.], 2015. P. 366–375.

Wein, Shira; Bonn, Julia. Comparing UMR and Cross-lingual Adaptations of AMR. In: PROCEEDINGS of the Fourth International Workshop on Designing Meaning Representations. Nancy, France: Association for Computational Linguistics, 2023. P. 23–33.

Zerbinati, Michel Monteiro; Roman, Norton Trevisan; Di-Felippo, Ariani. A Corpus of Stock Market Tweets Annotated with Named Entities. In: Gamallo, Pablo et al. (Ed.). **Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1**. Santiago de Compostela, Galicia/Spain:

Association for Computational Linguistics, mar. 2024. P. 276–284. Disponível em: <https://aclanthology.org/2024.propor-1.28/>.

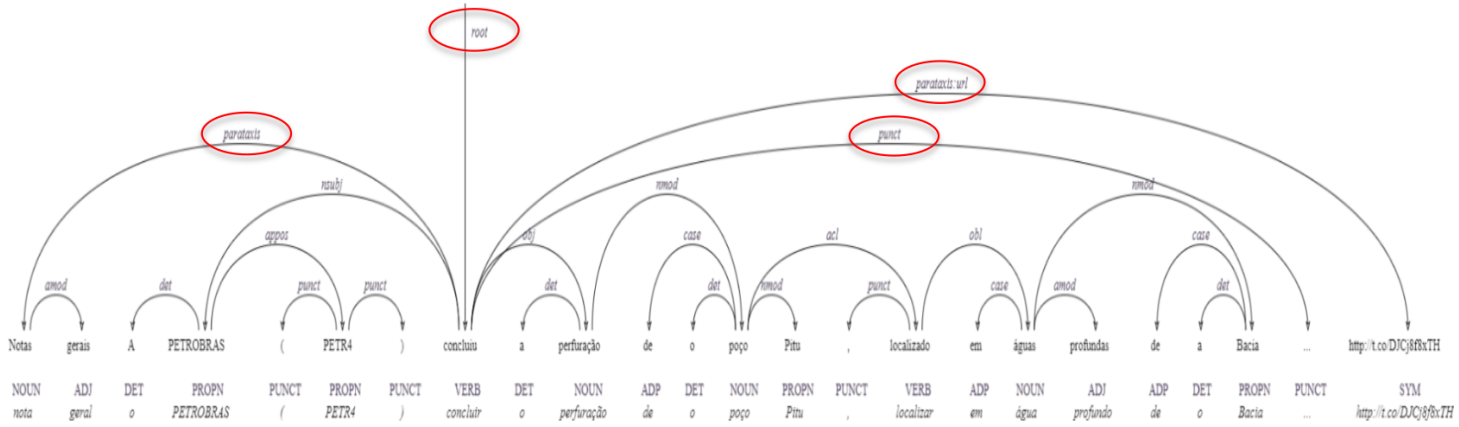
A Apêndice

A.1 Padrão 1

Formalização: notas gerais <sentença-truncada> ... <url>

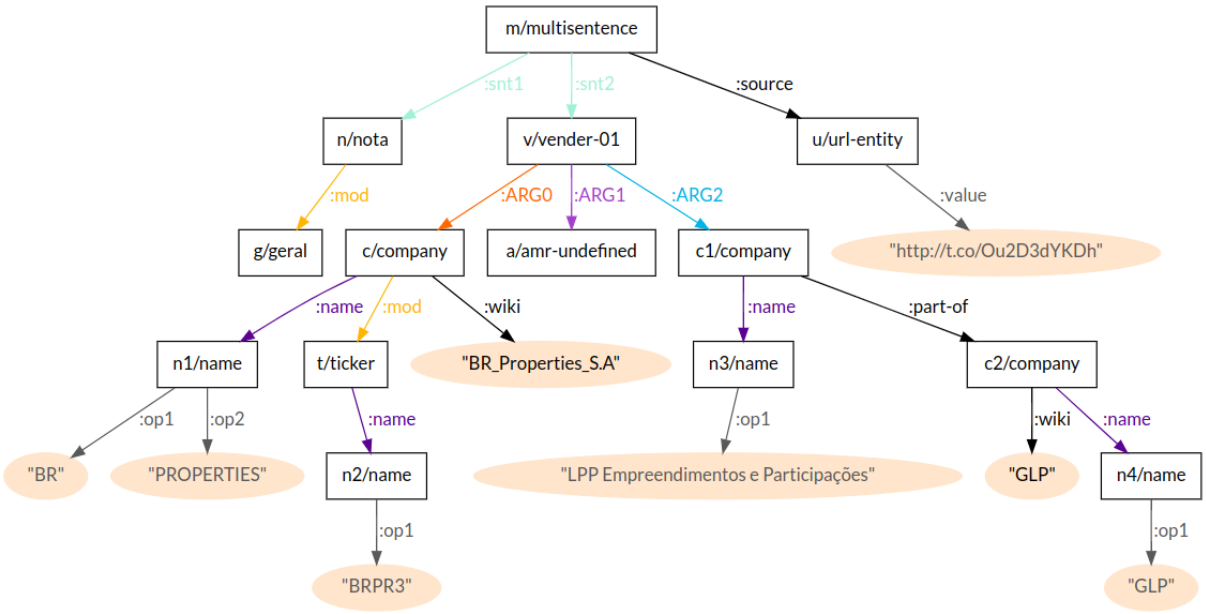
Exemplo: Notas gerais A BR PROPERTIES (BRPR3) vendeu a a LPP Empreendimentos e Participações , sociedade de o grupo GLP , a ... http://t.co/Ou2D3dYKDh

Figura A.1: Anotação-UD do Padrão 1.



Fonte: difelippo2024.

Figura A.2: Grafo AMR do Padrão 1.



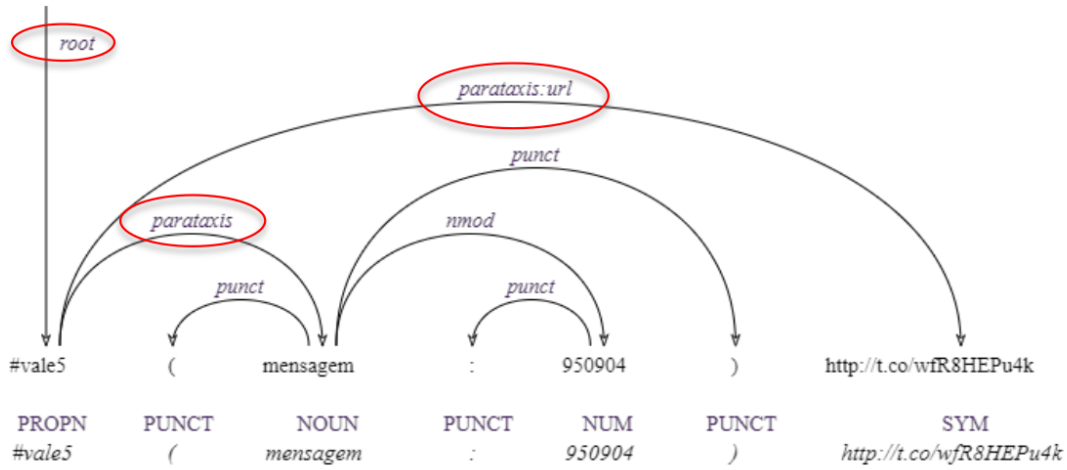
Fonte: O autor, 2025.

A.2 Padrão 2

Formalização: <hashtag-ticker>(mensagem: número)<url>

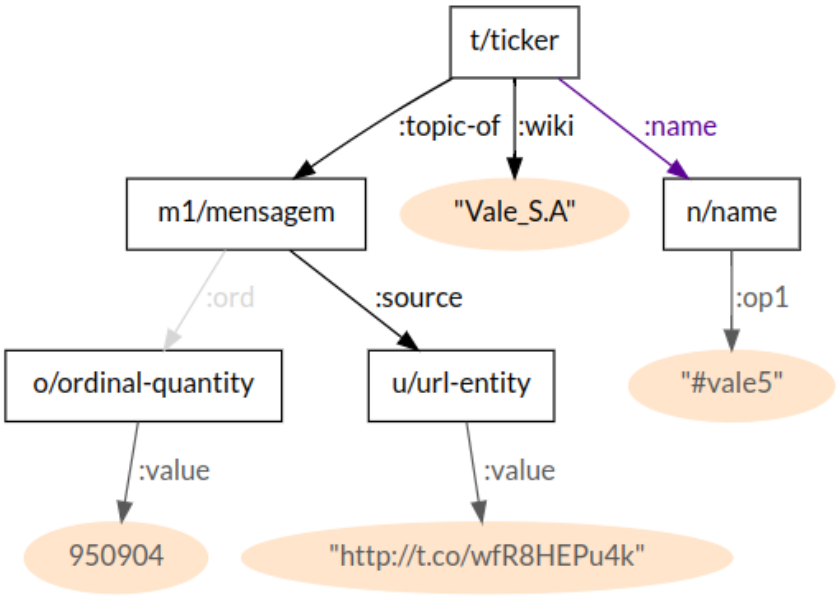
Exemplo: #vale5 (mensagem : 950904) http://t.co/wfR8HEPu4k

Figura A.3: Anotação-UD do Padrão 2).



Fonte: difelippo2024.

Figura A.4: Grafo AMR do Padrão 2.



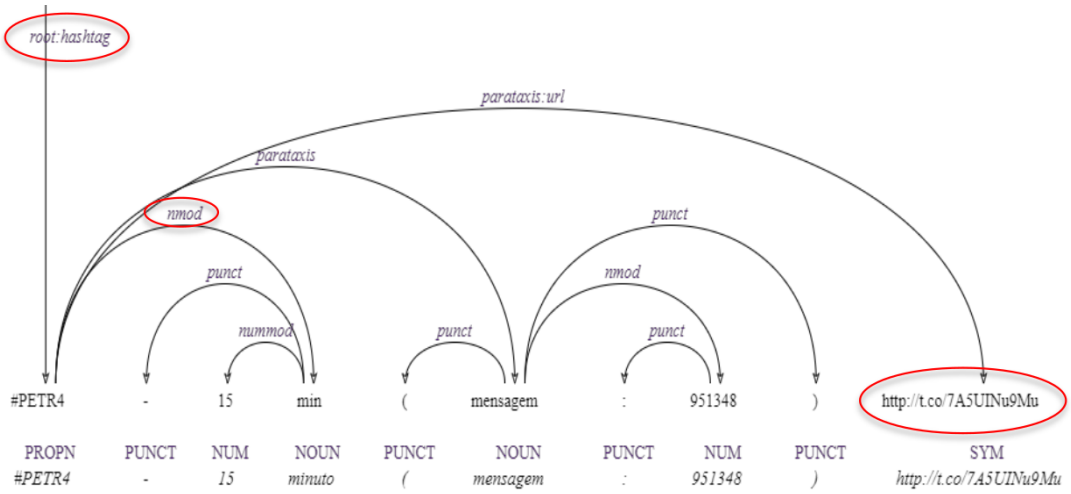
Fonte: O autor, 2025.

A.3 Padrão 3

Formalização: <hashtag-ticker><complemento>(mensagem: número)<url>

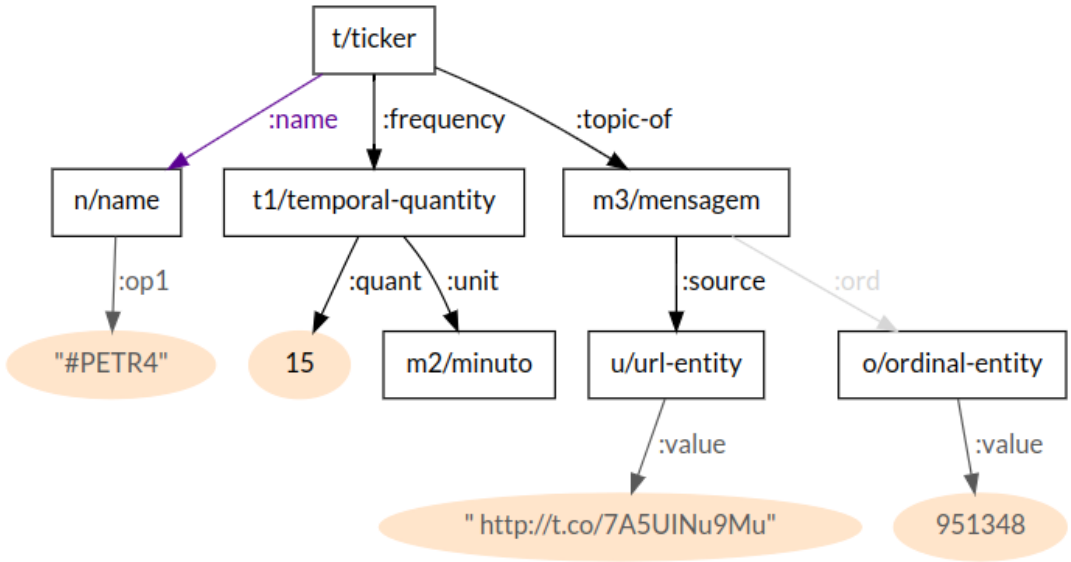
Exemplo: #PETR4 - 15 min (mensagem : 951348) http://t.co/7A5UINu9Mu

Figura A.5: Anotação-UD do Padrão 3.



Fonte: difelippo2024.

Figura A.6: Grafo AMR do Padrão 3.



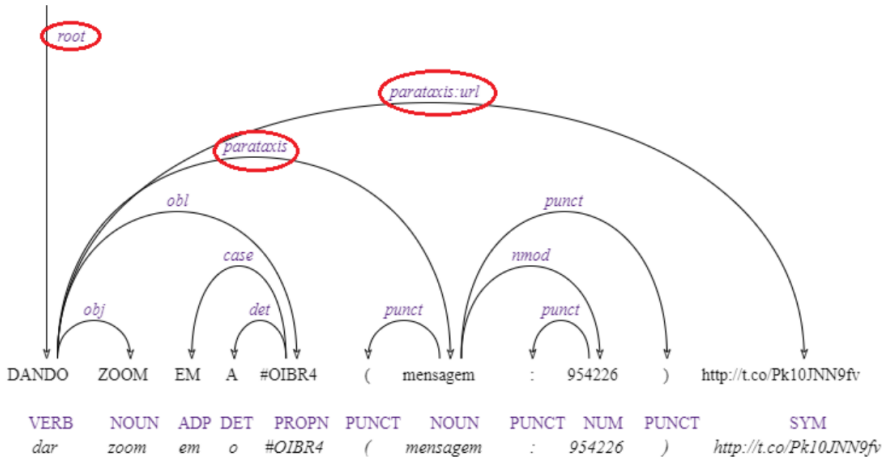
Fonte: O autor, 2025.

A.4 Padrão 4

Formalização: <sentença>(mensagem:número)...<url>

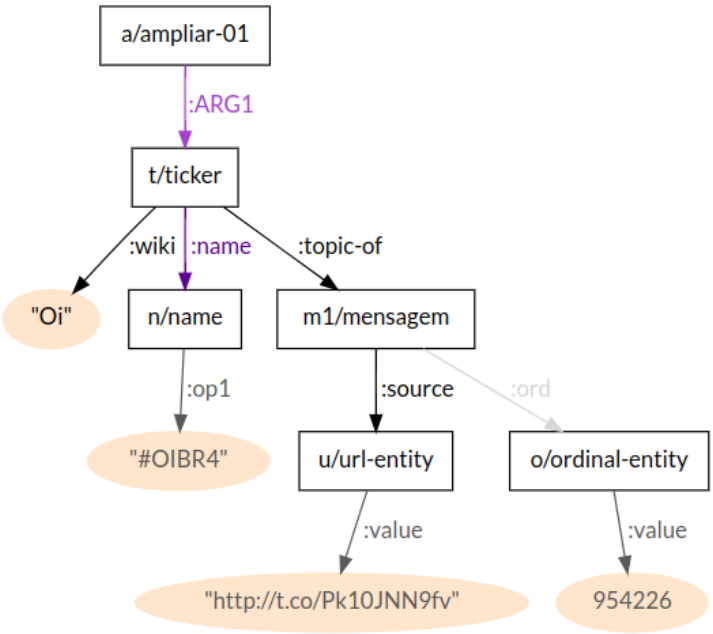
Exemplo: DANDO ZOOM EM A #OIBR4 (mensagem : 954226) http://t.co/Pk10JNN9fv

Figura A.7: Anotação-UD do Padrão 4.



Fonte: difelippo2024.

Figura A.8: Grafo AMR do Padrão 4.



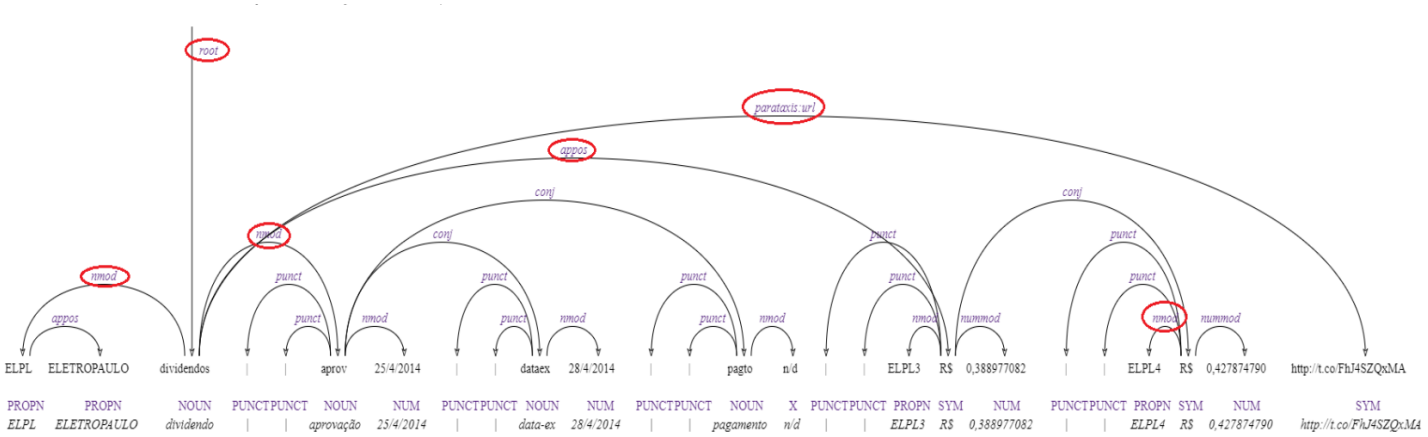
Fonte: O autor, 2025.

A.5 Padrão 5

Formalização: <ação&cia>dividendos<lista info&data><lista ticker&valor>||<url>

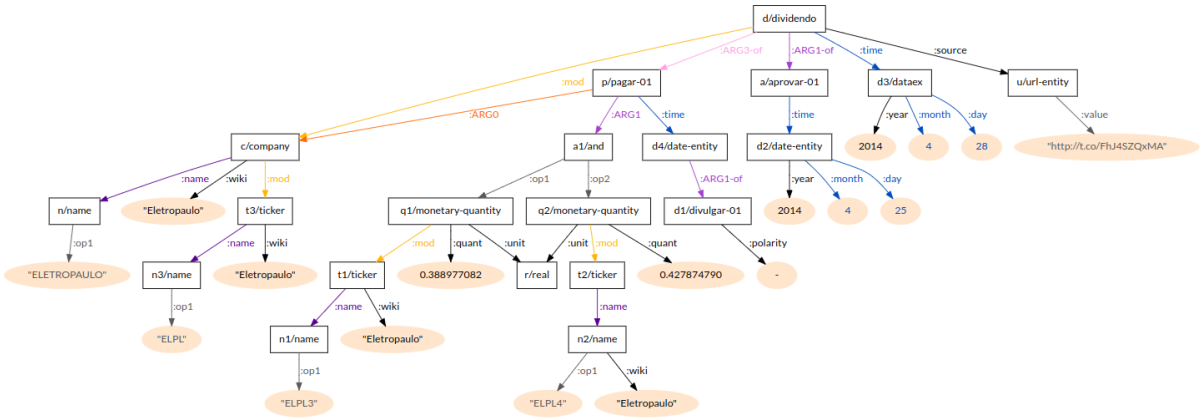
Exemplo: ELPL ELETROPAULO dividendos | | aprov 25/4/2014 | | dataex 28/4/2014 | | pagto n/d | | ELPL3 R$ 0,388977082 | | ELPL4 R$ 0,427874790 http://t.co/FhJ4SZQxMA

Figura A.9: Anotação-UD do Padrão 5.



Fonte: difelippo2024.

Figura A.10: Grafo AMR do Padrão 5.



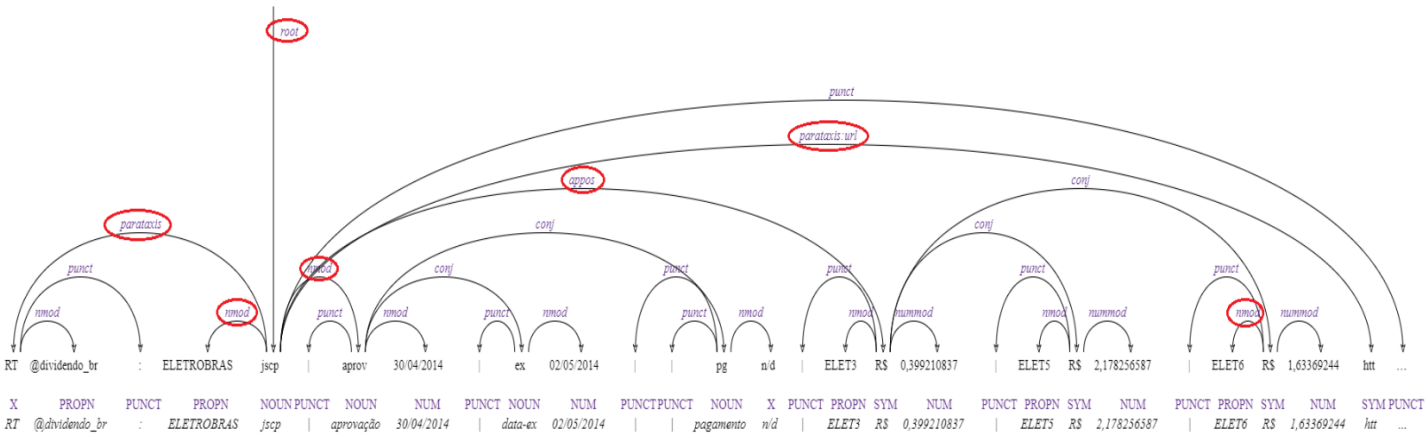
Fonte: O autor, 2025.

A.6 Padrão 6

Formalização: <RT @xxxx>:<ação&cia><dividendos/juros><lista info&data<lista ticker&valor>||<url>

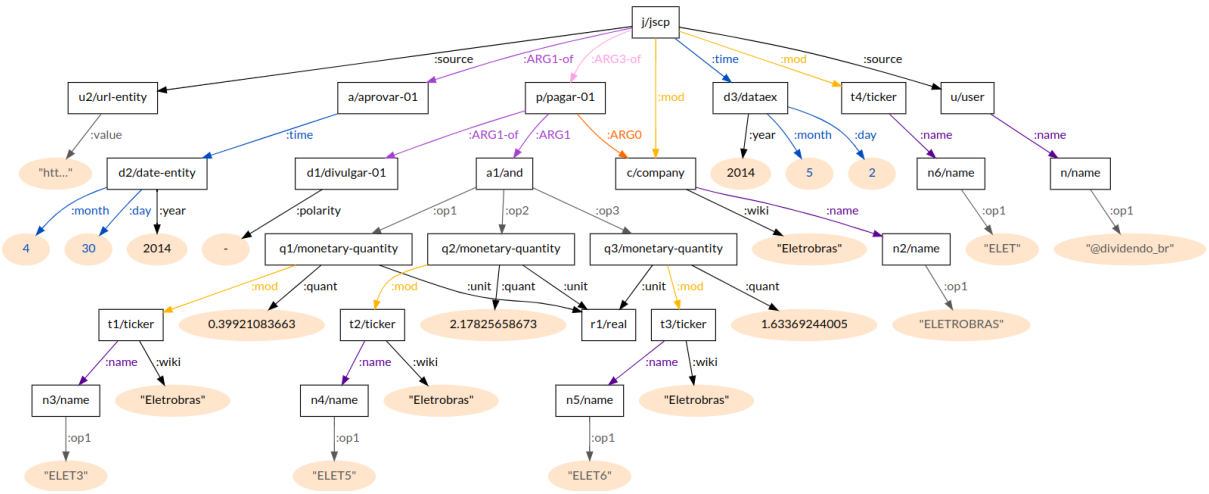
Exemplo: RT @dividendo_br : ELETROBRAS jscp | aprov 30/04/2014 | ex 02/05/2014 | | pg n/d | ELET3 R$ 0,399210837 | ELET5 R$ 2,178256587 | ELET6 R$ 1,63369244 htt

Figura A.11: Anotação-UD do Padrão 6.



Fonte: difelippo2024.

Figura A.12: Grafo AMR do Padrão 6.



Fonte: O autor, 2025.

A.7 Padrão 7

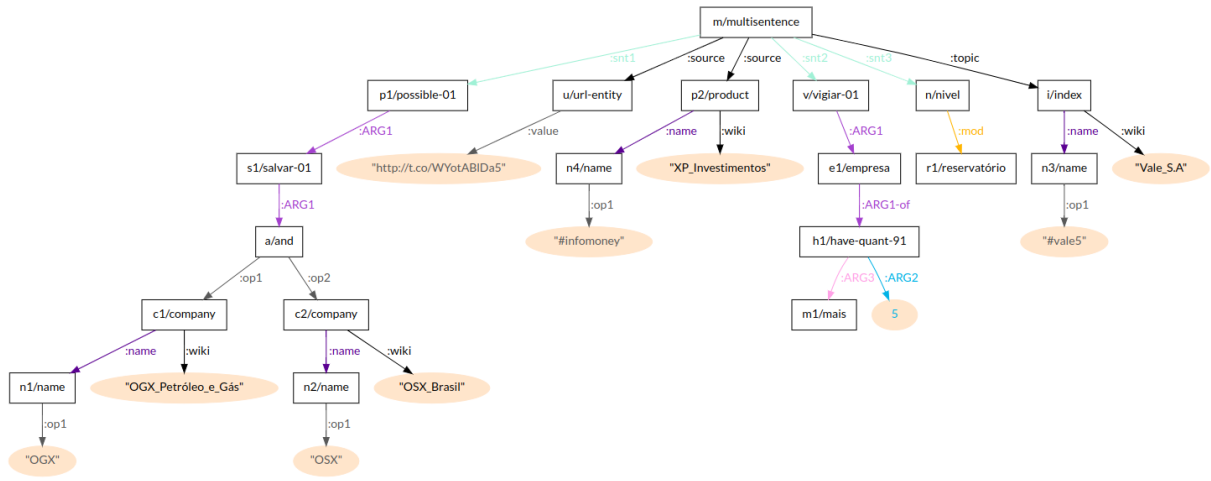
Formalização: <sentença> : <porção _truncada> ... <url> [hashtag opcional]

Exemplo: ‘ 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 A.13: Anotação-UD do Padrão 7.

Fonte: difelippo2024.

Figura A.14: Grafo AMR do Padrão 7.



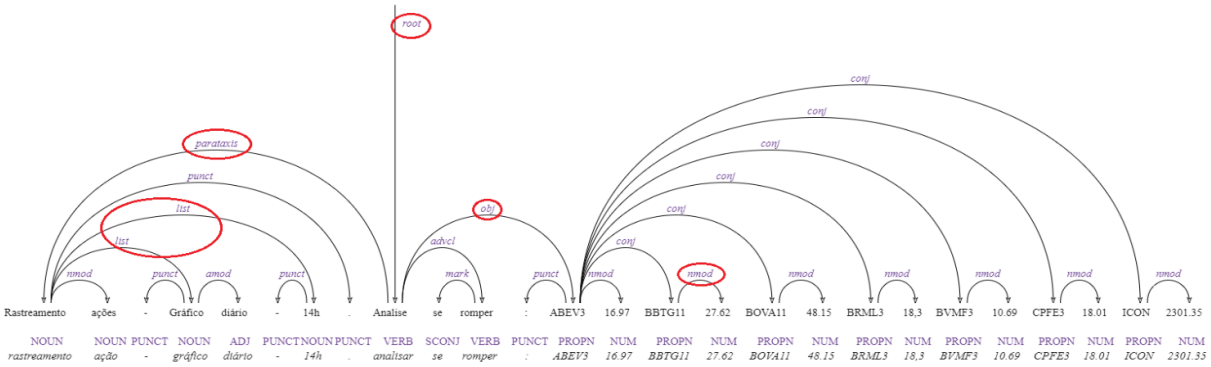
Fonte: O autor, 2025.

A.8 Padrão 8

Formalização: Rastreamento ações - Gráfico diário - 14h. Analise se romper : <lista ticker-valor>

Exemplo: 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 A.15: Anotação-UD do Padrão 8



Fonte: difelippo2024.

Figura A.16: Grafo AMR em PENMAN do Padrão 8

```
(m / multisentence
   :snt1 (r / rastrear-01
            :ARG1 (a / ação)
            :mod (g / gráfico
                    :frequency (r1 / rate-entity-91
                                    :ARG4 (t1 / temporal-quantity
                                              :unit (d / dia)
                                              :quant 1)))
            :time (t2 / temporal-quantity
                      :quant 14
                      :unit (h / hora)))
   :snt2 (a1 / analisar-01
             :condition (r2 / romper-03)
             :ARG1 (a2 / and
                       :op1 (t3 / ticker
                                :name (n1 / name
                                          :op1 "ABEV3")
                                :mod (m1 / monetary-quantity
                                         :quant "16,97")
                                :wiki "Ambev")
                       :op2 (t4 / ticker
                                :name (n2 / name
                                          :op1 "BBTG11")
                                :mod (m2 / monetary-quantity
                                         :quant "27,62")
                                :wiki "BTG_Pactual")
                       :op3 (t5 / ticker
                                :name (n3 / name
                                          :op1 "BOVA11")
                                :mod (m3 / monetary-quantity
                                         :quant "48,15")
                                :wiki "ETF_BOVA11")
                       :op4 (t6 / ticker
                                :name (n4 / name
                                          :op1 "BRML3")
                                :mod (m4 / monetary-quantity
                                         :quant "18,30")
                                :wiki "BRMalls")
                       :op5 (t7 / ticker
                                :name (n5 / name
                                          :op1 "BVMF3")
                                :mod (m5 / monetary-quantity
                                         :quant "10,69")
                                :wiki "B3_S.A.")
                       :op6 (t8 / ticker
                                :name (n6 / name
                                          :op1 "CPFE3")
                                :mod (m6 / monetary-quantity
                                         :quant "18,01")
                                :wiki "CPFL_Energia")
                       :op7 (t9 / ticker
                                :name (n7 / name
                                          :op1 "ICON")
                                :mod (m7 / monetary-quantity
                                         :quant "2301,35")
                                :wiki "Índice_Icon"))))
```

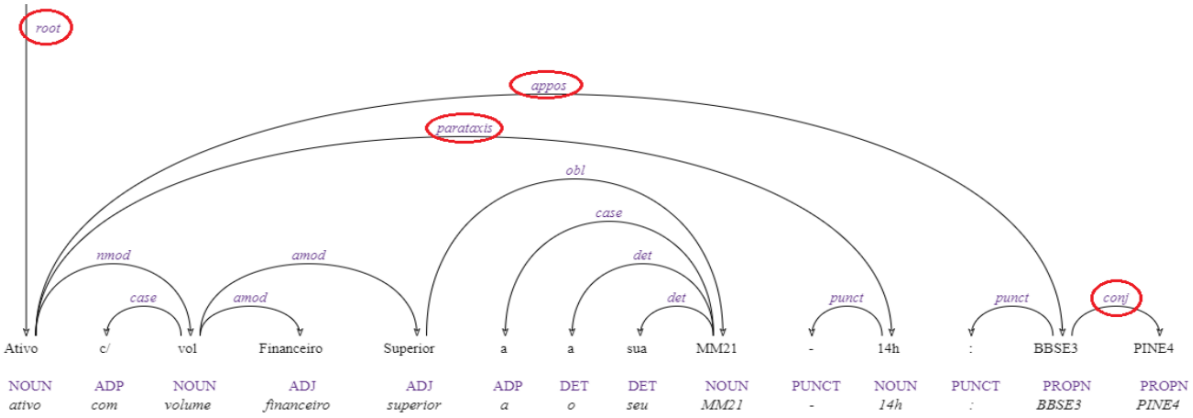
Fonte: O autor, 2025.

A.9 Padrão 9

Formalização: Ativo c/ vol Financeiro Superior a a sua MM21 - <hora> : <lista tickers>

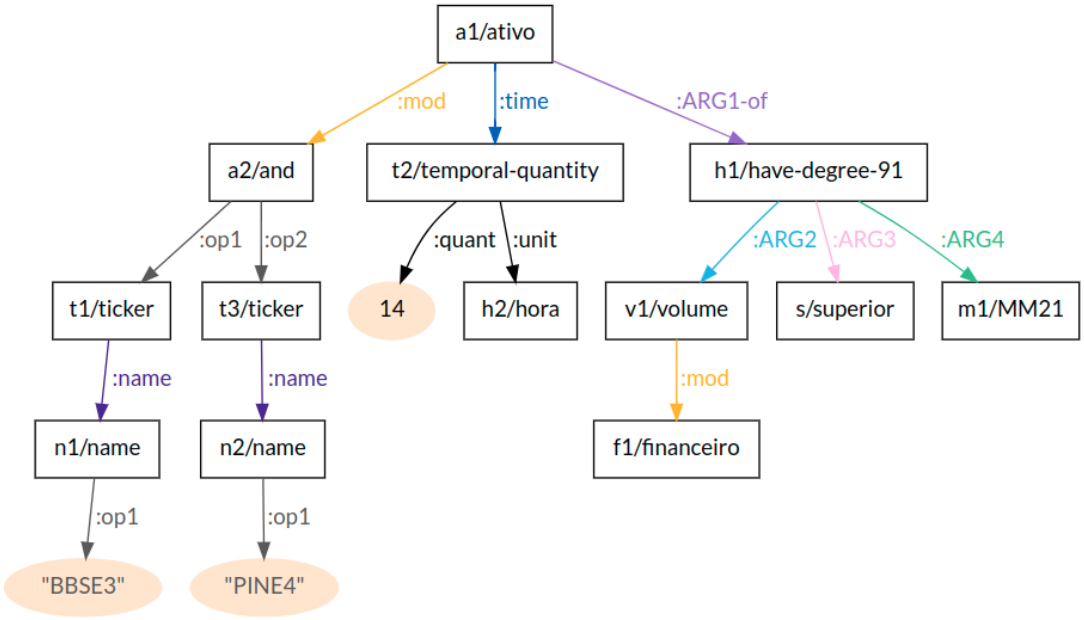
Exemplo: Ativo c/ vol Financeiro Superior a a sua MM21 - 14h : BBSE3 PINE4

Figura A.17: Anotação-UD do *tweet* em (19).



Fonte: difelippo2024.

Figura A.18: Grafo AMR do *Template 9*.



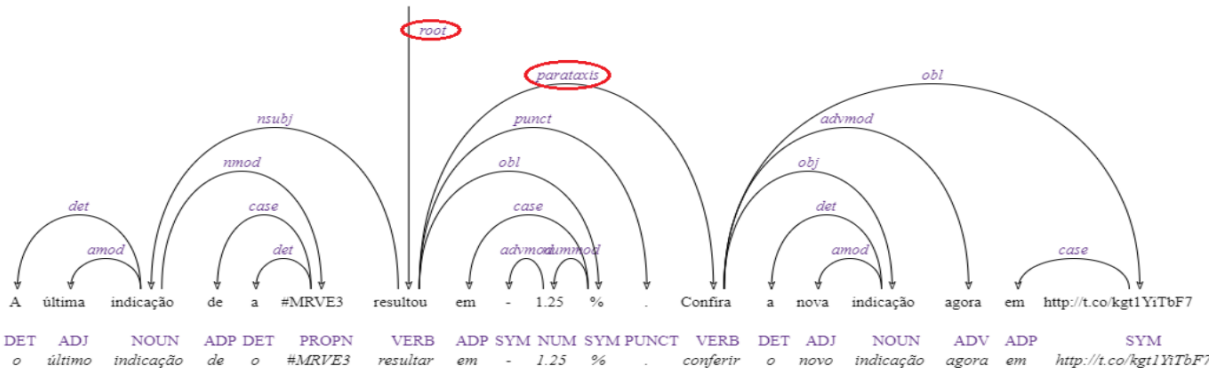
Fonte: O autor, 2025.

A.10 Padrão 10

Formalização: <sentença>. Confira a nova indicação agora em url

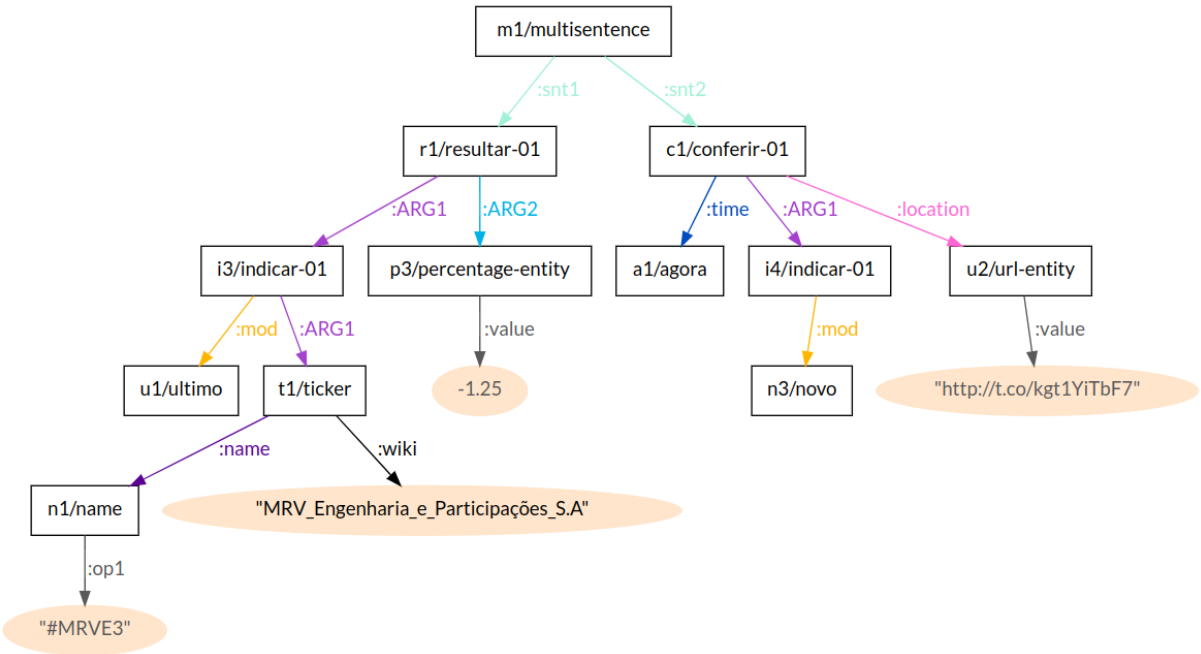
Exemplo: A última indicação de a #MRVE3 resultou em - 1.25 % . Confira a nova indicação agora em http://t.co/kgt1YiTbF7

Figura A.19: Anotação-UD do Padrão 10.



Fonte: difelippo2024.

Figura A.20: Grafo AMR do Padrão 10.



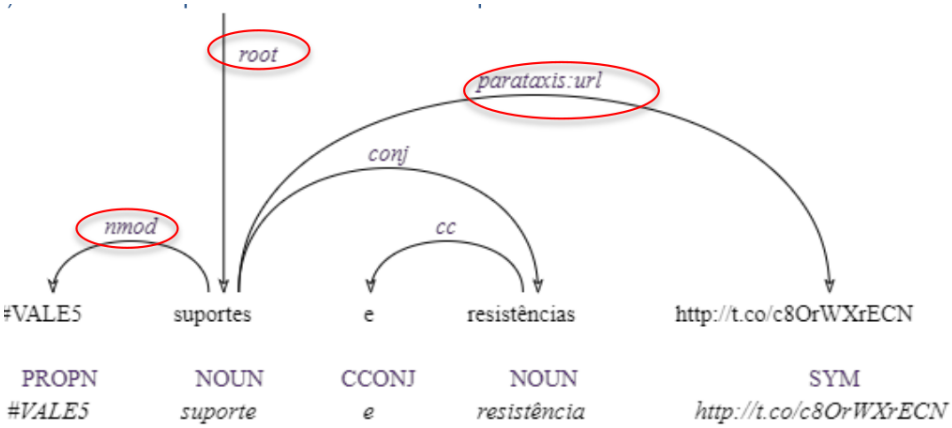
Fonte: O autor, 2025.

A.11 Padrão 11

Formalismo: <ticker><tema><url>

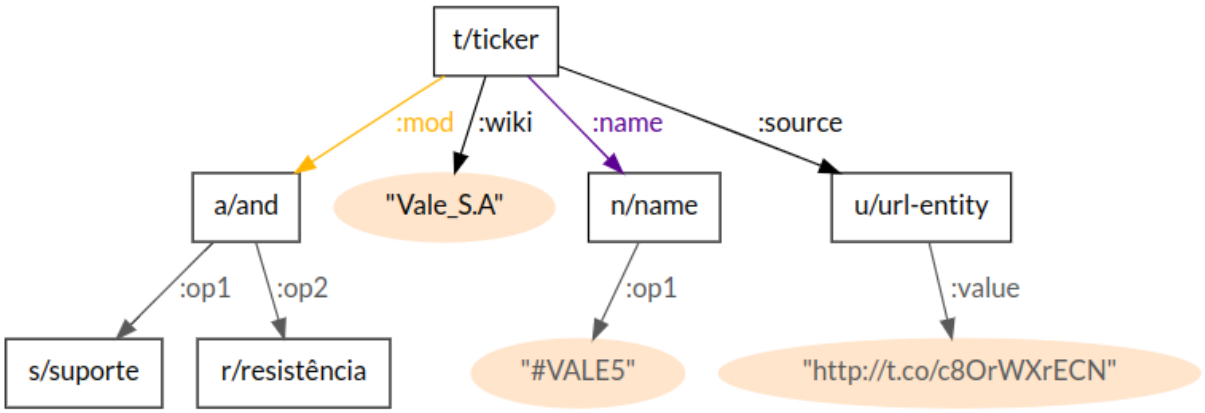
Exemplo: #VALE5 suportes e resistências http://t.co/c8OrWXrECN

Figura A.21: Anotação-UD do Padrão 11.



Fonte: difelippo2024

Figura A.22: Grafo AMR do Padrão 11.



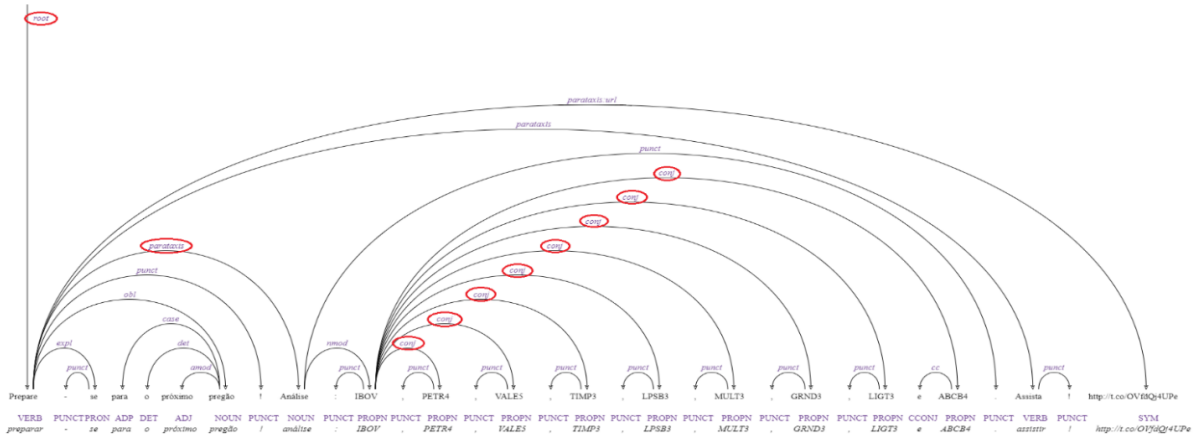
Fonte: O autor, 2025.

A.12 Padrão 12

Formalização: <prefixo> Prepare-se para o próximo pregão! <sufixo> <lista tickers>

Exemplo: Prepare-se para o próximo pregão! Análise: IBOV, PETR4, VALE5, TIMP3, LPSB3, MULT3, GRND3, LIGT3 e ABCB4. Assista! http://t.co/OVfdQj4UPe

Figura A.23: Anotação-UD do Padrão 12.



Fonte: difelippo2024

Figura A.24: Grafo AMR do Padrão 12.

```
(m / multisentence
   :snt1 (p1 / preparar-01
             :ARG1 (p2 / pregão
                       :mod (p3 / próximo))
             :mode imperative)
   :snt2 (a1 / analisar-01
             :ARG1 (a2 / and
                       :op1 (t1 / ticker
                                :name (n1 / name
                                          :op1 "IBOV")
                                :wiki "Ibovespa")
                       :op2 (t2 / ticker
                                :name (n2 / name
                                          :op1 "PETR4")
                                :wiki "Petróleo_Brasileiro_S.A.")
                       :op3 (t3 / ticker
                                :name (n3 / name
                                          :op1 "VALE5")
                                :wiki "Vale_S.A")
                       :op4 (t4 / ticker
                                :name (n4 / name
                                          :op1 "TIMP3")
                                :wiki "TIM_S.A")
                       :op5 (t5 / ticker
                                :name (n5 / name
                                          :op1 "LPSB3")
                                :wiki "LPS_Brasil_-_Consultoria_de_Imóveis_S.A.")
                       :op6 (t6 / ticker
                                :name (n6 / name
                                          :op1 "MULT3")
                                :wiki "Multiplan_Empreendimentos_Imobiliários_S.A.")
                       :op7 (t7 / ticker
                                :name (n7 / name
                                          :op1 "GRND3")
                                :wiki "Grendene")
                       :op8 (t8 / ticker
                                :name (n8 / name
                                          :op1 "LIGT3")
                                :wiki "Light_S.A.")
                       :op9 (t9 / ticker
                                :name (n9 / name
                                          :op1 "ABCB4")
                                :wiki "Banco_ABC_Brasil")))
   :snt3 (a3 / assistir-01
             :mode imperative
             :location (u / url-entity
                          :value "http://t.co/OVfdQj4UPe")))
```

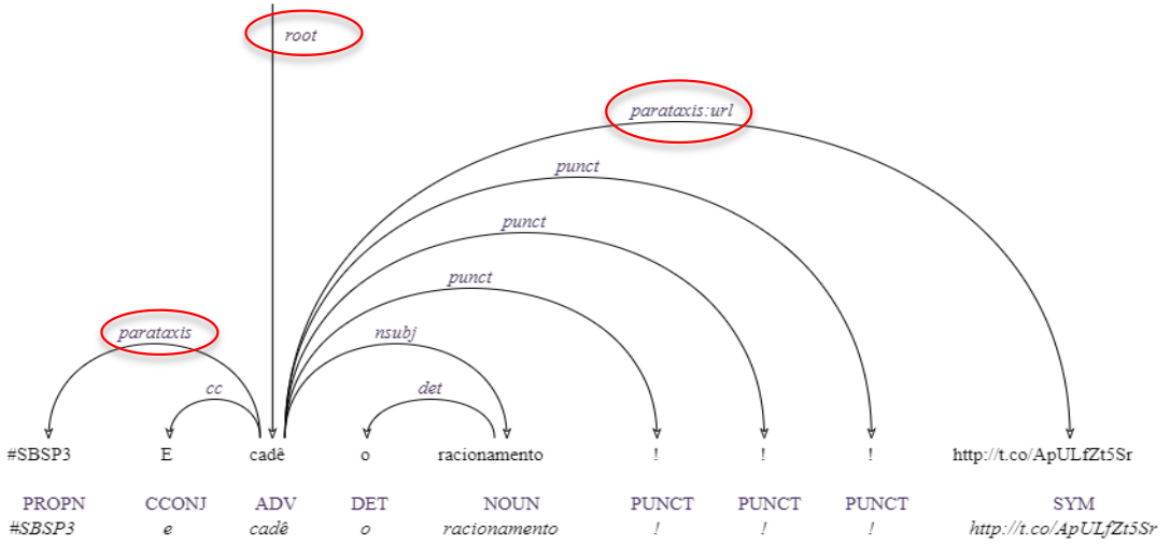
Fonte: O autor, 2025.

A.13 Padrão 13

Formalização: <prefixo> : <sentença> url

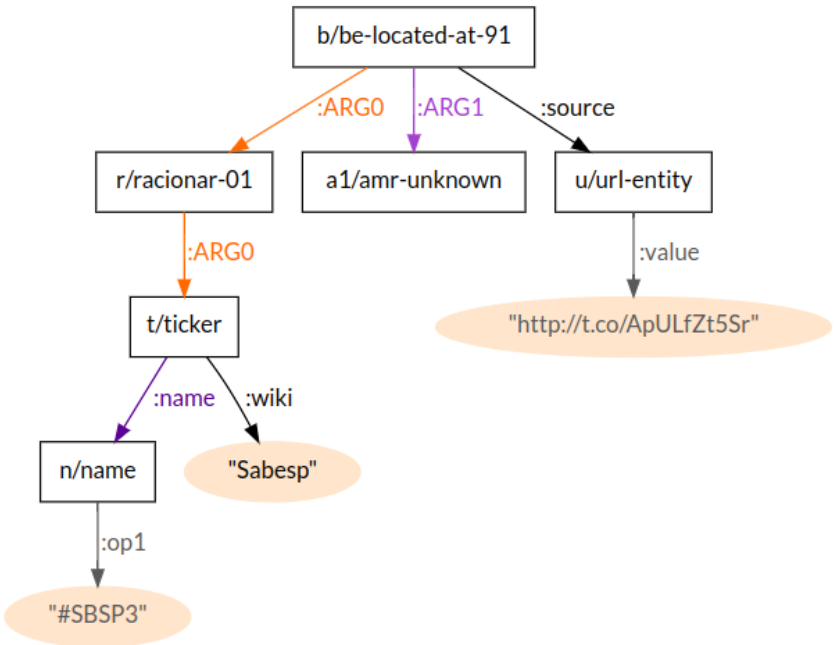
Exemplo: #SBSP3 E cadê o racionamento ! ! ! http://t.co/ApULfZt5Sr

Figura A.25: Anotação-UD do Padrão 13.



Fonte: difelippo2024.

Figura A.26: Grafo AMR do Padrão 13.



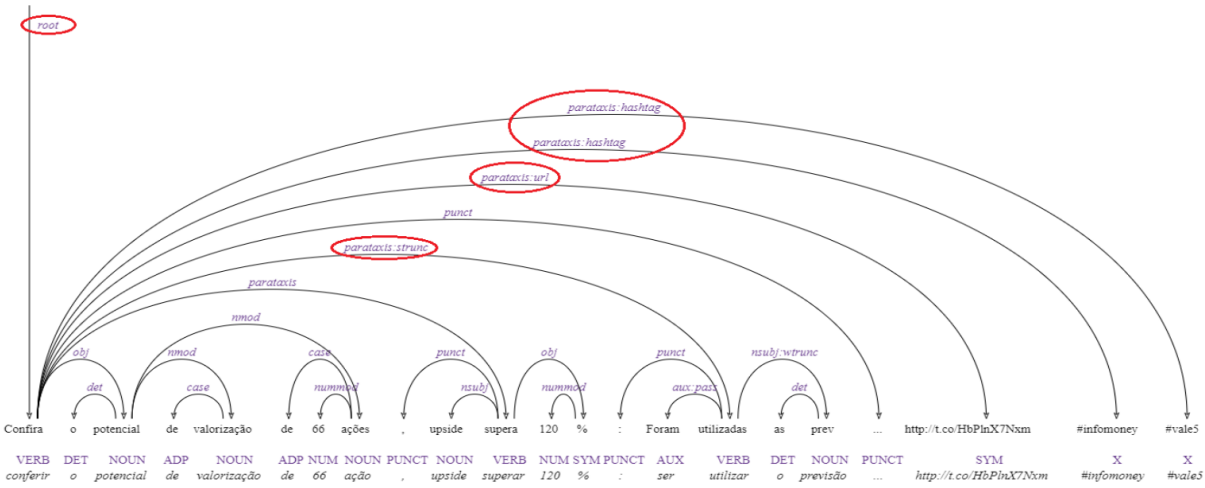
Fonte: O autor, 2025.

A.14 Padrão 14

Formalização: <sentença> : <sentença truncada> url <lista hashtags>

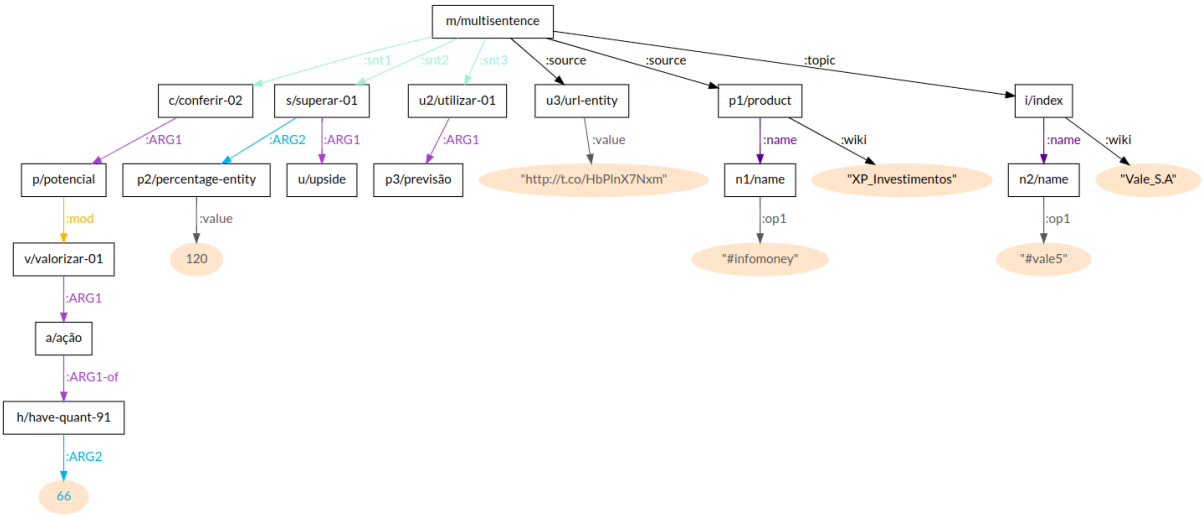
Exemplo: Confira o potencial de valorização de 66 ações , upside supera 120 % : Foram utilizadas as prev ... http://t.co/HbPlnX7Nxm #infomoney #vale5

Figura A.27: Anotação-UD do Padrão 14.



Fonte: difelippo2024.

Figura A.28: Grafo AMR do Padrão 14.



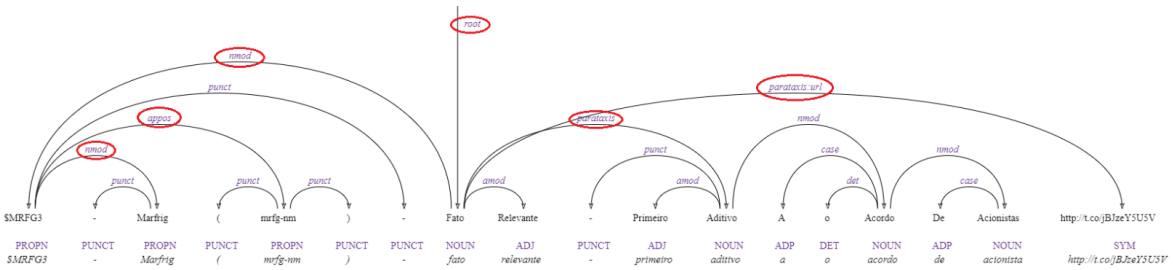
Fonte: O autor, 2025.

A.15 Padrão 15

Formalização: <cashtag> <empresa> (tipo de ação) - Fato Relevante - <fato> url

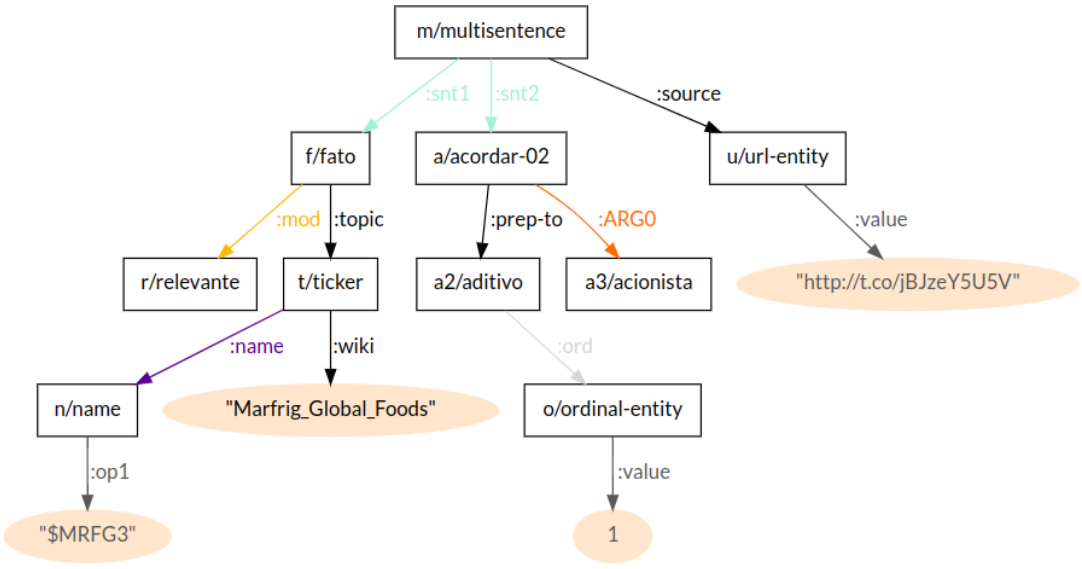
Exemplo: $MRFG3 - Marfrig (mrfg-nm) - Fato Relevante - Primeiro Aditivo A o Acordo De Acionistas http://t.co/jBJzeY5U5V

Figura A.29: Anotação-UD do Padrão 15.



Fonte: difelippo2024.

Figura A.30: Grafo AMR do Padrão 15.



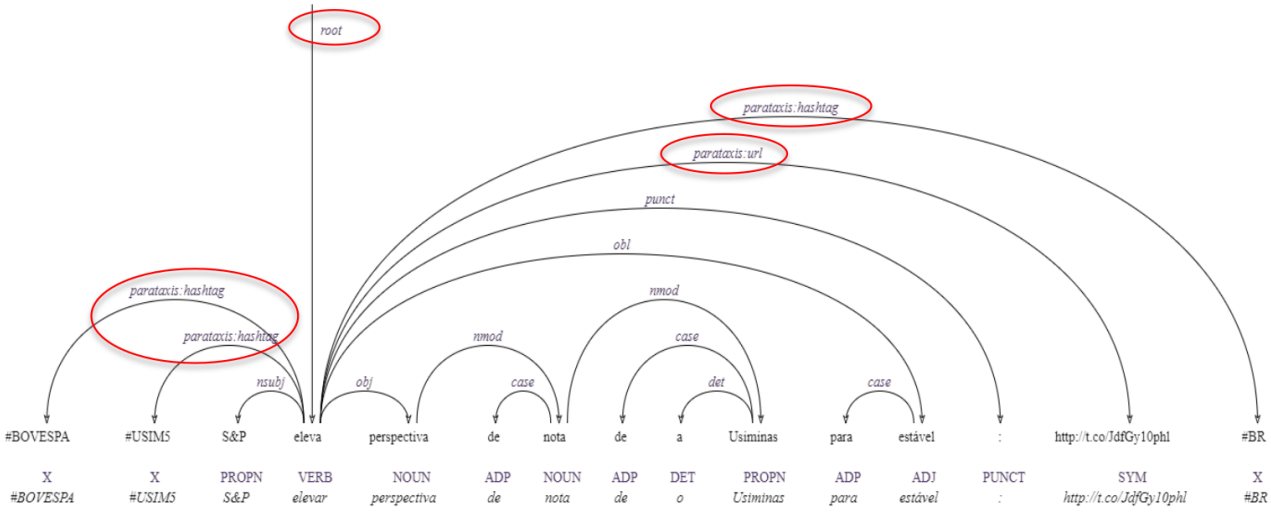
Fonte: O autor, 2025.

A.16 Padrão 16

Formalização: [prefixo opcional] <sentença> url

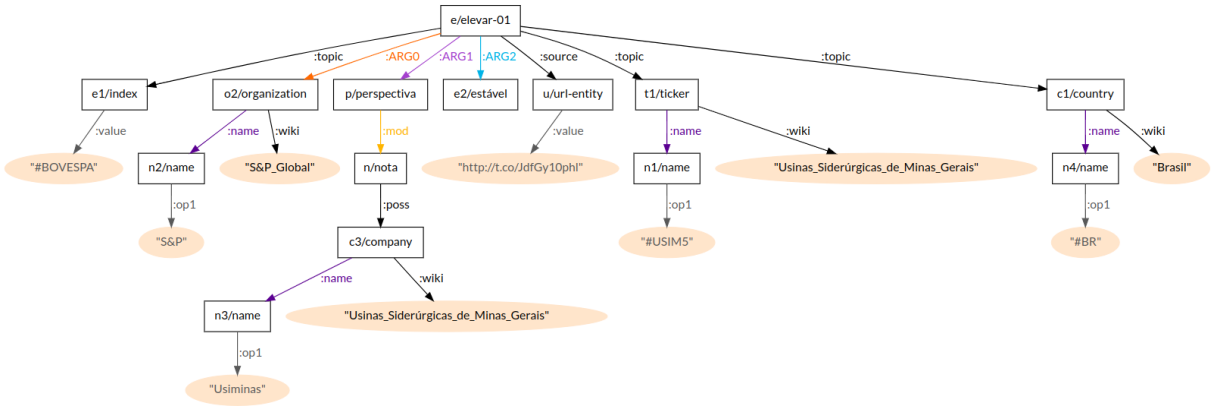
Exemplo: #BOVESPA #USIM5 S&P eleva perspectiva de nota de a Usiminas para estável : http://t.co/JdfGy10phl #BR

Figura A.31: Anotação-UD do Padrão 16.



Fonte: difelippo2024.

Figura A.32: Grafo AMR do Padrão 16.



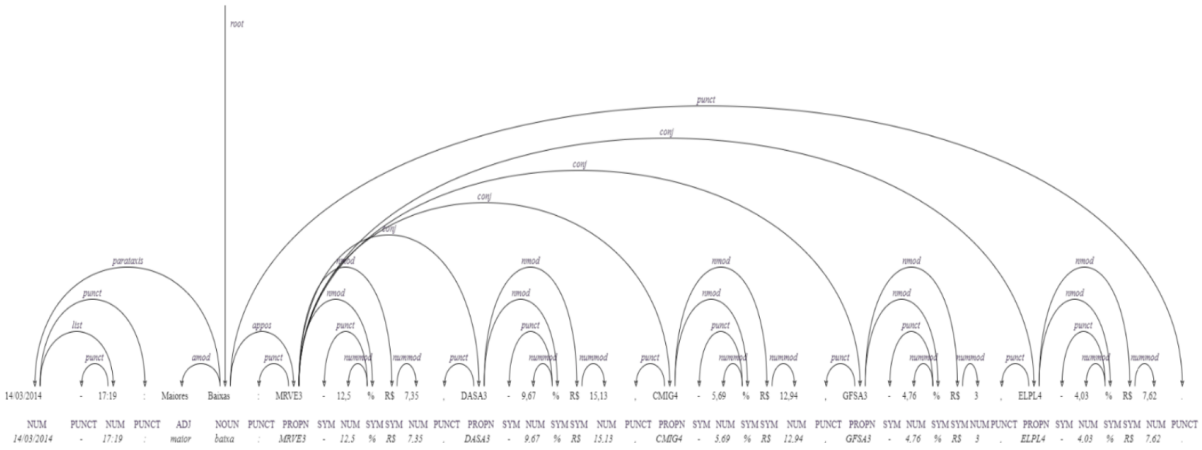
Fonte: O autor, 2025.

A.17 Padrão 17

Formalização: <lista ticker&[valor opcional]>

Exemplo: 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 A.33: Anotação-UD do Padrão 17..



Fonte: difelippo2024.

Figura A.34: Grafo AMR do Padrão 17.

```
(b / baixa
   :time (d / date-entity
            :year 2014
            :month 03
            :day 14
            :time "17:19")
   :degree (m1 / maior)
   :mod (a1 / and
            :op1 (t1 / ticker
                     :name (n1 / name
                               :op1 "MRVE3")
                     :wiki "MRV_Engenharia_e_Participações_S.A"
                     :value (p1 / percentage-entity
                                :value -12.5)
                     :ARG1-of (h1 / have-quant-91
                                  :ARG2 (m2 / monetary-quantity
                                            :quant "7,35"
                                            :unit (r / real))))
            :op2 (t2 / ticker
                     :name (n2 / name
                               :op1 "DASA3")
                     :wiki "Diagnósticos_da_America_S.A."
                     :value (p2 / percentage-entity
                                :value -9.67)
                     :ARG1-of (h2 / have-quant-91
                                  :ARG2 (m3 / monetary-quantity
                                            :quant "15,13"
                                            :unit r)))
            :op3 (t3 / ticker
                     :name (n3 / name
                               :op1 "CMIG4")
                     :wiki "Companhia_Energética_de_Minas_Gerais"
                     :value (p3 / percentage-entity
                                :value -5.69)
                     :ARG1-of (h3 / have-quant-91
                                  :ARG2 (m4 / monetary-quantity
                                            :quant "12,94"
                                            :unit r)))
            :op4 (t4 / ticker
                     :name (n4 / name
                               :op1 "GFSA3")
                     :wiki "Gafisa_S.A."
                     :value (p4 / percentage-entity
                                :value -4.76)
                     :ARG1-of (h4 / have-quant-91
                                  :ARG2 (m5 / monetary-quantity
                                            :quant "3,00"
                                            :unit r)))
            :op5 (t5 / ticker
                     :name (n5 / name
                               :op1 "ELPL4")
                     :wiki "Eletropaulo_Metropolitana_Eletricidade_de_São_Paulo_S.A."
                     :value (p5 / percentage-entity
                                :value -4.03)
                     :ARG1-of (h5 / have-quant-91
                                  :ARG2 (m6 / monetary-quantity
                                            :quant "7,62"
                                            :unit r)))))
```

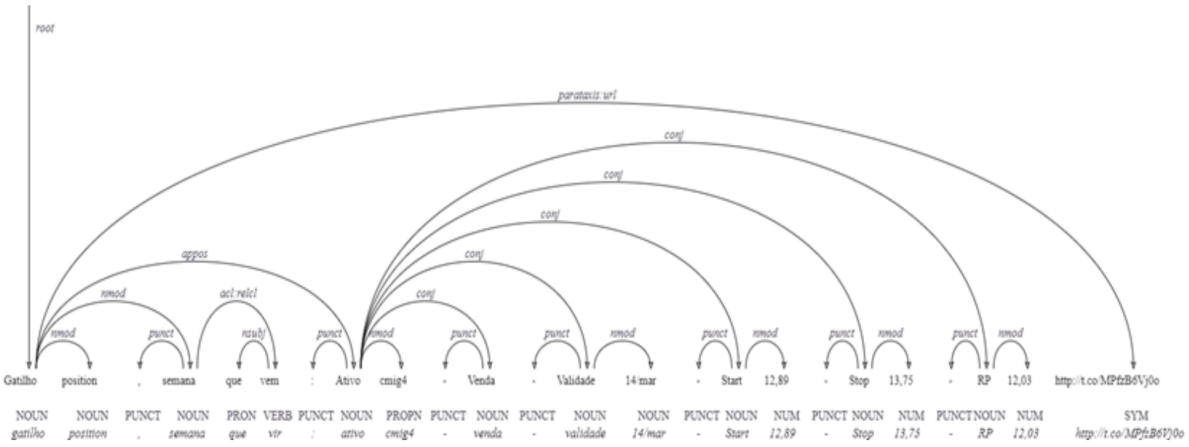
Fonte: O autor, 2025.

A.18 Padrão 18

Formalização: Gatilho position , semana que vem : Ativo <verbo> <validade-data> Start <valor> Stop <valor> RP <valor> <url>

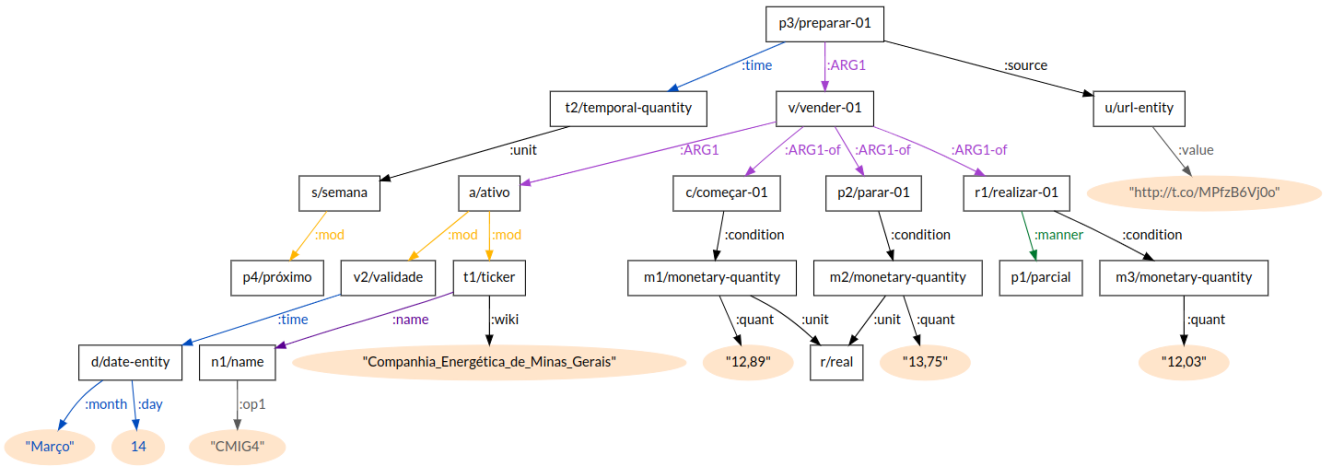
Exemplo: Gatilho position , semana que vem : Ativo cmg4 - Venda - Validade 14/mar - Start 12,89 - Stop 13,75 - RP 12,03 http://t.co/MPfzB6Vj0o

Figura A.35: Anotação-UD do Padrão 18.



Fonte: difelippo2024

Figura A.36: Grafo AMR do Padrão 18.



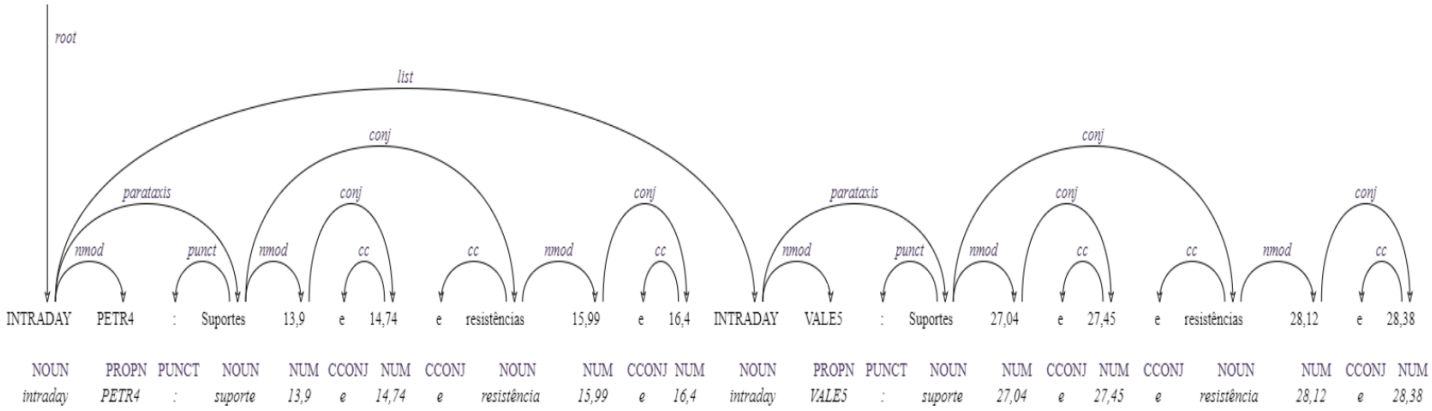
Fonte: O autor, 2025.

A.19 Padrão 19

Formalismo: INTRADAY <ticker>: Suportes <valor> e Resistências <valor> INTRADAY <ticker>: Suportes <valor> e Resistências <valor>

Exemplo: 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

Figura A.37: Anotação-UD do Padrão 19.



Fonte: difelippo2024

Figura A.38: Grafo AMR em PENMAN do Padrão 19.

```
(m / multisentence
   :snt1 (I1 / INTRADAY
             :mod (t1 / ticker
                      :name (n1 / name
                                :op1 "PETR4")
                      :wiki "Petróleo_Brasileiro_S.A."
                      :mod (a1 / and
                               :op1 (s1 / suporte
                                        :mod (a2 / and
                                                 :op1 (m1 / monetary-quantity
                                                          :quant "13,9"
                                                          :unit (r / real))
                                                 :op2 (m2 / monetary-quantity
                                                          :quant "14,74"
                                                          :unit r)))
                               :op2 (r1 / resistência
                                        :mod (a3 / and
                                                 :op1 (m3 / monetary-quantity
                                                          :quant "15,99"
                                                          :unit r)
                                                 :op2 (m4 / monetary-quantity
                                                          :quant "16,4"
                                                          :unit r))))))
   :snt2 (I2 / INTRADAY
             :mod (t2 / ticker
                      :name (n2 / name
                                :op1 "VALE5")
                      :wiki "Vale_S.A."
                      :mod (a4 / and
                               :op1 (s2 / suporte
                                        :mod (a5 / and
                                                 :op1 (m5 / monetary-quantity
                                                          :quant "27,04"
                                                          :unit r)
                                                 :op2 (m6 / monetary-quantity
                                                          :quant "27,45"
                                                          :unit r)))
                               :op2 (r2 / resistência
                                        :mod (a6 / and
                                                 :op1 (m7 / monetary-quantity
                                                          :quant "28,12"
                                                          :unit r)
                                                 :op2 (m8 / monetary-quantity
                                                          :quant "28,38"
                                                          :unit r)))))))
```

Fonte: O autor, 2025.

A.20 Padrão 20

Formalização: Ações ex-dividendos hoje : <lista-ticker>. As cotações históricas foram ajustadas . Saiba mais ! <url>

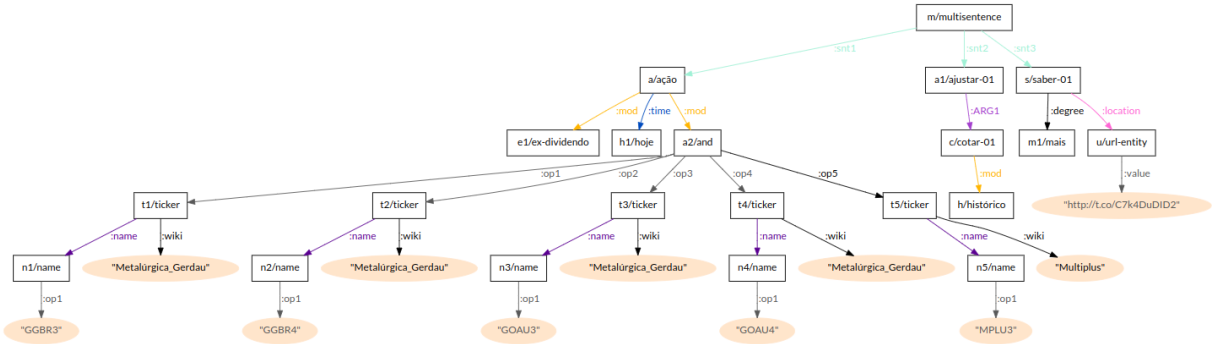
Exemplo: Ações ex-dividendos hoje : GGBR3 , GGBR4 , GOAU3 , GOAU4 e MPLU3 . As cotações históricas foram ajustadas . Saiba mais! http://t.co/C7k4DuDID2

Figura A.39: Anotação-UD do Padrão 20.

Fonte: difelippo2024

Figura A.40: Grafo AMR do Padrão 20

.



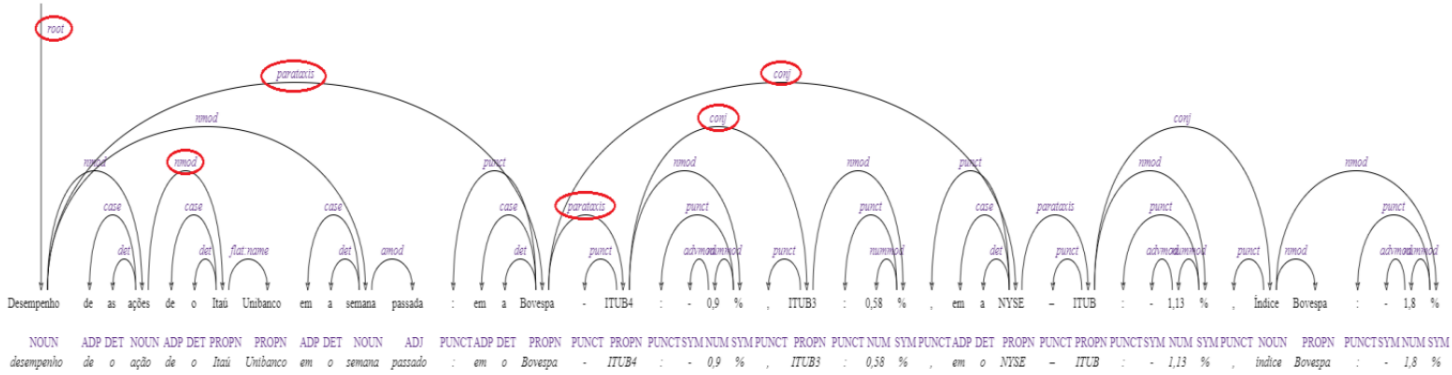
Fonte: O autor, 2025.

A.21 Padrão 21

Formalização: Desempenho de as ações <empresa> em a semana passada : [bolsa] <lista ticker&índice> , [bolsa <lista ticker&índice>]

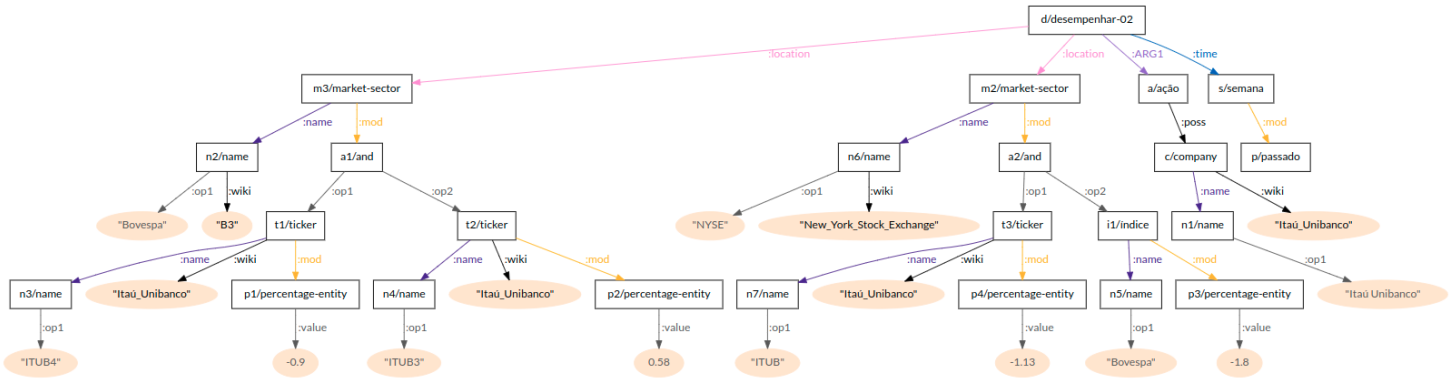
Exemplo: Desempenho de as ações de o Itaú Unibanco em a semana passada : em a Bovespa - ITUB4 : - 0,9 % , ITUB3 : 0,58 % , em a NYSE ITUB : - 1,13 % , Índice Bovespa : - 1,8%.

Figura A.41: Anotação-UD do Padrão 21.



Fonte: difelippo2024

Figura A.42: Grafo AMR do Padrão 21.



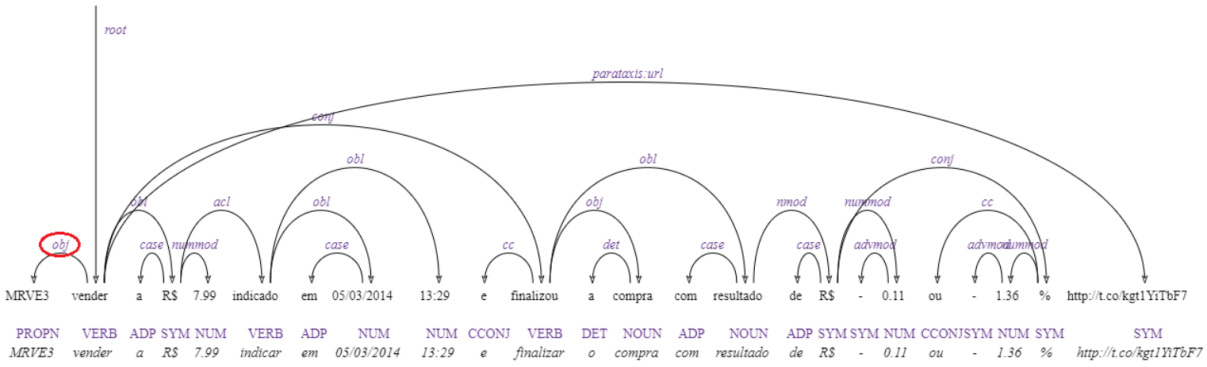
Fonte: O autor, 2025.

A.22 Padrão 22

Formalização: **<ticker> vender a <valor> indicado em <data-hora> e finalizou a compra com resultado de <valor> <url>**

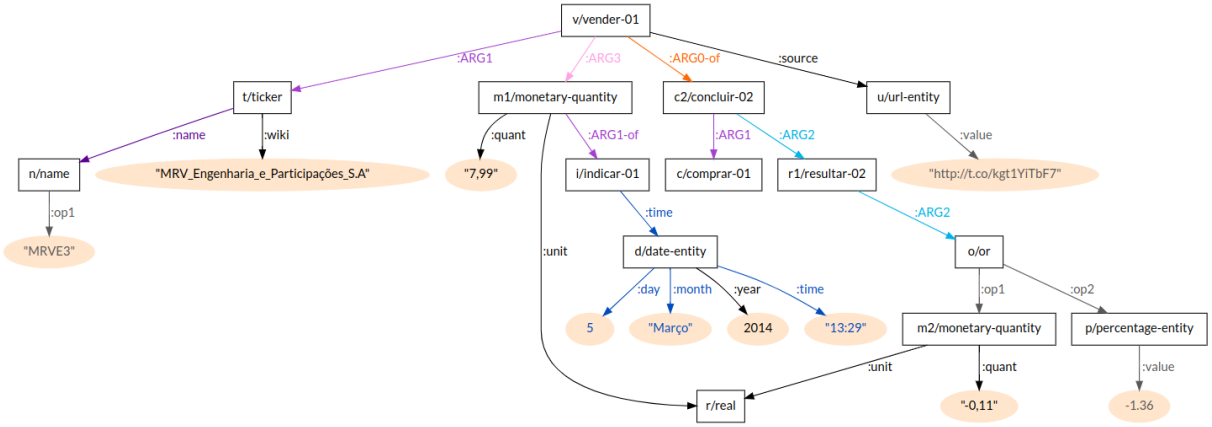
Exemplo: MRVE3 vender a R$ 7.99 indicado em 05/03/2014 13:29 e finalizou a compra com resultado de R$ - 0.11 ou - 1.36 % http://t.co/kgt1YiTbF7

Figura A.43: Anotação-UD do Padrão 22.



Fonte: difelippo2024

Figura A.44: Grafo AMR do Padrão 22.



Fonte: O autor, 2025.