Anotação de Narrativas Clínicas de acordo com as diretrizes das Dependências Universais

Adriana Pagano[1], Carlos A. S. Perini[1],

Cláudia Benevenute[2], Cristiano Colombo[2]

[1]Universidade Federal de Minas Gerais (UFMG)
Av. Antonio Carlos, 6627 – Belo Horizonte – MG – Brasil

[2]Instituto Federal do Espírito Santo (IFES)
Cachoeiro de Itapemirim – ES – Brasil
{apagano,perini}@ufmg.br,
{claudia.benevenute,cristiano.colombo}@ifes.edu.br

Abstract. *This is an ongoing study on Natural Language Processing of a corpus of Clinical Narratives in Brazilian Portuguese with two annotated versions: one by a machine and the other by humans. The frequency of POS and dependency relations of tokens in each version is calculated, and a corpus-driven analysis is performed, highlighting the corrections of the machine annotations made by the human. The comparison of these annotations allows the creation of treebanks that can be used to train new models using machine learning techniques and to improve various Natural Language Processing applications with corpora from the biomedical field. In addition, this comparison allows the analysis of the theoretical consistency of annotation to uncover the grammatical system of this type of corpus and to create annotation guides for Clinical Narratives in Brazilian Portuguese according to Universal Dependencies.*

Resumo. *Este é um estudo em andamento sobre Processamento de Língua Natural de um corpus de Narrativas Clínicas em português brasileiro com duas versões anotadas: uma pela máquina e outra por humanos. A frequência dos rótulos das classes de palavras e das relações de dependência dos tokens em cada versão é calculada e uma análise guiada pelo corpus é realizada, destacando as correções feitas pelos humanos nas anotações da máquina. A comparação dessas anotações permite a criação de treebanks que podem ser usados para treinar novos modelos usando técnicas de aprendizado de máquina e para aprimorar diversas aplicações de Processamento de Língua Natural com corpus da área biomédica. Além disso, essa comparação permite a análise da consistência teórica de anotação, a fim de identificar o sistema gramatical desse tipo de corpus e criar guias de anotação para Narrativas Clínicas em português brasileiro de acordo com as Dependências Universais.*

1. Introdução

As Narrativas Clínicas (NC) são relatos médicos produzidos em hospitais e centros de saúde, destinados a registrar o estado atual e a evolução do paciente durante

o tratamento [OLIVEIRA et al. 2022a]. Incluem sinais vitais, resultados de exames e procedimentos realizados, sendo redigidos rapidamente por profissionais de saúde [OLIVEIRA et al. 2022b]. Em formato digital, integram o Registro Eletrônico de Saúde do paciente, compondo seu histórico médico. Com o objetivo de desenvolver ferramentas de apoio ao acompanhamento médico e pesquisas na área, esses textos têm sido alvo de estudos no campo de Processamento de Língua Natural (PLN).

A anotação de NC exige conhecimento médico especializado e não pode ser adquirido rapidamente, tanto para elaborar diretrizes quanto para anotar [XIA and YETISGEN-YILDIZ 2012]. No entanto, experimentos demonstram que o treinamento médico isoladamente não garante alta concordância entre anotadores. É, portanto, fundamental envolver pesquisadores de PLN no processo de anotação desde o início, mesmo sem formação médica, para obter alta concordância entre os anotadores. Segundo [OLIVEIRA et al. 2022a], falta um corpus multidisciplinar para o progresso científico em PLN biomédico para o português brasileiro. Além disso, as NC demandam a elaboração detalhada de guias de anotação que garantam consistência nas anotações. Uma vez estabelecido o padrão-ouro[1], como no corpus THYME [STYLER et al. 2014], o uso de PLN e aprendizagem de máquina possibilita a recuperação automática de informações de textos biomédicos. A ausência desse padrão compromete aplicações importantes no campo clínico, como estudos retrospectivos e suporte à tomada de decisões clínicas [NÉVÉOL et al. 2018]. Neste estudo, utilizam-se as diretrizes das Universal Dependencies [MARNEFFE et al. 2021] e um corpus do projeto SemClinBR, enfrentando desafios específicos de textos hospitalares que incluem termos numéricos, símbolos, terminologia especializada e abreviações, conforme caracterizado por [DALIANIS 2018]. Havendo assim, a necessidade de desenvolver conjuntos de dados compartilhados e métodos padronizados permitindo comparações entre idiomas, além da criação de diretrizes estruturadas com informações linguísticas específicas, incentivando estudos sobre como as particularidades de cada língua podem contribuir para avanços metodológicos no PLN.

Este trabalho organiza-se em cinco seções, sendo a primeira, introdutória. Na seção 2, apresentam-se, de forma resumida, o projeto das DU, teoria usada para guiar as anotações. Na seção 3, apresenta-se o corpus de NC em português brasileiro. Na seção 4, há a metodologia para analisar o corpus dando especial atenção às etiquetas anotadas pelo modelo de linguagem e as correções humanas. Por fim, a seção 5 discute as considerações resultantes da análise dos dados.

2. O projeto das Dependências Universais

As DU apresentadas por [MARNEFFE et al. 2021], 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, as DU 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 das DU conta com 243 *tree-*

[1]Conjunto de dados cuidadosamente selecionado por especialistas humanos seguindo uma teoria de base para anotar corpus de modo consistente. [STYLER et al. 2014]

banks[2] anotados em 138 idiomas, e está em sua versão 2.11. Para a anotação nas DU, 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 (UPOS) 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.

3. O corpus

O SemClinBr [OLIVEIRA et al. 2022a] é um corpus anotado composto por textos clínicos em português brasileiro[3]. Ele é anotado no estrato semântico, 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 das DU [OLIVEIRA et al. 2022a]. [OLIVEIRA et al. 2022a] aponta que muitos desafios foram encontrados durante a anotação do DepClinBr devido às particularidades dos textos de NC. As principais características desses textos são palavras não reconhecidas por POS *taggers*, com muitos acrônimos e abreviações, com erros de ortografia, expressões numéricas, falta de pontuação, elipses, entre outros exemplificados por [OLIVEIRA et al. 2022b], p. 94. Além dessas características, os estudos de [DALIANIS 2018] descrevem particularidades dos registros de pacientes em diferentes idiomas, a quantidade de erros ortográficos em relação com outros tipos de texto, diferenças sintáticas, escolhas de palavras, abreviações, siglas, a negação entre outras expressões típicas do texto clínico.

Autores como [DALIANIS 2018] 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 NC pode ser vista como um caso especial de Desambiguação do Sentido da Palavra (WSD). [MOON et al. 2011], por sua vez, analisam a variação do uso de numerais em NC e os desafios colocados por eles especialmente na tarefa de extração de informação, alertando sobre a importância de 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 como em "spo2 = 93%". 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 teoricamente consistente para garantir o padrão-ouro.

4. Metodologia e resultados

O guia de anotação adotado durante as anotações de [DURAN et al. 2022], é projetado para ser suficientemente genérico, permitindo sua aplicação em diversas línguas ou gêneros discursivos. Esse guia baseia-se nas DU e possui, segundo [DURAN et al. 2022] "as diretrizes de anotação de etiquetas morfossintáticas adotadas no projeto, que visam instruir anotadores humanos no processo de anotação de corpus". O corpus utilizado, como citado anteriormente é um fragmento do SemClinBr, um corpus previamente existente, que foi construído e disponibilizado por [OLIVEIRA et al. 2022a].

[2]Corpus anotado linguisticamente com estruturas arbóreas para as relações sintáticas.

[3]O uso do SemClinBr foi aprovado pelo CEP da PUCPR, sob o registro 1.354.675. Mais detalhes em [OLIVEIRA et al. 2022a] sessão: *"Methods"*.

O corpus com 1000 sentenças foi pré-anotado automaticamente, segundo [OLIVEIRA et al. 2022b] pelo *toolkit* Stanza[4] com o modelo de treino UD-Bosque. A partir dessa anotação automática, linguistas fizeram a leitura e revisão sintática de cada sentença, usando a plataforma online Arborator[5].

As figuras 1a. e 2a. deste estudo são sentenças anotadas automaticamente antes da intervenção do anotador humano, enquanto as figuras 1b. e 2b. são as mesmas sentenças após a revisão dos linguistas. Observe a anotação da abrevição 'TX' em 1a. passa a ter uma *glossa* 'transplante' em 1b. A descrição de 'O' em 2a. foi corrigida da anotação automática do modelo como pronome acusativo e com a revisão humana passa a ter a glossa 'abordagem SOAP (Sigla para : Subjetivo-Objetivo-Avaliação-Plano)'.

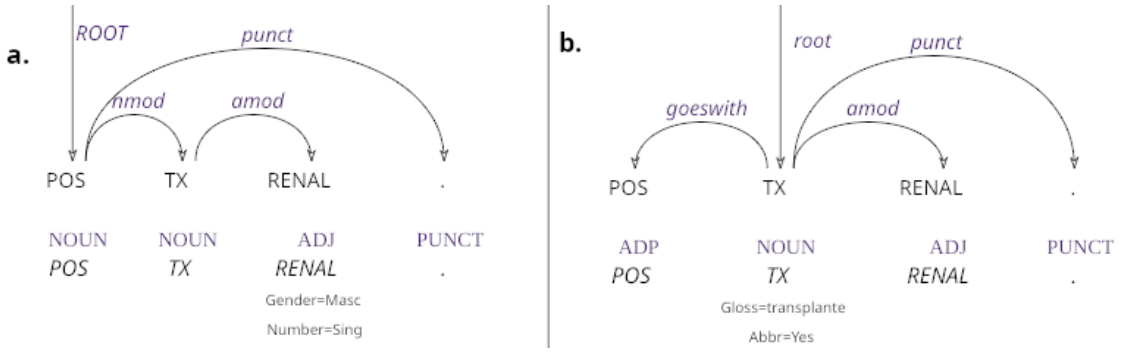


Figure 1. árvores de sentenças no Arborator.

Após revisão humana as sentenças foram exportadas da ferramenta Arborator em arquivos no formato CONLLU. As sentenças anotadas pela máquina e as mesmas sentenças revisadas por humanos possuem boa concordância com coeficiente Cohen Kappa para UPOS de 0.96 e para DepREL de 0.87. No entanto, para os exemplos apresentados, não se pode garantir uma boa concordância dada as particularidades das NC. Então, assim, procurou-se pelos detalhes, calculou-se as estatísticas de anotação desses dois copora para investigar os rótulos mais frequentes que foram corrigidos pelos anotadores humanos[6].

A abordagem guiada pelo corpus "é mais indutiva, de modo que as próprias construções linguísticas emergem da análise de um corpus" [BIBER 2015]. A busca pelo sistema linguístico desse corpus, por exemplo, verificar se as etiquetas POS mais frequentes anotadas pelo modelo estavam corretas seguindo um padrão teórico quando revisadas pelos humanos é apresentada por meio dos gráficos 1 a 4.

O gráfico 1 destaca as etiquetas[7] UPOS mais alteradas na pós-edição humana,

[4]*Toolkit* de PLN desenvolvido pelo grupo de PLN de Stanford, tem suporte para as DU e usa modelos de aprendizagem de máquina profundo com grande precisão.

[5]Disponível em: `https://arborator.icmc.usp.br/`

[6]composição da equipe de anotadores: um coordenador linguista, um doutorando linguista e dois graduandos em linguística

[7]As etiquetas UPOS, sigla para: Universal Part Of Speech tags, como NOUM, ADJ, PROPN e DepRel, sigla para: DEPendency RELation, como nmod, punct, conj são do padrão do projeto das UDs, a definição de cada uma está disponível em: `https://universaldependencies.org/treebanks/pt_bosque/index.html` para o português incluindo exemplos.

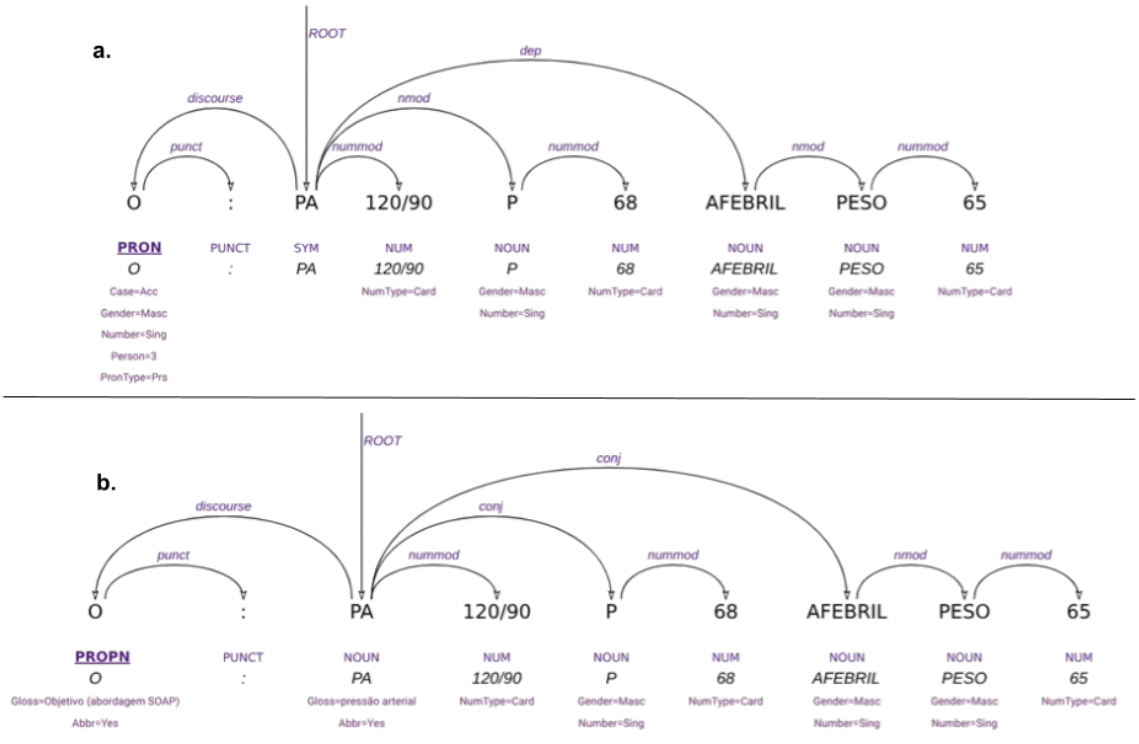
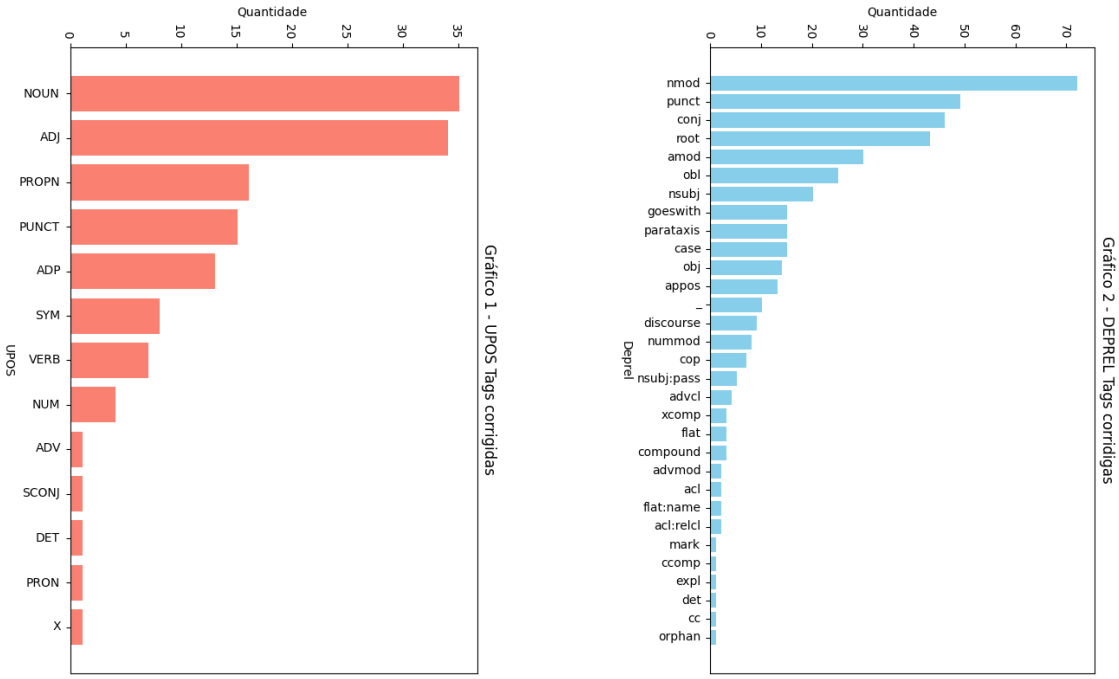


Figure 2. árvores de sentenças no Arborator.



com NOUN o rótulo mais corrigido, enquanto o gráfico 2 mostra as etiquetas DepRel mais alteradas com a revisão humana, sendo o primeiro lugar para nmod.

Os gráficos 3 e 4 mostram para quais rótulos foram anotados as etiquetas NOUN de UPOS e nmod das DepRel. A tendência das correções apresentam como os tokens anotados com NOUN foram alterados para ADJ e PROPN. Os nomes próprios (PROPN) é indicativo de Entidades Nomeadas (EN) que o modelo não conseguiu identificar. Ao passo que as alterações de nmod foram mais para punct e amod, mostrando um reflexo da correção NOUN para ADJ e provavelmente uma leve alucinação do modelo anotando com nmod no lugar de punct.

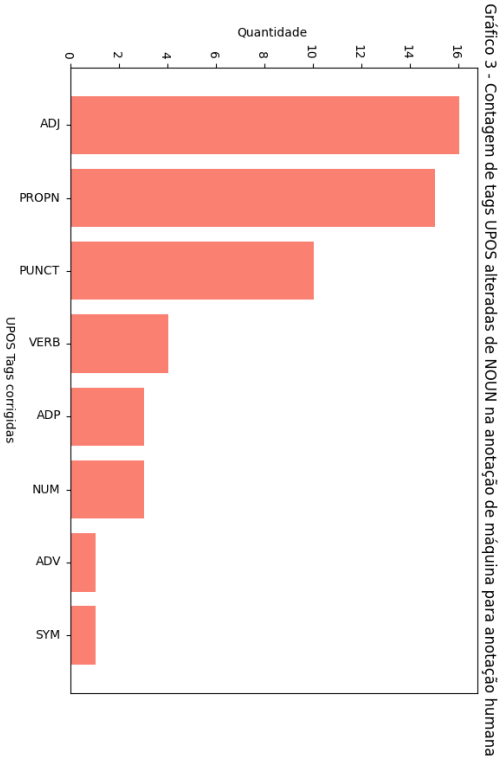


Gráfico 3 - Contagem de tags UPOS alteradas de NOUN na anotação de máquina para anotação humana

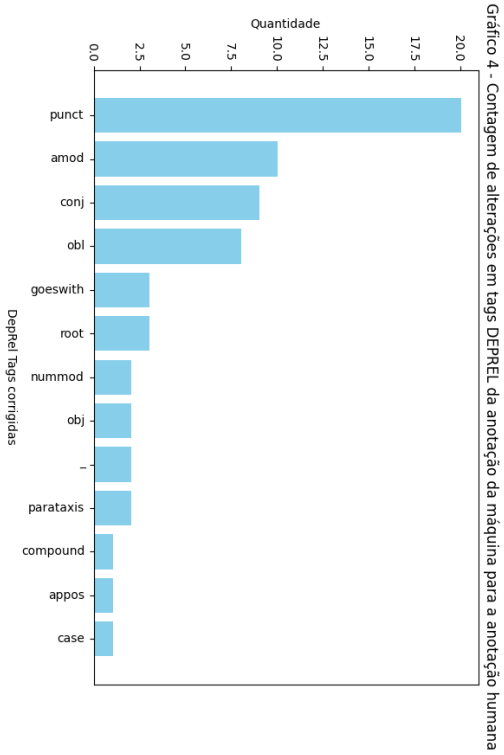


Gráfico 4 - Contagem de alterações em tags DEPREL da anotação da máquina para a anotação humana

5. Conclusão

Com o objetivo de produzir um corpus anotado de modo consistente, também chamado de padrão-ouro, as análises aqui apresentadas são muito pertinentes pois são direcionadas pela frequência dos rótulos, sejam eles UPOS ou DepRel. A análise baseada na frequência dos rótulos UPOS e DepRel é fundamental para a construção consistente de um corpus anotado. Embora o processo de anotação automática seja uma "caixa-preta" algorítmica [Ribeiro et al. 2016], a interpretação linguística dos dados quantitativos — como a correção de anotações (ex.: NOUN erroneamente classificados como PROPN) e a adição de camadas descritivas *enhanced* (ex.: expansão da descrição de abreviações[8]) - permite melhor descrever a gramática do corpus. Dessa forma, a frequência dos rótulos orienta a revisão manual, prioriza classes problemáticas e direciona esforços para atingir a anotação de qualidade no padrão-ouro para corpus de NC.

[8]a quantidade de abreviações colocadas em relação ao anotado pelas máquinas foi de 70 ocorrências contra nenhuma pela máquina.

References

BIBER, D. (2015). Corpus-based and corpus-driven analyses of language variation and use. In HEINE, B. and NARROG, H., editors, *The Oxford Handbook of Linguistic Analysis*. Oxford Academic, 2nd edition.

DALIANIS, H. (2018). Characteristics of patient records and clinical corpora. In *Clinical Text Mining*. Springer, Cham.

DURAN, M. S., NUNES, M. d. G. V., LOPES, L., and PARDO, T. A. S. (2022). Manual de anotação como recurso de processamento de linguagem natural: o modelo universal dependencies em língua portuguesa. *Domínios de Lingu@gem*, 16(4):1608–1643.

MARNEFFE, M. et al. (2021). Universal dependencies. *Computational Linguistics*, 47(2):255–308.

MOON, S., Pakhomov, S., Ryan, J., and Melton, G. B. (2011). Automated non-alphanumeric symbol resolution in clinical texts. *AMIA Annual Symposium Proceedings*, pages 979–986.

NÉVÉOL, A., DALIANIS, H., VELUPILLAI, S., et al. (2018). Clinical natural language processing in languages other than english: opportunities and challenges. *Journal of Biomedical Semantics*, 9(1):12.

OLIVEIRA, L. E. S., PETERS, A. C., DA SILVA, A. M. P., et al. (2022a). Semclinbr - a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

OLIVEIRA, L. F. A. d., OLIVEIRA, L. E. S. d., and MORO, C. (2022b). Challenges in annotating a treebank of clinical narratives in brazilian portuguese. In PINHEIRO, V., GAMALLO, P., AMARO, R., SCARTON, C., BATISTA, F., SILVA, D., MAGRO, C., and PINTO, H., editors, *Computational Processing of the Portuguese Language*, pages 90–100, Cham. Springer International Publishing.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

STYLER, W. F., BETHARD, S., FINAN, S., PALMER, M., PRADHAN, S., de GROEN, P. C., ERICKSON, B., MILLER, T., LIN, C., SAVOVA, G., and PUSTEJOVSKY, J. (2014). Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

XIA, F. and YETISGEN-YILDIZ, M. (2012). Clinical corpus annotation: Challenges and strategies. In *Proceedings of the 3rd Workshop on Building and Evaluating Resources for Biomedical Text Mining*, Istanbul. European Language Resources Association.