The incremental process of building an annotation scheme for clinical narratives in Portuguese: the contribution of human variation analysis

Ana Luísa Fernandes[1,2,5], Purificação Silvano[1,2,5], António Leal[2,5,6], Nuno Guimarães[1,2], Rita Rb-Silva[2,3,4], Luís Filipe Cunha[1,2], Alípio Jorge[1,2]

[1]INESC TEC, [2]University of Porto, [3]CI-IPOP, [4]RISE-Health, MEDCIDS, [5]CLUP
Porto, Portugal
[6]University of Macau, China

Correspondence: ana.l.fernandes@inesctec.pt

Abstract

The development of a robust annotation scheme and corresponding guidelines is crucial for producing annotated datasets that advance both linguistic and computational research. This paper presents a case study that outlines a methodology for designing an annotation scheme and its guidelines, specifically aimed at representing morphosyntactic and semantic information regarding temporal features, as well as medical information in medical reports written in Portuguese. We detail a multi-step process that includes reviewing existing frameworks, conducting an annotation experiment to determine the optimal approach, and designing a model based on these findings. We validated the approach through a pilot experiment where we assessed the reliability and applicability of the annotation scheme and guidelines. In this experiment, two annotators independently annotated a patient's medical report consisting of six documents using the proposed model, while a curator established the ground truth. The analysis of inter-annotator agreement and the annotation results enabled the identification of sources of human variation and provided insights for further refinement of the annotation scheme and guidelines.

1 Introduction

Manual annotation is a cornerstone of both linguistic research and natural language processing (NLP) (cf. e.g., Snow et al., 2008; Bhardwaj et al., 2010; Flickinger et al., 2017), enabling the research of linguistic phenomena and providing "gold labels" for training and assessing models in multiple NLP tasks (Pustejovsky and Stubbs, 2012; Pustejovsky et al., 2017; Levi and Shenhav, 2022). In addition to supporting data-driven approaches, manual annotation contributes to formalizing linguistic theories by offering a structured framework for empirical validation (Hovy and Lavid, 2010). Developing a comprehensive annotation scheme is critical to ensure that the annotation is systematic, consistent, interoperable, and comprehensive. A well-designed scheme enables the accurate representation of complex linguistic phenomena grounded in theory while maintaining practical applicability for annotators (Beck et al., 2020). When the data pertains to highly specialized subject matter, such as medical discourse, or involves the intersection of distinct domains, such as linguistics and medicine, the demands on scheme design increase substantially. In such cases, the annotation scheme and corresponding guidelines must be particularly precise and detailed to ensure accurate interpretation. This complexity challenges scheme designers and places additional cognitive and interpretive burdens on annotators (Graham and van der Meer, 2015). Among the additional challenges in annotating clinical narratives is the significant heterogeneity of the content and writing styles of medical reports, which vary not only across healthcare institutions (Zhu et al., 2023), but also between different departments or services within the same hospital. These texts are often written in a free and spontaneous manner, reflecting an inherent diversity of topics and concepts specific to the medical domain. Moreover, clinical texts differ substantially from non-clinical texts due to the highly technical and specialized nature of the field, as well as the frequent use of abbreviations, which significantly increases the complexity of their processing (Moharasan and Ho, 2019). Additionally, biomedical terminology is inherently complex, and it is common for certain terms to have different meanings depending on the context in which they are used. This further underscores the need for clear and context-sensitive annotation guidelines (Irrera et al., 2024).

A critical aspect of the annotation process is the assessment of both the effectiveness of the annotation scheme and the annotators' understanding of the guidelines. Successful annotation depends on the clarity, coherence, and comprehensiveness of

the documentation, as well as the annotators' training and familiarity with the scheme (Artstein and Poesio, 2008). Well-developed guidelines — featuring explicit definitions and illustrative examples — are essential for achieving reliable and accurate annotations (Pustejovsky and Stubbs, 2012). The validation of annotation schemes typically involves a combination of pilot studies, iterative guideline refinement, and qualitative analyses of problematic cases. The annotation process generally entails collecting judgments from multiple annotators for each data instance, a practice widely recognized for enhancing annotation quality (Snow et al., 2008). A commonly used metric to assess the quality of the annotation is inter-annotator agreement (IAA), which provides a quantitative assessment of annotation consistency (Artstein and Poesio, 2008). High IAA scores suggest clear and effective guidelines, whereas low agreement may stem from a variety of causes (Artstein, 2017; Basile et al., 2021; Bayerl and Paul, 2024), often revealing ambiguities or conceptual difficulties that require further attention.

Analyzing sources of annotation disagreement is determinant in improving annotation frameworks, providing valuable information on areas where guidelines may need clarification or extension (Artstein and Poesio, 2008; Hovy and Lavid, 2010). Although human variation in clinical annotation is natural, it is generally undesirable because, for example, the annotation can be used to develop information extraction algorithms for clinical research, where data must be unambiguous. Therefore, ambiguity must be eliminated, and disagreement in the annotation should be minimal or ideally nonexistent. Nevertheless, analyzing such variation in earlier stages of the annotation process can serve as a valuable diagnostic tool, revealing limitations or ambiguities in the current annotation design and accompanying guidelines. Observing patterns of annotator disagreement helps refine the guidelines and ultimately contributes to reducing annotation errors (Finlayson and Erjavec, 2017; Beck et al., 2020).

The primary objective of this paper is to describe a methodology to develop and validate an annotation scheme. We focus specifically on strategies aimed at minimizing human variation throughout the annotation process. To this end, we present a case study involving the design of an annotation scheme for medical reports written in European Portuguese. Our main contributions are as follows: (1) a methodological proposal for the design and validation of annotation schemes; (2) a case study illustrating the role of human variation analysis in refining annotation schemes and guidelines; (3) an annotation scheme for representing both linguistic and medical information in European Portuguese medical reports.

The paper is structured as follows. Section 2 reviews related work. Section 3 presents the case study, beginning with a description of the annotation scheme (3.1), followed by the results of the evaluation and a discussion (3.2) of how the findings informed improvements to the scheme and guidelines (3.2.2). The paper concludes with final remarks and directions for future work (4).

2 Related work

The development and validation of annotation schemes is a labor-intensive and demanding task. Yet, it is essential for both linguistic research and NLP applications. Over the past four decades, annotation strategies have evolved significantly. Since the early 1990s, when annotation became central to training machine learning models and practices were mostly improvised (Ide, 2017), there has been substantial progress toward systematizing and formalizing annotation methodologies.

A considerable body of work has focused on establishing principled standards for creating and validating annotation schemes. For example, Graham and van der Meer (2015) propose a seven-step annotation process. This process begins with selecting and preparing data, followed by formulating labels and attributes grounded in linguistic theory, and drafting the annotation scheme and accompanying guidelines. Subsequent steps include piloting the scheme on a sample dataset, evaluating the outcomes through IAA, and revising the scheme and guidelines if needed. The process concludes with large-scale annotation, periodic evaluations, and, finally, model training. A comparable approach is presented by Pustejovsky et al. (2017) through the MATTER annotation cycle (Model, Annotate, Train, Test, Evaluate, Revise), which emphasizes the iterative nature of annotation development. A key component of this cycle is the MAMA loop (Model-Annotate-Model-Annotate), whereby annotation schemes are continually tested and refined.

Designing a robust annotation scheme is inherently complex and critical for producing high-quality annotated datasets. As emphasized by Finlayson and Erjavec (2017), this process should be

multi-phased, collaborative, and supported by appropriate tools. Additionally, the complexity of annotation tends to increase with the level of linguistic detail involved (Flickinger et al., 2017).

Once the scheme is designed, it is necessary to rigorously evaluate the annotation scheme and its guidelines. Among various evaluation approaches, IAA agreement remains one of the most widely adopted and recognized. Artstein (2017) points out that IAA is not just a measure of reliability; it is also a tool for refining annotation schemes and understanding how annotators interpret them. Artstein and Poesio (2008) conceptualize IAA as an indicator of annotation "trustworthiness". Commonly used metrics for measuring IAA include Cohen's kappa (Cohen, 1960), Krippendorff's alpha (Krippendorff, 2004), and simple percentage agreement. Bhardwaj et al. (2010) introduce Anveshan (Annotation Variance Estimation), a framework designed to evaluate patterns of annotator agreement and disagreement. This framework includes IAA agreement analysis and outlier detection based on annotation values.

However, reporting IAA results alone is often insufficient. Additional contextual information is necessary for meaningful interpretation. Bayerl and Paul (2024) advocate for including essential metadata to ensure transparent assessment of agreement, such as annotator expertise (e.g., novices, domain experts, scheme developers). Furthermore, Bayerl and Paul (2024) identify factors that can influence IAA agreement such as the annotation domain, the number of categories in the annotation scheme, the number and expertise of annotators, the training provided to annotators, the purpose of the annotation task, and the specific agreement metrics used. From a different perspective, Basile et al. (2021) challenge the idea of a singular "correct" annotation. They identify three primary sources of disagreement — annotator-related, data-driven, and context-dependent — and argue for embracing disagreement within evaluation frameworks, promoting the use of multiple annotations and adaptive metrics.

Analyzing the sources of annotator disagreement can be a productive strategy for improving annotation schemes and guidelines. Teruel et al. (2018) and Hovy and Lavid (2010) demonstrate that such analysis can lead to greater clarity in annotation instructions and scheme structure. Likewise, Levi and Shenhav (2022) advocate for breaking down annotation tasks into distinct layers to effectively isolate and address sources of disagreement. Dickinson and Tufis (2017) highlight the value of "iterative enhancement" — a process that involves identifying errors to accelerate annotation and improve its quality. This iterative process often results in enhanced guidelines and refined annotation schemes. Beck et al. (2020) discuss five different sources of problems in annotations: ambiguities and variations in the data, uncertainty among the annotators, errors, and biases. According to the authors, failing to address these issues can have undesirable consequences for different phases of the annotation process, while resolving them can yield more robust scientific results.

While the majority of the reviewed studies emphasize important aspects to consider in the development and validation of annotation schemes, they rarely provide a detailed, step-by-step account of the entire annotation process. In contrast, our work aims to fill this gap by offering a comprehensive framework for structuring the annotation workflow. Specifically, we highlight the critical role of analyzing human variation as a means to iteratively refine both the annotation scheme and the accompanying guidelines.

3 A case study

In this section, we present the methodology developed to design and validate our annotation scheme, as outlined in Figure 1.

The proposed approach is structured into four distinct phases, each comprising multiple steps that guide the annotation process from conception to evaluation. To illustrate the practical application of our methodology, we conduct a case study in which we implement and assess an annotation scheme tailored to extract both grammatical and medical information embedded within clinical narratives. The source material includes admission reports, discharge summaries, and general clinical notes. This annotation scheme serves as the foundation for constructing an annotated corpus of medical records written in European Portuguese, specifically from patients diagnosed with Acute Myeloid Leukemia (AML), a relatively understudied condition, being the extraction of structured data from clinical narratives essential to support and facilitate research efforts. Additionally, the proposed annotation scheme and the resulting annotated dataset will enable a detailed investigation of the semantic characteristics of medical records, particularly for

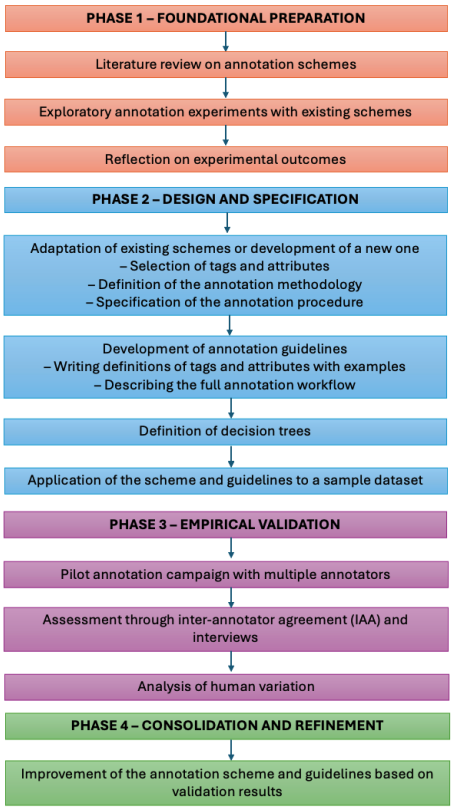

Figure 1: The proposed methodology for the development and validation of the annotation scheme.

temporal features.

Subsection 3.1 details the methodology employed in the development of the annotation scheme, while Subsection 3.2 discusses the procedures used to validate the scheme.

3.1 The development of the annotation scheme and guidelines

The initial step of Phase 1 involved a comprehensive review of the literature to identify existing frameworks for annotating clinical reports with morphosyntactic, semantic, and medical information[1]. Over the years, several proposals have focused on the annotation of grammatical information — particularly entities and temporal relations — as well as the integration of clinical information via medical ontologies (e.g., Roberts et al., 2009; Styler IV et al., 2014; Oliveira et al., 2022; Nunes et al., 2024).

Given our objective to represent both the temporal properties and key medical aspects of clinical reports in European Portuguese, we prioritized annotation schemes that provided robust frameworks for these two dimensions. For grammatical information, the Text2Story annotation scheme offered a comprehensive and multilayered proposal for capturing various temporal features in textual data. This scheme (Silvano et al., 2021; Leal et al., 2022) was developed in alignment with the ISO 24617 standard (International Organization for Standardization, 2012), and was originally applied to annotate morphosyntactic and semantic elements in European Portuguese news articles. Its temporal layer builds upon ISO TimeML (ISO-24617-1, 2012), a widely adopted standard with demonstrated applicability across diverse contexts, and includes adaptations tailored to the specificities of Portuguese. The Text2Story annotation scheme has several key advantages over alternative frameworks such as PropBank, Abstract Meaning Representation, and Penn Treebank since these are characterized as closed systems, with predefined structures and fixed category sets that constrain their flexibility and limit their applicability across diverse domains or layers of annotation. In contrast, ISO 24617, from which ISO TimeML is one part, offers a more open and modular architecture, supporting the integration of multiple layers of annotation. Additionally, ISO 24617 was conceived as an interoperable standard, designed to accommodate a range of theoretical models and natural languages, allowing for its adaptation, with minimal modifications, to different linguistic and contextual settings.

Concerning medical information, our review highlighted two annotation schemes — i2b2 (Sun et al., 2013) and MERLOT (Campillos et al., 2018) — as particularly relevant. Both were specifically designed for the medical domain and have demonstrated promising results in producing large-scale, complex clinical annotations, along with achieving high IAA scores. The selection of these schemes was based on a preliminary analysis that considered not only the coverage of relevant clinical categories but also the robustness of the models. Subsequently, practical annotation experiments were conducted using these frameworks to evaluate their performance in annotating our specific corpus. For this preliminary comparative analysis, six pseudonymized admission reports from patients treated at IPO-Porto, Portugal, were manually annotated using three different annotation schemes. The results demonstrated that the Text2Story annotation scheme was more effective in capturing morphosyntactic and semantic information. However,

[1]For a more detailed review of the annotation schemes designed for clinical narratives, the reader is referred to (Fernandes et al., 2025)

it was inadequate for representing domain-specific medical content. Conversely, while the i2b2 and MERLOT schemes facilitated the annotation of relevant clinical concepts, the labels employed were overly broad and lacked the specificity required for fine-grained semantic representation in the medical domain. The summary of the results of this comparison can be found in Table 5 in the Appendix A[2].

Following this initial evaluation, it became clear that none of the existing annotation schemes could be adopted without substantial modification. To further investigate the identified limitations and inform the development of a more suitable scheme, we analyzed a broader corpus of 100 pseudonymized clinical narratives from IPO-Porto, comprising admission reports, discharge summaries, and general clinical notes. This extended analysis was conducted in collaboration with a medical specialist from IPO-Porto to identify the essential clinical information that should be captured in the annotation process.

Grounded on the results of our analysis, we commenced Phase 2 - Design and Specification of the annotation scheme and guidelines. For grammatical information, we concluded that the Text2Story scheme provided a comprehensive set of labels for encoding the morphosyntactic and semantic properties of events and temporal expressions. In addition to entity structures (events and temporal expressions), the Text2Story scheme — consistent with the ISO TimeML standard — also includes link structures such as Temporal Links (TLinks), which support the representation of temporal relations among events. The selection of domain-specific medical labels was guided by the UMLS Metathesaurus ontology (Bodenreider, 2004), providing a systematic and internationally recognized framework. The definitions of the medical labels presented in this work were also informed by the contributions of Leite (2024), whose research on the same corpus proposed a preliminary set of clinically relevant categories validated by a specialized physician. Several of these categories were retained, while others were adapted or refined to better suit the present annotation goals.

Building on this foundation, a set of domain-specific tags was introduced to support the structured representation of medically relevant information. These include Sign or Symptom, Personal History (Past Medical History, Comorbidity or Undefined), Intercurrence, Examination, Examination Result, Principal Diagnosis, Characterization of the Disease, Medical Procedure, Treatment, Drug Administration Route, and Treatment Response. Adding these tags solved the problem of overly broad categories present in other schemes. Additionally, a decision tree was developed for selecting domain-specific medical labels to ensure consistency and accuracy in the annotation process, minimizing ambiguities and enhancing the replicability of results. Since the annotation of clinical narratives involves interpreting medical terms in different contexts, the hierarchical structure of the decision tree helps guide annotators in selecting the most appropriate labels, reducing inter-annotator variability. This enhancement appears to be particularly advantageous for both annotators with a medical background and those without. For the former, familiarity with this method, widely used in clinical settings to support decision-making (Bae, 2014), facilitates a more intuitive and effective adoption of the annotation scheme. For the latter, the decision tree serves as a structured guide that aids in understanding the annotation criteria, reducing the need for extensive prior knowledge of medical terminology and promoting greater standardization in the annotation process. Once the initial version of the annotation model was defined, it was iteratively tested and refined using the annotated data until it was capable of representing all relevant information present in the clinical records. Throughout this iterative process, comprehensive annotation guidelines were developed. These guidelines include detailed descriptions of each annotation phase, definitions and attributes for all labels, illustrative examples drawn from the dataset, and clarifications for complex or ambiguous cases encountered during annotation. This version of the scheme and guidelines can be found in the GitHub repository.

3.2 Assessment of the annotation scheme and guidelines

Phase 3 of our proposal involves the validation of the annotation scheme and its guidelines, with a focus on evaluating its consistency, reliability, and interpretability. As discussed in Section 2, IAA is a widely accepted strategy for assessing the quality of annotation guidelines and the clarity of the annotation model itself.

To carry out this evaluation, we conducted a

[2]A detailed analysis of the results from these experiments, and a thorough justification of the selection of the most suitable scheme will be the subject of future publication.

small-scale experiment involving two linguistics students with prior experience in annotation tasks. The INCEpTION tool (Klie et al., 2018) was configured with our proposed annotation scheme, and the annotators were provided with both the scheme and its accompanying guidelines. They were instructed to annotate a set of synthetic clinical reports, which included one group consultation note, three discharge reports, and one general report concerning a patient diagnosed with AML. These reports were generated by a specialist physician from IPO-Porto to ensure clinical relevance and realism. The reports can be found in the GitHub repository.

In addition to the IAA analysis, we implemented a curation-based evaluation strategy to further assess the validity and practical applicability of the annotation scheme and guidelines. The curator, who held a background in both linguistics and pharmaceutical sciences, reviewed the annotated documents to identify common annotation errors and challenges faced by the annotators. This process facilitated the detection of inconsistencies, such as the assignment of divergent labels to semantically similar events, which were often traced back to ambiguities or insufficient clarity in the annotation guidelines. Such findings were instrumental in refining both the scheme and its documentation, thereby improving the overall robustness and reliability of the annotation process.

Subsequently, we computed IAA metrics, which are reported in the following section. The agreement was quantified using Cohen's Kappa and Krippendorff's Alpha, two well-established statistical measures for evaluating reliability (Artstein, 2017). Values closer to 1 indicate stronger agreement and, by extension, a more reliable annotation scheme. Furthermore, treating the curator's annotations as the reference (or "gold standard"), we also measured the annotation distance between each annotator and the curator to assess alignment with expert judgment.

Finally, we conducted a detailed qualitative analysis of the sources of disagreement, to understand the underlying factors contributing to human variation in annotation. These findings provided insights that informed subsequent refinements to both the annotation scheme and the supporting guidelines.

3.2.1 The analysis of IAA and curation

The analysis of IAA and curation outcomes provides valuable insights into the effectiveness and clarity of the annotation scheme and its accompanying guidelines. As shown in Table 1, the identification of text spans corresponding to events and time expressions and temporal links (TLinks) between events, events and time expressions, and between time expressions achieved substantial agreement, as indicated by Cohen's kappa values (Landis and Koch, 1977). Notably, agreement between individual annotators and the curator is higher than that observed between annotators, for both text spans and TLinks. In particular, the agreement between Annotator 1 (ANN1) and the curator for text span identification reached the threshold for almost perfect agreement, suggesting strong alignment with the curation standard.

Table 1: IAA (initial pilot) on span and relation annotations (exact match criteria) between ANN1, ANN2, and the curator, based on the curated reference.

type	annotators	krippendorff_alpha	cohen_kappa
relation	ANN2, Curator	0.761	0.760
	ANN1, Curator	0.754	0.754
	ANN1, ANN2	0.614	0.614
span	ANN2, Curator	0.741	0.742
	ANN1, Curator	0.910	0.910
	ANN1, ANN2	0.682	0.684

A closer examination of the divergences between annotators and the curator regarding text span annotation reveals two primary sources of disagreement: (i) cases in which both annotators recognize the same event or temporal expression but differ in the extent of the annotated span; and (ii) cases in which only one annotator identifies the event or temporal expression.

In the first category, although both annotators consistently identify the same underlying event — typically marked by the same nuclear noun — discrepancies arise due to variations in the delimitation of the annotated span. These differences are attributable to factors such as: (a) the inclusion or omission of leading or trailing whitespace; (b) divergent judgments on whether to annotate the full nominal phrase, including modifiers or complements, versus only its nucleus (e.g., [antecedentes relevantes] 'relevant antecedents' vs. [antecedentes] 'antecedents'); (c) inclusion of quantifiers (e.g., [duas consolidações] 'two consolidations' vs. [consolidações] 'consolidations'); (d) the presence or absence of prepositions introducing the expression (e.g., [em remissão completa] '(in) complete remission' vs. [remissão completa] 'complete remission'); and (e) the presence of multiple semantic units within a single span, such as "cariótipo normal" ('normal karyotype'), which one

annotator treats as a single markable, while the other annotates "cariótipo" ('karyotype') and "normal" ('normal') as separate events.

The second category comprises 22 instances in which one annotator identified a markable that the other did not. These omissions often stem from challenges in interpreting domain-specific language and document structure. For instance, in one recurring case, the term "resumo" ('summary') — used to introduce a retrospective overview of the patient's clinical history — is annotated as a General Event Class by one annotator, while the other omits it, possibly not recognizing its functional role. Similar inconsistencies are observed with specialized medical terminology unfamiliar to one or both annotators. Terms such as "blastos" ('blasts') and "piperacilina-tazobactam" are annotated as events by one annotator, while the other does not annotate them. The same applies to acronyms and abbreviations from the medical domain (e.g., "7+3", "NPM1+", "FLT3+", "EV"), which are variably interpreted either as temporal expressions or domain-specific events.

Finally, several cases of disagreement can be attributed to differences in grammatical interpretation. For example, in the phrase "fez indução" ('did induction'), one annotator treats "fez" ('did') as a main verb and accordingly annotates it as an event, while the other classifies it as a light verb, and instead identifies "indução" ('induction') as the semantic nucleus, thereby excluding "fez" from annotation. Such differences highlight the challenges posed by complex syntactic constructions and further underscore the importance of clear, unambiguous annotation guidelines.

Turning to the analysis of inter-annotator agreement (IAA) on event attributes, as presented in Table 2, the results reveal considerable variability in agreement levels across different attributes. Agreement values between Annotators 1 (ANN1) and 2 (ANN2) range from fair ($\kappa = 0.22$ for Aspect) to almost perfect ($\kappa = 0.95$ for Part of Speech).

The low agreement for the Aspect attribute suggests potential issues in the clarity or interpretation of the guideline's definition. The current description — "The grammatical category that expresses the way an event is structured internally and unfolds over time (over an interval or in a moment), taking into account whether its duration is indeterminate or whether it has boundaries" — may have inadvertently introduced confusion. Although the Aspect attribute is intended to reflect grammatical aspect, its definition appears to overlap conceptually with lexical aspect, which is covered under the Class and Event Type attributes. This ambiguity likely contributed to the lower agreement for Aspect, especially when compared to the higher levels observed for Class ($\kappa = 0.56$) and Event Type ($\kappa = 0.68$), suggesting that annotators found it easier to identify lexical rather than grammatical aspectual properties.

The agreement for Verb Form is also relatively low ($\kappa = 0.37$), which is somewhat unexpected. This attribute involves the recognition of non-finite verb forms — typically a straightforward task for annotators with linguistic expertise. Interestingly, this agreement value is lower than that observed for Tense ($\kappa = 0.78$), despite the latter also involving morphological identification, albeit of finite verb forms. This discrepancy may indicate that the annotation of non-finite forms introduces ambiguities not present in the identification of tense.

As anticipated, the Part-of-Speech attribute yielded the highest agreement ($\kappa = 0.95$), reflecting the annotators' strong background in linguistics and the relative simplicity of identifying major word classes. In contrast, Polarity achieved only substantial agreement ($\kappa = 0.60$), which is somewhat surprising given that polarity identification is similarly considered a relatively simple classification task. This suggests that further clarification or refinement of the annotation criteria for Polarity may be beneficial.

With respect to the Specialized Event Class attribute, the agreement between annotators was substantial ($\kappa = 0.73$). Considering that the annotators have domain expertise in linguistics rather than medicine, this level of agreement suggests that the annotation manual's definitions and examples drawn from the clinical domain are generally accessible and comprehensible. Nevertheless, these results also point to opportunities for refinement, particularly in enhancing the clarity of domain-specific guidelines to further support non-expert annotators.

As for Time spans, the results are very diverse: the agreement values between annotators are less than chance agreement regarding "Temporal Function" (because one of the annotators did not perform this annotation), but are perfect and almost perfect regarding Time Type as revealed by Table 3.

Table 4 presents the results of IAA for temporal relation annotations across varying threshold lev-

Table 2: IAA scores (initial pilot) on event attributes between ANN1, ANN2, and the curator, based on the curated reference.

type	annotators	krippendorff_alpha	cohen_kappa
aspect	ANN1, ANN2	0.227	0.252
	ANN2, Curator	0.460	0.440
	ANN1, Curator	0.126	0.145
class	ANN1, ANN2	0.568	0.566
	ANN2, Curator	0.789	0.786
	ANN1, Curator	0.769	0.767
event	ANN1, ANN2	0.683	0.680
	ANN1, Curator	0.816	0.814
	ANN2, Curator	0.851	0.848
polarity	ANN1, ANN2	0.606	0.606
	ANN1, Curator	0.920	0.920
	ANN2, Curator	0.608	0.607
pos	ANN1, ANN2	0.959	0.959
	ANN2, Curator	0.889	0.889
	ANN1, Curator	1.000	1.000
specialized	ANN1, ANN2	0.731	0.730
	ANN2, Curator	0.792	0.792
	ANN1, Curator	0.820	0.819
tense	ANN1, ANN2	0.787	0.783
	ANN1, Curator	1.000	1.000
	ANN2, Curator	0.705	0.703
vform	ANN1, ANN2	0.379	0.375
	ANN1, Curator	0.462	0.429
	ANN2, Curator	0.690	0.667

Table 3: IAA results (initial pilot) for time expression attributes between ANN1, ANN2, and the curator, based on the curated reference.

type	annotators	krippendorff_alpha	cohen_kappa
temporal function	ANN1, ANN2	-0.326	0.063
	ANN2, Curator	-0.389	0.049
	ANN1, Curator	0.523	0.520
time type	ANN1, ANN2	1.000	1.000
	ANN2, Curator	1.000	1.000
	ANN1, Curator	0.904	0.902

els. As the threshold increases from 0 to 3, both the number of matched temporal links (TLinks) and the proportion of those matches that include agreement on the relation type (e.g., Before, After, Overlap) also increase. This suggests that applying more relaxed matching criteria — specifically regarding the span boundaries — improves alignment between annotators. Consequently, the percentage of agreement on TLink attributes rises from 26.7% at threshold 0 to 31.9% at thresholds 2 and 3. At threshold 0, among a total of 212 TLinks established between events, events and time expressions, and between time expressions, annotators agreed on the TLink in 41% of the cases, and only in 26% of the cases (56 out of 87) did they agree on the TLink attribute. However, when filtered to exclude the cases where annotators disagreed on the TLink attribute and considering only the 56 cases of agreement, the proportion of agreement significantly increases to 64.4%. Although further detailed analysis is required to identify the underlying causes of disagreement, these results point to the complexity of annotating temporal relations and suggest that clearer annotation guidelines may be necessary to ensure more consistent labeling. Additionally, these findings underscore the importance of further training for annotators to enhance reliability in this domain.

Table 6 in the Appendix A presents the distribution of label annotations in the initial pilot study after curation, while Table 7 shows the distribution of attributes for the specialized events in the same pilot study.

Table 4: Results of IAA between annotators in TLINKs and TLINKs attributes (initial pilot).

threshold	#TLink matches	#matches in TLink type	% agreement TLink matches	% agreement matches in TLink type	% agreement matches in TLink type (filtered)
0	87	56	0.414	0.267	0.644
1	103	64	0.490	0.305	0.621
2	109	67	0.519	0.319	0.615
3	110	67	0.524	0.319	0.609

3.2.2 Improvement of the annotation scheme and guidelines

The analysis of the curation results and IAA presented in Section 3.2 highlighted several issues that required clarification in the annotation scheme and its associated guidelines, particularly concerning the definition of markables. Although a detailed definition for markables was already provided in the guidelines, we decided to refine the instructions by specifying that markables should not include whitespace before or after the span, nor punctuation marks such as commas. Additionally, the statistical analysis revealed the need for further clarification regarding the annotation of noun complements and modifiers, as well as quantifiers. Specifically, when an event is accompanied by a temporal complement or modifier, such as "quadro recente" (‘recent case’), the modifier should be annotated with the Time label and receive the attributes defined by TIDES 2005 (Ferro et al., 2005). To facilitate this, an open field labeled Value was introduced. Furthermore, in cases where events are preceded by quantifiers, such as "duas consolidações" (‘two consolidations’), the quantifier should not be annotated as part of the event but should instead be captured in the quantification field.

Concerning lexicalized and semi-lexicalized expressions, although the guidelines already specified that the entire expression should be marked — including prepositions — we decided to include the example "em remissão completa" (‘in complete re-

mission'), as it is a recurrent expression in medical reports.

Another issue pertained to the annotation of abbreviations. For instances such as "O FLT3 foi +" ('the FLT3 was +'), where the symbol "+" represents the event 'positive', a mechanism was needed to ensure proper annotation. To address this, an open field called Observations was introduced, enabling the abbreviation to be annotated as an event with its full form recorded in that field.

With polarity, we clarified that events preceded by negative quantifiers, such as "nada" ('nothing'), or by negative verbs, such as "deixar de + infinitive" ('to stop + infinitive'), should also be annotated with a negative polarity attribute.

Some annotation errors arose due to the annotators' lack of medical knowledge. Although the decision tree assists in the selection of domain-specific labels, we believe that the annotation process would be further facilitated if annotators received brief training on the specific disease reported in the medical records — in this case, Acute Myeloid Leukemia. Familiarity with domain-specific concepts would enable annotators to better identify and apply the relevant labels. To this end, we incorporated a short video presentation, accessible via QR code, created by a specialist physician at IPO-Porto.

In addition to analyzing the curation results and IAA, we conducted interviews with annotators to identify the main difficulties encountered during the annotation process. The aim was to refine the annotation scheme and improve its applicability. One issue that was raised was related to the label General Event Class, which included an attribute called Class. This terminology caused ambiguity, complicating the annotation process. To resolve this, the scheme was reorganized, renaming General Event Class to General Event, while retaining the name of the Class attribute. To maintain terminological consistency, the label Specialized Event Class was also renamed to Specialized Event. Another issue highlighted by the annotators was the redundancy in annotating events within the Specialized Event Class, which required dual labeling with both Specialized Event Class and General Event Class. This redundancy arose because certain attributes, such as Polarity and Part of Speech, were only defined for the General Event Class. To address this, these attributes were integrated directly into the Specialized Event Class, eliminating the need for dual labeling. However, attributes exclusive to the General Event Class were not incorporated, as events in the Specialized Event Class typically correspond to nouns and adjectives, which only receive Polarity and Part-of-speech attributes. Another challenge reported by annotators was related to inter-document annotation. Annotators experienced difficulty identifying which relationships should be established between different medical reports for the same patient. To address this, the guidelines were clarified to specify how events and expressions should be linked across multiple reports. It was established that the Doctime (date of report creation) should always be connected to both the previous and subsequent report dates. Events in the reports should only link to the previous report via TLINK Identity when pertinent to the understanding of the patient's story. Additionally, two new attributes, Admission Date and Discharge Date, were introduced for dates. When a report is written during a hospitalization period, the Doctime of that report should be linked to both the Admission Date and Discharge Date of the corresponding report. When the Doctime corresponds to the Discharge Date, only the latter should be assigned.

Figure 2 in the Appendix A shows the annotation of a corpus excerpt using the latest version of the annotation scheme. The final version of the scheme and the corresponding guidelines can be accessed in the GitHub repository.

4 Final remarks

In this work, our main goal was to describe the incremental process of developing and validating an annotation scheme, along with its corresponding guidelines, capable of integrating both linguistic and medical domain information in an inter-document annotation. The results of the annotation and curation phases enabled improvements to both the scheme and the guidelines through an iterative refinement process. Developing an annotation scheme requires ongoing efforts toward improvement. With that in mind, we intend to further explore issues related to the identification of grammatical features and to develop a question–answer system that facilitates the selection of domain-specific labels, even for annotators without prior knowledge of the field.

Acknowledgments

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | https://doi.org/10.54499/LA/P/0063/2020. The authors also acknowledge the support of the StorySense project (DOI 10.54499/2022.09312.PTDC).

References

R. Artstein. 2017. Inter-annotator agreement. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 297–313. Springer Netherlands.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jong-Myon Bae. 2014. The clinical decision analysis using decision tree. *Epidemiology and Health*, 36:e2014025.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Petra Saskia Bayerl and Karin I. Paul. 2024. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*.

Christin Beck, Hannah Booth, Mennatallah El-Assady, and Miriam Butt. 2020. Representation problems in linguistic annotations: Ambiguity, variation, uncertainty, error and bias. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 60–73, Barcelona, Spain. Association for Computational Linguistics.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A framework for analysis of multiple annotators' labeling behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55, Uppsala, Sweden. Association for Computational Linguistics.

O. Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic Acids Research*, 32:D267–D270.

L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, and A. Névéol. 2018. A french clinical corpus with comprehensive semantic annotations: Development of the medical entity and relation limsi annotated text corpus (merlot). *Language Resources and Evaluation*, 52:571–601.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Markus Dickinson and Dan Tufis. 2017. Iterative enhancement. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 257–276. Springer.

Ana Luísa Fernandes, Purificação Silvano, Nuno Guimarães, Rita Rb-Silva, Tahsir Ahmed Munna, Luís Filipe Cunha, António Leal, Ricardo Campos, and Alípio Jorge. 2025. Human experts vs. large language models: Evaluating annotation scheme and guidelines development for clinical narratives. In *Proceedings of the Text2Story 2025 – Eighth International Workshop on Narrative Extraction from Texts*, pages 149–160, Lucca, Italy. CEUR-WS.org.

Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, and George Wilson. 2005. Tides 2005 standard for the annotation of temporal expressions. Available online at: `http://www.timeml.org/timex2/`.

Mark A. Finlayson and Tomaž Erjavec. 2017. Overview of annotation creation: Processes & tools. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 167–192. Springer.

Dan Flickinger, Stephan Oepen, and Emily Bender. 2017. Sustainable development and refinement of complex linguistic annotations at scale. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 353–377. Springer.

Yvonne Graham and Jacques van der Meer. 2015. Interannotator agreement for qualitative data analysis in research: Methods and strategies. *Qualitative Research Journal*, 15(3):1–18.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *International Journal of Translation Studies*, 22:13–36.

Nancy Ide. 2017. Introduction: The handbook of linguistic annotation. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 1–10. Springer, Dordrecht.

International Organization for Standardization. 2012. *ISO 24617:2012 - Language Resource Management – Semantic Annotation Framework*. ISO.

Orazio Irrera, Simone Marchesin, and Gianmaria Silvello. 2024. Metatron: Advancing biomedical annotation empowering relation annotation and collaboration. *BMC Bioinformatics*, 25(1):1–41.

ISO-24617-1. 2012. Language resource management - semantic annotation framework (semaf) - part 1: Time and events (semaf-time, iso-timeml). Standard, Geneva, CH.

J.-C. Klie, M. Bugert, B. Boullosa, R. Eckart de Castilho, and I. Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2nd edition. Sage Publications, Thousand Oaks, CA.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

A. Leal, P. Silvano, E. Amorim, I. Cantante, F. Silva, A. Jorge, and R. Campos. 2022. The place of iso-space in text2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 61–70. European Language Resources Association.

M. A. Leite. 2024. Ontology-based extraction and structuring of narrative elements from clinical texts. Mater´s thesis, Universidade do Porto.

Eitan Levi and Shaul R. Shenhav. 2022. A decomposition-based approach for evaluating inter-annotator disagreement in narrative analysis. *arXiv preprint*.

G. Moharasan and Tu-Bao Ho. 2019. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research*, 3(2):220–244.

M. Nunes, J. Boné, J. C. Ferreira, P. Chaves, and L. B. Elvas. 2024. Medialbertina: An european portuguese medical language model. *Computers in Biology and Medicine*, 182:109233.

L. E. S. e Oliveira, A. C. Peters, A. M. P. da Silva, C. P. Gebeluca, Y. B. Gumiel, L. M. M. Cintho, D. R. Carvalho, S. Al Hasan, and C. M. C. Moro. 2022. Semclinbr—a multi-institutional and multi-specialty semantically annotated corpus for portuguese clinical nlp tasks. *Journal of Biomedical Semantics*, 13(1):13.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 31–63. Springer, Dordrecht.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O'Reilly Media, Inc.

Angus Roberts, Robert J. Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42 5:950–66.

P. Silvano, A. Leal, F. Silva, I. Cantante, F. Oliveira, and A. Jorge. 2021. Developing a multilayer semantic annotation scheme based on iso standards for the visualization of a newswire corpus. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 1–13, Groningen, The Netherlands (online). Association for Computational Linguistics.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

W. Sun, A. Rumshisky, and O. Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46:S5–S12.

Marta Teruel, Cristian Cardellino, Federico Cardellino, Laura Alonso Alemany, and Serena Villata. 2018. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

E. Zhu, Q. Sheng, H. Yang, Y. Liu, T. Cai, and J. Li. 2023. A unified framework of medical information annotation and extraction for chinese clinical text. *Artificial Intelligence in Medicine*, 142:1–12.

A Appendix

Table 5: Comparison of the analyzed annotation frameworks

Feature	Text2Story	i2b2	Merlot
Medical domain coverage	-	+	++
Morphosyntactic and grammatical domain coverage	+++	+	+
Existence of the TLINK before_overlap (captures temporal info "recently")	-	+	-
Existence of the TLINK identity (captures coreference of same event)	+	-	-

Table 6: Distribution of annotation labels in the corpus of the initial pilot.

Label	Count
Specialized Events	100
General Events	64
Times	22
TLinks	228

Table 7: Distribution of Specialized Event tags

Category	Count
Personal History	3
Sign or Symptom	17
Examination	12
Examination Result	11
Principal Diagnosis	5
Treatment	19
Intercurrence	10
Characterization of the Disease	11
Treatment Response	10
Drug Administration Route	2

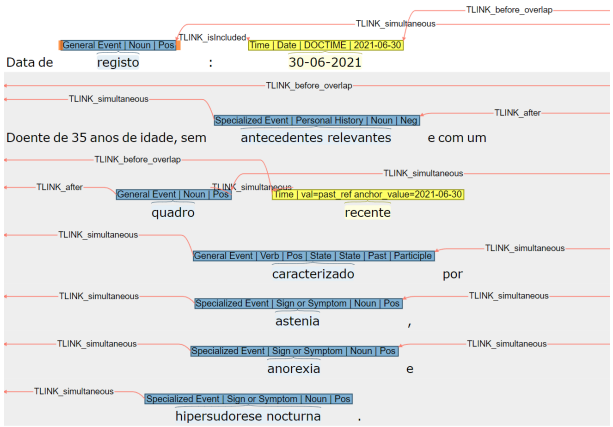

Figure 2: Annotation of an excerpt from a medical report using the latest version of the annotation scheme. Events are marked in blue and temporal expressions in yellow. The annotated excerpt illustrates the identification of various attributes associated with both events and temporal expressions, as well as the temporal relations between events and between events and temporal expressions. "Registration date: 06/30/2021. The patient is a 35-year-old with no relevant medical history, presenting with recent symptoms of asthenia, anorexia, and night sweats".