REVISTA DA **ABRALIN**

RELATÓRIO DE PESQUISA

Anotação de expressões com numerais em narrativas clínicas

Carlos Antônio de Souza PERINI
Universidade Federal de Minas Gerais (UFMG)

Ana Luisa dos Anjos Resende GUIMARÃES
Universidade Federal de Minas Gerais (UFMG)

OPEN ACCESS

EDITADO POR
- Marta Deysiane Alves Faria Sousa (UFS)
- Adriana Pagano (UFMG)
- Jorge Baptista (UALg)

AVALIADO POR
- Roana Rodrigues (UFS)
- Alisson Hudson Veras Lima (IFAL)

SOBRE OS AUTORES
- Carlos Antônio de Souza Perini Conceptualização, Análise Formal, Metodologia, Escrita – rascunho original e Escrita – análise e edição, Curadoria de Dados e Investigação.
- Ana Luisa dos A. R. Guimarães Conceptualização, Análise Formal, Metodologia, Escrita – rascunho original e Escrita – análise e edição.

DATAS
- Recebido: 20/11/2022
- Aceito: 28/06/2023
- Publicado: 09/09/2024

COMO CITAR
Perini, C. A. S.; Guimarães, A. L. A. R. (2024). Anotação de expressões com numerais em narrativas clínicas. *Revista da Abralin*, v. 22, n. 2, p. 215-248, 2024.

RESUMO

Este trabalho tem por objetivo analisar as expressões que contêm numerais em textos de narrativas clínicas a fim de identificar os potenciais desafios para o seu processamento computacional e de elaborar diretrizes para sua anotação de acordo com o projeto Universal Dependencies (UD). Utilizando um corpus de 1.000 narrativas clínicas, os *tokens* compostos por pelo menos um numeral em formato de algarismo foram selecionados e classificados de acordo com o formato de sua apresentação e sua eventual anotação seguindo as diretrizes das UD. Ocorrências de *tokens* pertencentes às dez classes mais frequentes no corpus foram estudadas e orientações para a anotação dessas classes foram elaboradas. Essas orientações foram registradas e posteriormente serão utilizadas para a compilação de um guia de anotação para projetos de treebanks de narrativas clínicas.

ABSTRACT

This paper aims to analyze expressions containing numerals in clinical narrative texts in order to identify potential challenges for their computational processing and to elaborate guidelines for their annotation according to the Universal Dependencies (UD) project. Using a corpus of 1,000 clinical narratives, *tokens* composed of at least one numeral in numeral format were selected and classified according to the format of their presentation and their eventual annotation following the UD guidelines. Occurrences of *tokens* belonging to the ten most frequent classes in the

corpus were studied and guidelines for the annotation of these classes were elaborated. These guidelines were recorded and will later be used to compile an annotation guide for clinical narrative treebank projects.

PALAVRAS-CHAVE
Narrativas clínicas. Dependências universais. Diretrizes de anotação.

KEYWORDS
Clinical narratives. Universal dependencies. Annotation guidelines.

Introdução

Os sistemas de informação de saúde estão cada vez mais integrados, permitindo centralizar e anexar prontuários eletrônicos de todo o histórico médico dos pacientes no formato digital e viabilizar melhorias contínuas para o atendimento e acompanhamento médico. "A acessibilidade aos detalhes dos dados do paciente disponíveis nos sistemas de saúde é fundamental para melhorar o processo de assistência médica e avançar na pesquisa clínica" (XIA; YETISGEN-YILDIZ, 2012, p.1). Segundo Oliveira et al. (2022, p. 1)

> [...] o alto volume de pesquisas com foco na extração de informações de pacientes de prontuários eletrônicos levou a um aumento na demanda por corpora anotados, que são um recurso precioso tanto para o desenvolvimento quanto para a avaliação de algoritmos de Processamento de Língua Natural (PLN).

A busca por informações nesses sistemas é importante para melhorar o atendimento e acompanhamento médico, o PLN ajuda na recuperação da informação de textos não estruturados. As aplicações típicas de PLN segundo Jurafsky e Martin (2008) são: reconhecimento de fala, modelo de linguagem (uso de estatísticas para prever a probabilidade de sequência de palavras), extração de informações, sumarização de texto, diálogos e chatbots, reconhecimento de entidade nomeada, geração de linguagem natural, pré-processamento de texto, análise morfológica, *part-of-speech tagging*, análise sintática, entre outros. E, no caso clínico, de acordo com Névéol et al. (2018, p.1) "as principais aplicações clínicas da PLN incluem a assistência a profissionais de saúde com estudos retrospectivos e tomada de decisões clínicas".

A criação de corpus visando tarefas de PLN deve estar pautada pela integração das tarefas (anotação morfossintática correlacionada com a anotação semântica para a extração de informações).

Dentre as tarefas, a anotação de funções sintáticas é especialmente relevante ao desenvolvimento de *parsers*[1] robustos, sobretudo de dependências entre constituintes de uma sentença. Para que *parsers* de dependência possam funcionar em línguas tipologicamente distintas e permitir a comparabilidade entre elas, foi desenvolvida a iniciativa UD - *Universal Dependencies*[2].

A anotação de relações de dependência requer a utilização e adequação das normas UD ao tipo de corpus anotado. Um tipo de texto desafiador para a anotação de dependências são as chamadas narrativas clínicas, textos produzidos sob pressão de tempo e em situações envolvendo risco de vida.

As narrativas clínicas demandam decisões sobre anotação devido a sua própria constituição, mas também devido a processos externos, tais como a desidentificação e anonimização de dados, e ao pré-processamento computacional que o corpus requer.

Como todo tipo de corpus, as narrativas clínicas demandam a elaboração detalhada de um guia de anotação que documente as decisões e garanta a consistência das anotações com impacto na concordância entre anotadores.

Para que informações possam ser extraídas automaticamente desses registros digitais e que as aplicações de PLN se realizem, é preciso descrever esses textos e construir modelos de linguagem que processam e anotam automaticamente novos textos do mesmo tipo. Essa descrição significa anotar as palavras de um texto dando a elas informações adicionais, como categorias morfológicas, relações sintáticas, classificação semântica com rótulos de entidades nomeadas ou de uma ontologia que podem ajudar o modelo de PLN a identificar melhor o texto e a realizar tarefas específicas, como a anotação automática, com mais precisão.

A anotação de corpus clínico requer uma atenção especial. Xia et al. (2012, p. 4-5) afirmam que

> [...] o conhecimento médico é imprescindível tanto para a elaboração das diretrizes de anotação quanto para a própria anotação, e não pode ser adquirido rapidamente. Como resultado, temos que confiar muito em especialistas médicos. Chamamos esse tipo de anotação de 'anotação especializada'[3] [...] experimentos mostraram que o treinamento médico por si só não é suficiente para obter uma alta concordância entre os anotadores, e os pesquisadores de PNL devem se envolver no processo de anotação o mais cedo possível, apesar de não terem treinamento médico.[4]

---

[1] Os *parsers* tem como objetivo analisar a estrutura sintática da frase e gerar uma representação formal dessa estrutura.

[2] http://universaldependences.org

[3] Tradução livre de: "medical expertise is a must for both design of the annotation guidelines and annotation itself, and it cannot be acquired quickly. As a result, we have to heavily rely on medical experts. We call this kind of annotation 'expert annotation'", (XIA; YETISGEN-YILDIZ, 2012, p. 4-5).

[4] Tradução livre de: "Our experiments show that medical training alone is not sufficient for achieving high inter-annotator agreement, and NLP researchers should get involved in the annotation process as early as possible despite their lack of medical training." (XIA; YETISGEN-YILDIZ, 2012, p. 4-5).

Segundo Oliveira et al. (2022, p. 1), "a falta de um corpo clínico multidisciplinar fora do âmbito da língua inglesa, especialmente no português brasileiro, é gritante e impacta fortemente o progresso científico no campo da PLN biomédica."[5]

Em projetos de anotação de textos é muito importante que se tenha uma teoria com um conjunto de normas para que se faça as análises. Neste estudo, o conjunto de diretrizes geral é o das UD e o corpus de 1.000 sentenças de narrativas clínicas extraído do projeto SemClinBR[6]. No caso de textos produzidos no ambiente hospitalar, tais como notas de evolução de enfermagem ou sumários de alta hospitalar, características específicas são um desafio para as anotações. Uma dessas características é a existência de termos que possuem numerais e outros símbolos em sua escrita, como em fórmulas químicas (como o oxigênio, O2), datas (como 25/12), quantidades volumétricas (como 100ml), temporais (como 10 min) e quantidades de outras unidades (como pressão arterial: 13/9 mmHg) entre outros. Examinar essas expressões com numerais é importante para elaborar orientações adequadas à sua anotação e pela relevância do significado de precisão que constroem na comunicação clínica, seja para a dosagem de medicamentos como para o tempo de internação, número de sessões, entre outras situações.

Dentre os desafios de anotação apresentados pelo texto clínico, Névéol et al. (2018, p. 8), apontam em seus estudos que, a partir dos dados das narrativas clínicas, conseguem desenvolver métodos poderosos para abordar tarefas clínicas de interesse da PNL, como desidentificação em prontuários eletrônicos, reconhecimento de entidade clínica, normalização e contextualização. Há a necessidade de desenvolver métodos e conjuntos de dados compartilhados, permitindo a comparação de abordagens dentro e entre idiomas e encorajam a criação de diretrizes de publicação mais estruturadas que incorporem informações sobre linguagem e métodos, recomendando estudos e análises de como as especificidades das línguas podem contribuir para avanços metodológicos.

No corpus deste estudo, os desafios estão nas expressões que contêm algarismos, que podem incluir também letras, sinais de pontuação ou símbolos. Além de constituírem um desafio de anotação em termos de classe de palavra e relação sintática, essas expressões requerem regras de tokenização cuidadosas pois decisões de descrição nos níveis linguísticos inferiores (morfologia) impactam níveis superiores como sintaxe e semântica.

Será assim explorada a ocorrência de números (no formato de algarismos) em um corpus de narrativas clínicas e indagamos formas de anotação em UD que permitam capturar a estrutura sintática com potencial para tarefas subsequentes, tais como a extração automática de informações, com resultados que podem ser interpretados e utilizados para diversos propósitos tanto na área médica como nos estudos linguísticos em geral e no PLN em particular.

Este artigo está dividido em 5 seções, incluindo esta introdução. Na próxima seção, apresentamos uma breve revisão da literatura disponível sobre anotação de narrativas clínicas e anotação em

---

[5] Tradução livre de: "The absence of a multipurpose clinical corpus outside the scope of the English language, especially in Brazilian Portuguese, is glaring and severely impacts scientific progress in the biomedical NLP field." (OLIVEIRA et al., 2022, p. 1).

[6] https://github.com/HAILab-PUCPR/SemClinBr

UD. A seção 3 descreve como o corpus foi conformado e a metodologia utilizada para analisar e categorizar as expressões com algarismos encontradas no corpus bem como a sua anotação sintática. Em seguida, a seção 4 traz os resultados dessa análise, apresentando uma classificação possível dessas expressões, expondo dados quantitativos sobre o corpus estudado e, como principal resultado deste estudo, descrevendo as orientações de anotação das expressões com numerais mais frequentes no corpus, elaborando diretrizes para sua anotação de acordo com as UD. Também são discutidos os desafios encontrados durante a execução da pesquisa e sua relação com trabalhos anteriores. Por fim, a seção 5 traz as considerações finais.

1. Revisão da literatura

As narrativas clínicas são um tipo de relato produzido na área médica, especialmente dentro de hospitais e centros de saúde, direcionado ao registro do estado presente de um paciente e da sua evolução ao longo de seu tratamento, incluindo também sinais vitais durante os atendimentos, resultados de exames e procedimentos pelos quais o paciente passou (OLIVEIRA et al., 2022). Esses textos costumam ser redigidos pelos profissionais da saúde de forma rápida, durante ou entre atendimentos médicos.

Quando em formato digital, podem passar a fazer parte do Registro Eletrônico de Saúde do paciente, que é composto por todo o seu histórico médico, seus dados, exames e outras informações relevantes à sua saúde. De posse desse conteúdo e visando criar ferramentas para auxiliar o acompanhamento médico e os estudos na área, muitas pesquisas têm sido conduzidas no campo de PLN. A principal tarefa estudada é a extração de informação desses registros médicos, especialmente a extração de entidades nomeadas (palavras ou expressões chave para o entendimento do texto e que podem ser agrupadas em categorias, como, no contexto clínico, nomes de pessoas, de exames, de medicamentos e outros). Uma das opções para potencializar os resultados dessa e de outras tarefas é associar a camada de anotação das entidades nomeadas com uma camada de anotação sintática.

Diversos trabalhos como Styler et al. (2014), Hanauer et al. (2019), Moon et al. (2011), Oliveira et al. (2022), Névéol et al. (2018), Xia et al. (2012) têm explorado os desafios da anotação de narrativas clínicas. Fan et al. (2013) apontam características típicas das narrativas clínicas em línguas como o inglês e o alemão, enfatizando o fato de as narrativas serem constituídas por sentenças incompletas ("ill formed sentences"), as quais demandam a elaboração de guias de anotação precisos e regras de segmentação adequadas a esse tipo de texto. Kara et al. (2018) apontam características típicas de narrativas clínicas em alemão, destacando a alta dependência de conhecimento de domínio para a compreensão dos textos, sua complexidade sintática, os processos de redução de informações e consequente elipse de elementos da oração (verbos e outros).

É importante, então, desenvolver ferramentas de PLN adaptadas ao contexto clínico, pois ele apresenta diversas particularidades na linguagem usada (como termos próprios, abreviaturas, siglas e símbolos específicos). Essas ferramentas, por sua vez, dependem de corpora anotados, que podem

ter essa anotação nas mais diversas camadas, a depender do objetivo em questão, mas que se beneficiam ainda mais da associação dessas diferentes camadas de anotação.

O SemClinBr (OLIVEIRA et al., 2022) é um corpus anotado composto por textos clínicos em português brasileiro. Ele é anotado em uma camada semântica, abrangendo as entidades encontradas e as relações entre elas. A partir dele, há também o trabalho em andamento do desenvolvimento do DepClinBr, que anota o mesmo corpus sintaticamente de acordo com as diretrizes do projeto UD (OLIVEIRA et al., 2022). Oliveira et al. (2022) aponta que muitos desafios foram encontrados durante a anotação do DepClinBr devido às particularidades dos textos de narrativas clínicas. As principais características desses textos são ilustradas no Quadro 1 conforme encontrado pelos autores.

Característica	**Exemplo**
Palavras não reconhecidas por POS *taggers* e lematizadores	Hidantalizado
	Facietomia
	flutter atrial
Amplo uso de acrônimos e abreviações	mantendo monitorização p 81, pa 103/74, sat o2 97% po le + ladf em 01/12/15
	colar cervical c/ queixas de dor
Erros de ortografia e digitação	em repoiso em o leito
	outras inetrcorrências
Expressões numéricas	ssvv a as 05:45 h pa = 133/74 mmhg , fc = 114 bpm, spo2 = 93%
Falta de pontuação Uso diferenciado de símbolos	glasgow: 9 mantendo monitorização p 81, pa 103/74, sat o2 97% dois ave um há 1 ano e outro 2 anos hma: inchaço, principalmente em o peridodo de a manha e em mmii (++++/++++), piora de o quadro ha 15 dias, estagio ii de irc solicitada svd + coleta de gasometria arterial # 61 # professora a as 11:30hs: realizado endoscopia digestiva + broncoscopia, sem intercorrência durante o exame e transporte.
	antes 2 carteiras/dia
Coordenação	apresenta curativo + tala gessada em mse, referindo algia moderada, edema distal, mobilidade diminuida, apresentou 1 episodio de emese, sendo medicada, diurese presente, segue cuidados

Comentários entre parênteses	refere cx em a bexiga devido a tumor (sic) em 2013
	hmp : pais falecidos po ca (em a o soube especificar)
	controle glicêmico com glicemia em jejum abaixo de 100 (não costuma anotar)
Redução	# retorno 7 dias
Elipse	apresentou problemas
	aceitando pouco a dieta vo

QUADRO 1 – Características dos textos de narrativas clínicas.
Fonte: Oliveira et al. (2022). Traduzido pelos autores.

Observando-se o Quadro 1, vê-se que, em grande parte das características citadas, os exemplos contêm algum tipo de algarismo, seja ele sozinho ("p 81", "retorno 7 dias") ou acompanhado de letras, sinais de pontuação ou outros símbolos ("103/7", "o2", "11:30hs"). Nas narrativas clínicas, a ocorrência de algarismos sozinhos (como numerais) ou dentro de siglas, fórmulas químicas, escritas de horários e datas e em outras apresentações são muito frequentes e podem tornar-se mais um desafio não só para o processamento do texto e sua tokenização, como também para as decisões de anotação morfossintática desses termos.

Autores como Moon et al. (2011) abordam, por exemplo, o uso e processamento automático dos símbolos usados em textos clínicos. Concluíram que a interpretação dos símbolos '+', '–', '/' e '#' e seu contexto circundante em narrativas clínicas pode ser vista como um caso especial de Desambiguação do Sentido da Palavra (WSD). Hanauer et al. (2018), por sua vez, analisam a variação do uso de numerais em narrativas clínicas e os desafios colocados por eles especialmente na tarefa de extração de informação, alertando sobre a importância de se levar essa variedade em consideração no desenvolvimento das ferramentas para tal fim. No entanto, não foram encontrados estudos que tratassem da combinação desses tipos de caracteres ou dos numerais juntamente com letras. Sabe-se que, para a extração de informação ser confiável e precisa, deve-se garantir que a tokenização do texto esteja correta e que sua anotação, tanto sintática quanto semântica, esteja adequada.

O projeto Universal Dependencies (UD), ou Dependências Universais em português, apresentado por Nivre (2015), é uma proposta de anotação morfossintática baseada em estudos de tipologia linguística com o fim de desenvolver um sistema de anotação consistente entre as línguas, e que possa ser utilizado com eficiência em tarefas de PLN que possam ser aplicadas a diversos idiomas (Marneffe et al., 2021). Dispondo de uma série de diretrizes de anotação gerais, a UD é uma iniciativa multilíngue aberta à comunidade científica que segue em constante evolução e desenvolvimento para aprimorar suas estratégias e disponibilizar cada vez mais material de livre uso e consulta para pesquisadores da área. Atualmente, o projeto UD conta com 243 *treebanks* anotados em 138 idiomas, e está em sua versão 2.11.

Para a anotação em UD, a palavra é a unidade básica de análise, e são utilizadas 37 etiquetas para a anotação de relações sintáticas (*deprel*), 17 etiquetas para classes de palavras (POS) e 24 etiquetas para atributos morfológicos (*features*). Essas etiquetas podem, em geral, ser utilizadas para todos os idiomas, mas são passíveis de adaptações de acordo com as particularidades de cada língua.

A Nomenclatura Gramatical Brasileira (NGB), publicada em 1959 pelo Ministério da Educação e Cultura e que está em vigor até hoje, define que os numerais são pertencentes a uma classe própria. Dessa forma, eles são classificados em numerais cardinais, ordinais, multiplicativos e fracionários, e podem ser flexionados em gênero e número. Para a UD, numerais também constituem uma classe de palavras exclusiva e indicada, na maior parte dos casos, pela etiqueta NUM. As etiquetas prototipicamente aplicáveis à anotação de numerais pela UD em português brasileiro para relações de dependência são *numod* (modificador numérico, geralmente de caráter quantificador) e *amod* (modificador adjetivo, utilizado para numerais ordinais), e para classe de palavras NUM (número) e ADJ (adjetivo).

No entanto, ao se observar o uso dos numerais, em especial em sua forma de algarismo, vê-se que eles podem aparecer em diversos tipos de apresentação, com muitos significados e, em alguns casos acompanhados de outros tipos de caracteres. Duran, Lopes e Pardo (2021) apontam que "Na escrita, um *token* de numeral pode conter, além de vários algarismos, sinais de pontuação, como vírgula, ponto e barra: 1,07; 127.234; 4/7" (p. 3). Além disso, os numerais podem ser encontrados, por exemplo, na descrição de datas no formato "DD/MM/AAAA", gerando um *token* que não possui apenas a característica quantitativa comum à classe dos numerais. Nesses casos, eles podem receber etiquetas de relações de dependência e de POS diferentes das prototípicas. As expressões de data, por exemplo, recebem uma etiqueta POS de NUM quando escritas no formato mencionado anteriormente (como 28/01/1959), mas serão consideradas como NOUN se apresentarem a abreviatura do mês com o ano em algarismos (como jan/59). Esse é um dos grandes desafios encontrados quando se precisa anotar sintaticamente expressões que são constituídas de números e outros caracteres em um mesmo *token*.

Essa questão mostra-se de grande relevância para o PLN como um todo, pois os algoritmos utilizados para a tokenização de um texto podem encontrar dificuldades para interpretar esse conjunto de caracteres (algarismos acompanhados de letras, sinais de pontuação e símbolos) como uma unidade individual. Assim, faz-se necessário um estudo aprofundado do texto que está sendo trabalhado e do formato das ocorrências de numerais grafados apenas com algarismos (e não por extenso) para que se conheça os desafios que podem ser enfrentados no processamento computacional do corpus.

2. Metodologia

O corpus utilizado nesta pesquisa é um fragmento do corpus SemClinBr[7], construído e disponibilizado por Oliveira et al. (2022), composto por narrativas clínicas cedidas por três hospitais brasileiros. Esse corpus compreende 1.000 notas clínicas[8] com propósito de anotação semântica usando terminologia da ontologia *Unified Medical Language System* (UMLS) (HUMPHREYS; MCCRAY; LINDBERG, 1993) que incluem sumários de alta hospitalar, notas de enfermagem, registros ambulatoriais e notas de evolução de pacientes dos domínios de cardiologia, nefrologia e endocrinologia. O corpus no total tem 9.409 sentenças, resultando em um número de 143.206 *tokens* (abrangendo palavras, algarismos, símbolos e pontuações). A média de *tokens* por sentença foi de 15,22 *tokens*, sendo que o número máximo encontrado em uma sentença foram 250 *tokens* e o mínimo, 1 *token*. Da totalidade de *tokens*, 11.500 continham algum algarismo em qualquer posição, o que compreende aproximadamente 8% do corpus total.

Alguns dos relatos são estruturados em tópicos seguindo o formato SOAP (Subjetivo – relacionado ao que o paciente relata ter passado e estar sentindo –, Objetivo – relacionado a informações concretas sobre o estado do paciente –, Avaliação – correspondendo a detalhes da avaliação médica – e Plano – o plano de ação traçado pelo profissional de saúde em relação ao paciente), enquanto outros são compostos apenas de um texto livre, não estruturado, normalmente redigido de forma rápida durante ou entre atendimentos.

No estudo de Oliveira et al. (2022), quando compuseram o corpus do SemClinBr[9], as notas foram transcritas, processadas computacionalmente e anonimizadas, de forma a eliminar quaisquer dados identificadores dos profissionais e pacientes envolvidos[10]. O corpus foi então anotado semanticamente de acordo com os tipos e grupos semânticos definidos no Sistema Unificado de Linguagem Médica (UMLS), possuindo 65.117 entidades e 11.263 relações anotadas. Os 1.000 arquivos em formato XML compõem o corpus anotado do SemClinBR e de cada arquivo foi extraído o relato. Daí foi feito um corpus com 1.000 relatos. Desses 1.000 relatos, recortou-se as expressões com numerais, as quais foram classificadas. As indicações das etiquetas utilizadas para a anotação semântica de cada

---

[7] O uso dos textos do SemClinBr foi aprovado pelo Comitê de Ética em Pesquisa (CEP) da PUCPR, sob o registro de número 1.354.675. Os procedimentos de coleta e transcrição estão especificados em Oliveira et al. (2022) sessão: "*Methods*".

[8] O corpus do SemClinBR é composto por 1.000 narrativas, cada uma em um arquivo XML com as anotações semânticas vinculadas com a ontologia UMLS. Para este estudo foram extraídas somente as sentenças sem as etiquetas semânticas e composto um arquivo com 1.000 linhas, cada linha uma narrativa clínica. Esse arquivo foi processado com uso de expressões regulares (regex) das quais derivou-se as classes. A primeira *regex* foi buscar números, depois, manualmente foram feitas as classes.

[9] Os dados foram obtidos a partir de duas fontes diferentes: (1) um corpus de 2.094.929 registros de um conjunto de hospitais no Brasil gerados entre 2013 e 2018 e (2) um corpus proveniente de um hospital universitário com base em registros no período entre 2002 e 2007, que responde por 5.617 dos lançamentos.

[10] Detalhes de como são feitas as transcrições para o processamento computacional das notas do corpus SemClinBR podem ser conferidas em: Oliveira et al. (2022).

um foram desprezadas porque o interesse neste estudo é verificar a anotação morfossintática seguindo as UD das expressões com numerais presente nesse corpus.

Para realizar a categorização das classes de *tokens* com algarismos encontrados nas narrativas do SemClinBR, um *script* em Python foi utilizado para extrair apenas o texto dos relatos. O resultado dessa extração foi um arquivo em formato TXT contendo todas as narrativas, o qual seria utilizado para as próximas etapas de processamento e extração dos dados.

Em seguida, o texto das narrativas foi tokenizado em sentenças e em *tokens* utilizando outro *script* em Python. Para essa tokenização, foram usadas funções da biblioteca de PLN Natural Language Toolkit (NLTK)[11]. A opção pelo uso do NLTK se deu para que fosse possível observar os problemas de tratamento de *tokens* (com algarismos) resultantes de uma tokenização convencional, sem passar por um processamento específico de domínio. Os *tokens* resultantes foram, então, filtrados para selecionar apenas aqueles que possuíam algum algarismo em qualquer posição e acompanhado (ou não) de outros caracteres (por exemplo, 18:00, spo2, ++4/+4, 2/2hrs, 14).

Por fim, foi realizada uma análise dos *tokens* com algarismo filtrados (os quais serão referenciados pela sigla TCA deste ponto em diante) a fim de encontrar padrões de escrita, de significado e de função sintática entre eles, visando sua subsequente anotação sintática de acordo com as diretrizes das UD. Esses padrões foram utilizados para estabelecer as classes dos tipos de TCA presentes nos textos. De forma similar à realizada com os símbolos em Moon et al. (2022), foi criada uma tabela apresentando cada classe, sua aplicação nas narrativas clínicas, exemplos retirados do corpus e as características do formato desses *tokens*. A partir das categorias definidas na etapa anterior, os padrões encontrados foram utilizados para gerar expressões regulares, ainda em Python, para classificar cada TCA de acordo com sua classe, ainda explorando a metodologia seguida por Moon et al. (2022). Foi gerada uma tabela contendo cada TCA[12], a sentença em que ele se encontrava e a classe a que ele pertencia, conforme mostrado no Quadro 2.

[11] Disponível em https://www.nltk.org/ .

[12] Os dados com mais detalhes sobre classes e quantidade de dados analisados serão explicados na seção seguinte.

ID	Sentença	*Token*	Classe	Tipo de número
0	18:00: PACIENTE RETORNOU DO CC LÚCIDO, ORIENTADO, COMUNICATIVO; MANTEM AVP COM STP.	18:00	Classe 9	9 - horário
1	REFERE TER APRESENTADO 01 EPISÓDIO DE EMESE NO CC.	1	Classe 17	17 - número puro
2	18:30: DESACORDADO, FAZ ABERTURA OCULAR ESPONTÂNEA, AGITADO.	18:30	Classe 9	9 - horário
3	13/10/2006, 15:40 HORAS: PERMANECE EM COMA, COM PUPILAS MÉDIO-FIXAS, MANTENDO CATETER DE DVE.	13/10/2006	Classe 5	5 - data
5	VM VIA TOT COM FIO2-50%, PCV, PP 25cmH20, PEEP 2cmH20, F15mrpm, SPO2 99%.	FIO2-50	Classe 1	1 - abreviatura/sigla
6	VM VIA TOT COM FIO2-50%, PCV, PP 25cmH20, PEEP 2cmH20, F15mrpm, SPO2 99%.	25cmH20	Classe 2	2 - unidade de medida
10	VM VIA TOT COM FIO2-50%, PCV, PP 25cmH20, PEEP 2cmH20, F15mrpm, SPO2 99%.	99	Classe 17	17 - número puro
11	Apresenta dextro de 88mg/dl, medicada CPM.	88mg/dl	Classe 2	2 - unidade de medida
12	08:20h: Glasgow 14, pupilas isocóricas fotorreagentes, curativo cefálico tipo capacete limpo e seco, SNE com dieta a 30ml/h, respiração espontânea, em névoa úmida 4l/, SPO2 95%, acesso venoso periférico em jugular externa D com plano básico de hidratação, monitorizado, estável hemodinamicamente, MV+ com roncos esparsos, abdome plano, flácido, indolor à palpação, RHA+, SVD com diurese efetiva, evacuação ausente, extremidaes bem perfundidas, higienizado e mobilizado no leito, pele íntegra.	08:20h	Classe 9	9 - horário
15	08:20h: Glasgow 14, pupilas isocóricas fotorreagentes, curativo cefálico tipo capacete limpo e seco, SNE com dieta a 30ml/h, respiração espontânea, em névoa úmida 4l/, SPO2 95%, acesso venoso periférico em jugular externa D com plano básico de hidratação, monitorizado, estável hemodinamicamente, MV+ com roncos esparsos, abdome plano, flácido, indolor à palpação, RHA+, SVD com diurese efetiva, evacuação ausente, extremidaes bem perfundidas, higienizado e mobilizado no leito, pele íntegra.	4l/	Classe 19	19 - não classificado
16	08:20h: Glasgow 14, pupilas isocóricas fotorreagentes, curativo cefálico tipo capacete limpo e seco, SNE com dieta a 30ml/h, respiração espontânea, em névoa úmida 4l/, SPO2 95%, acesso venoso periférico em jugular externa D com plano básico de hidratação, monitorizado, estável hemodinamicamente, MV+ com roncos esparsos, abdome plano, flácido, indolor à palpação, RHA+, SVD com diurese efetiva, evacuação ausente, extremidaes bem perfundidas, higienizado e mobilizado no leito, pele íntegra.	SPO2		1 - abreviatura/sigla
17	08:20h: Glasgow 14, pupilas isocóricas fotorreagentes, curativo cefálico tipo capacete limpo e seco, SNE com dieta a 30ml/h, respiração espontânea, em névoa úmida 4l/, SPO2 95%, acesso venoso periférico em jugular externa D com plano básico de hidratação, monitorizado, estável hemodinamicamente, MV+ com roncos esparsos, abdome plano, flácido, indolor à palpação, RHA+, SVD com diurese efetiva, evacuação ausente, extremidaes bem perfundidas, higienizado e mobilizado no leito, pele	95		17 - número puro

	íntegra.			
18	Maria Firmino, 79 anos.	79		17 - número puro
19	# Sepse pulmonar em D8 tazocin (paciente não recebeu por 2 dias Atb).	D8		6 - dia de tratamento
20	# Sepse pulmonar em D8 tazocin (paciente não recebeu por 2 dias Atb).	2		17 - número puro
21	# AVE há 2 anos (hemiparesia à E).	2		17 - número puro
22	# O: paciente em BEG, hipocorada 2+/4, hidratada, eupnéica, afebril.	2+/4		3 - avaliação de edema

QUADRO 2 – Planilha gerada para identificação da sentença no corpus, do *token* com algarismo encontrado nela e da classe a que ele pertence.
Fonte: elaborado pelos autores.

As expressões regulares foram elaboradas de forma que pudessem abarcar a maior variedade possível desses tipos de apresentações e de forma precisa. De posse da tabela gerada, foi possível extrair dados estatísticos referentes à ocorrência de cada tipo de TCA, os quais são apresentados na seção Resultados, gráfico 1.

Por fim, as dez primeiras classes de TCA mais frequentes foram identificadas e sentenças que continham exemplos de *tokens* pertencentes a elas foram selecionadas no corpus. Essas sentenças foram analisadas cuidadosamente a fim de chegar à melhor decisão sobre a anotação dos TCA de acordo com as diretrizes das UD e com as particularidades encontradas no domínio médico e na língua portuguesa. Essas decisões foram registradas e as árvores foram anotadas utilizando a ferramenta online gratuita Arborator Quick[13], desenvolvida por Guibon et al., 2020.

A partir da análise dos *tokens* com algarismos obtidos através do *script* em Python, 19 classes foram estabelecidas para agrupá-los, levando em consideração o formato de sua apresentação, seu uso nas narrativas e sua anotação sintática de acordo com o modelo das UD. Por exemplo, *tokens* que expressam horários foram agrupados em uma mesma classe por apresentarem formato similar (números sozinhos ou separados por dois-pontos e seguidos de abreviaturas da palavra "horas"), por serem utilizados para o mesmo objetivo (indicar horários de diversas situações no contexto clínico) e serem anotados da mesma forma (prototipicamente como *obl* para *deprel* e como NUM ou NOUN para POS, a depender da aparição ou não das letras dentro do *token*).

Pode-se resumir a metodologia nesses passos: O subcorpus utilizado para esta análise foi resultado de um processamento feito com expressões regulares que extraem os algarismos numéricos. De cada expressão com algarismo numérico encontrada, manualmente foi-se definindo as classes.

[13] Disponível em https://arborator.ilpga.fr/q.cgi.

Essas classes foram ordenadas pela quantidade de expressões que cada uma das classes ocorria no corpus[14]. Dessas, as 10 primeiras foram eleitas para a análise de dados nas diretrizes UD com a redação de orientações de anotação para cada situação.

3. Resultados e análises

O Quadro 3 caracteriza e exemplifica cada uma das 19 classes.

Classe	**Formato do *token***	**Uso no contexto de narrativas clínicas**	**Exemplos[15]**
1	Letras e algarismos em siglas ou acrônimos, podendo incluir barra ("/") em símbolos de unidade de medida isolados.	Fórmulas químicas de substâncias utilizadas no tratamento	O2 (oxigênio)
		Acrônimos	FiO2 (fração inspirada de oxigênio)
		Siglas para número de gestações e partos, classificação de doenças, nome de remédios e tipos de exames	G3 (terceira gestação) DM2 (diabetes mellitus tipo 2)
		Símbolos de unidade de medida	cmH2O g/cm²
2	Algarismo e símbolo da unidade de medida juntos no mesmo *token*.	Dosagem de medicamentos que o paciente utiliza, está utilizando ou irá utilizar	30ml/h 20mg 1cp
		Medida de pressão	70mmHg 5cmH2O
		Frequência cardíaca	97bpm
		Temperatura (grau Celsius)	35,9ºC
		Tamanho, volume ou peso de órgãos	2x2cm
		Peso, altura e idade do paciente	112kg 1,62m
3	Símbolo "+" acompanhado do numeral e da barra ("/").	Avaliação de intensidade de edema no corpo do paciente	+3/+
		Avaliação de nível de palidez do paciente	2+/4
		Avaliação de gravidade de sopro cardíaco	+4/+6
4	Símbolos "+" ou "-" acompanhados de um número decimal.	Representa o grau de determinados problemas de visão do paciente.	+1,00 -1,00
5	Algarismos separados por barra ("/") ou hífen ("-") indicando data, podendo conter a abreviatura de meses por escrito. Aparecem em formatos como DD/MM/AAAA, DD/MM/AA, mes/AAAA, DD-MM-AAAA e outros.	Data da chegada ou internação do paciente.	07/01/2012
		Data da coleta dos dados do paciente apresentados na narrativa.	06-09-2007
		Data de atendimentos médicos, procedimentos ou exames que foram ou serão realizados pelo paciente.	30/06 jan/15
6	Abreviações D ou DI acompanhadas de um algarismo.	Contagem de dias desde a internação do paciente.	D7 18ºDI

[14] Como definida mais adiante na seção resultados.

[15] Em exemplos que possuem mais de um *token*, o TCA está destacado em negrito.

	Pode apresentar o símbolo "+" entre D e o algarismo, na forma cardinal ou ordinal.		D+2
7	Algarismo acompanhado da barra ("/") e, na maioria dos casos, a abreviatura de "horas": h, hs, hrs ou hr. Os números mais frequentes são 2/2, 4/4, 6/6, 8/8 e 12/12.	Indicação da frequência de administração do medicamento ao paciente, geralmente com relação a horas.	2/2hrs 6/6h 12/12
8	Algarismo associado ao símbolo de ângulo ou temperatura (°).	Medida da angulação orientada para a cabeceira da cama do paciente.	45°
		Medida da temperatura do paciente.	37,4°
9	Algarismos separados por ":" acompanhados ou não da abreviatura de "horas": h, hs, hrs ou hr; ou algarismos sozinhos acompanhados pelas abreviaturas de "horas". Aparece em formatos como HH:MM, HH:MMh, HH:MMhrs, HHh e outros.	Horário da chegada ou internação do paciente.	00h
		Horário da coleta dos dados do paciente apresentados na narrativa.	01:55HS
		Marcação do horário de procedimentos ou exames a serem realizados pelo paciente.	13:00
10	Algarismos separados por hífen ou barra.	Indicação de período de tempo incerto.	**6-7** anos
		Classificação em uma escala em um intervalo.	APGAR **1/2**
		Indicação de intervalo em alguma medida ou resultado de exame	CR **0,88-0,9**
11	Em listas, algarismo acompanhado de hífen sem espaçamento.	Listar medicamentos utilizados pelo paciente.	1- 2-
12	Sigla acompanhada de algarismo, geralmente com hífen.	Representa o número do quarto, ala, leito ou outro tipo de acomodação em que se encontra o paciente	UTI-3 UI-2 LEITO-206-2
13	Letras T, L ou S acompanhadas de algarismo.	Número de vértebra lombar	L4-L5
		Número de vértebra torácica	T11
		Número de vértebra da região sacral	S1
14	Dois conjuntos de dois ou três algarismos separados pelos caracteres "x", "X" ou "/".	Medição da pressão arterial	130X70 137x75 150/100
15	Algarismo seguido de barra e número múltiplo de 5 (como 20, 30, 40, 80, 200).	Medição da acuidade visual do paciente	20/20
16	Algarismo e letra ou símbolo indicando número ordinal.	Indica o dia da internação ou dia após cirurgia (pós-operatório)	**2°** DI **7º** PO
		Indica o número da consulta ou cirurgia	**1ª** CONSULTA
		Indica o número do dedo dos pés ou das mãos envolvido no tratamento	**3°** quirodáctilo
17	Algarismo indicando número cardinal (inteiro ou decimal).	Indicam quantidades variadas (dosagem de medicamentos, dias, horas, quantidade de consultas, número de dias etc)	**7** ML/H **5** dias
		Avaliação em escalas	Glasgow **9**
		Dados do paciente (como peso, idade, altura)	**79** anos

		Resultados de exames	Cr - **0,7**
18	Sigla R acompanhada de algarismo Número 9999	Sigla ou número utilizado para substituir dados identificadores no corpus. Não aparece naturalmente nas narrativas clínicas. A sigla R acompanhada de um algarismo é utilizada para indicar residentes diferentes e o número 9999, para substituir o número de registro do profissional no Conselho Regional de Enfermagem.	R2 9999
19	Algarismo acompanhado de diversos tipos de caracteres, letras, símbolos ou pontuações de forma não padronizada.	Podem apresentar diversos significados na narrativa. São *tokens* variados que não foram identificados como pertencentes a nenhuma das outras categorias. Requerem tratamento manual.	esquerdo.4 38+5 95.MV+

QUADRO 3 – Categorias de tipos de *tokens* com algarismo encontrados em corpus de narrativas clínicas.
Fonte: elaborado pelos autores

Os *tokens* agrupados na classe 19 foram, em grande parte, resultantes de problemas de tokenização e de grafia do texto original. Outras instâncias lidavam com casos muito específicos e de baixa frequência no corpus, de forma que não foi possível generalizar nas expressões regulares usadas para a classificação. Esses são casos que precisam ser avaliados manualmente antes do processamento do corpus. Um exemplo desse tipo de *token* é "UBSR1MARINAR2FERNANDO", em que a falta de espaçamento entre as palavras não permitiu identificar os *tokens* de anonimização "R1" e "R2" que estão inseridos entre os nomes. Outro TCA fora do padrão encontrado foi "50mgFurosemida"; mais uma vez, a falta de espaçamento impediu a tokenização da expressão numérica da dose de um medicamento e de seu nome.

Das 19 classes, selecionou-se as 10 classes mais frequentes[16]. Dessas 10, extraiu-se frases de cada classe para a anotação em UD conforme o Quadro 4. A quantidade total de frases anotadas neste estudo em andamento são 33. Assim, encontram-se, no Quadro 4, as 33 frases que estão anotadas com as etiquetas UD na subseção 4.1[17], aqui estão numeradas de 01 a 33 e ordenadas pelo *ranking* de frequência da classe no corpus, seguidas do número da sua respectiva classe e o número total de frases de cada classe possui.

N.	**Frases**	***Ranking***	**Classe**	**Total de frases**
01	"MANTENDO DOPAMINA EM 56 ML/H"	1	17	6
02	"Paciente , 79 anos"			
03	"Uso: Atenol 50 - 1xd."			

[16] As 10 classes de maior frequência são as classes nessa ordem da mais frequente para a menos frequente: 17, 2, 9, 19, 5, 1, 7, 14, 16 e 18. O gráfico 1 tem ilustrado o quantitativo percentual de cada uma.

[17] subseção onde se encontram também as árvores de cada uma das 33 frases bem como as decisões de anotação de cada uma.

04	“Cr - 0,7 ; glic - 140 ; ureia - 33”			
05	“Lab (07/05/15): Hemograma sem alterações Cr 0,9”			
06	“Ramsay 5, pupilas isomióticas.”			
		2	2	0[18]
07	“18h: Paciente permanece com hipotensão”	3	9	7
08	“01:00: DM + Anemia.”			
09	“EVOLUIU PARA PCR ÁS 11:30 HORAS”			
10	“ÓBITO ÀS 12H48MIN.”			
11	“PEÇO URINA 24H.”			
12	“1cp a cada 12h”			
13	“Afebril há 24h.”			
14	“POSITIVO PAR ISQUEMIA7-5-15”	4	19	2
15	“DRC ESTAGIO 5FALENCIA ENXERTO RENAL”			
16	“CX AGENDADA PARA 21/02 PELA MANHÃ.”	5	5	5
17	“16/05/2007, 08:56h: Sonolento, pouco responsivo.”			
18	“Retorno dia 11/10/16.”			
19	“Nova bx em dez/14”			
20	“EXAMES (03/02/2015) TSH 3.”			
21	“Suporte de O2 sob máscara de névoa úmida.”	6	1	3
22	“USA LST, ANLO, HCTZ, PT4, MTF”			
23	“DM2 diagnosticado há 6 anos.”			
24	“Ticlopidina 250 mg 12/12h”	7	7	5
25	“Levotiroxina 50 mcg 1x ao dia”			
26	“Está fazendo hidratação 3X/dia .”			
27	“- Enalapril 5 mg 12/12 horas.”			
28	“Diltiazem 60 mg (2x/d)”			
29	“PA: 150/75 mmhg.”	8	14	2
30	“18:00 PA = 150/80 RESP = 16”			
31	“# 3º DI UTI.”	9	16	1
32	“Enfermeria Florence coren 9999”	10	18	2
33	“R1 Vital Brasil”			
	Total			**33**

QUADRO 4 – As 33 expressões com algarismo numérico extraídas do corpus e organizadas por classe.
Fonte: elaborado pelos autores.

A partir do uso de expressões regulares elaboradas para identificar as classes do Quadro 1, foi possível computar o número de ocorrências encontradas para cada uma delas e a frequência correspondente no corpus SemClinBr. Esses dados são apresentados no Gráfico 1.

[18] *Tokens* que apresentam o algarismo junto do símbolo de unidade de medida não se mostram produtivos para uma anotação sintática.

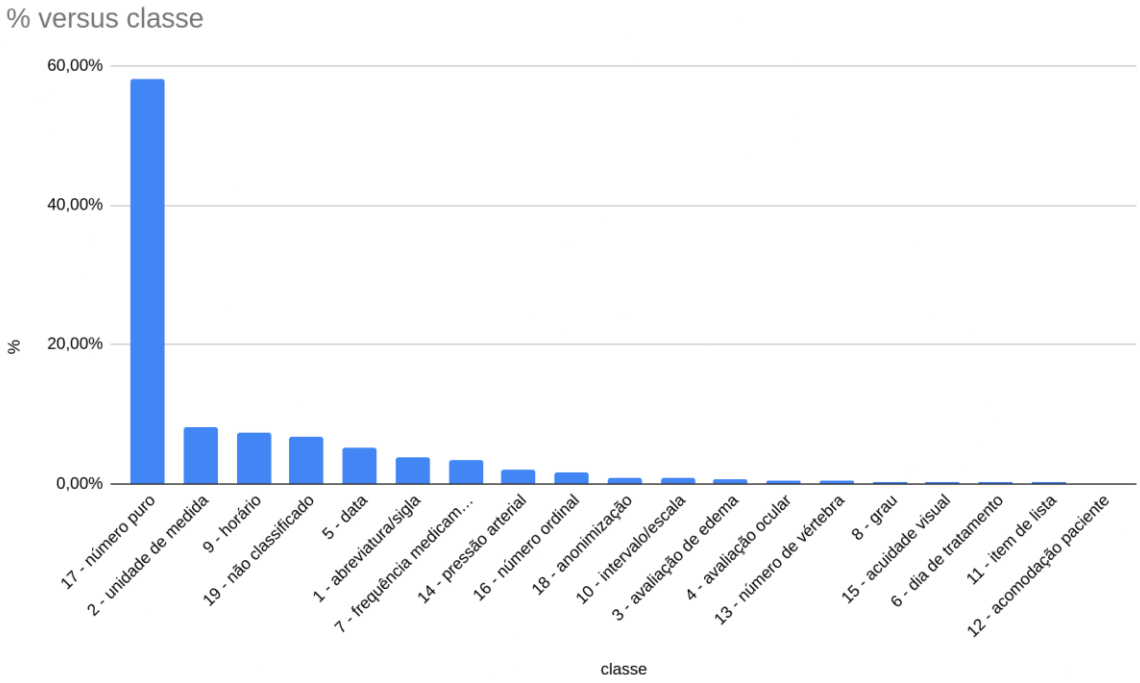


GRÁFICO 1 – Frequência da ocorrência de *tokens* por classe[19] contendo algarismos do corpus SemClinBr.
Fonte: elaborado pelos autores.

A partir dos dados da Gráfico 1, pode-se perceber que mais da metade dos TCA são os da classe 17, que são os *tokens* compostos apenas de algarismos não acompanhados de outros caracteres (como pontuação, símbolos e letras), compreendendo 58% do total encontrado no corpus. Em geral, esses *tokens* não causam problemas no momento da anotação.

Em seguida, vale ressaltar a ocorrência dos *tokens* da classe 2 (algarismos acoplados com símbolos de unidades de medida) na segunda posição, atingindo 8,10% do total de TCA. Essa é uma característica típica do domínio de narrativas clínicas, já que os textos costumam lidar com informações de dosagem de medicamentos e outras medições. É um tipo de *token* que, na anotação em UD, pode levantar discussões sobre sua natureza numérica (com uma classificação POS de NUM) ou nominal (com uma classificação POS de NOUN). É importante que a decisão de anotação leve em consideração o objetivo de uso do corpus para definir qual a melhor forma de identificar esse tipo de *token* e a informação que ele constrói. Sendo o objetivo do uso do corpus anotado o de treinar modelos de linguagem para anotação automática de outras narrativas clínicas, quanto mais informações de anotação, quanto mais detalhes anotados, havendo mais camadas (morfológica, sintática, semântica, com ontologias) o modelo produzido por redes neurais com um corpus assim anotado irá anotar com maior precisão. Para o corpus selecionado, dando atenção especial aos *tokens* numéricos,

[19] As classes estão definidas no Quadro 3.

a decisão de anotação é a de encontrar o máximo de informações possíveis que o *token* numérico constrói dentro da expressão nas várias camadas dos estratos microlingúísticos. Além disso, a anotação feita deve ser aquela acordada entre os anotadores.

É interessante notar, também, a frequência de ocorrência de *tokens* das classes 9 (que abrange *tokens* com diversas formas de expressão de horário) com 7,25%, e 5 (os formatos de expressão de data), com 5,10%. Como as narrativas clínicas referem-se ao estado do paciente em dado momento (seja nas notas de evolução ou no sumário de alta), é esperado encontrar informações que localizam temporalmente o relato. Uma classe que também se refere à localização temporal é a classe 6, que corresponde à indicação do dia de tratamento; no entanto, ela foi uma das menos frequentes (0,17%). Todavia, isso significa apenas que o *token* no formato específico aqui estudado (com as siglas D ou DI seguidas de um numeral) é pouco frequente, mas a informação pode aparecer de formas diferentes no corpus, inclusive com o uso de números ordinais (classe 16) separado da sigla (como em "8° DI por CA") e com o uso de números cardinais também separados de outras palavras (como em "1 dia de UTI"). Isso mostra a variedade de expressões que podem ser usadas para se tratar do conceito temporal. Outro dado importante que se destaca no Gráfico 1 é que a quarta classe[20] de *tokens* com algarismo mais frequente encontrada foram de casos que não estavam dentro dos padrões estabelecidos para nenhuma das outras categorias. A partir de uma análise mais detalhada, viu-se que são *tokens* que sofreram problemas de tokenização (devido a falta de espaçamento entre palavras e pontuações), de grafia e de digitação, e casos muito particulares que dificultam a generalização em expressões regulares e requerem categorias muito específicas. É necessário que haja uma revisão manual desses casos para que eles sejam tratados em uma etapa de pré-processamento do corpus ou sejam realocados para novas categorias ou categorias adaptadas.

3.1. Orientações para anotação em UD

Cada uma das classes de *tokens* com algarismos apresentadas no Quadro 3 possuem significados e características sintáticas próprios (os quais foram consideradas no momento da categorização). Portanto, a anotação em UD de cada um deles se dará de forma adequada a essas características. Nas próximas subseções, estão dispostas as decisões tomadas para anotar os TCA pertencentes às dez classes mais frequentes identificadas na subseção 5.2, acompanhadas de árvores para exemplificação dessa anotação. As árvores foram feitas a partir de exemplos retirados do corpus SemClinBr, que podem ser apenas segmentos de sentenças selecionados para destacar a expressão com algarismo ou uma sentença completa. As decisões de anotação foram tomadas com base no Manual de Anotação de Relações de Dependência – versão revisada e estendida (DURAN, 2022), no Manual de Anotação de POS *tags* (DURAN, 2021), nas diretrizes do projeto de anotação apresentado em Oliveira et al.

[20] Classe 19: Algarismo acompanhado de diversos tipos de caracteres, letras, símbolos ou pontuações de forma não padronizada.

(2022) e estudando as particularidades de textos de narrativas clínicas, também considerando a relevância das informações que podem ser extraídas desses textos.

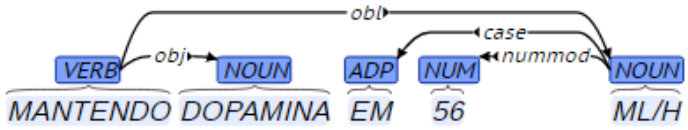
A seguir são apresentadas as classes pela ordem da frequência do Gráfico 1, a saber: 17, 2, 9, 19, 5, 1, 7, 14, 16 e 18.

Posição 1 - Classe 17: Números cardinais (inteiros ou decimais) apenas com algarismo ou o separador decimal

POS: Sempre serão classificados como NUM.

deprel: Geralmente, ocorrem como *nummod* (em casos clássicos de natureza quantitativa) ou como *nmod* (em classificação de escalas e quando são promovidos a *head* pois o nominal que o acompanha está elíptico). Também podem ser *root* (ou *head*) em relações de cópula intermediadas pelos símbolos ":", "=" ou "-".

- *nummod* (casos clássicos de natureza quantitativa)



.

FIGURA 2 – *nummod* na sentença "MANTENDO DOPAMINA EM 56 ML/H"
Fonte: elaborada pelos autores.

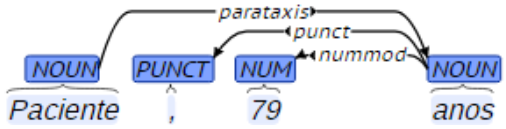


FIGURA 3 – Exemplo de *nummod* no segmento "Paciente , 79 anos"
Fonte: elaborada pelos autores.

- *nmod* (no caso de expressão de dosagem de um medicamento quando o símbolo da unidade de medida está elíptico)

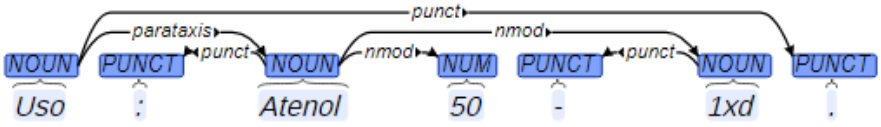


FIGURA 4 – Exemplo de *nmod* no segmento "Uso: Atenol 50 - 1xd."
Fonte: elaborada pelos autores

- *root* em relações de cópula (resultados de exames intermediados ou não por "=", ":" ou "-")

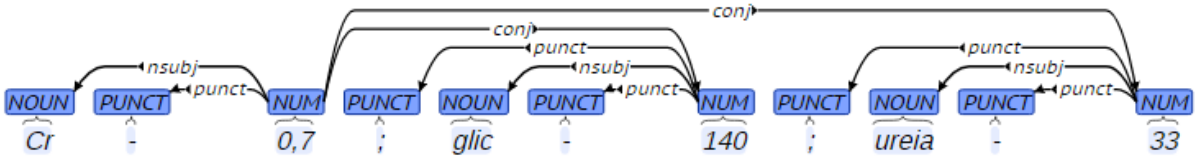


FIGURA 5 – Exemplo de *root* no segmento "Cr - 0,7 ; glic - 140 ; ureia - 33"
Fonte: elaborada pelo autor

- *head* em relações de cópula (resultados de exames intermediados ou não por "=", ":" ou "-")

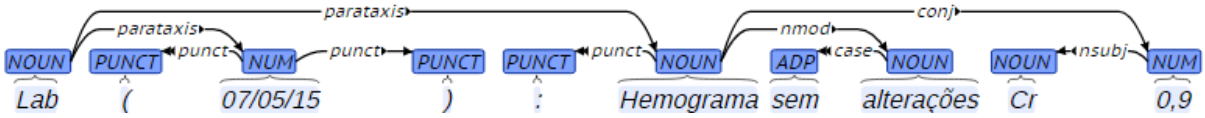


FIGURA 6 – Exemplo de *head* no segmento "Lab (07/05/15): Hemograma sem alterações Cr 0,9"
Fonte: elaborada pelos autores

- *nmod* (classificação de escalas)

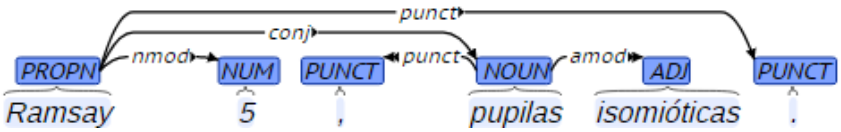


FIGURA 7 – Exemplo de *nmod* no segmento "Ramsay 5, pupilas isomióticas."
Fonte: elaborada pelas autoras

Posição 2 - Classe 2: Algarismo e símbolo da unidade de medida juntos no mesmo *token*

Tokens que apresentam o algarismo junto do símbolo de unidade de medida não se mostram produtivos para uma anotação sintática. Portanto, a orientação é que haja um pré-processamento do corpus para separar o numeral e a unidade de medida em *tokens* diferentes.

Posição 3 - Classe 9: Algarismos separados por ":" acompanhados ou não da abreviatura de "horas": h, hs, hrs ou hr; ou algarismos sozinhos acompanhados pelas abreviaturas de "horas"

POS: Serão considerados como NUM se o único caractere diferente de algarismos for o dois-pontos (p. ex. 18:00) e como NOUN se for um *token* alfanumérico (p. ex. 14h).

deprel: Em geral, *tokens* no formato HH:MM serão *nummod* se o *token* de "horas" ou suas abreviaturas como "h", "hr" ou "hrs" estiverem presentes como outro *token* (sendo esse *token* o seu *head*). Se uma das diversas abreviaturas da palavra "horas" estiver no mesmo *token*, o *token* pode admitir as relações de *obl* (geralmente com um *head* verbal, mas também ocorrendo com um *head* nominal se o verbo estiver elíptico), *nmod* (quando fizer parte de nome de exames e quando for um complemento nominal na indicação de algum tipo de frequência) ou *obj* (quando precedido pelo verbo "haver").

- *obl* (*head* verbal)

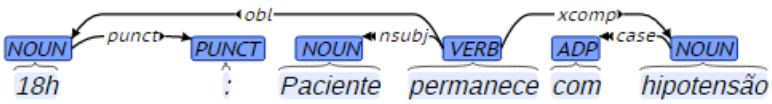


FIGURA 8 – Exemplo de *obl* na sentença "18h: Paciente permanece com hipotensão"
Fonte: elaborada pelos autores

- *obl* (*head* nominal)

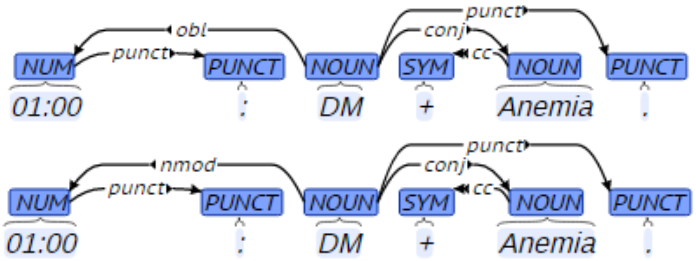


FIGURA 9 – Exemplo de *obl* no segmento "01:00: DM + Anemia."
Fonte: elaborada pelos autores

Vale notar que, na anotação da Figura 9, o símbolo "+" recebeu a relação *cc*, de coordenação conjuntiva, de acordo com o apontado por Moon et al. (2011) sobre o uso desse símbolo no contexto das narrativas clínicas.

- *nummod* (seguido pelo *token* "horas")

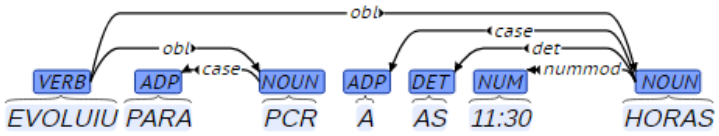


FIGURA 10 – Exemplo de *nummod* na sentença "EVOLUIU PARA PCR ÀS 11:30 HORAS"
Fonte: elaborada pelos autores

- *nmod*

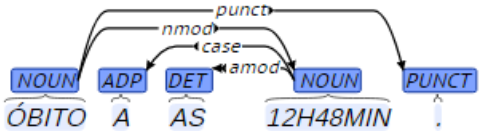


FIGURA 11 – Exemplo de *nmod* no segmento "ÓBITO ÀS 12H48MIN."
Fonte: elaborada pelos autores

- *nmod* (nome de exame)

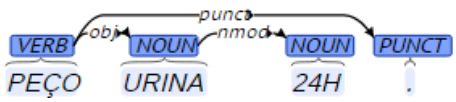


FIGURA 12 – Exemplo de *nmod* na sentença "PEÇO URINA 24H."
Fonte: elaborada pelos autores

- *nmod* (com frequência de administração de medicamento)

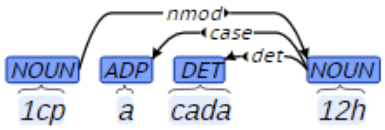


FIGURA 13 – Exemplo de *nmod* no segmento "1cp a cada 12h"
Fonte: elaborada pelos autores

- *obj* (precedido pelo verbo "haver")

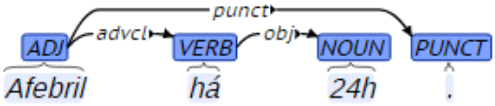


FIGURA 14 – EXEMPLO DE *OBJ* NA SENTENÇA "AFEBRIL HÁ 24H."
FONTE: ELABORADA PELOS AUTORES

Posição 4 - Classe 19: Algarismo acompanhado de diversos tipos de caracteres, letras, símbolos ou pontuações de forma não padronizada

POS: Sempre serão anotados como NOUN.

deprel: Podem admitir diversos tipos de relações, a depender da estrutura da sentença e decisões de anotação. Seguem alguns exemplos de possíveis ocorrências e suas respectivas relações de dependência.

- *obl* (complemento de adjetivo)

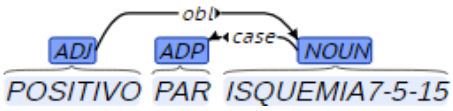


FIGURA 15 – Exemplo de *obl* no segmento "POSITIVO PAR ISQUEMIA7-5-15". Por problemas de transcrição da narrativa, não houve espaço entre "isquemia" e "7-5-15", colocando-os em um *token* único com consequente anotação como NOUN.
Fonte: elaborada pelos autores

- *nmod* (complemento nominal)

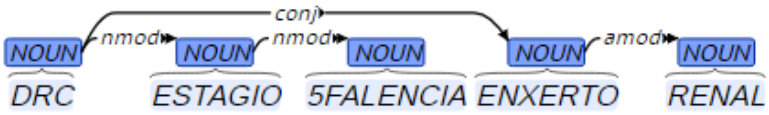


FIGURA 16 – Exemplo de *nmod* no segmento "DRC ESTAGIO 5FALENCIA ENXERTO RENAL". Por problemas de transcrição da narrativa, não houve espaço entre o número "5" e a palavra"falência", colocando-os em um *token* único com consequente anotação como NOUN.
Fonte: elaborada pelos autores

Posição 5 - Classe 5: Algarismos separados por barra ("/") ou hífen ("-") indicando data, podendo conter a abreviatura de meses por escrito

POS: Podem ser interpretados como NUM, se o único caractere diferente de algarismos for a barra (p. ex. 26/05); ou como NOUN, se for um *token* que mistura o nome do mês por escrito e o ano em algarismos (p. ex. maio/2016).

deprel: Por ser um adjunto que trata de uma localização temporal precisa (por exemplo, "no dia 10/04"), normalmente indicará uma relação de *obl*, mas podem admitir outras relações, como *nmod* e *parataxis*.

- *obl* (com verbo explícito)

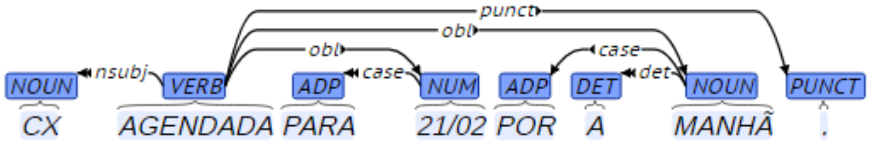


FIGURA 17 – Exemplo de *obl* na sentença "CX AGENDADA PARA 21/02 PELA MANHÃ."
Fonte: elaborada pelos autores

- *obl* (complemento do adjetivo)

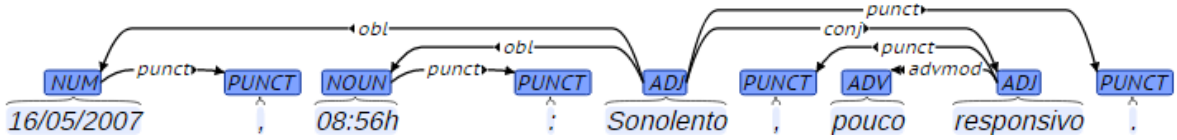


FIGURA 18 – Exemplo de *obl* no segmento "16/05/2007, 08:56h: Sonolento, pouco responsivo."
Fonte: elaborada pelos autores

- *nmod* (complemento nominal)

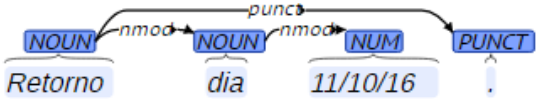


FIGURA 19 – Exemplo de *nmod* no segmento "Retorno dia 11/10/16."
Fonte: elaborada pelos autores

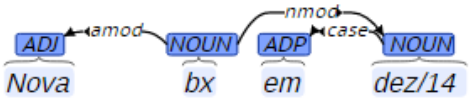


FIGURA 20 – Exemplo de *nmod* no segmento "Nova bx em dez/14"
Fonte: elaborada pelos autores

- *parataxis* (quando indicado entre parênteses)

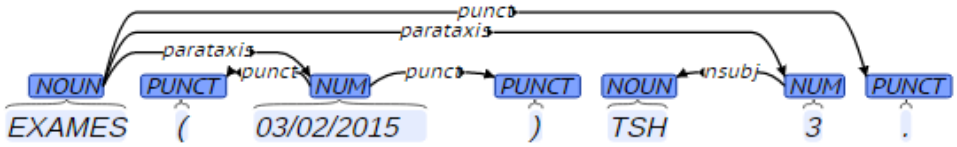


FIGURA 21: Exemplo de *parataxis* no segmento "EXAMES (03/02/2015) TSH 3."
Fonte: elaborada pelos autores

Posição 6 - Classe 1: Letras e algarismos que formam siglas ou acrônimos, podendo incluir barra ("/") no caso de símbolos de unidade de medida isolados

POS: Como correspondem a um encurtamento de um ou mais substantivos, serão sempre considerados como NOUN.

deprel: Costumam receber as relações *nmod*, *nsubj* ou ser *head* em outras relações.

- *nmod* (complemento nominal)

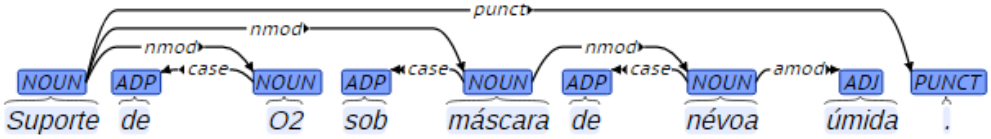


FIGURA 22 – Exemplo de *nmod* no segmento "Suporte de O2 sob máscara de névoa úmida."
Fonte: elaborada pelos autores

- *head* (em relação de *conj*)

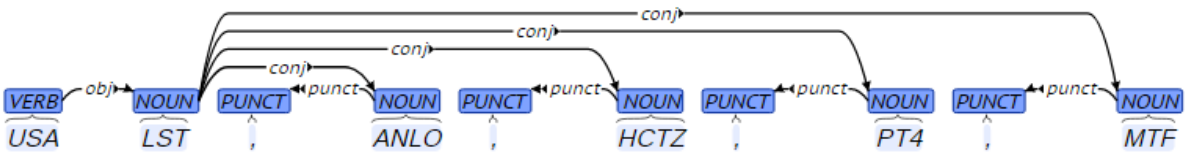


FIGURA 23 – Exemplo de *head* na sentença "USA LST, ANLO, HCTZ, PT4, MTF"
Fonte: elaborada pelos autores

- *nsubj* (no caso de um nome de doença)

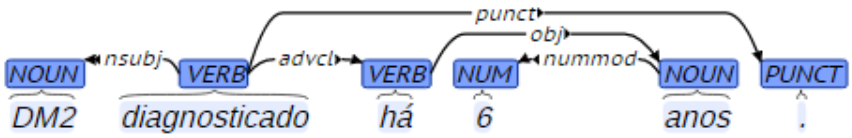


FIGURA 24 – Exemplo de *nsubj* na sentença "DM2 diagnosticado há 6 anos."
Fonte: elaborada pelos autores

Posição 7 - Classe 7: Algarismo acompanhado da barra ("/") e, na maioria dos casos, a abreviatura de "horas": h, hs, hrs ou hr

POS: Serão considerados NOUN, se seguidos de "h", "hr", "hrs" ou "x" (sigla para "vezes") no mesmo *token*; ou NUM, se possuírem apenas os algarismos separados pela barra.

deprel: Podem receber relações de *nmod*, se o *head* for um nominal; *obl*, se o *head* for verbal, e *nummod* caso acompanhe a palavra "horas" (ou suas abreviaturas) como um *token* separado. Também admite relação de *parataxis* se informado como complemento entre parênteses.

- *nmod* (complemento nominal)

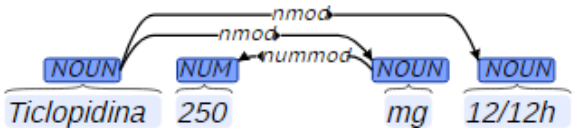


FIGURA 25 – Exemplo de *nummod* no segmento "Ticlopidina 250 mg 12/12h"
Fonte: elaborada pelos autores

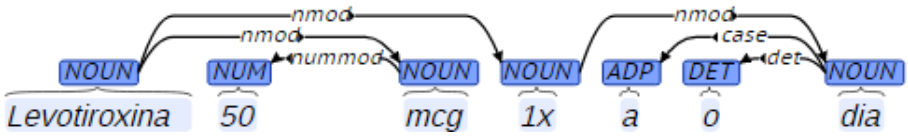


FIGURA 26 – Exemplo de *nmod* no segmento "Levotiroxina 50 mcg 1x ao dia"
Fonte: elaborada pelos autores

- *obl* (complemento verbal)

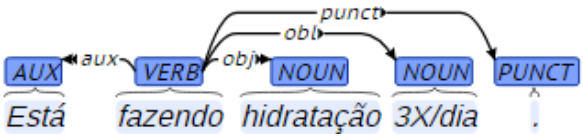


FIGURA 27 – Exemplo de *obl* na sentença "Está fazendo hidratação 3X/dia ."
Fonte: elaborada pelos autores

- *nummod* (modificador numérico)

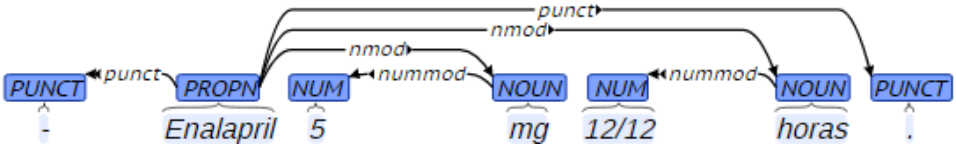


FIGURA 28 – Exemplo de *nummod* no segmento "- Enalapril 5 mg 12/12 horas."
Fonte: elaborada pelos autores

- *parataxis* (quando indicado entre parênteses)

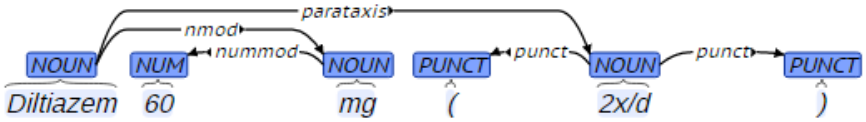


FIGURA 29 – Exemplo de *parataxis* no segmento "Diltiazem 60 mg (2x/d)"
Fonte: elaborada pelos autores

Posição 8 - Classe 14: Dois conjuntos de dois ou três algarismos separados pelos caracteres "x", "X" ou "/"

POS: Sempre serão anotados como NUM.

deprel: Podem receber relações de *nummod*, quando acompanhado da unidade de medida de pressão arterial em um *token* diferente; e *root* (ou *head*) se não acompanhado da unidade de medida e em relações de cópula intermediadas ou não pelos símbolos ":", "=" ou "-".

- *nummod* (modificador numérico)

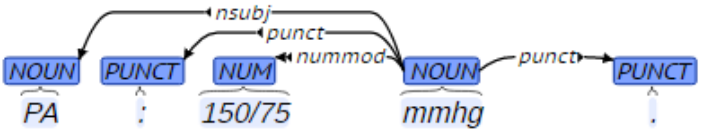


FIGURA 30 – Exemplo de *nummod* no segmento "PA: 150/75 mmhg."
Fonte: elaborada pelos autores

- *root* (em relação de cópula)

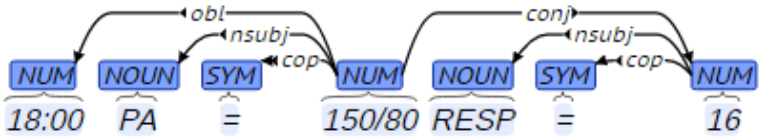


FIGURA 31 – Exemplo de *root* no segmento "18:00 PA = 150/80 RESP = 16"
Fonte: elaborada pelos autores

Posição 9 - Classe 16: Números ordinais indicados por algarismo e letra ou símbolo

POS: Sempre serão classificados como ADJ.

deprel: Sempre receberão a relação de *amod*.

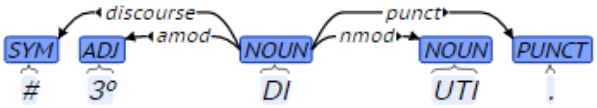


FIGURA 32 – Exemplo de *amod* no segmento "# 3o DI UTI."
Fonte: elaborada pelos autores

Posição 10 - Classe 18: Sigla R acompanhada de algarismo; número 9999

POS: Serão considerados como NOUN, se forem *tokens* alfanuméricos (como R1), e como NUM, se forem compostos apenas de algarismos (como 9999).

deprel: Podem assumir as relações de dependência *nmod*, *root* ou ser o *head* em outras relações. Seguem alguns exemplos de possíveis ocorrências e suas respectivas relações de dependência.

- *nmod* (acompanhado de seu *head* nominal)

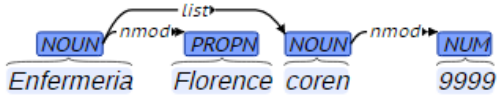


FIGURA 33 – Exemplo de *nmod* no segmento "Enfermeria Florence coren 9999"
Fonte: elaborada pelos autores

- *root* (em segmentos que não possuem verbo)

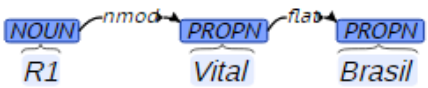


FIGURA 34 – Exemplo de *root* no segmento "R1 Vital Brasil"
Fonte: elaborada pelos autores

Em suma, as decisões de anotação de etiquetas de POS e relações de dependência (deprel) para as dez classes de *tokens* contendo algarismo mais comuns encontradas no corpus SemClinBr são as dispostas no Quadro 5 a seguir, adaptado do Quadro 1.

Classe	**Formato do *token***	**Uso no contexto de narrativas clínicas**	**Exemplos**[21]	**Anotação em UD**
1	Letras e algarismos em siglas ou acrônimos, podendo incluir barra ("/") em símbolos de unidade de medida isolados.	Fórmulas químicas de substâncias utilizadas no tratamento	O2 (oxigênio)	**POS:** NOUN **deprel:** *nmod*, *nsubj*, ou *head* em outras relações
		Acrônimos	FiO2 (fração inspirada de oxigênio)	
		Siglas para número de gestações e partos, classificação de doenças, nome de remédios e tipos de exames	G3 (terceira gestação) DM2 (diabetes mellitus tipo 2)	
		Símbolos de unidade de medida	cmH2O g/cm²	
2	Algarismo e símbolo da unidade de medida juntos no mesmo *token*.	Dosagem de medicamentos que o paciente utiliza, está utilizando ou irá utilizar	30ml/h 20mg 1cp	Não serão anotadas neste formato; devem ser tokenizadas em um pré-processamento do texto antes da anotação.
		Medida de pressão	70mmHg 5cmH2O	
		Frequência cardíaca	97bpm	
		Temperatura (grau Celsius)	35,9ºC	
		Tamanho, volume ou peso de órgãos	2x2cm	
		Peso, altura e idade do paciente	112kg 1,62m	
14	Dois conjuntos de dois ou três algarismos separados pelos caracteres "x", "X" ou "/".	Medição da pressão arterial	130X70 137x75 150/100	**POS:** NUM **deprel:** *nummod*, *root*, ou *head* em outras relações
5	Algarismos separados por barra ("/") ou hífen ("-") indicando data, podendo conter a abreviatura de meses por escrito. Aparecem em formatos como DD/MM/AAAA,	Data da chegada ou internação do paciente.	07/01/2012	**POS:** NUM ou NOUN **deprel:** *obl*, *nmod*, ou *parataxis*
		Data da coleta dos dados do paciente apresentados na narrativa.	06-09-2007	
		Data de atendimentos médicos, procedimentos ou exames que foram ou serão realizados pelo paciente.	30/06 jan/15	

[21] Em exemplos que possuem mais de um token, o TCA está destacado em negrito.

	DD/MM/AA, mes/AAAA, DD-MM-AAAA e outros.			
7	Algarismo acompanhado da barra (“/”) e, na maioria dos casos, a abreviatura de “horas”: h, hs, hrs ou hr. Os números mais frequentes são 2/2, 4/4, 6/6, 8/8 e 12/12.	Indicação da frequência de administração do medicamento ao paciente, geralmente com relação a horas.	2/2hrs 6/6h 12/12	**POS:** NUM ou NOUN **deprel:** *nmod*, *obl*, *nummod*, ou *parataxis*
9	Algarismos separados por “:” acompanhados ou não da abreviatura de “horas”: h, hs, hrs ou hr; ou algarismos sozinhos acompanhados pelas abreviaturas de “horas”. Aparece em formatos como HH:MM, HH:MMh, HH:MMhrs, HHh e outros.	Horário da chegada ou internação do paciente.	00h	**POS:** NUM ou NOUN **deprel:** *nummod*, *obl*, *nmod*, ou *obj*
		Horário da coleta dos dados do paciente apresentados na narrativa.	01:55HS	
		Marcação do horário de procedimentos ou exames a serem realizados pelo paciente.	13:00	
16	Algarismo e letra ou símbolo indicando número ordinal.	Indica o dia da internação ou dia após cirurgia (pós-operatório)	**2°** DI **7º** PO	**POS:** ADJ **deprel:** *amod*
		Indica o número da consulta ou cirurgia	**1ª** CONSULTA	
		Indica o número do dedo dos pés ou das mãos envolvido no tratamento	**3°** quirodáctilo	
17	Algarismo indicando número cardinal (inteiro ou decimal).	Indicam quantidades variadas (dosagem de medicamentos, dias, horas, quantidade de consultas, número de dias etc)	**7** ML/H **5** dias	**POS:** NUM **deprel:** *nummod*, *nmod*, *root*, ou *head* em outras relações
		Avaliação em escalas	Glasgow **9**	
		Dados do paciente (como peso, idade, altura)	**79** anos	
		Resultados de exames	Cr - **0,7**	
18	Sigla R acompanhada de algarismo Número 9999	Sigla ou número utilizado para substituir dados identificadores no corpus. Não aparece naturalmente nas narrativas clínicas. A sigla R acompanhada de um algarismo é utilizada para indicar residentes diferentes e o número 9999, para substituir o número de registro do profissional no Conselho Regional de Enfermagem.	R2 9999	**POS:** NOUN ou NUM **deprel:** *nmod*, *root*, ou *head* em outras relações
19	Algarismo acompanhado de diversos tipos de caracteres,	Podem apresentar diversos significados na narrativa. São *tokens* variados que não foram	esquerdo.4 38+5 95.MV+	**POS:** NOUN

	letras, símbolos ou pontuações de forma não padronizada.	identificados como pertencentes a nenhuma das outras categorias. Requerem tratamento manual.		**deprel:** diversas, a depender do conteúdo do *token*

QUADRO 5 – Resumo das decisões de anotação em UD para as dez classes de *tokens* contendo algarismos mais frequentes encontradas no corpus SemClinBr.

4. Considerações finais

Observando-se a frequência das diversas formas de apresentação de *tokens* com algarismos e das classes estabelecidas, percebeu-se que o quarto tipo mais frequente foram aqueles que não puderam ser classificados. Isso mostra a importância de uma análise aprofundada de *tokens* compostos por números presentes em corpora de narrativas clínicas, e a necessidade de um cuidado com esse tipo de *token*, já que não são previsíveis nem facilmente tratáveis, como exposto por Oliveira L. F. et al. (2022). Com esses problemas de tokenização, a anotação de um corpus de narrativas clínicas se mostra mais complexa, necessitando de uma verificação manual e profundo conhecimento do corpus tratado.

Os resultados dessa análise em português brasileiro também convergem com o apontado pelos trabalhos de Fan et al. (2013) e Kara et al. (2018). Além disso, como aponta Hanauer et al. (2019) sobre o estudo das variações de numerais em textos clínicos, o tratamento de *tokens* com algarismos também é essencial para garantir uma base confiável para a extração de informação desses textos.

Segundo as UD e as diretrizes de projetos no Brasil, *tokens* no formato híbrido (que intercalam letras, números, sinais de pontuação e/ou símbolos) receberão a etiqueta POS de NOUN (DURAN; LOPES; PARDO, 2021); mas, no domínio clínico, a etiqueta NUM pode ser mais útil para extração de informação. Por isso, tomou-se a decisão de que certos tipos de *tokens*, como aqueles que contêm o algarismo acompanhado de símbolos de unidade de medida deveriam ser separados em um pré-processamento do corpus antes da anotação. Dessa forma, *tokens* únicos como “80mg” passarão a ser anotados como os *tokens* “80” (com etiqueta POS de NUM) e “mg” (com etiqueta POS de NOUN).

O presente trabalho se propôs a analisar a ocorrência de *tokens* com algarismos em textos de narrativas clínicas visando avaliar seus desafios tanto para a elaboração de ferramentas para tarefas de PLN quanto para sua anotação sintática segundo o projeto UD. Com os resultados alcançados, foi possível contribuir para a elaboração de diretrizes de anotação em UD de dez classes desses *tokens* no domínio estudado e identificar possíveis problemas de tokenização que podem ser encontrados nesse tipo de texto. Tudo isso colabora para o avanço nos estudos em Processamento de Língua Natural e no desenvolvimento de ferramentas de PLN aplicáveis ao português brasileiro na área médica, especialmente aquelas que tomam como base a anotação sintática em UD. Os próximos passos incluem a análise das nove classes de *tokens* com algarismos remanescentes neste estudo para a determinação de suas diretrizes de anotação, bem como a elaboração de diretrizes gerais para posterior anotação e revisão do corpus DepClinBr.

Informações complementares

Avaliação e resposta dos autores

Avaliação: https://doi.org/10.25189/rabralin.v22i2.2127.R

Editores

Marta Deysiane Alves Faria Sousa
Afiliação: Universidade Federal de Sergipe
ORCID: https://orcid.org/0000-0003-1297-2037

Adriana Pagano
Afiliação: Universidade Federal de Minas Gerais
ORCID: https://orcid.org/0000-0002-3150-3503

Jorge Baptista
Afiliação: Universidade do Algarve - INESC-ID Lisboa
ORCID: https://orcid.org/0000-0003-4603-4364

RODADAS DE AVALIAÇÃO

Avaliador 1: Roana Rodrigues
Afiliação: Universidade Federal de Sergipe
ORCID: https://orcid.org/0000-0002-7748-8716

Avaliador 2: Alisson Hudson
Afiliação: Instituto Federal de Alagoas
ORCID: https://orcid.org/0000-0001-6597-6547

AVALIADOR 1

No artigo intitulado “ANOTAÇÃO DE EXPRESSÕES COM NUMERAIS EM NARRATIVAS CLÍNICAS”, os autores apresentam a problemática da anotação (segundo as diretrizes da UD) de numerais, em formato de algarismo, em um corpus de narrativas clínicas e sugerem orientações para a sua anotação (em POS e deprel).

O trabalho apresenta contribuições tanto para tarefas de PLN, quanto para os estudos descritivos do português brasileiro (PB), com reflexões sobre a relevância, complexidade, proposta de classificação e de anotação dos numerais no corpus de análise. Além disso, os dados são descritos de maneira muito didática, facilitando a compreensão das etapas da pesquisa – e sua análise – para pessoas familiarizadas ou não com as diretrizes da UD.

A seguir, pontuo algumas sugestões que devem ser discutidas e avaliadas pelos autores:

Em 2.1, aparece pela primeira vez a menção ao trabalho de Moon et al. (2011). Como se trata de um trabalho apresentado em outros momentos do texto e tomado como referência, poderia haver mais (ou algum) detalhe(s) sobre o realizado naquela investigação. Em algum lugar no texto a referência está equivocada como Moon et al. (2022).

Na seção 3, Metodologia, é apresentado o passo a passo para a extração de casos de TCA do corpus. No entanto, toda a seção gera vários questionamentos: a quais "classes" os autores se referem? Quantas classes existem? Essas classes foram sugeridas por Moon et al. (2011)? Qual a dimensão do corpus de análise e do trecho do corpus analisado? Quantas ocorrências de TCA foram levantadas? Sabe-se que na seção 4, Resultados, todas essas perguntas são respondidas, mas foi um incômodo ler uma seção de Metodologia que não apresentava dados precisos do levantamento feito. Sendo assim, é importante: (i) a reorganização dos dados, contemplando, já na Metodologia, as informações acima apresentadas; ou (ii) inserir uma nota de rodapé, ao início da seção de Metodologia, informando que os dados mais detalhados sobre classes e quantidade de dados analisados serão explicados na seção seguinte.

Na seção 3.3, em lugar de uma Figura da planilha, sugere-se a reprodução da referida planilha em formato de Quadro.

Em 4.1, no Quadro 2, sugere-se a inclusão de duas colunas (transformando o Quadro em Tabela): Número de ocorrências | %. Com isso, é possível excluir a Tabela 1, da seção 4.2. Na verdade, sugere-se fortemente transformar a Tabela 1 em um gráfico de barras horizontais, o que torna a visualização e leitura dos dados quantitativos mais acessíveis e interessantes (no Gráfico, os dados podem/devem seguir a ordem das Classes propostas (de 1 a 19), sendo possível visualizar mais facilmente quais são as 10 classes mais frequentes). Além disso, não parece ser necessária a existência da nota de rodapé 8.

Em 4.3, os autores afirmam que as árvores foram feitas com base em exemplos retirados do corpus. *É muito importante informar no artigo quantos casos foram realmente analisados em cada classe para a efetiva elaboração da proposta de anotação realizada. Seria possível quantificar e relacionar a proposta de anotação ao número de casos analisados em cada classe?

AVALIADOR 2

Gostaria de parabenizar os autores pela discussão empreendida e pelo ótimo trabalho. Envio minhas sugestões anexadas em documentos em PDF para que possam ser apreciadas e, se julgarem importantes, possam ser adicionadas na versão a ser novamente submetida.

Ao meu ver, o texto é extremamente importante, mas há pontos cruciais de escrita científica que merecem ser contempladas e que estejam mais claras ao leitor, sobretudo no que trata sobre pontos específicos de uma pesquisa que poderia ser replicada.

Reforço que o texto é deveras interessante e que merece ser publicado, ficando a minha indicação de que os pontos comentados possam ser analisados, a fim de melhorar o resultado final, que é de mérito dos autores.

Meus parabéns pela pesquisa e pelo trabalho empreendido.

Conflito de Interesse

Os autores não têm conflitos de interesse a declarar.

Protocolo e Pré-Registro de Pesquisa

Os autores avaliaram os roteiros da Equator Network e concluíram que nenhum dos roteiros apresentados eram aplicáveis ao presente trabalho. A pesquisa conduzida não foi pré-registrada em um repositório institucional independente.

REFERÊNCIAS

DURAN, M.S.; LOPES, L.; PARDO, T.A.S. Descrição de numerais segundo modelo Universal Dependencies e sua anotação no português. **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2021)**, v. 13, p. 344–352, 29 nov. 2021. DOI: https://doi.org/10.5753/stil.2021.17814. Acesso em: 13 set. 2022

DURAN, M.S. Manual de Anotação de PoS tags: Orientações para anotação de etiquetas morfossintáticas em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). **Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação**, Universidade de São Paulo. São Carlos-SP, Outubro, 55p. 2021. Disponível em: https://drive.google.com/le/d/1BddPswn-_Ioo-A5GsldA1cO1kqbcCahb/view?usp=sharing. Acesso em 16 nov. 2022

DURAN, M.S. Manual de Anotação de Relações de Dependência - Versão Revisada e Estendida: Orientações para anotação de relações de dependência sintática em Língua Portuguesa, seguindo as diretrizes da abordagem Universal Dependencies (UD). **Relatório Técnico do ICMC 440. Instituto de Ciências Matemáticas e de Computação**, Universidade de São Paulo. São Carlos-SP, Outubro, 166p. 2022. Disponível em: https://drive.google.com/le/d/1ile8Wfxu1qdrZOmLGqkvVuQ4fXvHgVMo/view?usp=sharing. Acesso em 16 nov. 2022

FAN, J. et al. Syntactic parsing of clinical text: guideline and corpus development with handling ill-formed sentences. **Journal of the American Medical Informatics Association : JAMIA**, v. 20, n. 6, p. 1168–1177, 1 nov. 2013. DOI: http://doi.org/10.1136/amiajnl-2013-001810. Acesso em: 20 nov. 2022.

HANAUER, D. A. et al. Complexities, variations, and errors of numbering within clinical notes: the potential impact on information extraction and cohort-identification. **BMC Medical Informatics and Decision Making**, v. 19, n. S3, abr. 2019. DOI: 5 https://doi.org/10.1186/s12911-019-0784-1. Acesso em: 20 nov. 2022.

HUMPHREYS, B. L.; MCCRAY, A. T.; LINDBERG, D. A. B. The Unified Medical Language System. **Methods of Information in Medicine**, v. 32, n. 04, p. 281–291, 1993.

JURAFSKY, D.; MARTIN, J. **Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing**. Upper Saddle River, NJ: Prentice Hall, 2008.

KARA, E. et al. A Domain-adapted Dependency Parser for German Clinical Text. Vienna, Austria: **Proceedings of the 14th Conference on Natural Language Processing (KONVENS 2018)**, set. 2018. Disponível em: <https://konvens.org/proceedings/2018/PDF/konvens18_02.pdf>. Acesso em: 21 nov. 2022.

NIVRE, J. (2015). Towards a Universal Grammar for Natural Language Processing. In: Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2015. **Lecture Notes in Computer Science, vol 9041. Springer, Cham.**
https://doi.org/10.1007/978-3-319-18111-0_1. Acesso em: 13 set. 2022.

MARNEFFE, M. et al. Universal Dependencies. **Computational Linguistics 2021**; 47 (2): 255–308. doi: https://doi.org/10.1162/coli_a_00402. Acesso em: 13 set. 2022.

MOON, S. R.; PAKHOMOV, S.; RYAN, J. et al. Automated non-alphanumeric symbol resolution in clinical texts. **AMIA ... Annual Symposium proceedings. AMIA Symposium**, 2011, p. 979-986. Acesso em: 13 set. 2022.

NÉVÉOL, A., DALIANIS, H., VELUPILLAI, S. et al. Clinical Natural Language Processing in languages other than English: opportunities and challenges. **J Biomed Semant 9**, 12 **(2018)**. https://doi.org/10.1186/s13326-018-0179-8. Acesso em 17 maio 2023.

OLIVEIRA, L.E.S., PETERS, A.C., DA SILVA, A.M.P. et al. SemClinBr - a multi-institutional and multi-specialty semantically annotated corpus for Portuguese clinical NLP tasks. **J Biomed Semant 13, 13 (2022)**. https://doi.org/10.1186/s13326-022-00269-1. Acesso em: 13 set. 2022.

OLIVEIRA, L. F. A. et al. Challenges in Annotating a Treebank of Clinical Narratives in Brazilian Portuguese. **Lecture Notes in Computer Science**, v. 15, p. 90–100, 2022. DOI: https://doi.org/10.1007/978-3-030-98305-5_9. Acesso em: 7 out. 2022.

XIA, F.; YETISGEN-YILDIZ, M. Clinical corpus annotation: Challenges and strategies. In: WORKSHOP ON BUILDING AND EVALUATING RESOURCES FOR BIOMEDICAL TEXT MINING, 3., 2012, Istanbul. **Proceedings** [...] Istanbul: European Language Resources Association, 2012. Disponível em: http://faculty.washington.edu/melihay/publications/LREC_BioTxtM_2012.pdf. Acesso em: 27 jun. 2023.