







Challenges in Annotating a Treebank of Clinical Narratives in Brazilian Portuguese

Lucas Ferro Antunes de Oliveira¹(✉) , Adriana Pagano² ,
Lucas Emanuel Silva e Oliveira¹ , and Claudia Moro¹ 

¹ Pontifical Catholic University of Paraná, Curitiba, PR, Brazil
lucas.ferro@pucpr.edu.br, {lucas.emanuel,c.moro}@pucpr.br

² Federal University of Minas Gerais, Belo Horizonte, MG, Brazil
apagano@ufmg.com

Abstract. Dependency parsing can enhance the performance of Named Entity Recognition (NER) models and can be leveraged to boost information extraction. NER tasks are essential to deal with clinical narratives, but models for Brazilian Portuguese dependency parsing are scarce, even less for clinical texts and its specificities. This paper reports on the development of a treebank of clinical narratives in Brazilian Portuguese and the drafting of guidelines. Based on a corpus of 1,000 clinical narratives manually annotated with semantic information, split into 12,711 sentences, we identified some characteristics of these texts that differ from traditional domains and have a deep impact on the annotation process, such as extensive use of acronyms and abbreviations, words not recognized by POS taggers, misspelling, special use of some symbols, different uses for numerals, heterogeneity of sentence sizes, and coordinated phrases without any punctuation. We developed a document to describe the annotation types and to explain how difficult cases should be treated to ensure consistency, including examples that could be found in this kind of texts. We created a Tag versus Frequency relation to justify some of the characteristics and challenges of the corpus. The corpus when completely annotated will be made available to the entire scientific community that performs research with clinical texts.

Keywords: Clinical narratives · Treebank development · Annotation guidelines

1 Introduction

Clinical text is a rich source of information to be tapped for the purposes of problem-solving related to quality of care, clinical decision support and safe information flow among all the parties involved in healthcare. Clinical texts present great challenges in their handling, even greater challenges than the texts found in medical literature. In these texts, health professionals describe the patient's

entire health history, vital signs, relevant clinical findings and procedures performed. Because they are written in a patient care environment, where there is a time pressure involved, they are more susceptible to errors. Furthermore, they are texts that do not have a formal structure, have high variability and heterogeneity, and may contain highly complex vocabulary [7].

Information extraction in this domain is usually carried out by means of Named Entity Recognition (NER), relation extraction, temporal expression identification, negation detection and terminology mapping [2, 15]. These tasks rely on models built on corpora, which can be enriched by annotation at different linguistic levels. In the case of dependency parsing, morphology and syntax annotations can enhance the performance of NER models and can be leveraged to boost information extraction [1].

Although several annotated corpora of clinical texts are available in languages such as English, French, German, English, Spanish and Chinese among others, corpora for Portuguese are very scarce (see [8]). Moreover, there are few resources that can be used to adapt general domain corpora to clinical text in Portuguese.

In this paper, we report on a syntactically-annotated corpus of clinical narratives and describe the lessons learned in annotating dependency relations, considering the particularities and complexities of the clinical text. We report our experience and findings within a collaborative institutional setting for guideline development and corpus annotation. The guidelines and annotations will be made available for download, with instructions on how to load and visualize the syntactic parses.

This article is organized as follows. Section 2 briefly presents available corpora in the biomedical domain and related research on clinical notes. Section 3 describes the design and development of our corpus annotation, including the scheme, guidelines, and the training of annotators. Section 4 shows the results achieved in the annotation process with statistical information. Section 5 discusses implications of our statistics and results and summarizes the faced challenges. Finally, Sect. 6 concludes the paper and considers possible future directions.

2 Related Work

In the field of medical NLP research, size and availability of clinical text corpora is dependent on the type and the language of a corpus. While raw text corpora are large in size and available [3, 4, 19], annotated corpora are much smaller in size and scarcer in under-represented languages. Among annotated corpora, part-of-speech (POS) [14, 18] and named entity recognition (NER) [9] are the most common type of annotation. Corpora of clinical texts annotated for syntactic structure (treebanks) are still emergent and largely concentrated on English, Spanish and Chinese. In the case of Portuguese, to the best of our knowledge there is only a single corpus in European Portuguese [6] and one in Brazilian Portuguese [11] annotated for NER, and explored in [17]. Another initiative contemplates the development of a morphological corpus (i.e., POS tags) for

clinical texts in Portuguese, which was recently applied in [13]. No treebank of clinical text, which includes dependency relations, is available in Portuguese, which motivated the creation of our corpus.

A common issue raised by all treebank development projects are the challenges posed by the very nature of clinical texts [10], which we will detail in the following section together with the decisions made in our project to meet those challenges.

3 Materials and Methods

This section presents the main characteristics of the texts compiled in our corpus, the creation and evolution of annotation guidelines, and the corpus development process.

3.1 Data Preparation

The texts used to create our corpus were retrieved from SemClinBr [11], a corpus of 1,000 clinical narratives manually annotated with semantic information¹. As SemClinBr aimed to support NLP tasks in several medical specialties, it compiled clinical narratives from distinct medical areas (e.g., cardiology, nephrology, orthopedy).

The 1,000 narratives have been split into 12,711 segments relying on periods and blank spaces as indicators for boundaries. 177 sentences having more than 50 tokens were set aside for future annotation. The time to annotate sentences varies according to their size; therefore it was necessary to separate the large sentences to minimize the time invested in the annotation until a sufficiently large corpus for training was obtained. As a first stage in our annotation project, we selected 2,955 sentences (approximately 25%).

In order to create our treebank, we pre-annotated our corpus using the Stanza toolkit [16]. The model Stanza uses by default for Portuguese is the pt-bosque model trained, using UD-Bosque², a treebank containing annotated news texts in Brazilian and European Portuguese following the Universal Dependencies (UD) guidelines, which is the standard adopted in our project. The output CoNLL-U files were imported in ArboratorGrew³ to be annotated by two independent annotators and curated by a third one. We developed our own guidelines for annotation drawing on the UD principles and documentation⁴ and available guidelines for Portuguese by Souza et al. (2020) (unpublished work).

¹ The use of SemClinBr texts was approved by the Ethics Committee in Research (CEP) of PUCPR, under register nº. 1,354,675.

² https://universaldependencies.org/treebanks/pt_bosque/index.html.

³ <https://arboratorgrew.elizia.net/#/>.

⁴ <https://universaldependencies.org>.

3.2 Corpus Characteristics

The corpus is composed of several types of clinical narratives, such as nursing notes, discharge summaries, and ambulatory notes, which required to be tokenized differently. The tokenization process yielded sentences with large variation in size, ranging from 2 and 3 token segments to over 260 token ones. Segments with 50 or more tokens proved very difficult to annotate in our annotation interface due to the need to use horizontal scroll bar to draw arcs between distant tokens.

Table 1. Segments illustrating size variation

Size	Example
Long (78 tokens)	Às 15:27: CLIENTE ALERTA, CONSCIENTE, PUPILAS ANISOCORICAS E) D, COMUNICATIVA, DISCRETA DISPNEIA AOS ESFORÇOS, COM SUPORTE DE OXIGENOTERAPIA EM CATETER NASAL A 2 /L, MANTENDO OXIMETRIA DE PULSO SAT: 96 %, REPOUSO NO LEITO, ACEITANDO POUCO A DIETA, ACESSO PERIFÉRICO EM MSE SALINIZADO, ABDOMEN GLOBOSO E FLÁCIDO, APRESENTANDO HIPEREMIA EM REGIÃO SACRA, MANTENDO COLCHÃO DE AR, ELIMINAÇÕES PRESENTES
Medium (18 tokens)	sem deficit de força, sem rigidez em membros, sem bradicinesia, sem alterações de pares cranianos
Short (3 tokens)	Marevan 5mg

Table 1 shows three tokenized sentences that are very different in size and in capitalization. The longest sentence have 78 tokens and all words are upper cased, while the medium is composed of 18 and all tokens are lower cased, and the shortest one is 3 token long and capitalized.

It is possible to notice that in the first example in Table 1, there are some different kinds of usage of numbers. “15:57:” appears as an event, with the following tokens describing the state of a patient. The number 2 and the number 96, appearing in “[...] NASAL A 2 /L [...]” (“nasal at 2 /L”) and “[...] PULSO SAT: 96 %[...]” (“pulse sat: 96%”), respectively, mean number measurements, quantity, dosage. These examples illustrate the problems posed by numbers in our corpus and their joint use with symbols.

Table 1 also shows that some sentences have subject (“CLIENTE” “client”) and predicate with non finite verbs (“MANTENDO” “maintaining”, “ACEITANDO” “accepting”, and “APRESENTANDO” “presenting”), while others (examples 2 and 3 in Table 1) have neither subject nor verb. Each case demands

a decision as regards its “root”, which was carried out following the UDs guidelines. Table 2 shows further recurrent characteristics of our corpus and illustrates them.

Table 2. Corpus characteristics and examples

Characteristic	Example
Words not recognized by POS taggers/lemmatizers	Hidantalizado
	Facietomia
	flutter atrial
Extensive use of acronyms and abbreviations	mantendo monitorização p 81, pa 103/74, sat o2 97% po le + ladf em 01/12/15
	colar cervical c/ queixas de dor
Misspelling	em repouso em o leito
	outras inetcorrências
Numerical expressions	ssvv a as 05:45 h pa = 133/74 mmhg , fc = 114 bpm, spo2 = 93%
	glasgow: 9
	mantendo monitorização p 81, pa 103/74, sat o2 97%
No punctuation	dois ave um há 1 ano e outro 2 anos
Special use of symbols	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 + bronco- scopia, sem intercorrência durante o exame e transporte.
antes 2 carteiras/dia	
Coordination	apresenta curativo + tala gessada em mse, referindo algia moderada, edema distal, mobilidade diminuida, apresentou 1 episodio de emese, sendo medicada, diurese presente, segue cuidados
Parenthetical comments	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)
Reduction	# retorno 7 dias
Ellipsis	apresentou problemas
	aceitando pouco a dieta vo

Characteristics in Tables 1 and 2 have a deep impact on the annotation process.

Words wrongly tagged and lemmatized by the Bosque model (news texts) have to be manually corrected, delaying the annotation process. Most sentences have abbreviations and acronyms; Therefore an abbreviation list had to be created aiming to help annotators with the healthcare terminology and vocabulary.

Additionally, a frequent problem found in the narratives is reiterated misspelled words, misuse of symbols and lack of punctuation to mark off sections. This may be accounted for by the fact that health professionals need to produce text quickly and often resort to “copying and pasting” from previous notes.

Symbols are used for a variety of purposes. Sometimes they can be used as a markup in the text, indicating the beginning of a section (e.g. “#”); other times they can be used to signal coordination as equivalent to “and” (e.g. “+”); further they can indicate operations with numbers as “+” (“plus”) or “-” (“minus”) and result or comparisons like “=” or “<” and “>”. Some punctuation marks are very frequent in clinical text, as is the case for colons and dashes, and they demand an interpretation regarding their function (explanation, rephrasing, equation, etc.)

Bearing in mind all these characteristics of clinical narratives, we developed a document to describe the annotation types and explain how complex cases should be treated to ensure consistency. We adopted an iterative process of annotation, evaluation, and refinement of the annotation scheme and guidelines. Batches of 200 sentences were made available to the annotators and upon their annotation, their results were analyzed to improve the guidelines. The updated guidelines led to significantly better results on subsequent batches.

3.3 Decisions Made in the Annotation Process

The corpus was annotated for POS, lemma and dependency relations following the Universal Dependencies (UDs) guidelines as adopted by the Brazilian and European Portuguese treebanks adhering to UD. In addition, available guidelines for clinical text in other languages (see Related Work) were consulted in order to decide how to handle particular specs of clinical text in Brazilian Portuguese.

As Table 1 illustrates, our corpus consists of a significant number of segments that can be characterized as fragments. There are instances of medical dosage, description of patient status, etc., which are mostly isolated noun phrases. In these cases, those phrases were annotated having the main noun as root. Most sentences are subject-less sentences, particularly those referring to the patient. Incomplete sentences due to line break or transcription errors were annotated as they were, without any insertion or editing.

A major decision had to do with how to handle ellipsis of major syntactic functions in our corpus. Annotating elliptical sentences implicates deciding whether a sentence will be analyzed considering elided functions or not, which

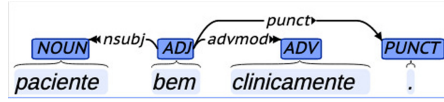


Fig. 1. Annotation presupposing a copula relation

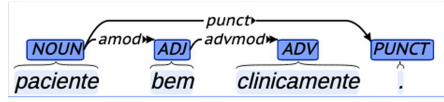


Fig. 2. Annotation of noun modified by adjective

has been shown to have a significant impact on the final results [5]. For instance, “paciente bem clinicamente” (“patient well clinically”) can be annotated either as having “bem” (“well”) or “paciente” (“patient”) as root, as shown below.

In Fig. 1, an elliptical copula relation is presupposed. In Fig. 2, the noun “paciente” (“patient”) is taken as the main head being modified by an adjective. We opted for the latter in order to avoid relying on annotators’ interpretation of elided functions.

Variation for the same type of information was also an issue regarding prepositional phrases, particularly because comparable functions exhibited different realizations. This was particularly noticeable for temporal expressions. Thus, the time at which entries were recorded in the narratives and notes could be equally expressed as “às 01:30 horas” (“at one hour thirty”), “às 01:30” (“at one thirty”), “01:30 horas” (“one hour thirty”), “01:30” (“one thirty”), and “01h30” (“one hour thirty”). To ensure consistency in annotation and aid temporal information retrieval, all forms were annotated as an oblique relationship to the main root.

Clinical text contains abbreviations indicating dose or procedures. We adopted a principle to annotate the abbreviation based on the Part-of-Speech category of the words as they would be written in their full-form (e.g. “rx” -> “raio x” (“x-ray”) (NOUN)).

As a result of corpus anonymization, names of physicians and nurses were rendered as “Doutor Vital Brasil” (“Doctor Vital Brasil”) and “Enfermeira Florence Nightingale” (“Nurse Florence Nightingale”). These names were annotated as Proper Nouns in the corpus.

In the case of terms quoted from the Portuguese version of the International Classification of Diseases (ICD), their English counterparts were checked so that the annotation would be comparable in both languages.

4 Results

We annotated 2,901 sentences out of 2,955 available ones. 54 sentences remained unannotated due to errors in their CoNLL-U files that prevented the file to be

Table 3. Frequency of annotated dependency and POS tags

DEP tag	Frequency	DEP tag	Frequency	DEP tag	Frequency	POS tag	Frequency
punct	7818	xcomp	248	aux	23	NOUN	12564
case	6094	flat	214	ccomp	20	PUNCT	7919
nmod	4460	advcl	206	iobj	14	ADP	6141
conj	4414	discourse	179	dep	6	ADJ	5779
amod	4402	parataxis	153	subj:pass	4	VERB	2774
root	2928	compound	143	csubj	3	DET	1894
obl	2092	mark	108	nsubj:agent	2	NUM	1329
det	1876	nsubj:pass	107	subj	2	CCONJ	839
obj	1217	cop	73	advmod:part	1	ADV	688
cc	966	flat:name	68	csubj:pass	1	SYM	471
nummod	915	fixed	53	obl:pass	1	PROPN	354
advmod	552	aux:pass	47	orphan	1	X	151
acl	520	acl:relcl	31	reparandum	1	PRON	148
goeswith	476	flat:foreign	31	vocative	0	SCONJ	112
appos	465	obl:agent	25	dislocated	0	AUX	94
nsubj	287	expl	24			PART	15
						INTJ	0

uploaded in ArboratorGrew, the graphic interface used for annotation. Out of 2,901 sentences, 824 (28.4% of annotated sentences) were less or equal to 5 token long. The average number of tokens per sentence is 14.23.

Table 3 shows the frequency of each Dependency tag and Part-of-Speech tag annotated in the 2,901 sentences.

As regards inter annotator agreement, the highest results were 99.78 for the POS-Tagging and 97.23 for the Dependency Parsing. The lowest scores were 88.22 and 69.69 for POS-Tagging and Dependency Parsing, respectively.

5 Discussion

Regarding POS-tagging, the high number of “NOUN” tags can be accounted for the narratives mentioning conditions, drugs, medical procedures and facilities, as “CC” (“Centro Cirúrgico” or “Surgery Center”), and medical exams details and descriptions. This proves an important characteristic of the clinical narratives, which describe patients’ health conditions.

Approximately 48.6% of all annotated sentences had no “VERB” tags (1,409 sentences); however the number of “VERB” tags is high, which can be accounted for this POS operating in dependency relationships such as modifiers (“acl” and “acl:relcl”). This is probably explained by the specific characteristics of each section of the clinical text. For example, in defining comorbidities and current health status, the focus will likely be on “coding” the patient’s diagnoses, without using verbs. On the other hand, when the continuity of care is informed, verbs are usually used, as seen in [12] (e.g., visit the cardiologist in 10 d).

The high frequency of the POS tag “PUNCT” can be accounted for by the different uses of punctuation in the corpus, as is also the case of the tag “SYM”.

As regards the tag “X”, this can be accounted for by parsing errors due to typos. This is illustrated by the example in Fig. 3. Here the word “não” had been mistyped as “nao” in the clinical narrative and was wrongly parsed as “em a o”. Following UD, the three tokens were annotated with the “goeswith” relation and the POS of the correct word is assigned to the first token, the remaining two receiving the X tag.

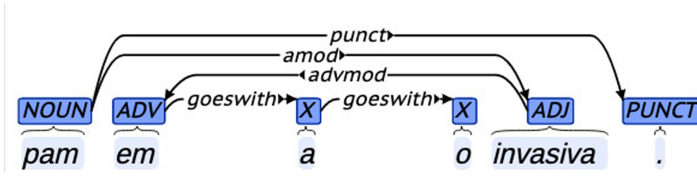


Fig. 3. Annotation of typo and incorrectly parsing

Regarding dependency relations, the most frequent tag is “punct”, which can be accounted for the high number of punctuation marks in our corpus, followed by “case”, which is typically realized by the “ADP” POS tag. Worth highlighting are the tags “nmod” and “conj”. The former can be linked to the high number of NOUN tags, while the latter reflects a key characteristic of our corpus sentences, which consist of several segments coordinated between each other.

Dependency tags with frequency below 20 have been set aside to be reviewed in the next stage of our project, in order to check whether they are rare categories or should be re-annotated.

We understand that the main contribution of this article lies in the solutions found in carrying out the annotation of clinical texts, which can help other researchers to understand the particularities and complexities of clinical texts, allowing them to build additional resources for the medical informatics community. The corpus has an innovative character, being the only corpus in Brazilian Portuguese with clinical texts annotated with morphological and syntactic information.

6 Conclusion

In this paper, we reported the main challenges in annotating clinical narratives with morphological and syntactic information. Additionally we describe the preliminary results obtained in the first stage of our annotation. We annotated 2,901 sentences that compose a treebank in the CoNLL-U format following the Universal Dependencies guidelines. We achieved high agreement between annotators and overall good results for the present stage in our project. Further steps include reviewing annotated segments to improve our model, annotating the remaining segments, and making the corpus and guideline available for the scientific community. Upon finalizing our treebank, this will be used to train another pipeline to be used to enhance the semantic NLP tasks of NER and information extraction targeted by our project.

References

1. Bretonnel Cohen, K., Demner-Fushman, D.: Biomedical Natural Language Processing. John Benjamins (2014). <https://www.jbe-platform.com/content/books/9789027271068>
2. Dalianis, H.: Basic building blocks for clinical text processing. In: *Clinical Text Mining*, pp. 55–82. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-78503-5_7
3. Dalianis, H., Hassel, M., Henriksson, A., Skeppstedt, M.: Stockholm EPR corpus: a clinical database used to improve health care. In: *Swedish Language Technology Conference*, pp. 17–18 (2012)
4. Hao, T., Rusanov, A., Boland, M.R., Weng, C.: Clustering clinical trials with similar eligibility criteria features. *J. Biomed. Inf.* **52**, 112–120 (2014)
5. Jiang, Z., Zhao, F., Guan, Y.: Developing a linguistically annotated corpus of Chinese electronic medical record. In: *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 307–310. IEEE (2014)
6. Lopes, F., Teixeira, C.A., Oliveira, H.G.: Contributions to clinical named entity recognition in Portuguese. In: *BioNLP@ACL* (2019)
7. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Med. Inf.* **17**(01), 128–144 (2008)
8. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., Zweigenbaum, P.: Clinical natural language processing in languages other than english: opportunities and challenges. *J. Biomed. Semantics* **9**(1), 1–13 (2018)
9. Ogren, P.V., Savova, G.K., Chute, C.G., et al.: Constructing evaluation corpora for automated clinical named entity recognition. In: *LREC*, vol. 8, pp. 3143–3150 (2008)
10. Oinam, N., Mishra, D., Patel, P., Choudhary, N., Desai, H.: A treebank for the healthcare domain. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pp. 144–155 (2018)
11. Oliveira, L., et al.: Semclinbr-a multi institutional and multi specialty semantically annotated corpus for Portuguese clinical NLP tasks. In: *CoRR* (2020)
12. Oliveira, L.E.S., de Souza, A.C., Nohama, P., Moro, C.M.C.: A novel method for identifying continuity of care in hospital discharge summaries. In: Zhang, Y.-T. (ed.) *The International Conference on Health Informatics*. IP, vol. 42, pp. 284–287. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-03005-0_72
13. de Oliveira, L.F.A., e Oliveira, L.E.S., Gumiel, Y.B., Carvalho, D.R., Moro, C.M.C.: Defining a state-of-the-art POS-tagging environment for Brazilian Portuguese clinical texts. *Res. Biomed. Eng.* **36**(3), 267–276 (2020). <https://doi.org/10.1007/s42600-020-00067-7>
14. Pakhomov, S.V., Coden, A., Chute, C.G.: Developing a corpus of clinical notes manually annotated for part-of-speech. *Int. J. Med. Inf.* **75**(6), 418–429 (2006)
15. Percha, B.: Modern clinical text mining: a guide and review. *Ann. Rev. Biomed. Data Sci.* **4**(1), 165–187 (2021). <https://doi.org/10.1146/annurev-biodatasci-030421-030931>, pMID: 34465177
16. Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D.: Stanza: a Python natural language processing toolkit for many human languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (2020). <https://nlp.stanford.edu/pubs/qi2020stanza.pdf>

17. Schneider, E.T.R., et al.: BioBERTpt - a Portuguese neural language model for clinical named entity recognition. In: Proceedings of the 3rd Clinical Natural Language Processing Workshop, pp. 65–72. Association for Computational Linguistics, November 2020. <https://doi.org/10.18653/v1/2020.clinicalnlp-1.7>, <https://aclanthology.org/2020.clinicalnlp-1.7>
18. Tateisi, Y., Tsujii, J.: Part-of-speech annotation of biology research abstracts. In: LREC (2004)
19. Wu, S.T., Liu, H., Li, D., Tao, C., Musen, M.A., Chute, C.G., Shah, N.H.: Unified medical language system term occurrences in clinical notes: a large-scale corpus analysis. *J. Am. Med. Inf. Assoc.* **19**(e1), e149–e156 (2012)