Can ISO 24617-1 go clinical? Extending a General-Domain Scheme to Medical Narratives

Ana Luísa Fernandes
CLUP / Porto, Portugal
INESC TEC / Porto, Portugal
University of Porto / Porto, Portugal
ana.l.fernandes@inesctec.pt

Purificação Silvano
CLUP / Porto, Portugal
INESC TEC / Porto, Portugal
University of Porto / Porto, Portugal
purificacao.silvano@inesctec.pt

António Leal
CLUP / Porto, Portugal
University of Porto / Porto, Portugal
University of Macau / Macau, China
antonioleal@um.edu.mo

Nuno Guimarães
INESC TEC/ Porto, Portugal
University of Porto / Porto, Portugal
nuno.r.guimaraes@inesctec.pt

Evelin Amorim
INESC TEC/ Porto, Portugal
University of Porto / Porto, Portugal
evelin.f.amorim@inesctec.pt

Abstract

The definition of rigorous and well-structured annotation schemes is a key element in the advancement of Natural Language Processing (NLP). This paper aims to compare the performance of a general-purpose annotation scheme — Text2Story, based on the ISO 24617-1 standard — with that of a domain-specific scheme — i2b2 — in the context of clinical narrative annotation; and to assess the feasibility of harmonizing ISO 24617-1, originally designed for general-domain applications, with a specialized extension tailored to the medical domain. Based on the results of this comparative analysis, we present Med2Story, a medical-specific extension of ISO 24617-1 developed to address the particularities of clinical text annotation.

1 Introduction

Developing robust and coherent annotation schemes is key to the advancement of Natural Language Processing (NLP). These schemes provide formalized frameworks that define which linguistic or domain-specific phenomena are to be annotated, and how such information should be consistently represented across datasets. By standardizing the annotation process, they ensure that labeled data is meaningful and interpretable to downstream algorithms (Pustejovsky and Stubbs, 2012).

Throughout the years, several annotation frameworks have been developed providing structured labels and attributes for morphosyntactic (Marcus et al., 1993; Marneffe et al., 2021), semantic roles (Palmer et al., 2005; Jindal et al., 2022; Baker et al., 1998), coreference (Hovy et al., 2012), temporal (Pustejovsky et al., 2003) and discourse relations (Mann and Thompson, 1988; Prasad et al., 2018) information. Additionally, multi-layer annotation schemes that can cover different linguistic phenomena (Basile et al., 2012; Bos et al., 2017; Silvano et al., 2021; Bonn et al., 2024) have been proposed, thus enabling a more overarching linguistic representation. Concurrently, the growing demand for annotated schemes has heightened the need for standardization and interoperability. Initiatives such as the ISO 24617 — Semantic Annotation Framework (ISO TC37/SC4, 2012) support the development of reusable annotation models, thereby promoting consistency and facilitating comparative evaluation across datasets (Ide and Romary, 2006). For the most part, these annotation schemes are domain-general, designed to capture linguistic structure and meaning in any type of text. Nevertheless, some domains, such as the medical field, require more specialized annotation approaches. Due to the complexity and specificity of clinical language and concepts, task-specific annotation schemes are essential. These schemes are designed to capture entities such as medical conditions, medications (Sun et al., 2013), negation and uncertainty (Vincze et al., 2008), and temporal information (Uzuner et al., 2011; Roberts et al., 2021). Such domain-focused

Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21), pages 41–52

schemes are crucial for enabling effective information extraction in clinical settings, ultimately supporting decision-making and research in healthcare.

Choosing between a general-purpose and a domain-specific annotation scheme is a critical design decision that significantly affects the quality, applicability, and transferability of annotated datasets. Each approach offers distinct advantages and limitations, depending on the project's objectives, the nature of the source texts, and the intended downstream applications.

This paper pursues two main objectives: (1) to compare the performance of a general-purpose annotation scheme with that of a domain-specific scheme in the context of annotating clinical narratives; and (2) to explore the feasibility of harmonizing ISO 24617-1, a general-domain scheme, with a specialized medical branch. To that end, we introduce Med2Story, an extension of ISO 24617 tailored to the medical domain.

The creation of Med2Story will enable the systematization of data relevant to different domains: in linguistics, by supporting the study of issues such as the aspectual properties of event-denoting nouns; in computational research, by facilitating the training of models for the extraction of medical information; and in medicine, by promoting the detection of patterns and the transformation of unstructured data into structured data that is important to clinical research.

The remainder of this paper is structured as follows. Section 2 presents an overview of general-purpose and domain-specific annotation schemes. Section 3 describes the experimental setup, including the methodology, dataset, annotation schemes, and key findings. Section 4 introduces the Med2Story annotation framework. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Related work

Over the past several decades, numerous annotation schemes have been developed to address the representation of grammatical and domain-specific information in textual data. Within the scope of this study, we distinguish between *general-purpose annotation schemes*, which aim to capture linguistic structures and meaning in a domain-agnostic manner, and *domain-specific annotation schemes*, which are tailored to encode specialized knowledge relevant to particular fields. General-purpose schemes tend to offer broader linguistic coverage, often requiring detailed linguistic expertise for accurate annotation. In contrast, domain-specific schemes are typically narrower in focus and demand specialized domain knowledge (medical, economic) for effective annotation.

While many annotation schemes have concentrated on individual linguistic levels, such as morphological, syntactic, semantic, or pragmatic features, there have also been efforts to develop comprehensive, multilayer frameworks that encompass several of these dimensions. Among these, the Universal Dependencies (UD) framework (Nivre, 2016; Marneffe et al., 2021) stands out for its typologically-informed approach to morphosyntactic annotation, enabling cross-linguistic comparison. In the realm of semantic annotation, frameworks such as the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) and, more recently, the Uniform Meaning Representation (UMR) (Jens E. L. et al., 2021) focus on modeling multiple layers of meaning. UMR extends AMR to facilitate document-level semantic annotation, incorporating semantic roles, temporal relations, and discourse structures. Another important contribution is the ISO 24617-Semantic annotation framework (SemAF), which includes multiple modules for the annotation of temporal, referential, spatial, quantificational, and semantic-role-related information, among others. This framework offers a language-agnostic, interoperable and theoretical neutral architecture, allowing for its adaptation across languages with minimal modification. The Text2Story annotation scheme (Silvano et al., 2021; Leal et al., 2022), developed in compliance with ISO 24617 standards, is also a multilayer framework applied to the annotation of morphosyntactic and semantic information in European Portuguese texts.

Turning to domain-specific annotation, particularly within the biomedical and clinical domain, the focus of the present study, there have been efforts to develop annotation schemes that encode both domain-relevant and grammatical information. For instance, Albright et al. (2013) created an annotation scheme with syntactic and semantic layers alongside medical concepts. González-Moreno et al. (2025) annotated a dataset of Spanish clinical records with semantic groups. The MERLOT corpus (Campillos et al., 2018) comprises 500 French clinical narratives annotated for linguistic, seman-

tic, and structural features. The i2b2 annotation guidelines (Sun et al., 2012) include clinical and temporal annotations. Oliveira et al. (2022) developed SemClinBr, comprising 1,000 clinical notes in Brazilian Portuguese with semantic annotations, while, for European Portuguese, Lopes et al. (2019) compiled a set of 281 clinical case texts annotated for medical entities. Despite these advances, most existing clinical annotation resources exhibit several limitations. As noted by Zhu et al. (2023), inconsistencies are common, and comprehensive annotation encompassing both domain-specific and grammatical features is lacking.

Both general-purpose and domain-specific annotation schemes possess distinct strengths and limitations, which we assessed though the experiment described in the following section.

3 The experiment

3.1 The methodology

To assess the efficacy of both general-domain and domain-specific annotation schemes in capturing temporal information within medical reports, an experimental study was conducted utilizing two distinct schemes: the Text2Story scheme (Silvano et al., 2021; Leal et al., 2022) and the i2b2 scheme (Sun et al., 2013). As outlined in Section 2, the former is predicated on ISO standards 24617, whereas the latter was expressly designed for annotating clinical texts in English. The selection of the Text2Story annotation scheme was motivated by its comprehensive nature, interoperability, language-agnostic capabilities, and the potential for integrating, in a harmonized fashion, annotations across multiple semantic layers — such as referential, semantic roles, and spatial information, although the current focus is solely on its temporal module. Conversely, the i2b2 scheme was chosen due to its extensive validation and demonstrated capacity to encode not only specific medical information but also temporal features inherent in medical reports.

Following the selection of the annotation scheme, six pseudonymized clinical reports from patients diagnosed with Acute Myeloid Leukemia (AML) written in European Portuguese were annotated using the two distinct schemes. Subsequently, the annotation outputs were systematically analyzed to evaluate the respective strengths and limitations of each approach. Based on this comparative analysis, the most effective strategy for annotating both grammatical structures and domain-specific information in medical reports was identified.

3.2 The dataset

The dataset used in this study consists of six pseudonymized medical reports written in European Portuguese, originating from multidisciplinary group consultations involving six patients diagnosed with AML and followed at the Portuguese Oncology Institute in Porto, Portugal (IPO-Porto). Access to these documents was granted by the IPO-Porto Ethics Committee, and the research project was conducted within the framework of a data management plan approved by the institute (Rb-Silva and Karimova, 2021). The reports exhibit a complex temporal structure, as they incorporate relevant clinical history, diagnostic tests performed and their results, the patient's clinical trajectory leading up to the AML diagnosis, and the proposed treatment plan. The length of the reports varies, reflecting the amount of information available for each patient. The documents analyzed range from 115 to 316 words, with an average length of 210 words. This variation was intentional, as it allows for the investigation of whether text length influences the temporal complexity of the medical narrative and the annotation process.

3.3 The annotation

The annotation tool used in this study was the BRAT Rapid Annotation Tool (BRAT), developed by Stenetorp et al. (2012). Regarding the annotators, an analysis of inter-annotator performance differences conducted by Roberts et al. (2008) showed that a combination of linguistic and clinical expertise among annotators tends to result in higher annotation quality. The authors also argue that a document should not be annotated by a single annotator, as individual annotation may reflect several issues, including annotator-specific idiosyncrasies, occasional errors, and consistently lower performance. Based on these findings, the annotation team in this study consisted of one annotator and two curators with different expertise in the field of semantics. The annotator had a background in linguistics and pharmaceutical sciences, while the curators had extensive experience in linguistics. The annotation process followed a two-tier methodology: the initial annotations were carried out by the annotator and then reviewed by one of the curators. To ensure consistency and address ambiguities, weekly meetings were held with all team members to discuss challenging cases and refine annotation guidelines.

Six pseudonymized clinical reports were annotated according to two schemes: the i2b2 annotation scheme (Sun et al., 2013) and the temporal layer of the Text2Story scheme (Silvano et al., 2021; Leal et al., 2022). In both cases, the annotation followed a two-phase approach: in the first phase, events and temporal expressions were identified and annotated; in the second phase, temporal relations between these elements were established. This methodological separation between the annotation of entities and the annotation of relations was designed to enhance coherence and reliability. Annotating events and temporal expressions separately allowed for a clearer definition of the narrative's core elements prior to the relational analysis. As observed during the process, annotating events and relations simultaneously could compromise consistency across documents due to evolving annotation criteria. Therefore, the staged approach was essential for ensuring uniformity in the application of annotation standards.

3.4 The annotation schemes

3.4.1 Text2Story — an ISO-based general annotation scheme

The temporal layer of the Text2Story annotation scheme is based on the ISO 24617-1:2012 standard, Language Resource Management – Semantic Annotation Framework (SemAF) – Part 1: Time and Events (SemAF-Time, ISO-TimeML) (ISO-24617-1, 2012). This layer encompasses the annotation of events, temporal expressions, and temporal relations among these elements.

The **EVENT** category is defined as any eventuality that occurs or happens, as well as states or circumstances with temporal relevance — that is, elements directly associated with a temporal reference or change throughout the text. According to the scheme, events are classified into the following categories:

Occurrence: a situation that takes place or occurs;

State: a situation in which something holds or is considered true, with temporal relevance;

Reporting: an action by which an entity (person or organization) declares, narrates, or reports a situation;

Perception: a situation involving the physical perception of another situation;

Aspectual: an event focusing on a specific aspect of another event (e.g., beginning, end, or continuation);

Intensional Action: an event that introduces another event as an argument within an intentional context;

Intensional State: a state that introduces another event as an argument within an intentional context.

Each event is further annotated with attributes that specify its semantic and morphosyntactic properties, including:

Type: the type of event (state, process, and transition);

Tense: the grammatical verb tense (past, present, and future);

Aspect: verbal aspect (e.g., perfective, imperfective, progressive);

Polarity: polarity value (positive or negative);

Vform: verb form (gerundive, infinitive, and participle);

Mood: verb mood (subjunctive, future, conditional, and imperative);

Part of Speech: grammatical category (e.g., verb, noun, adjective);

Modality: expressed modality (e.g., epistemic, deontic).

TIMEX3 refers to temporal expressions representing time units. TIMEX3 expressions are annotated with one of the following tags: Date, Time, Duration, and Set. Additionally, the scheme includes the tag PUBLICATION_TIME, which marks the moment when the text was published.

Temporal relations are represented through **TLINK**, which describes links between two events, between two temporal expressions, or between an event and a temporal expression. Possible relations include: *before*, *after*, *includes*, *is_included*, *identity*, *begins*, *ends*, *begun_by*, and *ended_by*.

3.4.2 i2b2 — a specialized annotation scheme

The i2b2 temporal annotation scheme, also based on the ISO-TimeML standard, comprises the annotation of events (EVENT), temporal expressions (TIMEX3), and temporal relations (TLINK) among these elements.

EVENT refers to events mentioned or described in clinical narratives that are relevant to reconstructing the patient's clinical timeline. These events include, among others, symptoms, diseases, treatments, tests, and actions related to admission, transfer, or discharge from clinical departments. The scheme defines several types of events, namely:

Problem: Includes patient complaints, symptoms, diseases, and diagnoses;

Test: Refers to clinical (laboratory or physical) tests and their results;

Treatment: Covers medications, surgeries, and other clinical procedures;

Clinical_Department: Used to mark the clinical units to which the patient was admitted;

Evidential: Verbs expressing demonstration, reporting, or evidence are annotated as EVENTs, since, in clinical contexts, the source of information can be as important as the information itself;

Occurrence: This is the default EVENT type and is used for all other clinically relevant events that occur to the patient.

In addition to event type, EVENT may also be annotated for polarity (positive or negative) and modality, the latter being categorized as: factual, hypothetical, hedged, conditional, possible, or proposed.

TIMEX3 refers to temporal expressions indicating dates, times, durations, and frequencies. The scheme also includes the SECTIME tag, which records the creation date of the clinical report.

TLINK denotes temporal relations between EVENT and TIMEX3, and can assume the following values: *before*, *after*, *begun_by*, *ended_by*, *simultaneous*, *overlap*, and *before_overlap*.

3.4.3 The findings and discussion

The annotation schemes were successfully applied to the corpus under analysis; however, several limitations were identified throughout the process and will be discussed below.

Regarding the i2b2 annotation scheme, one of the main obstacles concerned the annotation of entities that, while clinically relevant, did not constitute eventualities. Entities such as clinical departments, hospital institutions, or drugs were annotated as events, which introduced difficulties in establishing temporal relations with actual events. According to the i2b2 guidelines, anything relevant to the patient's clinical timeline is considered an event: "In a medical record, anything that is relevant to the patient's clinical timeline is an event" (problem, test, treatment, clinical_department, evidential, and occurrence) (Sun et al., 2012). By including non-eventive entities as events, the grammatical and semantic integrity of the annotation was compromised. Some illustrative examples include the following:

(1) *"Por degradação do estado geral, com icterícia, náuseas e vómitos frequentes, foi internada no Hospital X" [Due to general condition deterioration, with jaundice, nausea, and frequent vomiting, the patient was admitted to Hospital X].*

(2) *"Decide-se propor o doente para tratamento de quimioterapia com idarrubicina e citarabina" [The patient was proposed for chemotherapy treatment with idarubicin and cytarabine].*

In example (1), "Hospital X" was annotated as an event (clinical_dept), although it did not constitute a semantic event. In (2), the expressions "chemotherapy", "idarubicin", and "cytarabine" were annotated as events (treatment), despite their semantic differences. "Chemotherapy" is an eventuality (a treatment that occurs), but the drugs "idarubicin" and "cytarabine" are participants, not events. In such cases, simultaneity TLINK were used to establish temporal relations. However, this did not adequately reflect the temporal structure underlying the described clinical situation.

This approach led to a loss of semantic and morphosyntactic information, since annotations that did not represent eventualities were treated as such. As mentioned in the previous section, the authors of the Text2Story scheme, following the ISO-24617-1 standard, define an event as an eventuality that happens or occurs, or a state or circumstance that is temporally relevant — that is, directly related to a temporal expression or change throughout the text. According to the same authors, a participant is the named entity that plays a relevant role in the described event or state. This distinction between events and participants allows for a more precise and granular representation of information. As a matter of fact, we observed that for more semantically accurate and grammatically rich annotation, entities should be explicitly represented as participants or as events.

Additionally, the i2b2 scheme treats as events only clinical concepts, clinical departments, occurrences, and evidential events, thus excluding stative eventualities. This limitation led to relevant information loss, as in example (3):

(3) *"Apresentava ainda conglomerados adenopáticos no retroperitoneu alto interessando sobretudo o compartimento pericelíaco, estendendo-se ao hilo hepático e à região pericefalopancreática" [The patient also presented with lymph node conglomerates in the upper retroperitoneum, mainly affecting the periceliac*

compartment, extending to the hepatic hilum and pericephalopancreatic region].

The verbs "affecting" and "extending" express states but were not annotated as events unless forcefully included under "occurrence", the default event type. These should be annotated as *state* events, in line with ISO-TimeML. Ambiguity also existed between the *problem* and *test* labels, as illustrated by example (4):

(4) *"Apresentava dilatação das vias biliares intra-hepáticas por provável compressão extrínseca no hilo, sem lesões hepáticas focais e um derrame pleural esquerdo diminuto" [The patient presented with dilation of the intrahepatic bile ducts due to probable extrinsic compression at the hilum, without focal liver lesions, and a small left pleural effusion].*

Expressions like "extrinsic compression", "focal liver lesions", and "pleural effusion" were annotated as *test* because they are exam results. However, they could also be considered clinical complications or pathological manifestations — thus *problems*. We observed that there was the need for a clearer distinction between the test and its result by introducing more specific tags. Another issue was the lack of support for discontinuous annotation. Annotation guidelines require events to be continuous sequences of text. For instance, in an example like (4), "focal liver lesions" must be annotated entirely, although ideally "lesions" and "focal" should be marked, excluding "liver" (an anatomical location). Such anatomical detail should be represented in a separate layer for more informative annotation.

Also problematic is the requirement to link all events to SECTIME (report creation date). This is not always necessary or relevant, particularly when a temporal relation can be inferred transitively. Enforcing this redundant link can overload and obscure the annotation.

Temporal relations posed challenges as well. The scheme enforces symmetry (*before*/*after*) but lacks mirror relations for *before_overlap*. For example, in (5), a *before_overlap* relation was needed between "recurrent infections" and "headaches", violating the guideline that TLINK should be annotated from right to left. An *after_overlap* relation would resolve this.

(5) *"Quadro recente caracterizado por hipersudorese, infeções de repetição e mais recentemente cefaleias" [Recent condition characterized by hyperhidrosis, recurrent infections, and more recently, headaches].*

Cases were also found where events and temporal expressions refer to the same point in time, as in (6).

(6) *"Assintomático até janeiro de 2017, altura em que inicia queixas de dor pélvica" [Asymptomatic until January 2017, when pelvic pain began].*

The word "when" refers to "January 2017", but the scheme lacks an *identity* TLINK to express this equivalence. Simultaneity annotation does not fully capture the relation. We suggest introducing an *identity* TLINK type.

As for the temporal dimension of the Text2Story annotation scheme, although it allowed for relevant morphosyntactic and semantic annotation, it lacked specificity for clinical annotation. We concluded that new, domain-specific labels were needed. Additionally, we noticed that aspectual annotation was really complex for nominal events, mainly for non-derived nominal events. This was not a scheme limitation, but one in the literature on aspectual classification of nouns. As a result, a great amount of disagreements among annotator and curators when labeling aspectual class for nominal events was observed.

Furthermore, though guidelines indicated that events should only be annotated with negative polarity when preceded by "not", we noticed that it would be necessary to include implicit negation, such as with the preposition "without", in (7).

(7) *"Doente sem antecedentes relevantes" [Patient without relevant history].*

As noted in subsection 3.4.1, the scheme allows verbs, nouns, adjectives, and prepositions as events, but not relative pronouns, which can be relevant in examples like (8):

(8) *"Fez um hemograma que mostrou a presença de leucocitose" [A blood count was done which showed leukocytosis].*

Therefore, the inclusion of *relative pronoun* as a valid POS tag would improve the annotation scheme.

In our experiment, we also noticed that some cases required more precise temporal relations (cf. (9)).

(9) *"Quadro recente caracterizado por hipersudorese, infeções de repetição e mais recentemente cefaleias" [Recent condition characterized by hy-*

perhidrosis, recurrent infections, and more recently, headaches].

No TLINK adequately captured "more recently". This gap would be solved with the inclusion of more specific temporal relations, such as *immediately_before* or, similar to the i2b2 scheme, *before_overlap*, as well as their respective mirror relations, *immediately_after* and *after_overlap*.

As noted, clinical texts are often written freely, with a variety of topics and medical concepts, complicating systematic annotation, as illustrated by example (10).

(10) *"Decide-se propor o doente para tratamento de quimioterapia com idarrubicina e citarabina, associado a tratamento intratecal" [The patient was proposed for chemotherapy with idarubicin and cytarabine, along with intrathecal treatment].*

ISO 24617-1 suggests annotating "intrathecal treatment" as one event. However, "intrathecal" indicates the administration route. We concluded that annotating "treatment" as an event of type *treatment*, and "intrathecal" as *route of administration* would be better, preserving necessary semantic granularity.

Regarding the quantitative analysis of the temporal structure of clinical reports, Table 1 and 2 in the Appendix A (also available in the paper GitHub repository [1]) present the frequencies of events, their respective attributes, and the temporal relations identified in both annotation schemes.

In the annotation conducted using the Text2Story scheme, the most frequent aspectual class was *state*, which was expected, since clinical history, diagnoses, and diseases are typically expressed as states. It was also observed that most of the annotated events were nouns. The polarity of events was predominantly positive, with negative occurrences being rare and mostly restricted to expressions such as "no relevant medical history".

In terms of event type, the most frequent class was *transition*, which is justified by the presence of significant clinical changes in the texts. With respect to temporal attributes, the most common tense was *pretérito perfeito [simple past]*, compatible with the retrospective nature of many clinical descriptions (e.g., "O hemograma *mostrou* leucocitose" [The blood count *showed* leukocytosis]). As for the *vform* (verbal form), the most frequent value was *participle*, often appearing in passive or descriptive constructions, such as "Quadro clínico *caracterizado* por hipersudorese" [Clinical presentation *characterized* by excessive sweating]. Concerning the *mood* attribute, only two instances in the conditional and one in the subjunctive were recorded.

With respect to TLINK, the most frequently annotated relation was *simultaneous*, with a significantly higher prevalence than other relations. This finding aligns with the informative structure of clinical reports, where multiple symptoms or conditions tend to occur or be described as happening simultaneously within the same temporal episode.

As for the annotation using the i2b2 scheme, the most frequent categories were *PROBLEM* and *TEST*, as the former includes medical history, diseases, and diagnoses, while the latter covers clinical examinations and their results. As observed with the Text2Story scheme, the predominant polarity was positive, and the most common temporal relation was also *simultaneous*, for the same reasons mentioned above.

It is also worth noting that a total of 323 events were annotated using the Text2Story scheme, compared to only 188 events annotated with the i2b2 scheme. This discrepancy can be attributed to the i2b2 scheme's lower capacity to represent semantically oriented events (e.g., states), which results in a significant loss of information. This limitation is also reflected in the number of temporal relations established: 1845 TLINK in the Text2Story scheme, versus only 1418 TLINK in the i2b2 scheme.

4 The Text2Story medical branch – Med2Story

After the comparative analysis between the i2b2 and Text2Story annotation schemes, the decision was made to develop an extension of the ISO-based Text2Story framework[2]. This decision was motivated by Text2Story's effectiveness in capturing morphosyntactic and grammatical phenomena, in contrast to its limitations in representing specialized clinical knowledge. On the other hand, while i2b2 includes medical domain categories, it was found to be overly broad and insufficiently granular, reducing the accuracy of clinical annotation.

[1] https://github.com/analuisacardosofernandes/Can-ISO-24617-1-go-clinical-

[2] The detailed methodology of the design and validation of the extension is described in (Fernandes et al., 2025)

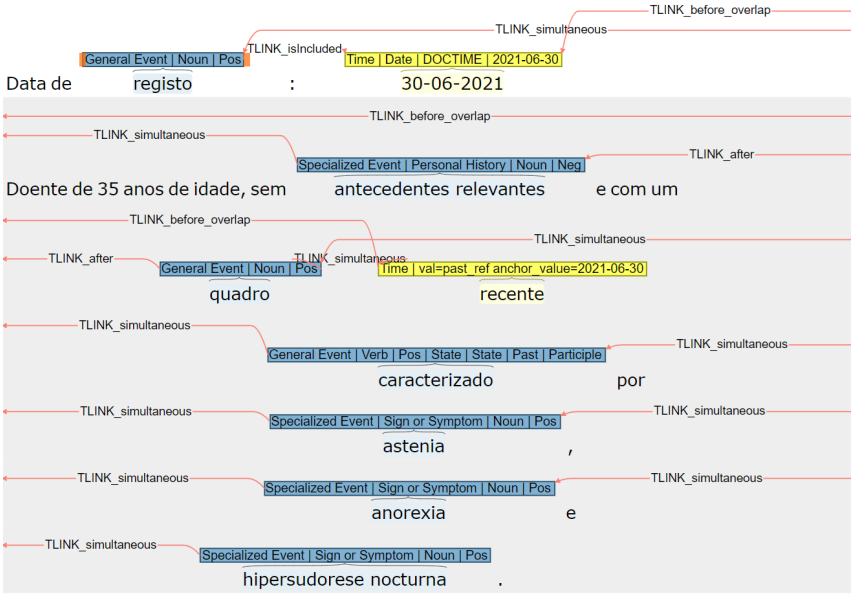

Figure 1: Annotation of an excerpt from a medical report using the Med2Story scheme. Events are marked in blue and temporal expressions in yellow. The annotated excerpt illustrates the identification of various attributes associated with both events and temporal expressions, as well as the temporal relations between events, between events and temporal expressions and between temporal expressions. "Registration date: 30/06/2021. The patient is a 35-year-old with no relevant medical history, presenting with recent symptoms of asthenia, anorexia, and night sweats".

Furthermore, i2b2 lacks systematic support for annotating morphosyntactic and semantic linguistic features, making it less suitable for deeper linguistic analyses.

The first development step consisted in defining a set of labels that could rigorously capture relevant clinical information. A fundamental distinction was introduced between two types of events: *general event* and *specialized event*. Events classified as general event retained the original Text2Story attributes — namely, *class*, *type*, *tense*, *aspect*, *polarity*, *vform*, *modality*, and *POS* — suitable for linguistic description of clinical narratives. Specialized events incorporated medical domain-specific attributes, allowing for a more detailed and meaningful representation of clinical content. All events were annotated as general events, and only those conveying clinical content were additionally annotated as specialized events.

The selection of clinical labels was conducted in collaboration with a hematologist from IPO-Porto, who participated in validating the clinically relevant categories. This phase was based on the analysis of a corpus of 40 pseudonymized medical reports from patients diagnosed with AML, including discharge summaries, general reports, and consultation notes.

As for nominal events, only the *POS* and *polarity* attributes were annotated under the general event layer, since current literature does not yet offer viable solutions for aspectual annotation of nouns.

The definition of medical domain labels was guided by the principles of the UMLS Metathesaurus ontology (Bodenreider, 2004), widely recognized as a systematic reference for organizing biomedical terminology. Additionally, contributions from Leite (2024), whose research on the same corpus proposed a preliminary set of clinically validated categories, were considered. Some of these categories were retained, while others were adapted or refined to align with the goals of this scheme. The final set of medical categories included: *Sign or Symptom*, *Personal History* (with subcategories: *Past Medical History*, *Comorbidity*, and *Undefined*), *Intercurrence*, *Examination*, *Examination Result*, *Principal Diagnosis*, *Characterization of the Disease*, *Medical Procedure*, *Treatment*, *Drug Administration Route*, and *Treatment Response*. These categories addressed two key gaps in previous schemes: the lack of clinical categories in ISO-based Text2Story and the excessive generality of i2b2, as exemplified by the use of the generic label *test* to annotate both examinations and their results.

Regarding annotation scope, only occurrences, states, or circumstances with temporal relevance were annotated as events, following the approach of Text2Story and ISO 24617-1. Entities such as medications, organs, institutions, or healthcare services were not annotated at the event layer, but instead in the referential layer, as participants in events.

Two modifications were introduced to the grammatical attributes: *relative pronoun* under *POS*; and the extension of negative polarity to include cases of implicit polarity.

As for temporal expressions, their annotation was not addressed in depth due to its complexity and the need for a more robust framework based on ISO 24617-1. Instead, Text2Story's guidelines were followed, with the addition of two specific attributes: *Admission Time* (date of patient admission), and *Discharge Time* (date of hospital discharge).

The attribute *DOCTIME* (report creation date) was retained. When discharge date and report creation date coincide, only the *Discharge Time* label should be used.

Finally, regarding TLINK, we followed the guidelines of the Text2Story annotation scheme. TLINK are established between events, between events and temporal expressions, and between temporal expressions. Their annotation proceeds from the last event in the linear order of discourse to the first, thereby ensuring relational consistency across annotations. This rule applies only in cases where transitivity can be verified. In general, transitivity is preserved, which enables the temporal localization of all events. Moreover, a direct link is established between each event and the temporal expression that situates it within the discourse timeline. The i2b2 approach, which systematically links all events with anchors such as *DOCTIME*, *Admission Time*, or *Discharge Time* was not adopted, as this practice proved redundant and of limited informational value. Instead, it is proposed that only events with no explicit or definite temporal relation be linked to those anchors.

Additionally, the TLINK *after_overlap* and their mirror *before_overlap* were introduced, ensuring that links are consistently established left-to-right, or from the event to the temporal expression, in alignment with the other TLINK.

Figure 1 shows an annotation example using the Med2Story scheme. The complete scheme, guidelines, decision tree, and Appendices are available in the associated GitHub repository.

5 Conclusion

In this study, we set out to compare the performance of a general-purpose annotation scheme — Text2Story, based on ISO standard 24617-1 — with that of a domain-specific scheme, i2b2, in the context of annotating clinical narratives.

The results show that the Text2Story annotation scheme is applicable to this type of text. However, it proves to be insufficiently informative with respect to domain-specific medical categories, highlighting the need to create new specialized tags. On the other hand, its capacity to represent morphosyntactic and semantic information is notably robust.

As for the i2b2 scheme, although it enables the annotation of medical information, its tags are overly broad, and it provides limited detail at both the morphosyntactic and semantic levels. Given these limitations, we developed a medical-specific extension of the ISO 24617-1 scheme, called Med2Story, designed to meet the requirements of the clinical domain.

From a conceptual and structural perspective, the ISO standard is robust and comprehensive, allowing for proper integration of domain-specific aspects related to the medical field. Although certain tags and attributes could be refined to represent information more precisely — such as the introduction of the TLINK after_overlap/before_overlap — ISO 24617-1 is designed to accommodate extensions for more specialized information, namely through the inclusion of tags derived from the medical ontology.

In future research, the proposed scheme will be applied to a set of medical reports in European Portuguese. Building on this, we intend to create a parallel dataset by translating these reports into other languages, with the aim of evaluating the applicability and robustness of the approach across different linguistic contexts. Furthermore, we propose extending Med2Story with additional annotation layers, particularly referential annotation, to enhance its descriptive and analytical scope.

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F. Styler IV, Colin Warner, Jena D. Hwang, Jinho D. Choi, Dmitriy Dligach, Rodney D. Nielsen,

and James Martin. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930. Open Access.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop (LAW VII)*, pages 178–186, Sofia, Bulgaria.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. Developing a large semantically annotated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3196–3200, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Johan Bos, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. 2017. *The Groningen Meaning Bank*.

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(2):571–601.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Rb-Silva, Luís Filipe Cunha, and Alípio Jorge. 2025. Enhancing an annotation scheme for clinical narratives in portuguese through human variation analysis. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX)*, Vienna, Austria.

A. González-Moreno, A. Ramos-González, I. González-Carrasco, et al. 2025. A clinical narrative corpus on nut allergy: annotation schema, guidelines and use case. *Scientific Data*, 12:173.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2012. Ontonotes: A large training corpus for enhanced processing. In *Handbook of Linguistic Annotation*.

Nancy Ide and Laurent Romary. 2006. Representing linguistic corpora and their annotations. In *LREC*.

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

ISO TC37/SC4. 2012. Language resource management—semantic annotation framework (semaf). International Organization for Standardization.

Van Gysel Jens E. L., Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Kunstliche Intelligenz*, 35:343–360.

Ishan Jindal, Alexandre Rademaker, Michał Ulewicz, Linh Ha, Huyen Nguyen, Khoi-Nguyen Tran, Huaiyu Zhu, and Yunyao Li. 2022. Universal proposition bank 2.0. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC)*, pages 1700–1711, Marseille, France.

A. Leal, P. Silvano, E. Amorim, I. Cantante, F. Silva, A. Jorge, and R. Campos. 2022. The place of isospace 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.

Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. 2019. Contributions to clinical named entity recognition in Portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, Florence, Italy. Association for Computational Linguistics.

William C. Mann and Sandra A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Marie-Catherine Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308.

Joakim Nivre. 2016. Universal dependencies: A cross-linguistic perspective on grammar and lexicon. In *Proceedings of the Workshop on Grammar and Lexicon (GramLex)*, pages 38–40, Osaka, Japan.

L. E. S. 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.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Rashmi Prasad, Bonnie Webber, and Alan Lee. 2018. Discourse annotation in the PDTB: The next generation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 87–97, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Pustejovsky, José Castaño, Robert Ingria, and Roser Saurí. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *New Directions in Question Answering*, volume 3, pages 28–34.

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

R. Rb-Silva and Y. Karimova. 2021. aMILE: Application of text mining to clinical reports of patients with acute myeloid leukemia.

A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, and I. Roberts. 2008. Semantic annotation of clinical text: The clef corpus. In *Proceedings of Building and Evaluating Resources for Biomedical Text Meaning: Workshop at LREC*.

Kirk Roberts, Dina Demner-Fushman, Joseph M Tonning, and Graciela Gonzalez. 2021. Annotated clinical text corpora: A systematic review. *Journal of the American Medical Informatics Association*, 28(9):1931–1941.

Purificação Silvano, António Leal, Fátima Silva, Inês Cantante, Fatima Oliveira, and Alípio Mario 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.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of EACL 2012 Demonstrations*, pages 102–107.

Weiyi Sun, Anna Rumshisky, and Özlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46:S5–S12.

Weiyi Sun, Anna Rumshisky, Ozlem Uzuner, Peter Szolovits, and James Pustejovsky. 2012. *2012 i2b2 Clinical Temporal Relations Challenge Annotation Guidelines*. i2b2 National Center for Biomedical Computing. Adapted from the THYME project guidelines by Will Styler, Guergana Savova, Martha Palmer, and James Pustejovsky.

Özlem Uzuner, Brett R South, Sheng Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Veronika Vincze, György Szarvas, Richárd Farkas, György Mora, and József Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(S11):S9.

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 Frequency of Annotation by Scheme

Table 1: Quantification of Events and Temporal Links – i2b2 Annotation Scheme

Class	
Events	188
Occurence	17
Clinical_Dept	9
Problem	63
Test	63
Evidential	8
Treatment	28
Polarity	
Positive	183
Negative	5
TLINK	
Overlap	131
Before_Overlap	129
Simultaneous	908
Before	48
After	149
Begun_By	46
Ended_By	7

Table 2: Quantitative analysis of Events and Temporal Links – Text2Story Annotation Scheme

Class	
Events	323
Occurence	110
State	110
Reporting	4
I_Action	10
POS	
Noun	198
Verb	88
Adjective	23
Noun	198
Polarity	
Positive	314
Negative	9
Event_Type	
Transition	71
Process	7
State	32
Tense	
Past	34
Imperfect	6
Present	22
Aspect	
Perfective	28
Imperfective	7
Vform	
Participle	19
Infinitive	7
Gerundive	9
Mood	
Conditional	2
Subjunctive	1
Modality	
Poder	2
TLINK	
Includes	52
Is_Included	231
Identity	162
Simultaneous	1137
Before	74
After	167
Begun_By	15
Ended_By	7