DANTEStocks-AMR em Construção: Avanços e Desafios na Anotação Semântica de Tweets Financeiros

Gabriel Ceregatto[1,2], Ariani Di Felippo [1,2,3]

[1]Núcleo Interinstitucional de Linguística Computacional (NILC)
[2]Programa de Pós-Graduação em Linguística (PPGL/UFSCar)
[3]Departamento de Letras – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13565-905 – São Carlos – SP – Brasil

ceregattog@gmail.com, ariani@ufscar.br

Abstract. *Abstract Meaning Representation (AMR) is a formalism widely used in Natural Language Processing (NLP) to represent the meaning of utterances as directed graphs. This work presents the pioneering annotation of DANTEStocks-AMR, a corpus of 4,048 financial tweets in Portuguese, previously annotated with Universal Dependencies (UD). The AMR graphs are being built semi-automatically by adapting both the original English guidelines and those for Portuguese, considering the corpus's specificities and leveraging UD annotations. The paper discusses corpus-specific features that required adjustments to the AMR model and presents statistical data from DANTEStocks-AMR (v. 1), with one quarter of the tweets annotated.*

Resumo. *O formalismo Abstract Meaning Representation (AMR) é amplamente utilizado no Processamento de Língua Natural (PLN) para representar o significado de enunciados em grafos orientados. Este trabalho se insere na tarefa pioneira de anotação do corpus DANTEStocks-AMR, composto por 4.048 tweets do mercado financeiro já anotados com Universal Dependencies (UD). Os grafos estão sendo construídos de forma semiautomática, com adaptações das diretrizes originais do inglês e daquelas já definidas para o português às especificidades do corpus, com apoio da anotação-UD. O artigo discute características do corpus que exigiram ajustes no modelo AMR, e apresenta estatísticas do DANTEStocks-AMR (v.1), com um 1/4 dos tweets anotados.*

1. Introdução

Diante da importância do Twitter/X como fonte de dados e opinião pública, *corpora* anotados de *tweets*/*posts* têm sido desenvolvidos em diferentes línguas com o objetivo de criar ferramentas para o processamento automático desse gênero de "conteúdo gerado por usuário" (CGU). Isso se deve pelas características da linguagem não-canônica usada na plataforma, marcada por ortografia e sintaxe irregulares, pontuação assistemática, uso de marcas específicas da rede, além de um estilo frequentemente econômico, fragmentado e coloquial. Tais aspectos dificultam o desempenho de modelos treinados em textos formais e exigem abordagens linguísticas específicas [Sanguinetti *et al*., 2023].

Em português, destaca-se o DANTEStocks [Di Felippo, Roman, 2025], *corpus* pioneiro de 4.048 *tweets* do mercado financeiro com anotação padrão-ouro segundo o modelo *Universal Dependencies* (UD) [de Marneffe *et al*. 2021]. A anotação-UD morfossintática (*part-of-speech* ou PoS tags) e sintática (dependências) já subsidiaram o desenvolvimento de um *tagger* [Silva *et al*. 2021] e dois *parsers*, sendo um dedicado aos

tweets [Di Felippo *et al*. 2024a] e outro multigênero [Di Felippo *et al*. 2024b]. Quanto à semântica, o *corpus* possui anotações de emoção [Silva *et al*., 2020] e entidades nomeadas [Zerbinatti *et al*., 2024], as quais podem subsidiar o desenvolvimento de reconhecedores de entidades nomeadas e classificadores de emoção. No entanto, essas ferramentas geralmente realizam uma interpretação superficial dos enunciados (*shallow semantic analysis*).

Para avançar na interpretação de *tweets*, tem-se realizado a anotação do DANTEStocks segundo o modelo *Abstract Meaning Representation* (AMR) [Banarescu *et al*., 2013], que representa o significado de um enunciado em um grafo conceitual, composto por nós (conceitos) e arestas (relações semânticas entre eles). Com isso, busca-se enriquecer o *corpus* com representações que possibilitem a exploração de métodos voltados à análise semântica profunda (*deep semantic analysis*).

Atualmente, há diversos *corpora* com anotação-AMR padrão-ouro em várias línguas, inclusive o português [Anchiêta; Pardo, 2018; Sobrevilla Cabezudo; Pardo, 2019; Seno *et al*., 2022; Inácio *et al*., 2023]. Tais recursos vêm sendo construídos de forma manual ou semiautomática, possibilitando a investigação de diversos métodos de *parsing* AMR. Para a anotação dos grafos, conta-se atualmente com um conjunto de diretrizes publicamente disponíveis, definidas a partir de dados em inglês padrão ou formal. Nesse sentido, a anotação AMR em outras línguas e/ou em gêneros que não seguem a norma escrita tradicional requer adaptações às *guidelines* originais, a fim de contemplar fenômenos específicos dessas variações linguísticas.

Neste artigo, analisam-se particularidades estruturais do DANTEStocks que exigiram ajustes no modelo AMR, e apresentam-se as estatísticas do DANTEStocks-AMR (v.1), correspondente a 25% dos tweets anotados. Na Seção 2, os principais construtos do modelo AMR são descritos. Na Seção 3, apresentam-se trabalhos correlatos à tarefa de anotação AMR. Na Seção 4, descreve-se o DANTEStocks e a metodologia de anotação AMR. Na Seção 5, discutem-se alguns fenômenos estruturais do *corpus* e suas respectivas propostas de anotação-AMR. Na Seção 6, apresentam-se as estatísticas referentes ao estágio atual de anotação do *corpus*. Por fim, na Seção 7, considerações finais são feitas, destacando contribuições e limitações, além de trabalhos futuros.

2. O Modelo *Abstract Meaning Representation*

Trata-se de um paradigma de representação semântica que abstrai aspectos sintáticos da sentença (como ordem das palavras, categorias gramaticais, flexões morfológicas, funções sintáticas e palavras funcionais como artigos e preposições) e foca nos conceitos centrais e nas relações entre eles [Banarescu *et al*., 2013]. A representação-AMR de um enunciado é um grafo enraizado, direcionado e acíclico, composto por nós (conceitos) e arestas (relações semânticas entre os conceitos) (Figura 1). Os conceitos podem ser (i) itens lexicais (p.ex.: *boy*), (ii) estruturas predicado-argumento (p.ex.: *want-01*), (iii) ou palavras-chave do modelo, que indicam tipos especiais de entidade (p.ex.: *date-entity*, *world-region*, etc.), quantidades (p.ex.: *monetary-quantity*, *distance-quantity*, etc.) e conjunções lógicas (p.ex.: *and*, etc.). Na Figura 1, os conceitos são *want-01*, *boy*, *believe-01* e *girl* e as relações :ARG0 e :ARG1. As variáveis *w*, *b*, *b1* e *g* são identificadores únicos dos conceitos, que permitem representar correferências e manter a estrutura conectada do grafo. Uma representação AMR também pode ser codificada de outras duas formas (Figura 2): lógica de primeira ordem e notação PENMAN [Bateman *et al*., 1991].

Os conceitos representados por estruturas predicado-argumento, como want-01 e belive-01, são provenientes de um recurso externo, no caso, do repositório de frame files do PropBank [Palmer *et al*., 2005]. Diante do sentido de um verbo representado por um frame ou conjunto de argumentos numerados (Arg0 a Arg5) (*roleset*) (Figura 3), o anotador estrutura a representação AMR do enunciado. Para o português, esse repositório é Verbo-Brasil [Duran *et al.*, 2013].

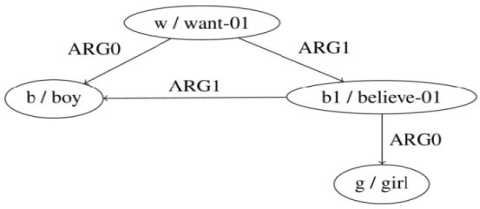

Figura 1. Grafo AMR da sentença *The boy wants the girl to believe him.*

```
∃ 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

Figura 2. Demais formatos de anotação AMR.

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

Figura 3. *Roleset* de *want.01*.

3. Trabalhos Relacionados

Entre os recursos AMR para o inglês, tem-se o AMR 3.0[1], com 59.255 mil sentenças de diferentes gêneros, o TLP [Banarescu *et al*., 2013], contendo 1.562 sentenças do livro *The Little Prince* de Saint-Exupéry, e o Bio-AMR [May, Priyadarshi, 2017], que compreende 6.452 sentenças de artigos científicos. Com relação a outras línguas, há iniciativas que se dedicam, sobretudo, à anotação-AMR de versões do livro *The Little Prince* em chinês, espanhol, vietnamita, coreano, turco e persa [Wein, Bonn, 2023]. Além desses recursos monolíngues, existe um *corpus* multilíngue, o AMR 2.0 – *Four Translations*[2].

Em português, o primeiro recurso anotado com AMR foi o AMR-LittlePrince [Anchiêta, Pardo, 2018], com 1.527 sentenças da tradução brasileira do referido conto. Esse recurso resulta do alinhamento automático com a versão inglesa e revisão manual. Outros dois *corpora* são o AMRNews, com 870 sentenças jornalísticas, e o OpiSums-PT-AMR, com 404 sentenças opinativas e resumos de comentários sobre livros e eletrônicos [Inácio *et al*., 2023]. Este último, aliás, é o mais próximo do DANTEStocks por também englobar um gênero CGU. Ambos foram anotados manualmente, com adaptações das *guidelines* originais ao português [Sobrevilla Cabezudo, Pardo, 2019]. O AMRScien-Br-

[1] https://catalog.ldc.upenn.edu/LDC2020T02
[2] https://catalog.ldc.upenn.edu/LDC2020T07

Corpus [Seno *et al*., 2022] reúne 200 sentenças de divulgação científica traduzidas do inglês via mapeamento de grafos AMR, seguidas de pós-edição humana.

4. O *Corpus* DANTEStocks e a Metodologia de Anotação AMR

Os 4.048 *tweets* (81.037 *tokens*) que compõem o DANTEStocks são sobre as 73 ações do Ibovespa[3] [Di-Felippo, Roman, 2025]. Os *posts* foram compilados em 2014 com base na ocorrência de ao menos um *ticker*[4] e possuem limite de 140 caracteres. Os *tweets* estão em sua forma original, isto é, sem segmentação em unidades estruturais menores (sentenças ou sintagmas) e normalização. Assim, o *corpus* possui uma mistura de linguagem padrão e não-padrão [Scandarolli *et al.*, 2023; Di Felippo *et al*. 2024a]. A não-canonicidade do *corpus* é marcada por ortografia e sintaxe irregulares, pontuação assistemática, uso de elementos próprios da plataforma (como *hashtags*, menções, etc.), estilo econômico, fragmentado e coloquial, além de elementos típicos do mercado financeiro como *cashtags* (p.ex.: $BBDC3) e *tickers*. A anotação-UD em nível sintático [Di-Felippo *et al*., 2024a] está ilustrada na Figura 4, uma vez que foi usada como ponto de partida para a interpretação dos *tweets* e representação AMR. Nesse nível, a anotação consiste em relações de dependência (deprels) entre palavras codificadas por setas direcionadas. A representação básica é uma árvore, em que uma palavra é a raiz (*root*) (p.ex.: "assinado"), e as demais dependem de outra palavra.

A anotação-AMR partiu da organização dos *tweets* em 3 grupos com base no tipo de linguagem [Barbosa, 2024]: (i) *tweets* com linguagem relativamente padrão, constituídos por uma ou mais sentenças bem estruturadas (1), (ii) *tweets* com padrões estruturais recorrentes (2), e (iii) *tweets* com estrutura variada ou "miscelânea" (3). Diante disso, a anotação-AMR iniciou com os *tweets* do conjunto (ii), posto que são desafiadores ao se distanciarem mais da linguagem formal. Ao total, 22 padrões distintos foram identificados, os quais compreendem 1.143 instâncias (e 1.128 *tweets* distintos).

(1) a. Sera k petr4 já entrou na baixa?
    b. PETR4 subiu na bolsa 13,50. Muito bem, surpreso com o resultado.
(2) a. #OIBR4 (mensagem: 956643) http://t.co/VD2ApxqWqR
    b. #BBAS3 Banco da Brasil (mensagem: 956467) http://t.co/75T8wtmEXw
(3) a. Tô de olho no HB esperando o MOMENTO HISTÓRICO de PETR4 na era PT. Falta $0,01 pra 13. E 13 é ... PT!
    b. R$ 13 ... que ironia hein? ,) #PETR4

Para cada um dos 22 padrões, de 2 a 3 instâncias aleatórias foram manualmente anotadas com o auxílio do editor metAMoRphosED [Heinecke, 2023]. Tais instâncias foram empregadas na anotação das restantes por meio do uso de um *Large Language Model* (LLM) (GPT-4o via ChatGPT (OpenAI)) e da técnica *few-shot prompting* [Brown *et al*., 2020; Liu *et al*., 2023; Wei *et al*., 2022], com posterior revisão manual.

Apresentam-se aqui os fenômenos linguísticos observados nos padrões que exigiram a definição de diretrizes de anotação específicas, por ainda não estarem contemplados nas *guidelines* disponíveis. Embora a AMR tenha como objetivo representar o significado dos enunciados abstraindo aspectos sintático, no caso dos *tweets* – que frequentemente apresentam estruturas fragmentadas, elipses e informalidade (cf.

[3] O Índice Bovespa é principal índice de ações da bolsa de valores brasileira, a B3 (Brasil, Bolsa, Balcão).
[4] Código alfanumérico cujas letras indicam a empresa e o número é o tipo da ação ("Petr4" representa as "ações preferenciais da Petrobras").

(1)-(3)) – a anotação-UD funcionou como um guia valioso. Ao fornecer ou sugerir relações estruturais ausentes ou ambíguas, a anotação sintática auxiliou na construção mais precisa das representações semânticas em AMR.

5. Fenômenos Linguísticos e Diretrizes Propostas

5.1. Multiplicidade de segmentos

Um dos fenômenos que caracteriza muitos padrões é a multiplicidade de segmentos. O *tweet* (4) é uma das 25 instâncias do *corpus* que apresentam o Padrão #1, definido formalmente como [Notas gerais <sentença-truncada>...<url>] [Barbosa, 2024]. O Padrão #1 indica que os *tweets* possuem 3 blocos de informação: (i) expressão fixa "Notas gerais", (ii) sentença principal truncada por elipse ("..."), e (iii) URL final que remete à fonte da informação. Essa estrutura é confirmada pela anotação UD, que indica independência sintática entre os blocos por meio da relação *parataxis* (Figura 4). Como a AMR foi projetada para a anotação monossentencial, as diretrizes originais não englobam estratégia para representar múltiplos segmentos. Para tanto, adaptou-se a estratégia de O'Gorman *et al* (2018), originalmente proposta para representar o conteúdo de um texto inteiro, concatenando as sentenças sob um novo nó raiz e unificando variáveis que são correferentes. Trata-se do emprego do construto *m/multisentence* no topo do grafo e da relação *:sntN* para conectar os segmentos ou blocos ao conceito-topo (Figura 5). A relação *:sntN*, no entanto, não é empregada para conectar URLs ao conceito-topo, pois o modelo AMR dispõe da relação *:source*, que é mais informativa para o caso.

(4) Notas gerais A BROOKFIELD (BISA3) não conseguiu reverter perdas e registrou prejuízo líquido consolidado de R$... http://t.co/4e7aORZ6LU

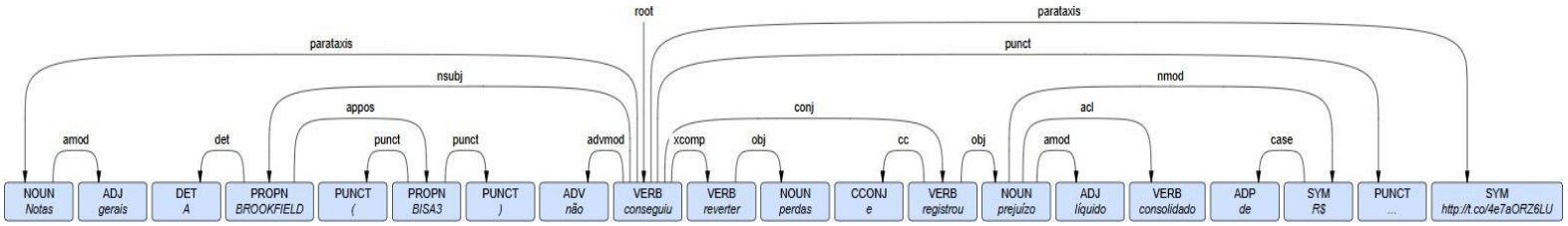

Figura 4. Anotação sintática-UD do *tweet*-exemplo (4).

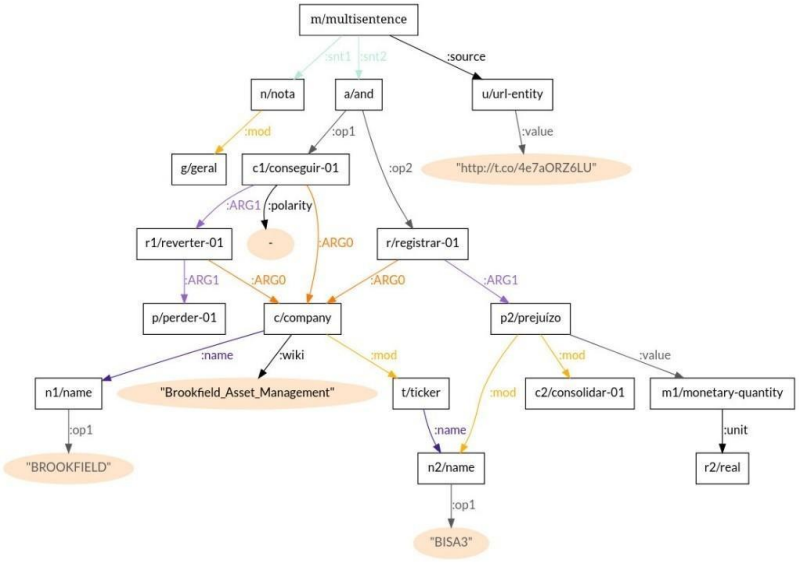

Figura 5. Anotação-AMR de múltiplos segmentos.

5.2. Locução adverbial “pelo menos”

Diante da necessidade de representar o significado da locução adverbial “pelo menos”, como em (5), de forma unificada no grafo AMR, optou-se por tratá-la como um conceito único por meio da abstração hifenizada “pelo-menos”, o que é ilustrado na representação PENMAN parcial do *tweet* (5) na Figura 6. Essa escolha se justifica pelo fato de a locução funcionar como uma unidade semântica, cujo sentido não é derivado da composição literal de seus constituintes (“por”+“o”+“menos”), mas sim da função pragmática que exerce, como indicar quantidade mínima. A anotação-UD reconhece esse comportamento unitário ao marcar os dependentes “o” (DET) e “menos” (NOUN) como *fixed* do núcleo “por”. Embora a criação de nós por hifenização não seja a estratégia mais alinhada aos princípios do modelo AMR, que busca representar diretamente os significados e não apenas a forma gráfica das expressões, essa escolha foi adotada como uma solução provisória. O objetivo é garantir consistência e viabilidade técnica na anotação de um corpus marcado por linguagem não-canônica e informal, como é o caso dos *tweets*.

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

```
:ARG1 (c1 / contratar-01
          :ARG1-of (h / have-value-91
                      :ARG2 (m1 / monetary-quantity
                                :value "90 bilhões"
                                :unit (r / real))
                      :degree (p / pelo-menos))
```

Figura 6. Anotação-AMR para a locução “pelo menos”.

5.3. Expressão “por ADJ nisto”

A expressão “poe longo (prazo) nisto” foi interpretada como indicando um grau elevado de duração, equivalente a “muito longo” ou “longo demais”. Com base nisso, adotou-se a mesma estratégia usada pelo modelo AMR para representar superlativos, que é o uso do frame reificado have-degree-91 como conceito central, conforme ilustrado na Figura 7. Para isso, foi necessário inferir o conceito implícito de intensidade (“muito”) e estruturar o grafo conectando os elementos da seguinte forma: o prazo (ARG1) é caracterizado como longo (ARG2), em um grau muito elevado (ARG3).

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

```
(t / ticker
   :wiki "Petrobras"
   :name (n / name
            :op1 "#PETR4")
   :topic-of (m1 / mensagem
                 :ord (o / ordinal-quantity
                         :value 955011))
   :mod (g / gráfico
           :mod (a / and
                   :op1 (h / have-degree-91
                           :ARG1 (p / prazo)
                           :ARG2 (l / longo)
                           :ARG3 (m2 / muito))
                   :op2 (p1 / prazo
                            :mod (c / curto))))
   :source (u / url-entity
              :value "http://t.co/cIh5ZtcFko"))
```

Figura 7. Representação PENMAN do grafo AMR para o *tweet* em (5).

5.4. Lista de Segmentos "Ticker+Porcentagem+Moeda"

Em alguns dos 22 padrões de Barbosa (2024) ocorre uma lista de segmentos compostos por "ticker+porcentagem+moeda". No *tweet* (7), por exemplo, há uma lista de 5 segmentos desse tipo justapostos. Para a representação dessa lista, tomou-se como diretriz a utilização da conjunção *a/and* e da relação *:opN* para cada um dos segmentos da lista. Na Figura 8, tem-se a representação PENMAN parcial do *tweet* (7), contendo os dois primeiros segmentos da lista ilustrando a diretriz.

(7) 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**.

```
(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)))
```

Figura 8. Anotação-AMR para lista de segmentos "ticker+porcentagem+moeda".

5.5. Truncamento Estrutural

Os truncamentos resultam da limitação de caracteres imposta pela plataforma. Quando estruturais, interferem diretamente na intepretação do *tweet* e na construção do grafo AMR. Para os casos em que um argumento do predicado está truncado, como o Arg1 de *v/vender-01* no *tweet* (8), a diretriz empregada para a anotação-AMR foi a utilização do conceito *a/amr-undefined*, como ilustrado na Figura 9, que exibe a representação PENMAN da anotação AMR do *tweet* (8).

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

```
:snt2 (v / vender-01
         :ARG0 (c / company
                  :name (n1 / name
                           :op1 "BR"
                           :op2 "PROPERTIES")
                  :mod (t / ticker
                          :name (n2 / name
                                   :op1 "BRPR3"))
                  :wiki "BR_Properties_S.A")
         :ARG1 (a / amr-undefined)
```

Figura 9. Representação PENMAN parcial do grafo AMR para o *tweet* em (8).

6. Fenômenos Linguísticos e Diretrizes Propostas

A partir da anotação das 1.143 instâncias (e 1.128 *tweets*), os 10 *rolesets* mais frequentes (Tabela 1) refletem o estilo direto do discurso financeiro. Isso também pode justificar o fato de *rolesets* relacionados a verbos modais (como *possible-01* e *recommend-01*) não estarem entre eles. O *frame have-degree-91*, por sua vez, destaca-se por ter sido empregado para representar o grau de valorização, desvalorização e expectativas de desempenho das ações. A Tabela 2 revela que as relações semânticas mais frequentes são *:op1*, :*name* e :*wiki*, o que se justifica pela alta incidência de entidades nomeadas no corpus, representadas por meio dessas relações. O destaque de :*mod* se justifica pelo emprego desta para relacionar ações e suas respectivas empresas, como em "ações da Petrobras subiram". A recorrência de :time reflete a centralidade da variável temporal na análise do mercado financeiro. Já as relações :*value*, :*unit* e :*quant* são comuns por expressarem quantias monetárias, porcentagens e unidades de medida. As relações argumentais centrais, como :ARG0 e :ARG1, não figuram entre as mais frequentes porque os *tweets* anotados, ao refletirem os padrões estruturais recorrentes, geralmente não apresentam estruturas sintáticas completas nas quais tais argumentos ocorrem.

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

Tabela 1: Os 10 *rolesets* mais frequentes.

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

Tabela 2. As 10 relações mais frequentes.

7. Considerações Finais e Trabalhos Futuros

A tarefa de anotação AMR do DANTEStocks se mostrou desafiadora, sobretudo quando se trata de *tweets* com os chamados "padrões estruturais (recorrentes)". Diz-se isso porque eles são muito fragmentados, sendo necessário recorrer à anotação-sintática-UD prévia do *corpus* como suporte para a representação semântica. Atualmente, as diretrizes aqui apresentadas e outras estão sendo compiladas em um manual de anotação AMR para tweets do mercado financeiro, que em breve estará publicamente disponível. Ademais, está em andamento a anotação manual, por um segundo anotador, de uma amostra das 1.143 instâncias já anotadas. Essa dupla anotação será utilizada para calcular o grau de concordância entre anotadores, com o objetivo de estimar a consistência das representações AMR e, assim, validar a qualidade da anotação e estabelecer um padrão-ouro para essa parcela do *corpus*. A anotação do restante do *corpus* — que inclui *tweets* com linguagem relativamente padrão e aqueles com estruturas mais variadas — é uma tarefa prevista para etapas futuras

Agradecimentos. Este trabalho foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Os autores deste trabalho agradecem ao Centro de Inteligência Artificial (C4AI-USP) e o apoio da Fundação de Apoio à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM Corporation.

Referências Bibliográficas

Anchiêta, R. and Pardo, T.A.S (2018) “Towards AMR-BR: A SemBank for Brazilian Portuguese Language”, In: *Proceedings of the 11th International Conference on Language Resources and Evaluation* (LREC), Miyazaki, Japan. ELRA.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Palmer, M., Schneider, N. and Xue, N. (2013). “Abstract Meaning Representation for Sembanking”, In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse* (LAW-ID), Sofia, Bulgaria, p. 178–186.

Barbosa, B.K.S. (2024). Descrição sintático-semântica de nomes predicadores em tweets do mercado financeiro em português. Dissertação de Mestrado, Universidade Federal de São Carlos, São Carlos/SP.

Bateman, J., Matthiessen, C., Nanri, K. and Zeng, L. (1991). “The re-use of linguistic resources across languages in multilingual generation components”, In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence* (IJCAI), V. 2, p. 966– 971.

Brown, T., *et al*. (2020). Language models are few-shot learners. In *Advances in neural information processing systems*, 33, pages 1877–1901.

de Marneffe, M-C., Dozat, T., Silveira, N., Hajič, J., Manning, C.D., McDonald, R. and Nivre, J. (2021). Universal Dependencies. In *Computational Linguistics*, 47(2), pages 1-54.

Di Felippo, A. and Roman, N.T. (2025). “DANTEStocks: a multi-layered annotated corpus of stock market tweets for Brazilian Portuguese”. In *Brazilian Journal of Applied Linguistics* (Corpus Linguistics: Studies and Applications), pages 1–23. *To Appear*

Di Felippo, A., Roman, N., Barbosa, B., and Pardo, T.A.S (2024b). “Genipapo - a multigenre dependency parser for Brazilian Portuguese”, In: *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology* (STIL), p. 257–266, Porto Alegre, RS, Brasil: SBC.

Di Felippo, A., Nunes, M.G.V., and Barbosa, B.K.S. (2024a). “A dependency treebank of tweets in Brazilian Portuguese: syntactic annotation issues and approach”, In: *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology* (STIL), p. 192–201, Porto Alegre, RS, Brasil: SBC.

Duran, M.S., Martins, J.P. e Aluísio, S.M. (2013) “Um repositório de verbos para a anotação de papeis semânticos disponível na web”, In: *Anais do 9º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (STIL), p. 168–172, Fortaleza, CE, Brasil: SBC.

Heinecke, J. (2023). “metAMoRphosED, a graphical editor for Abstract Meaning Representation”, In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics* (ISA), p. 27–32, Nancy, France: ACL.

Inácio, M.L., Cabezudo, M.A.S., Ramisch, R., Di Felippo, A. and Pardo, T.A.S. (2023) The AMR-PT corpus and the semantic annotation of challenging sentences from journalistic and opinion texts. In *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, 39(3). https://doi.org/10.1590/1678-460X202339355159

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Wang, K., Zhang, T. and Liu, Y. (2023). “Jailbreaking ChatGPT via prompt engineering: an empirical study”. *arXiv* preprint arXiv:2305.13860.

May, J.; and Priyadarshi, J. (2017). “SemEval-2017 Task 9: Abstract Meaning Representation parsing and generation”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation* (SemEval), p. 536–545. Vancouver, Canada.

O’Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K. and Palmer, M. (2018) “AMR beyond the sentence: the multi-sentence AMR corpus”, In: *Proceedings of the 27th International Conference on Computational Linguistics* (COLING), p. 3693–3702, Santa Fe, New Mexico, USA: ACL.

Palmer, M., Gildea, D. and Kingsbury, P. (2005) The Proposition Bank: an annotated corpus of semantic roles, In *Computational Linguistics*, 31(1), pages 71–106.

Plutchik, R. and Kellerman, H. (1986) “Emotion: Theory, Research and Experience”, New York: Academic Press.

Sanguinetti, M. C. *et al.* (2023). Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. In *Language Resources Evaluation*, 57, pages 493–544.

Scandarolli, C.L., A. Di Felippo, N.T. Roman, and Pardo, T.A.S. (2023). “Tipologia de fenômenos ortográficos e lexicais em CGU: o caso dos tweets do mercado financeiro”, In: *Anais do 14º Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (STIL), p. 240-248, Belo Horizonte, MG, Brasil: SBC.

Seno, E., H. Caseli, M. Inácio, R. Anchiêta, and Ramisch, R. (2022). Xpta: um parser AMR para o Português baseado em uma abordagem entre línguas. In *Linguamática*, 14, pages 49–68.

Silva, F.J.V., Roman, N.T. and Carvalho, A.M.B.R. (2020). Stock market tweets annotated with emotions. In *Corpora*, 15(3), pages 343–354.

Sobrevilla Cabezudo, M. A. and Pardo, T.A.S (2019) “Towards a General Abstract Meaning Representation Corpus for Brazilian Portuguese”, In: *Proceedings of the 13th Linguistic Annotation Workshop* (LAW), p. 236–244, Florence, Italy: ACL.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *arXiv* preprint arXiv:2201.11903.

Wein, S., and J. Bonn (2023). “Comparing umr and cross-lingual adaptations of AMR”, In: *Proceedings of the 4th International Workshop on Designing Meaning Representations* (DMR), p. 23–33, Nancy, France. ACL.

Zerbinati, M. M., Roman, N. T., and Di-Felippo, A. (2024). “A corpus of stock market tweets annotated with named entities”, In: *Proceedings of the 16th International Conference on Computational Processing of Portuguese* (PROPOR)*,* V. 1, p. 276–284, Santiago de Compostela, Espanha. ACL.