Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)

including of a contribution to
the Second Workshop on Multimodal Semantic Representation (MMSR II)

Workshop at IWCS 2025, September 24, Düsseldorf, Germany

Harry Bunt, editor

Proceedings of the 21st Joint ACL - ISO Workshop on
Interoperable Semantic Annotation (ISA-21)
including a contribution of the Second Workshop on Multimodal Semantic Representation (MSR II)

Workshop at IWCS 2025
Düsseldorf (Germany), September 24, 2025

Department of Cognitive Science and Artificial Intelligence
School of Humanities and Digital Sciences
Tilburg University, The Netherlands

ISBN: 978-90-74029-40-7

Table of Contents

ISA-21 Organising Committee

Harry Bunt (chair)	Tilburg University
Nancy Ide	Vassar College, Poughkeepsie, NY
Kiyong Lee	Korea University, Seoul
Volha Petukhova	Saarland University, Saarbrücken
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin

ISA-21 Programme Committee

Johan Bos	University of Groningen
Harry Bunt	Tilburg University
Stergios Chatzykiriakidis	University of Crete
Jae-Woong Choe	Korea University, Seoul
Robin Cooper	Gothenburg University
Rodolfo Delmonte	Università Ca' Foscari, Venezia
David DeVault	USC Institute for Creative Technologies, Playa Vista
Simon Dobnik	University of Gothenburg
Jens Edlund	KTH, Stockholm University
Alex Chengyu Fang	City University Hong Kong
Robert Gaizauskas	University of Sheffield
Koiti Hasida	Tokyo University
Elisabetta Jezek	Università degli Studi di Pavia
Nikhil Krsishnaswamy	University of Colorado, Boulder, Col.
Kiyong Lee	University of Korea, Seoul
Philippe Muller	IRIT, Université Paul Sabatier, Toulouse
Catherine Pelachaud	Université de Paris
Volha Petukhova	Universität des Saarlandes, Saarbrücken
Massimo Poesio	Queen Mary University, London
Laurent Prévot	Aix-Marseille University
Stephen Pulman	Apple Research UK
James Pustejovsky	Brandeis University, Waltham, MA
Laurent Romary	INRIA/Humboldt Universität Berlin
Manfred Stede	University of Potsdam
Matthew Stone	Rutgers University
Thorsten Trippel	University of Tübingen
Carl Vogel	Trinity College Dublin
Menno van Zaanen	North West University South Africa, Mahikeng
Annie Zaenen	Stanford University, Palo Alto
Heike Zinsmeister	Universität Hamburg

Preface

Welcome to the proceedings of the twenty-first edition of the series of joint ACL-ISO workshops on interoperable semantic annotation (ISA-21), this year organized as part of the International Conference on Computational Semantics (IWCS) 2025. We are thankful to the IWCS 2025 organizers for hosting this workshop. The submitted papers that were accepted for presentation at the workshop have been arranged in these proceedings in alphabetical order of the names of their first authors.

The Second Workshop on Multimodal Semantic Representations (MMSR II) was originally scheduled as another IWCS 2025 workshop, on the same day. Since this workshop featured only one accepted submission (plus two invited talks), and overlaps in content with ISA-21, it was decided to merge the MMSR II workshop into the ISA-21 workshop. The accepted paper of the MMSR II workshop is placed at the end of these proceedings.

We thank the members of the ISA-21 and MMSR II program committees for reviewing the submitted papers timely, and we thank the authors of accepted papers for revising their contributions timely taking the review comments into account.

Thank you!

The ISA-21 organisers,

Harry Bunt, Nancy Ide, Kiyong Lee, Volha Petukhova, James Pustejovsky, and Laurent Romary

Engagement and Non-Engagement: Two Notions at the Core of an Annotation Schema of Enunciative Strategies

Cyril Bruneau
Laboratoire MoDyCo
Université Paris Nanterre
c.bruneau@parisnanterre.fr

Delphine Battistelli
Laboratoire MoDyCo
Université Paris Nanterre
dbattist@parisnanterre.fr

Abstract

This study provides an annotation schema of a wide range of enunciative strategies underlying every enunciation process by which an enunciator actualizes a predicative content. We show that most of these enunciative strategies involve the enunciator in a relationship of *Engagement* (concerned with the notions of truth value and axiological/appreciative value) or *Non-Engagement* toward a stated predicative content. Our approach takes place in the French enunciative framework rooted in the work of Bally (1932). We explicitly compare our approach with that of Appraisal theory (Martin and White, 2003). We also illustrate the applications of our schema with a manual annotation experiment conducted on a corpus of French history textbooks. This experiment reveals interesting diachronic variations in the enunciator's modes of *Engagement* and *Non-Engagement*.

1 Introduction

In the lineage of what is called the enunciative approach of language notably rooted in the work of Bally (1932), we are interested in the various strategies of actualization of a predicative content that are mobilized by an enunciator when producing an utterance. A predicative content such as *be blue(the car)*, comprising a predicate (*be blue*) and its argument (*the car*), can be mobilized in an utterance in multiple ways (see ex. (1a) to (1g)), suggesting different actualizations and attitudes with regard to the same predicative content.

(1a) The car is blue
(1b) The car might be blue
(1c) I'm glad the car is blue
(1d) The car should not be blue
(1e) Is the car blue?
(1f) Do you know that the car is blue?
(1g) I heard the car is blue

The characterization of these enunciative operations is not addressed in NLP as a standalone problem, although several tasks make use of notions which are part of it, such as modality. We propose an operational schema for NLP of the full range of enunciative operations. Our schema is based on (Desclés, 2009), with several additions which are made explicit in Section 3. The characterization of these enunciative operations provides an analytical framework for observing and quantifying enunciative profiles across diverse corpora. Furthermore, enriching a corpus with such semantic and enunciative information can prove crucial for automatic detection tasks in NLP related to the enunciator's engagement, such as hate/toxic speech detection, opinion mining, ideological content analysis, uncertainty analysis, and more. After clarifying our theoretical framework in Section 2, we present a comprehensive annotation schema for enunciative operations in Section 3. Section 4 details the results of a manual annotation experiment conducted on a corpus using this schema.

2 The notion of engagement: a recurring issue in literature

2.1 Related works

The question of how an enunciator validates a predicative content and/or positions in relation to it has been approached in various ways, directly or not concerned with the notion of engagement. Works on **commitment** (see De Brabanter and Dendale (2008) for an extended presentation) examine the beliefs of a speaker that can be inferred from their discourse, aiming to determine whether an event (term covering a conceptual notion close to that of predicative content) is presented as *actual*, *non-actual*, or *uncertain*. These studies suggest that such inferences are not limited to assertions, the primary mode of commitment (De Marneffe et al.,

2019). For example, the utterance (1e), formulated as a question, leaves room for doubt regarding the truth value of the predicative content *be blue (the car)*, unlike example (1f), which — though also phrased as a question — demonstrates speaker commitment inferable from the verb "know" (Jiang and de Marneffe, 2019).

The enunciator's commitment can also be linked to the notion of **factuality** when it is related to the truth value of the predicative content (i.e. when presented as factual). In fact, Jiang and de Marneffe (2019) use commitment and factuality interchangeably. Works on factuality e.g. (Saurí and Pustejovsky, 2009) propose annotation schemas covering (i) the enunciator's certainty regarding an event's truth, (ii) its possibility (as in ex. (1b)), or (iii) its probability.

Other related notions are **modality** (see for example in NLP (Pyatkin et al., 2021)) and **evidentiality** (see for ex. (Su et al., 2010)). They seek to capture respectively the attitude of the enunciator toward their content, and the nature of the source of the information. Modality may convey judgments of the enunciator regarding for instance the uncertainty of the predicative content (as in (1b)), while evidentiality may distantiate the enunciator from the information by highlighting its source (1g). These notions thus play a role regarding the enunciator's commitment we can infer.

The notions of **stance** and **sentiment** as described in NLP might also be intuitively linked to this question. Stance detection aims at assessing the enunciator's favorability toward a predetermined target (*in favor* or *against* the target which is not necessarily mentioned in the predicative content) (Mohammad et al., 2016). More recent works rely on the SDQC tagset (*Support, Deny, Query, Comment*) (Gorrell et al., 2019; Evrard et al., 2020). Similarly, sentiment analysis aims to detect the polarity (*positive* as in (1c), *negative* as in (1d), *neutral*) of the enunciator's opinion toward an explicitly mentioned target (e.g., a person, an organization), sometimes including the identification of the target (Nakov et al., 2016).

Closer to our approach is the **Appraisal theory** (Martin and White, 2003), which meticulously classifies evaluative language into categories such as *Attitude*, *Engagement* and *Graduation*. This theory is *"concerned with the interpersonal in language, with the subjective presence of writers/speakers in texts as they adopt stances towards both the material they present and those with whom they communicate."*(ibid., p.1). The Appraisal theory models the evaluative operations available to an enunciator. These operations are not limited to the truth value of the statement but also refer to axiological and appreciative dimensions. We provide an overview of the Appraisal framework in Figure 5 in the appendix, to facilitate comparison with our approach.

2.2 Contributions and Distinctions

In this study, we provide a new typology for NLP of the enunciative operations an enunciator may deploy in order to actualize their predicative content into an utterance, following the conceptualization initiated in (Desclés, 2009), himself in the lineage of Bally's works. Desclés (2009) designates this set of operations by the term *"prise en charge"* (De Brabanter and Dendale, 2008; Coltier et al., 2009). At the core of these enunciator's operations lies the notion of **Engagement**, defined as the enunciator's capacity to engage or disengage with either the truth value or the axiological/appreciative value of the predicative content they articulate. What we define as engagement is close to the aforementioned notion of speaker commitment, except that its scope is not limited to (non)factual events. Our notion of engagement also coexists with the opposing notion of **Disengagement**, a common operation that enables the communication of uncertain or distanced information. The notion of **Non-Engagement** (a default category implying neither engagement nor disengagement) is also central to this schema. It captures the enunciator's apparent neutrality. The central articulation of these three notions enables contrasted analyses of the enunciative strategies in diverse corpora, as presented in Section 4, and constitutes an original contribution for NLP. Moreover, our definition of the enunciator's engagement relies on explicit linguistic markers. Thus, unlike works on commitment, the simple declaration of an event (presented as factual) without explicit markers of engagement (assertion, certainty...) will be classified as *Non-Engagement*.

Our approach also differs from works on stance and sentiment as we do not seek to capture the polarity of the utterance, nor do we focus on an explicit target within (or outside) the predicative content. Consequently, a strongly favorable or unfavorable opinion would be treated as strong engage-

ment, with no consideration for polarity or target, while a weakly favorable opinion accompanied by uncertainty might indicate disengagement. Stance detection and sentiment analysis are therefore very different tasks from ours, but they can be seen as complementary in certain use cases.

We share certain notions presented in Appraisal theory, which relate to the annotation schema we propose. The category *Attitude* partially aligns with our *Appreciative modality* category, as explained in section 3.2. The *Engagement* category, which aims to describe the enunciator's stance towards the positions referenced in the discourse, encompasses notions related to both engagement and potential disengagement, which are described in our approach under the categories of *Validation*, *Modality*, and *Representation of speech* (see section 3.2). However, significant differences can be outlined between Appraisal theory and our approach:

1. We do not consider the notions captured by the *Graduation* category in Appraisal theory as modifiers, but as indicators that characterize the categories of engagement we describe, particularly by clarifying the concepts of certainty/uncertainty and negotiability/non-negotiability.

2. Our conception of engagement only focuses on the predicative content expressed by the enunciator. Consequently, we do not consider the evaluation that is made of an interlocutor, or the enunciator of a discourse to which the primary enunciator refers.

3. Unlike stance detection approaches, we do not assume that an enunciator's engagement is necessarily in favor of or against a specific target. Additionally, we do not incorporate the notion of engagement *Polarity* into our schema.

4. As a result of 3, we propose the central concept of *Non-Engagement*, which describes a form of neutrality adopted by the enunciator. This concept enables us to differentiate between utterances presented as negotiable and other types of utterances that exhibit explicit engagement or disengagement. We believe this notion of Non-Engagement is crucial for analyzing certain corpora, as demonstrated in Section 4.

Modality and *Evidentiality* find their place in our typology among operations involving enunciator engagement / disengagement (with evidentiality overlapping both *Epistemic modality* and *Plausibility*, as detailed in the description of the *Plausibility* category) (see Section 3.2).

3 Annotation schema

3.1 Global view

Figure 1 represents the enunciative operations which can be mobilized by an enunciator. This typology is inspired by the one proposed in (Desclés, 2009), which has since been revised: (i) We added the *Representation of speech* categories as they are presented in (Authier-Revuz, 2020) in order to refine the category formerly referred to as *Reported Enunciation*; (ii) we added the notion of *Deictic Anchoring*; (iii) we added the *Appreciative Modality* present in (Desclés, 2009) but not in the final typology, as we do not want to limit the analysis to the truth value of the predicative content; and (iv) we changed the layout in order to make clear which categories fall under the scope of *Engagement/Disengagement*.

The primary enunciator (i.e. the initiator of the utterance act) is denoted as "E" on the left side of the schema. Each category represents an operation through which the enunciator actualizes a predicative content (that is, transforms a raw predicative content into a concrete utterance: see examples (1a) to (1g)). This actualization may entail a degree of validation of the content, which is precisely what this typology seeks to capture. Four ways - not necessarily exclusive - of actualizing a predicative content are distinguished: on the one hand, *Non-Engagement* and *Engagement / Disengagement* which are mutually exclusive; on the other hand, *Contextual frame of reference* and *Deictic anchoring* which can be added to one of the previous ones. These operations are analyzed at the clause level (independent clauses, which may be juxtaposed or coordinated). Each category can be associated to a clause within the utterance, and multiple categories may describe a single clause, except for a few categories excluding each other, which are presented in Section 3.2. The resulting manual annotations can thus be used to train a multilabel text classifier.

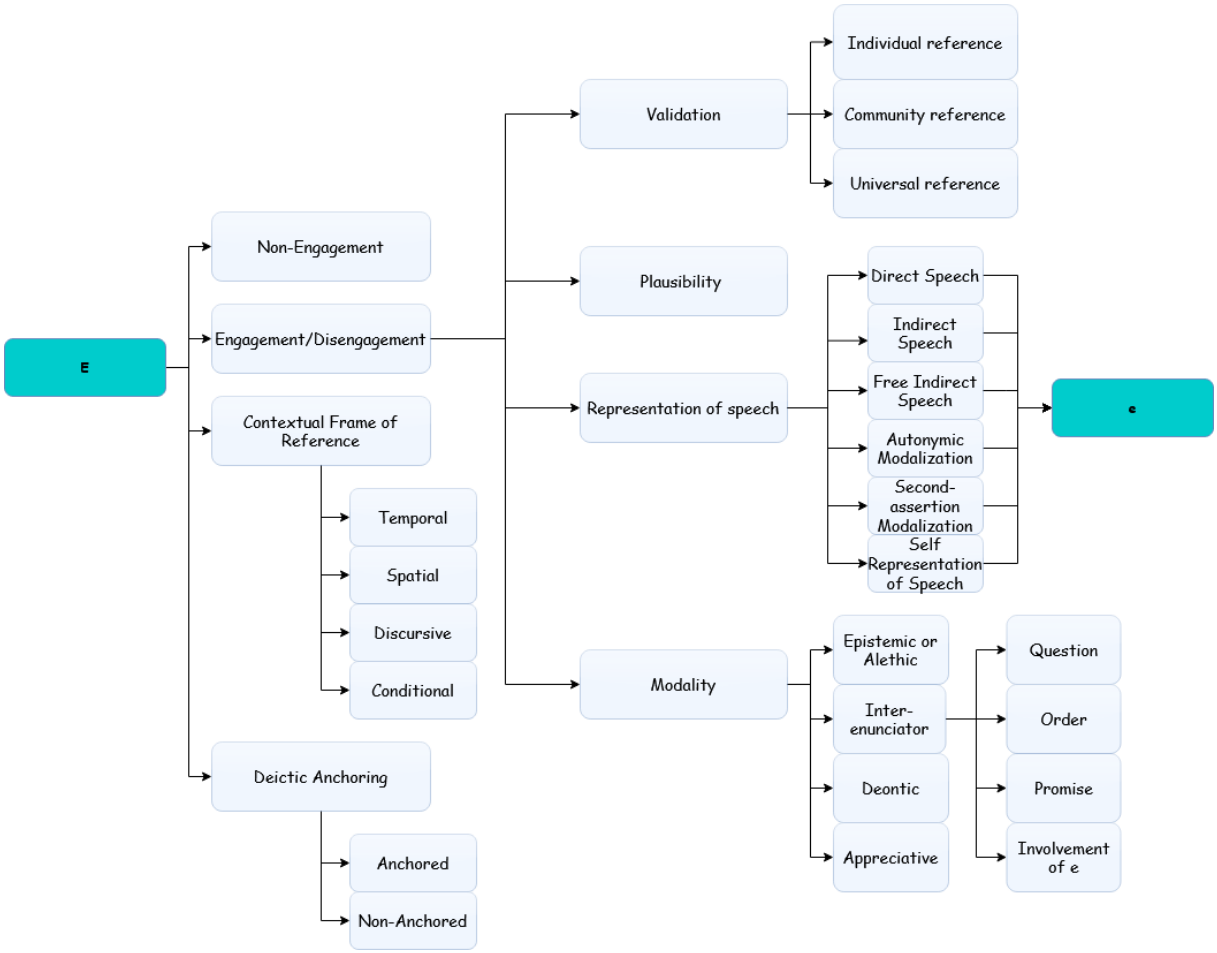


Figure 1: Annotation schema for enunciative operations[1]

3.2 Detailed description of the categories

We now detail the categories of our enunciative operations typology, with examples drawn from a corpus of French textbooks, which we present in Section 4.1. The categories under *Engagement / Disengagement* constitute the main part of the schema.

- The *Non-Engagement* category corresponds to the default mode of utterance: the enunciator does not express any particular engagement or disengagement, and their statement is perceived as *"negotiable"* with the interlocutor (Desclés, 2009) (see example (1a)). This category is often identified by the absence of markers from other categories. It is described in opposition to (and is thus incompatible with) any category classified under *Engagement / Disengagement*.

- *Contextual frame of reference* describes a type of utterance in which the truth value of the enunciator's predicative content is conditioned by an explicit context, which may be *Discursive*, *Conditional*, *Spatial*, or *Temporal*. These contextual frames of reference often serve to clarify the scope of the predicative content and are not considered markers of the enunciator's engagement or disengagement. Example (2) illustrates a temporal context via the adverbial *"from the late 15th century onward"*, example (3) a spatial context via *"in the capital"*.

  (2) Comment l'Europe s'ouvre-t-elle sur le monde à partir de la fin du XVe siècle ? / How did Europe open itself to the world from the late 15th century onward?

  (3) et commença dans la capitale la plus épouvantable tuerie dont nos annales fassent mention. / and began in the capital the most terrible slaughter recorded in our annals.

[1] Arrows between categories indicate a subdivision.

- The *Deictic anchoring* category is divided into two subcategories: *Anchored* and *Non-anchored*. It characterizes how the predicative content is related to the enunciative situation. *Anchored* clauses contain markers of the enunciator's presence and/or anchors that indicate a direct relationship with the enunciative context, such as the pronoun *"our"* in example (3). These markers do not signal the enunciator's engagement or disengagement.

- The *Validation* of the enunciator's predicative content is related to statements formulated as assertions: the enunciator presents their predicative content as *"non-negotiable"*. These utterances strongly reflect the enunciator's engagement regarding the truth value of the content. Three types of validation may be distinguished, depending on their referential framework - that is, the origin of the assertion -: the enunciator as an individual (*Individual*) (example (4) due to the adjective *"true"*), the enunciator as part of a broader community (*Community*) (example (3) where *"our"* stands for the French community), or a universal idea adopted by the enunciator (*Universal*, example (5)). Validation characterizes operations identified within the Appraisal framework as *monoglossic* engagement, or *heteroglossic* when associated with the notion of dialogic contraction, which can be linked to the criterion of non-negotiability.

  (4) Ils substituèrent le culte de la déesse Raison à celui du vrai Dieu / They replaced the worship of the true God with that of the goddess Reason.

  (5) L'envie s'attache toujours aux grands talents. / Envy always clings to great talents.

- *Plausibility* describes a mediated type of utterance in which the enunciator formulates a plausible hypothesis triggered by observed evidence (which is shared by co-enunciators) and/or by an inference from shared knowledge (example (6)). The inferential and hypothetical nature of the predicative content leads to a partial disengagement of the enunciator regarding the truth value: the statement is presented as *"negotiable"* (4). *Plausibility* only relies on abductive inference, which should not be confused with deductive inference (Desclés and Guentchéva, 2024). The former can be considered as a specific case of mediativity: it aims to infer the cause of an objectively observed situation. The latter seeks to deduce a consequence from an observed situation and corresponds, from an enunciative perspective, to what we describe under the category *Epistemic modality*.

  (6) Le siège de la Rochelle, où périrent plus de quarante mille catholiques, fut une preuve que le parti calviniste n'avait rien perdu de sa puissance. / The siege of La Rochelle, where over forty thousand Catholics perished, proved that the Calvinist faction had lost none of its power.

- The macro-category *Modality* encompasses all the attitudes an enunciator may adopt toward their predicative content. Except for the *Appreciative* and *Deontic* subcategories, *Modality* encompasses evaluations categorized within the Appraisal framework under *Expand*, which reflects a dialogic expansion. *Epistemic (or Alethic)* presents a predicative content within a framework of uncertainty. It constitutes a partial disengagement by the enunciator (ex. (7), (1b)). *Inter-enunciator* modality applies to *Questions*, *Orders*, and *Promises*, which require linking predicative content to another enunciator in order to acquire a truth value, as well as the involvement of the co-enunciator in the statement. Example (2) also illustrates a question. These two types of modality reflect the enunciator's partial disengagement regarding the truth value of the clause. However, promises may convey a judgment or intention of the enunciator on an axiological level, thus indicating engagement on the axiological value. *Deontic modality* characterizes a judgment expressed by the enunciator based on external codes and rules, such as institutional norms. It reflects the enunciator's engagement on the axiological value while simultaneously disengaging from the truth, as the predicative content does not receive a truth value (8). *Appreciative modality* (which includes bouletic, axiological, and appreciative dimensions) expresses an individual judgment by the enunciator, as well as their engagement on the axiological value, as in (3), (11). This category can be related to the *Attitude* defined in the Appraisal framework, whether it concerns affects, judg-

ments, or appreciations. These modalities are strong evidence of the enunciator's processes of engagement or disengagement.

(7) Il aurait dit à cette occasion "Paris vaut bien une messe ". / He reportedly said on that occasion "Paris is well worth a mass".

(8) Le sanctuaire de l' école doit être préservé des passions intéressées et des luttes stériles des partis. / The sanctuary of the school must be preserved from self-interested passions and the sterile struggles of partisan factions.

- The *Representation of speech* category encompasses all forms of reported speech, as well as *Self-Representation of speech* as developed by Authier-Revuz (2020). The six proposed categories of reported speech are distinguished by the type of enunciative anchoring employed by the primary and secondary enunciators ("unified" or "dissociated" anchoring, ibid.) and by the status of the reported speech (as the "source of speech" or as the "object of speech", ibid.). The six categories are as follows: *Direct speech*, *Indirect speech*, *Free indirect speech*, *Autonymic Modalization* (the act of borrowing lexical elements from another speech, as in example (9)), *Second-assertion modalization* (the act of reporting the source of an utterance as in (10)), and *Self-representation of speech* (the act of representing one's own speech in the process of being produced), as in (11) with *"I say it with regret"*. These six categories demonstrate the enunciator's engagement regarding the fact that another enunciator has uttered the reported speech. However, this initial engagement may be attenuated by a modalization of the reporting act, possibly leading to partial disengagement (see ex. (7)). The enunciator's engagement toward the predicative content of the reported speech is more difficult to evaluate. Although some categories of reported speech appear to favor a distancing between the primary enunciator and the reported speech (*Direct speech*, *Autonymic modalization*), particularly because they involve distinct enunciative anchoring, they do not constitute sufficient evidence to determine the primary enunciator's engagement. *Self-representation of speech* (11) is a special case, describing a mode of utterance in which the enunciator underlines their own speech, thereby confirming potential engagement or disengagement toward either value within the self-quoted speech. It is worth noting that a reflexive loop may emerge between the secondary enunciator ("e") and the primary ("E") at the origin of the utterance. Indeed, the reported speech may eventually be qualified by all the enunciative operations applicable to "E". As it describes various ways of representing another enunciator's speech, this category can be compared to the evaluations under the *Heteroglossic* category of Appraisal theory, although we do not adopt the distinction between *Contract* and *Expand* in the description of our subcategories, as our approach does not focus on the enunciator's stance.

(9) Pour construire la basilique Saint-Pierre-de-Rome, le pape vend des "indulgences". / To fund the construction of St. Peter's Basilica in Rome, the Pope sells "indulgences."

(10) Selon l'Église catholique, elles pardonnent les péchés. / According to the Catholic Church, they forgive sins.

(11) Et là, dans sa fureur, je le dis à regret, c'est le seul crime de ce héros, mais il est affreux, il fit massacrer 3000 personnes. / And there, in his fury, I say it with regret, for it is this hero's only crime, yet an awful one, he had 3,000 people slaughtered.

The schema incorporates a total of 24 distinct annotation labels. Except for a few incompatibilities between these labels (*Anchored* necessarily excludes *Non-anchored*, *Non-Engagement* excludes *Engagement / Disengagement*, and *Inter-enunciator modality* excludes Non-anchored) most of these enunciative mechanisms can apply additively to the same clause. An annotation guide (Bruneau and Battistelli, 2024) in French provides a more detailed description of the different categories, along with annotation examples and frequently encountered linguistic markers for each. In the following section, we present the results of a manual annotation of a corpus based on this schema, in order to illustrate the proportions of these enunciative categories that enunciators may employ in a specific context.

4 Exploring some manual annotations

4.1 Corpus description

In this section, we present the results of the manual annotation of a collected French corpus, comprising 858 clauses extracted from history lessons in eight textbooks. Four of them (referred to as "ancient") were published in the 19th century. They were referenced by the MoDoAp project[2], which draws on a larger corpus of digitized textbooks hosted by the French online library Gallica[3]. In addition, four textbooks published in the 21st century ("modern") were collected to support diachronic comparison in this study. The clauses are evenly distributed across 3 themes: (i) Christopher Columbus and the Age of Discovery, (ii) Martin Luther and the Reformation, and (iii) the French Revolution. The clauses composing this corpus are independent clauses (see Section 3.1). We designed a tool for the automatic segmentation of the corpus into independent clauses. Based on the syntactic analysis provided by Stanza (Qi et al., 2020), this tool implements a heuristic that consists in identifying verbs (or adjectives or nouns within verbal constructions) that have a syntactic dependency of the *conjunction* or *parataxis* type and are syntactically linked to a verbal (or adjectival or nominal within a verbal construction) root.

The complete annotated corpus is available online[4].

History textbooks constitute a specific textual genre, as the enunciator's engagement toward the truth and factuality of the described events is of particular interest, and as they are not axiologically neutral. Moreover, variations may exist among textbooks depending on factors such as educational level and the specific editorial competition of certain periods, which can be observed through the analysis of enunciative operations and the enunciator's engagement.

4.2 Annotation process

The annotation was conducted by two experts in enunciative linguistics. For each independent clause, a label (0 or 1) was assigned to each category in the annotation schema. Initially, a provisional annotation guide was provided to the annotators, detailing the different categories and relevant linguistic markers for French. A preliminary inter-annotator agreement score was computed on a larger corpus of 1000 clauses, to identify the categories leading to the most disagreement. A discussion between annotators was held to highlight sources of inconsistency, followed by a second round of annotation, leading to the inter-annotator agreement (IAA) results presented in Figure 2, with the categories related to *Engagement / Disengagement* highlighted in blue. Cohen's kappa (Cohen, 1960) was used to calculate agreement between the two annotators, applied in a binary manner for each category. During the preliminary annotation process, various difficulties emerged, leading to revisions of the annotation guidelines, including:

Category	IAA	Category	IAA
Non-Engagement	0.83	Appreciative	0.90
Val Individual	0.93	IE Question	1.0
Val Community	0.84	IE Order	1.0
Val Universal	1.0	IE Promise	0.0
Context Temporal	0.97	IE involv. e	0.97
Context Spatial	0.91	Plausibility	1.0
Context Conditional	1.0	Direct Speech	0.96
Context Discursive	1.0	Indirect Speech	0.88
Anchored	0.90	Free Indirect Speech	
Non-anchored	0.92	Second-assert. Modal	0.86
Epistemic	0.92	Autonymic Modal.	0.88
Deontic	0.96	Self repr. of Speech	1.0

Figure 2: IAA scores for each category

1. The need to distinguish between a spatio-temporal information contextually added to the clause (*Contextual frame of reference*) and that which is inherently part of the content. Spatio-temporal frames of reference were eventually considered as being exclusively adverbials, in opposition to spatio-temporal information playing the role of grammatical subjects or object complements (as the ones underlined in examples (12) and (13) respectively, which would not be annotated as *Contextual frame of reference*):

   (12) Rome est devenue la ville la plus puissante / Rome has become the most powerful city

   (13) Cette armée a conquis des régions de plus en plus lointaines / This army has conquered increasingly distant regions

[2] https://modoap.huma-num.fr
[3] https://gallica.bnf.fr
[4] https://github.com/CyrilBruneau/ISA-21

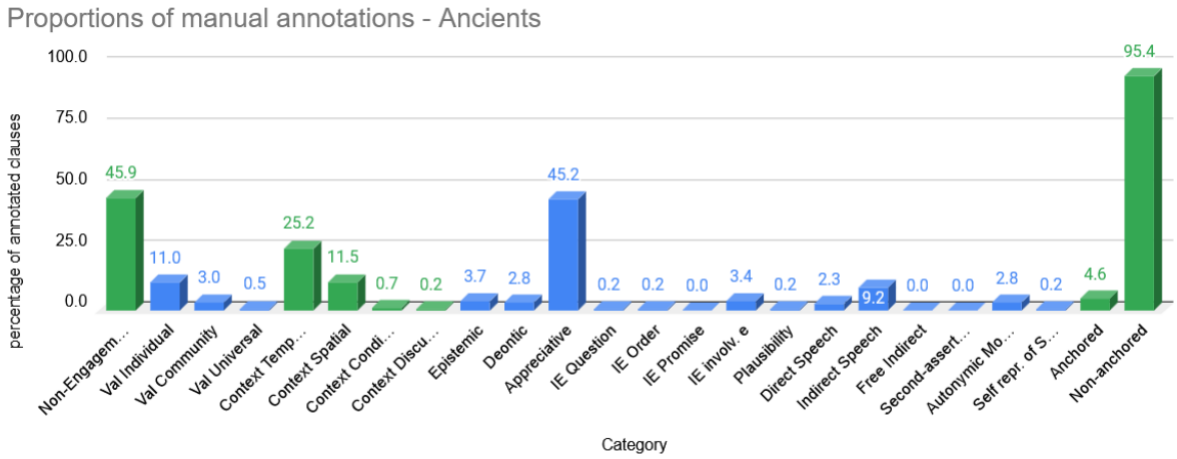


Figure 3: Proportions of the annotated categories - Ancient corpus

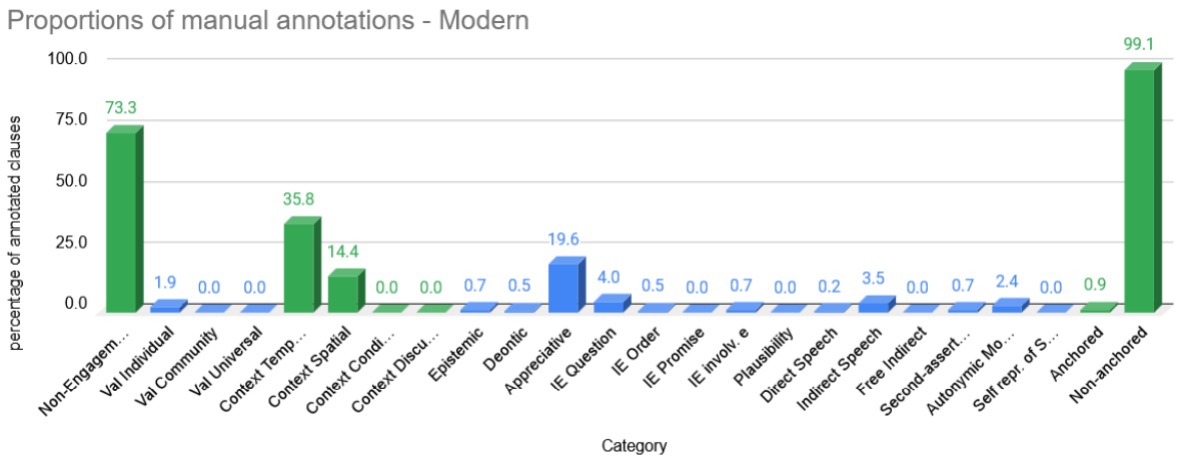


Figure 4: Proportions of the annotated categories - Modern corpus

2. The necessity of differentiating adjectives that trigger a subjective evaluation from the enunciator (*Appreciative modality*, see the ones underlined in example (14)) from more "objective" adjectives (15). To implement this distinction, we relied on the division between "classifying" and "non-classifying" adjectives established in (Maingueneau, 2020).

   (14) Louis XV fut un roi vicieux, égoïste / Louis XV was a vicious, selfish king

   (15) Les principales villes du royaume devinrent le théâtre de scènes analogues / The kingdom's major cities became the stage for similar scenes;

3. The wide variety of possible linguistic markers for *Appreciative modality*.

After the IAA calculations, a gold standard corpus was established at the intersection of the two sets of manual annotations provided by the annotators, comprising the clauses reflecting a full agreement. 142 clauses were removed from the 1000 initially annotated, leading to the 858-clause corpus we describe. We adopted this conservative gold standard, based on strict intersection, in order to ensure maximal reliability of the annotated corpus. This choice also preserves the original distribution of disagreements, which can be analyzed separately in future work to better understand the sources of annotation variability.

4.3 Annotation results

Figure 3 presents the proportions of manually annotated clauses for each category in the ancient textbooks subcorpus, while Figure 4 focuses on the modern textbooks. Blue categories are the ones under *Engagement / Disengagement* in the typology, in contrast with the green categories. Among the 24 categories shown in the annotation schema, only *Promises* and *Free indirect speech* have not been encountered in the overall corpus.

The two temporal series primarily differ in their proportions of *Non-Engagement* (73.3% for modern textbooks vs. 45.9% for ancient), which is explained by the greater presence of *Appreciative modality* (implying engagement) in ancient textbooks (45.2% compared to 19.6%). Individual enunciator engagement regarding truth-value (*Validation - Individual*) is significantly higher in ancient textbooks (11% of the clauses vs. 1.9% of the modern ones), although the majority of enunciator engagement across the entire corpus relies on *Appreciative modality*. *Non-anchored* clauses are strongly favored in this textual genre. *Spatial* and *Temporal* contexts are relatively frequent in both series, as expected in history lessons.

5 Conclusion

The annotation schema we propose offers a framework for capturing the wide range of enunciative strategies underlying every enunciation process by which an enunciator actualizes a predicative content. This schema focuses on two important global categories relevant for describing the ways an enunciator positions in relation to the truth value and the axiological value of a predicative content: *Engagement* - notion most frequently used in works that refer to Appraisal theory (e.g. (Zeng et al., 2024)) - and *Non-Engagement* - notably absent from Appraisal approach-. When a lot of *Non-Engagement* textual units are highlighted in a text (or in a corpus of texts), it underlines a phenomenon in which the enunciator "fades away". We illustrated this phenomenon here by comparing history lessons from two corpora. More generally, our frame of analysis allows for the description of enunciative profiles and compare them between diverse corpora. Future work may extend this approach to other genres and corpora, and leverage the schema's potential for training NLP models in tasks such as toxic speech detection, as in (Battistelli et al., 2023) and ideological content analysis.

References

Jacqueline Authier-Revuz. 2020. *La représentation du discours autre: principes pour une description*. De Gruyter.

Charles Bally. 1932. *Linguistique générale et Linguistique française, par Charles Bally*. Impr. des Presses universitaires le France.

Delphine Battistelli, Valentina Dragos, and Jade Mekki. 2023. Annotating social data with speaker/user engagement. illustration on online hate characterization in french. In *International Conference on Computing and Communication Networks*, pages 317–330. Springer.

Cyril Bruneau and Delphine Battistelli. 2024. Guide d'annotation manuelle de la prise en charge énonciative - Version V2 (Février 2024). Https://hal.science/hal-04541398v1/file/Guide$_{A}nnotation_{P}EC_{V}2_{0}2 - 2024.pdf$.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Danielle Coltier, Patrick Dendale, and Philippe De Brabanter. 2009. La notion de prise en charge: mise en perspective. *Langue française*, (2):3–27.

Philippe De Brabanter and Patrick Dendale. 2008. Commitment: The term and the notions. *Belgian Journal of Linguistics*, 22(1):1–14.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Jean-Pierre Desclés. 2009. Prise en charge, engagement et désengagement. *Langue française*, 162(2):29–53.

Jean-Pierre Desclés and Zlatka Guentchéva. 2024. Évidentialité, médiativité, modalité épistémique: une approche énonciative. In *SHS Web of Conferences*, volume 191.

Marc Evrard, Rémi Uro, Nicolas Hervé, and Béatrice Mazoyer. 2020. French tweet corpus for automatic stance detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6317–6322.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. Semeval-2019 task 7: Rumoureval 2019: Determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation: NAACL HLT 2019*, pages 845–854. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2019. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4208–4213.

Dominique Maingueneau. 2020. Chapitre 6. classifiance et non-classifiance. *Lettres Sup*, 2:107–132.

James R Martin and Peter R White. 2003. *The language of evaluation*, volume 2. Springer.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California. Association for Computational Linguistics.

Valentina Pyatkin, Shoval Sadde, Aynat Rubinstein, Paul Portner, and Reut Tsarfaty. 2021. The possible, the plausible, and the desirable: Event-based modality detection for language processing. *arXiv preprint arXiv:2106.08037*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Qi Su, Chu-Ren Huang, and Helen Kaiyun Chen. 2010. Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 10–17.

Jiamei Zeng, Min Dong, and Alex Chengyu Fang. 2024. Annotating evaluative language: Challenges and solutions in applying appraisal theory. In *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation@ LREC-COLING 2024*, pages 144–151.

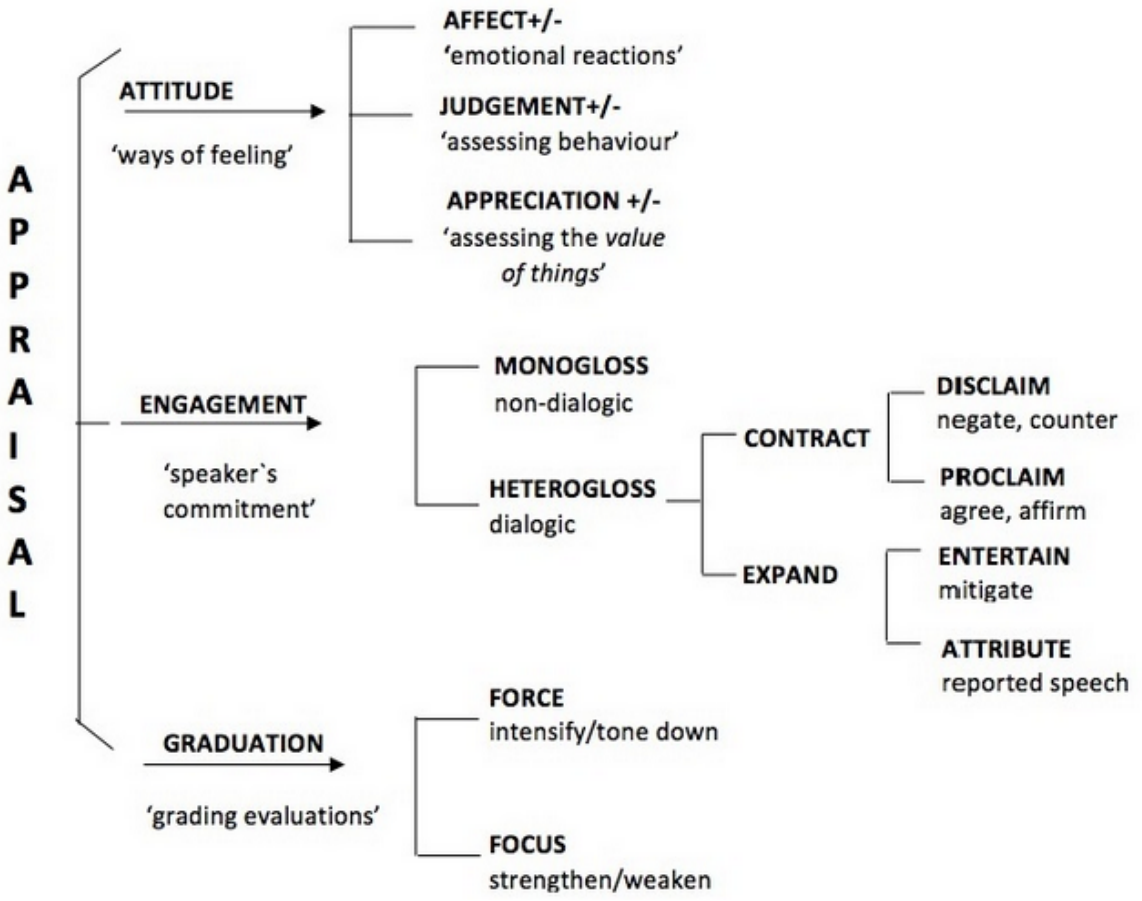


Figure 5: Appraisal Framework, adapted from (Martin and White, 2003, p. 38)

Revisiting the ISO-TimeML abstract syntax

Harry Bunt
Tilburg U., Netherlands
harry.bunt@uvt.nl

Alex Fang
City U. of Hong Kong
alex.fang@cityu.edu.hk

Kiyong Lee
Korea U. / Seoul
ikiyong@gmail.com

Volha Petukhova
Saarland U., Germany
V.V.Petukhova@gmail.com

Purificação Silvano
Univ. of Porto, Portugal
puri.msilvano@gmail.com

James Pustejovsky
Brandeis U., Waldham
jamesp@brandeis.edu

Abstract

This paper describes some of the ongoing work within the ISO preliminary work item PWI 254617-17, 'Interlinking of annotations'. This PWI investigates the possibilities and problems of combining annotations made with different annotation schemes. using the 'interlinking' approach (Bunt, 2024) applied to different parts of the multi-part standard ISO 24617, 'Semantic annotation framework'. This paper focuses on the combination of ISO-TimeML and QuantML at the level of abstract syntax. A new version is defined for the ISO-TimeML abstract syntax specification and how it relates to the concrete (XML-based) syntax as a basis for this combination. As a side-effect, some issues in the use of ISO-TimeML come to light that could be relevant for a possible future second edition of this standard.

1 Introduction

1.1 Background

Existing semantic annotation schemes are often focused on a specific type of semantic information, such as TimeML (Pustejovsky, 2003) on time and events, SpatialML (Mani et al., 2010) on spatial information, DAMSL (Allen & Core, 1997) and DIT++ (Bunt, 2007) on dialogue acts, and PDTB (Prasad et al, 2008; 2019) on discourse relations. The ISO Semantic Annotation Framework (ISO 24617, 'SemAF') was set up as a multi-part standard, with different parts focusing on different semantic domains.

Developing the SemAF standard as a set of separate sub-standards has proved useful, as it is better feasible to develop an annotation schema for a well-delineated semantic domain. The first two parts of SemAF, informally known as 'ISO-TimeML' and 'DiAML', are successful examples of the application of this approach, as the annotation of time and events is clearly separable from the annotation of dialogue acts. However, some of the semantic domains are not entirely disjoint; some semantic phenomena play a role in more than one sub-standard.

For example, the expression *"every Monday"* quantifies over mondays. Being a temporal expression, ISO-TimeML provides an annotation of this expression, including an indication of its quantifying character. ISO-TimeML has only a rudimentary treatment of quantification, however (Bunt & Pustejovsky, 2010), while it is the focus of SemAF part 12, QuantML This paper reports on activities within the ISO preliminary work item PWI 254617-17, Interlinking of annotations. This PWI investigates the possibilities and problems of combining annotations made with different annotation schemes, using the interlinking approach introduced in (Bunt, 2024). In particular this approach seems interesting for combining annotations made according to different parts of SemAF, which focus on different types of semantic information. On this approach, links are added between elements of different annotations for indicating that these elements correspond to the same entities mentioned in the primary data. This allows annotations of the same entities with different types of information, and therefore facilitates the merge of the semantic information in the respective annotations.

When considering the combination of annotations from different SemAF parts, we have to consider all three interrelated levels distinguished in the architecture of a SemAF scheme (see Fig. 1): (1) the concrete syntax, conventionally with an XML-based reference format, (2) the abstract syntax, expressing the semantically relevant information of the annotations in the form of pairs, triples, and other set-theoretical structures (and interrelated with the concrete syntax through encoding and decoding functions), and (3) the semantics of the annotations.

At the level of concrete syntax, interlinking consists of adding identity links between components of representations from different schemes, indicating that the same stretch of primary data is annotated

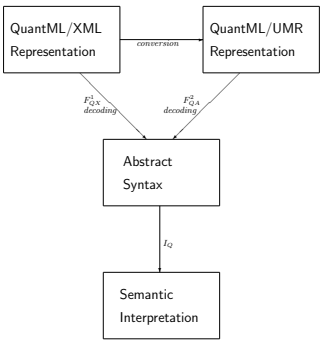


Figure 1: Levels and interrelations in SemAF annotation schemes.

from different points of view. At the level of abstract syntax, the structures of the interlinked annotations are combined into a single set-theoretical structure. At the level of semantics, finally, the semantic interpretation function describes the meaning of the joint abstract syntax expressions.

At the level of concrete syntax, the addition of identity links between two (or more) representations is a straightforward matter, although there may be some issues in the identification and use of markables, but the real challenges lie at the levels of abstract syntax and semantics. In particular, sitting in between the levels of concrete representation and semantics, the combination of annotations at the level of abstract faces a dual challenge.

On the one hand, the expressions at that level should have a systematic encoding-decoding relation to each of the respective concrete representations, and on the other hand they should capture the information contained in the combined annotations in a way that allows their joint semantic interpretation.

Since ISO-TimeML (ISO-24617-1:2012 Time and events) and QuantML (ISO 24617-12:2025 Quantification) are two of the best developed and most complex SemAF parts, a sensible strategy would seem to first explore the possibilities of combining their respective annotations, in particular at the level of abstract syntax. QuantML has a fully developed abstract syntax, but ISO-TimeML, being the oldest SemAF part, has an abstract syntax that is not fully specified and at some points lacks conceptual clarity.

This paper therefore revisits the ISO-TimeML abstract syntax, aiming to develop a full, conceptually clear specification for the concrete (XML-based) representations as they are. Section 2 takes a step in that direction. Since the abstract syntax is required to allow systematic decoding of concrete representations, the adequacy of any revised version can be tested by specifying the decoding function. Section 3 is therefore devoted to the mapping of concrete representations to expressions of the abstract syntax. Section 4 indicates the next steps towards fully specified interlinked ISO-TimeMML- and QuantML-annnotations. formulating a version of the abstract syntax and the semantics of ISO-TimeML in the same style as QuantML

1.2 ISO-TimeML

ISO-TimeML distinguishes three types of temporal objects: instants, dates, and periods. With respect to instants, the ISO 24617-1:2012 specification document notes that in reality, nothing happens in infinitesimally small time; every event or state that occurs in reality (or in someone's mind) requires more than zero time, although natural languages offer speakers the possibility to express themselves as if something occurs at a precise instant (as in *"I will call you at twelve oclock"*. Such an instant is often associated with the beginning of an event, as in this example. The explicit mentioning of the start of an event, as in *"I was sad when Mary started to cry"*, illustrates the same phenomenon. Punctual events are associated with precise instants, as in the example *"Gates will close at 9:25."*

The notion of a precise instant is similar to that of a point in mathematics. Euclid defined a point as a spatial entity that which has no parts. In other words, a point is an indivisible spatial object with zero length, breadth, and height. Natural language speakers refer to instants as points on a timeline, as intervals of zero length, even though they probably know that such intervals do not really exist. In everyday language, instants are referred to with the precision of minutes, as in *"Its five past twelve"*. It is therefore appropriate to consider such intervals as instants in the ISO-TimeML abstract syntax.

Fully specified references to instants consist of a (fully specified) date and time. A fully specified date contains the specification of (1) a year, (2) a month and (3) a day number, or (2) a week number and a (3) day name. In practice, reference to instants is often underspecified, such as *"Monday at two"*, intended to be understood as next Monday at two p.m. or as last Monday at two p.m. depending on the context (which also allows to infer the year and

the week). Underspecification is represented in ISO-TimeML by using the character 'X' in values of the @value attribute. (Examples below.)

Instants are annotated in the ISO-TimeML reference representation format by <TIMEX3> expressions with @type="TIME"; dates by expressions of type "DATE"; periods by expressions with @type="DURATION".

From a semantic point of view, year numbers, calendar month names, and day names function like proper names. Just like *"James"* refers to a contextually particularly salient person named James, *"Tuesday"* refers to the contextually most salient day named 'Tuesday'. Year numbers (*"1984"*) refer to certain time intervals independent of context, just like country names (*"Denmark", "Japan"*) refer to geopolitical regions independent of context.

The specification of the ISO-TimeML abstract syntax is best done with (a) the semantics in mind and (b) specifying the decoding function that relates it to the concrete syntax - which in turn calls for a precise specification of optional attributes and default values in the concrete syntax. In (Bunt, 2018) several forms of optionality are distinguished: (a) semantic, i.e. a certain type of annotation structure may contain such a component, but does not have to for being interpretable; (b) a component that does not have to be specified in the concrete syntax, since it has a default value in the abstract syntax; (c) a component in the concrete representation that has no semantic interpretation. These distinctions are useful for a clear formulation of the abstract syntax in relation to the concrete syntax and semantics.

The <EVENT>, <TIMEX3> and <TLINK> elements all pose problems for the distinction between required and optional attributes. For example, the @relatedToEvent attribute in <TLINK> is not applicable if a @relatedToTime value is specified, and vice versa. Also, the attributes @tense and @aspect in <EVENT> elements are applicable only if @pos="VERB", and @beginPoint is applicable only if @type="DURATION".

The possible values of the @relType attribute in <TLINK> elements specify temporal relations between events and/or temporal entities. The value "IDENTITY" is unusual in this respect, as it designates the identity of two events, rather than a temporal relation; it would seem to entail the temporal relation SIMULTANEOUS. It may be noted that SIMULTANEOUS, AFTER and BEFORE are all instances of the discourse relations Synchrony and Asynchrony, defined in the ISO standard for annotating discourse relations (ISO 24617-8:2016).

The conditional applicability of various attributes in elements of the concrete syntax means, in the 3-layer architecture of SemAF parts, that alements like <TIMEX3>, can correspond to several different structures in the abstract syntax.

'

2 Abstract Syntax

2.1 Overview

As in the case of other SemAF parts, the abstract syntax of ISO-TimeML has two components: (1) the specification of a store of primitive concepts, called the Conceptual Inventory, and (2) the recursive specification of the annotation structures that may be formed by combining primitive concepts or annotation structures to form set-theoretic structures like pairs and triples.

The structures defined by the abstract syntax come in two forms: (a) *entity structures*, i.e., structures that contain semantic information about a stretch of source data (a *markable*), and (b) *link structures*, which express semantic relations between two or more entity structures. An entity structure has the form of a pair ⟨markable, semantic information⟩; a link structure has the form ⟨'entity structure 1, entity structure 2, .. entity structure n⟩, semantic relation⟩. Entity structures are represented in the concrete syntax by XML elements that have a @target attribute whose value refers to the relevant stretch of source data.

2.1.1 Conceptual inventory

The minimal building blocks of ISO-TimeML annotation structures are constants. These fall into one of the five categories listed below. Constants that denote properties are unary predicates characterizing event types, event classes, tenses, aspects, polarity, and set-theoretic type. Natural numbers are used for capturing the information expressed in examples such as *"twice", "three times",* and *"double"*. Rational numbers are needed for examples like *"half a day"*.

1. Linguistic semantic properties: unary predicates, like 'occurrence', 'process', and 'past'.

2. Relations: binary predicates for expressing temporal relations, durations, numerical relations, subordination relations, and aspectual relations.

3. Named temporal concepts: calendar years, calendar months, calendar days, month numbers, week numbers, weekday numbers, and clock times.

4. Temporal units, like hours, days, weeks, months, and years.

5. Natural numbers and rational numbers.

2.1.2 Entity structures

The abstract syntax has entity structures for events and for temporal entities. An event structure is a 6-tuple <E, T_y, C, T, A, V>, consisting of an event predicate, an event type, an event class, a tense, an aspect, and a veracity.

Temporal entity structures fall into 6 categories: (1) instant, (2) date, (3) period, (4) set of any of these, (5) amount of time, and (6) a frequency. Items in these categories are all represented in the concrete syntax by <TIMEX3> expressions with different values of @type.

2.1.3 Instants

An instant structure, corresponding to a <TIMEX3> element of type TIME, is one of the following.

1. a pair ⟨ day, clock time⟩ Clock times are predicate constants designating a time on the clock, for example annotating ‘*“four p.m.*, which is represented in XML as a <TIMEX3> element with @value=T16:00. These predicate constants take the form of sequences of two numbers, followed by a colon symbol (‘:’) followed by another sequence of two numbers. The first two numbers are 00, 01, ...24 and the last two 00, 01, ...59 (as in 16:00).

2. a triple ⟨instant, time amount, begin/end relation⟩ (*“half an hour before midnight”*).

3. a triple ⟨event, time amount, begin/end relation⟩ (*“ten minutes after the explosion”*)

4. a single clock time.

5. a pair ⟨ date structure, clock time ⟩.

2.1.4 Dates

A date structure is any (complete or incomplete) specification of a time interval of a length of one day by means of concepts related to the calendar, corresponding to a <TIMEX3> element of type DATE and is one of the following.

1. a triple ⟨year, month name or number, day name or number⟩ (*“December 25, 2024”* or *“2024-52-3”*).

2. a pair ⟨year, month⟩ (*“December 2024”*) or or ⟨year, season⟩ (*Spring 2025”*).

3. a pair ⟨month, day number⟩ (*“December 25”*).

4. a pair ⟨week, day name⟩ (*“Friday next week”*).

5. a predicate constant denoting a year, a month, or a day.(*“1984”, “May”, “Sunday”, “labour day”, “leap day”, “the 25th”*).

Named temporal entities like *“Wednesday”* work as other proper names and definite descriptions; they refer to a contextually uniquely determined entity.

2.1.5 Periods

A period structure is one of the following structures, which specify a time interval that does not form a date.

1. a pair of two structures indicating the beginning and end points of a contiguous time interval, viz. ⟨instant, instant⟩ or ⟨date, date⟩, or ⟨period⟩(*“between two and five on January 1, 2025”, “from May through September”*).

2. a triple ⟨t, t_A, R⟩ where t is an instant structure, indicating the beginning or end of a period, t_A is a time-amount structure indicating the length of the period, and R is ‘before’ or ‘after’ (*“the week before Christmas”, “the week following May 1”*).

3. a triple ⟨e, t_A, R⟩ where e is an event structure, indicating the beginning or end of an event, t_A is a time-amount structure indicating the length of the period, and R is either “before” or“after” (*“two days before the attack”, “a month after the cease-fire”*).

2.1.6 Time-amount structures

A time-amount structure is a triple ⟨’numerical relation, rational number, temporal unit⟩ (*“less than two hours”*).

2.1.7 Frequency structures

A frequency structure is a natural number or a pair ⟨natural number, temporal unit⟩.

2.1.8 Quantification structures

A quantification structure corresponds to a <TIMEX3> element of type SET. From a semantic point of view, such elements contain information about three aspects of a quantification: (1) a quantifier in the sense of classical logic, expressed by @quant values like EVERY and SOME, (2) a domain that the quantifier ranges over, indicated by the @value attribute, and (3) repetitions of an event indicated by the optional attribute @freq.

For the abstract syntax this means that a quantification structure is one of the following, where a domain is a set of days, weeks, months, or years:

1. a pair ⟨domain, quantifier⟩.
2. a triple ⟨domain, quantifier, frequency⟩.

2.2 Link structures

ISO-TimeML has link structures for (1) anchoring events in time; (2) temporal ordering of events, (3) ordering of periods, dates or instants relative to each other; (4) measuring a time interval; (5) specifying subordination relations between events; and (6) indicating aspectual relations between events.

a. Temporal anchoring: a triple ⟨event structure, temporal entity structure, anchoring relation⟩. The anchoring relation corresponds to a natural language expression like *"at", "in", "during"*.

b. Temporal event relations: a triple ⟨event structure, event structure, temporal relation⟩. Temporal relations are predicate constants corresponding to natural language expressions like *"while", "after", "just before"*.

c. Intra-time relations: a triple ⟨temporal entity, temporal entity, temporal relation⟩.

d. Time measurement, corresponding to the use of MLINK in the concrete syntax: a pair ⟨event structure, time-amount structure⟩ or a pair ⟨period structure, time-amount structure⟩.

e. Subordination structures, corresponding to the use of SLINK in the concrete syntax: a triple ⟨event structure, event structure, subordination relation⟩.

f. An aspectual link structure, corresponding to the use of ALINK, is a triple ⟨event structure, event structure, aspectual relation⟩.

3 Completeness and semantic adequacy

3.1 Requirements on abstract syntax

The abstract syntax of a markup language should meet two fundamental requirements (ISO 224617-5:2016, Principles of semantic annotation), First, it should be *complete* in the sense that for every representation structure of the concrete syntax an abstract annotation structure is defined. In other words, a decoding function (see Fig. 1) is a total function. Second, every abstract annotation structure should have a well-defined semantics. Regarding the first requirement, in this paper we present a specification of the decoding function of ISO-TimeML. Regarding the second requirement, we indicate the direction in which the semantic interpretation will go. Notes from the PWI 24617-17 project containing more details which will be made available in future project reports and follow-up papers.

3.1.1 Decoding: events and participants

The decoding function dF computes the entity structure of the abstract syntax that contains the semantically relevant information in a given concrete representation, abstracting away from other than semantic elements. Since an entity structure provides semantic information about a certain stretch of primary data, it always has the form ⟨m,s⟩, where m is a markable and s is semantic information. The use of markables in entity structure allows us to attach different semantic information to different occurrences of the same source words. This provides an opening for dealing with lexical ambiguities. In this paper we are not concerned with lexical disambiguation and simplify the presentation of abstract annotation structures by suppressing markables in the abstract syntax.

3.1.2 Decoding events

The decoding function *dF* is defined for <EVENT> elements as follows.

```
dF(<EVENT xml:id=e1 target=m1
     pred=P1 type=T1 class=C1
     tense=t1 aspect=a1/>)
= ⟨dF(P1), dF(T1), dF(C1), dF(t1),
   dF(a1)⟩
```

Example: *"Mary laughed"*.

```
dF(<EVENT xml:id="e1" target="#m1
     pred="laugh" type="occurrence"
     class="process" tense="past"
     aspect="none"/>)
= ⟨laugh, occurrence, process, past⟩
```

3.1.3 Decoding <TIMEX3> elements

a. Instants

A complete specification of an instant is formed by the complete specification of a date, which is formed by (1) a year plus (2) a month and a day number, or a week number and a weekday plus (3) a clock time. In ISO-TimeML these components are represented as parts of the string that forms the value of the @value attribute in a <TIMEX3> element of type TIME. The decoding function, which extracts the components from such strings, is defined as follows.

Example: *"July 5, 2012, at 4 p.m."*

dF(<TIMEX3 xml:id=t1 target="#m1 type="TIME"
value="2012-07-05T16:00"/>)
= ⟨⟨ 2012, july, 5 ⟩, 16:00⟩

An instant can also be specified by describing its distance from another instant, as in *"Two hours before (December 31, 2024,) midnight"*. The definition of the decoding function for this type of specification is defined as follows:

dF(<TIMEX3 xml:id="t1" target="#m1"
type="DURATION" value="PkU"
beginPoint="#t2" endpoint="#t3"/>
<SIGNAL xml:id="s1" target="#m2" pred="R"/>
<TIMEX3 xml:id="t2" type="TIME"/>
<TIMEX3 xml:id="t3" target="#m3"
type="TIME" anchorTime="#t1"
value="yvwz-mn-d1Tij:kl"/>)
= ⟨dF(#t3), dF(#t1), dF(R)⟩
= ⟨dF(yvwz-mn-d1Tij:kl), dF(PkU), dF(R)⟩

Example: *"Ttwo hours before December 31, 2024, midnight."*

Markables: m1 = two hours, m2 = before,
m3 = December 31, 2024, midnight
dF(<TIMEX3 xml:id="t1" target="#m1"
type="DURATION" value="P2H"
beginPoint="#t2" endPoint="#t3"/>
<SIGNAL xml:id="s1" target="#m2"
pred="BEFORE"/>
<TIMEX3 xml:id="t2" type="TIME"
value="2024-12-31:T22:00"
anchorTime="#t1"/>
<TIMEX3 xml:id="t3" target="#m3"
type= "TIME"
value="2024-12-31:T24:00"/>)
= 2024, december, 31 ⟩, 24:00⟩, ⟨ 2, hour⟩, before⟩⟩

The XML representation used here follows the ISO 24617-1:2012 document, where DURATION expressions have both a value of the @value attribute, specifying the length of a time period, and values of the @beginPoint and @endPoint attributes. Specifying a temporal distance in this way may run into two problems: (1) if two of these three attributes have values, then the value of the third can be inferred, therefore assigning values to all three results in expressions which are either redundant or potentially inconsistent; (2) the expressive power is insufficient for representing durations such as *"less than two hours"*.

To resolve the latter problem, an attribute @length, could be introduced, whose value refers to the specification of an amount of time. To resolve the former pproblem, it would seem best to require only two of the three DURATION attributes to be specified, not all three.

Without changing the use of the @value attribute, it would seem preferable to use the <TIMEX3> and <TLINK> elements in combination with the relation I-BEFORE (immediately before).

b. Dates

The decoding of XML representations of dates is for the most part very similar to that of instants. A complete specification of a date consists either of the specification of a year, a month and a day number (as in *"December 25, 2024"*) or a year, a week number and a weekday number (as in *"2024-52-3"*). Such a structure denotes a specific, unique date in a context-independent fashion. As in the case of a complete explicit instant description, the decoding of the XML representation rests on the decoding of the value of the @value attribute.

The decoding of a fully specified date is as follows.
dF(<TIMEX3 xml:id="t1" target="#m1" type="DATE"
value="yvwz-mn-d1"/>)
= dF(yvwz-mn-d1)
= ⟨dF(yvwz), dF(mn), dF(d1)⟩

Example: *"July 5, 2012"*

dF(<TIMEX3 xml:id="t1" target= "#m1" type="DATE"
value="2012-07-05"/>)
=⟨ 2012, july⟩

c. Periods

A contiguous time interval can be defined by the specification of a begin- and an end point (*"From two to five", "From May through September"*).

Example: *"On New Years day I biked from ten to five."*

dF(<TIMEX3 xml:id="t1" target= "#m1" type=
"DURATION" beginPoint="#t2" endpoint="#t3"
value="P7H"/>
<SIGNAL xml:id="s1" target="#m4" pred="FROM/>
<SIGNAL xml:id="s2" target="#m6" pred="TO"/>
<TIMEX3 xml:id="t2" target="#m2" type="TIME"
value="2025-01-01T14:00"/>
<TIMEX3 xml:id="t3" target= "#m3" type="TIME"

value="2025-01-01T17:00"/>)
=⟨ dF(#t2), dF(#t3) ⟩
=⟨⟨january,1⟩,10:00⟩,⟨january,1⟩,17:00⟩⟩

(To link this interval to the biking event, two additional <TLINK> elements are needed as follows, where "e1" is the identifier of the event:
<TLINK eventID="e1" relatedToTime="#t2"
signalID="#s1" relType="BEGUN_BY"/>
<TLINK eventID="#e1" relatedToTime="#t3"
signalID="#s2" relType="ENDED_BY"/>
See also ISO 24617-1:2012, p. 84.))

A period can also be specified in a relative way by a beginning or an end point and the amount of time that separates that point from (a) a given instant or date (*"for two hours after midnight"* or *"the week beginning May 5"*) or (b) from the beginning or end of an event (*"two weeks before the attack"*).

Example: *"(for) two hours after midnight"*

dF(<TIMEX3 xml:id="t1" target= "#m2"
type="DURATION" beginPoint="#t2"
value="PT2H"/>
<SIGNAL xml:id="s1" target="#m4'
pred="AFTER"/>
<TIMEX3 xml:id="t2" target= "#m1" type=
"TIME" value="XXXX-XX-XXT00:00"/>)
= ⟨dF(#t2), dF(#t2), dF("AFTER")⟩
=⟨⟨2,hour⟩,00:00, after⟩

3.2.3 Time amounts

An amount of time is represented by a <TIMEX3> element of type DURATION that has neither a begin point nor an end point specified, but which has a @value attribute with a value of the form 'PnU', where'P' as before stands for 'period', 'n' for a real number, and 'U' for a temporal unit (like second, minute, hour, day,). Such a <TIMEX3> element does not identify a specific period and can be viewed as a case of underspecification, denoting the set of all periods of the specified length. Its use is to link an event to an amount of time through an <MLINK> element with @relType="MEASURES". The decoding of such a <TIMEX3> element is defined as follows.

dF(<TIMEX3 xml:id=ti target=mj
type=DURATION value=PnU
tense=t1 aspect="'NONE"/>)
= ⟨dF(n), dF(U)⟩

Example: *"John taught for three hours"*

dF(<EVENT xml:id="e1" target= "#m2" pred="teach"
class="OCCURRENCE" type="PROCESS"
<TIMEX3 xml:id="t1" target= "#m3" type=
"DURATION" value="P3H"/>
<MLINK eventID="e1" relatedToTime="#t1"
relType="MEASURES"/>)
= ⟨ dF(#e1), dF(#t1), dF(MEASURES) ⟩
= ⟨⟨ teach, occurrence, process, past,⟩
⟨ 3, hour ⟩, duration ⟩

3.2.4 Frequency structures

Example: *"twice a month"*

dF(<TIMEX3 xml:id="t3" type="SET" value=
"P1M" freq="2X"/>)
=⟨dF(XXXX-XX), dF(EVERY), dF(2X)⟩
=⟨month, all, 2⟩

For the treatment in ISO-TimeML of repetitions rather than frequencies, as in *"John kissed Mary twice"* see Section 4.

3.2.5. Quantification structures

<TIMEX3> expressions with type="SET" have a @value and a @quant attribute, and optionally a @freq attribute.

dF(<TIMEX3 xml:id="t1" type="SET"
value="PnU" quant="q1"/>)
= ⟨dF(PnU), dF(q1) ⟩
=⟨⟨dF(n), dF(U)⟩, dF(q1)⟩

Example: *"Every Monday"*

dF(<TIMEX3 xml:id="t1" type="SET" value=
"XXXX-WXX-1 quant="EVERY"/>)
= ⟨dF(XXXX-WXX-1), dF(EVERY)⟩
= ⟨monday, all⟩

3.2 Link structures

a. Event time relations

<TLINK> elements that specify information about the time of occurrence of an event have the attributes @eventID, @relatedToTime, and @relType. Their decoding is defined as follows:

dF(<TLINK eventD="#e" relatedToTime="#t"
signalID="#s" relType="R"/>
<SIGNAL xml:id="s" pred="R"
= ⟨dF(#e), dF(#t), dF(R)⟩

b. Temporal discourse relations

<TLINK> elements that specify information about the temporal relation between two events have the attributes @eventID and @relatedToEvent, plus a @relType attribute whose value represents the relation. Decoding:

dF(<TLINK eventD= "#e1" relatedEvent="#e2"
signalID="#s" relType="R"/>

<SIGNAL xml:id="*s*" pred="R1"/>
= ⟨dF(#e1), dF(#e2), dF(R)⟩

c. Relations between temporal entities
Relations between two times, dates, or periods, possibly quantified, as in *"twenty minutes every Monday"*, are represented by <TLINK> elements with the attributes @timeID, @relatedToTime, and @relType. Their decoding is like in the above cases a and b.

4 Next steps and issues for further study

Revisiting the ISO-TimeML abstract syntax, we are in fact applying the CASCADES method for developing or improving an annotation schema (Pustejovsky, Bunt & Zaenen, 2017). This means that the specification of an annotation scheme consists of four consecutive stages: (1) establishment of a metamodel, (2) - (3) specification of concrete and abstract syntax and the decoding function that connects them, (4) definition of semantic interpretation. Feedback loops go back from any stage to any previous stage.

In reverse engineering mode, one can start at any of the four stages and follow steps forward or backward to ensure inter-stage consistency. In the case of revisiting the ISO-TimeML abstract syntax, we take the existing concrete syntax and the metamodel on which it is loosely based for granted. Starting at the abstract syntax stage (3), the next step forward is the specification of the semantics; the most relevant feedback step is ensuring the consistency between abstract and concrete syntax, which is accomplished by specifying a decoding function.

Regarding the next step forward, we have started the definition of a revised compositional semantics for the expressions of the (revised) abstract syntax as outlined above. Inspired by the specification of the semantics of QuantML (see Bunt, 2023), this semantics takes the form of a recursive function that interprets annotation structures as second-order DRSs (Kamp & Reyle, 1993).

Regarding the next step backward, while specifying the decoding function we have noted some unclear aspects, limitations and gaps in the concrete syntax as specified in ISO 24617-1:2012 and in the guidelines for its use. This could be of interest for a possible future update of the specification and the guidelines. Some these issues are the following.

1. In ISO-TimeML, periods are assumed to be contiguous. This is not always realistic, in view of sentences such as *"I studied from nine to five"* - there must have been interruptions in this interval. This calls for the introduction of possibly discontinuous intervals.

2. <TIMEX3> elements of type DURATION have a length which is specified by the @value attribute. This suggests that the duration of an event is conceived as a period of a certain length. Although technically possible in many cases, this does not seem general enough. It is not clear how an example like *"In Hong Kong it's 7 hours later than in Amsterdam* could be annotated.

3. The @quant attribute in <TIMEX3> elements of type SET is limited in the variety of generalized quantifiers that it allows. Examples like *"More than 3000 students protested"* cannot be annotated. See also Bunt & Pustejovsky (2010) for a discussion of the annotation of quantification over times and events.

4. Similarly, the use of values like "2X" for the @value attribute in order to represent *"twice"* does not permit to represent 'generalized' repetitions and frequencies like *"more than three times", "at least twice a week"*.

5. Some attributes in <EVENT> are either syntactic or lexical in nature. The attribute @pos (part-of-speech) is obviously not semantic; @tense and @aspect are only partly semantic, as seems to be reflected in the fact that they apply only to events expressed by verbs. Moreover, for verbs the values of the attributes @type and @class can typically be obtained from the lexical information of the verb and do not need to be annotated.

6. A similar issue concerns the use of <SIGNAL> elements. As the examples in Section 3.3 illustrate, <TLINK> elements that express a temporal relation have @signalID as one of their attributes, whose value refers to a <SIGNAL> element. This attribute seems semantically superfluous, since its @pred value, which specifies the temporal relation, is also expressed in the value of the @relType attribute in the <TLINK> element.

7. An <MLINK> element always has @relType = "MEASURES". Since this is always the

case, it would seem superfluous to annotate it as such.

8. According to the ISO 24617-1 document, the possible values of the @value attribute for dates, times, pertiods, frequencies and quantifications are taken from the TIDES scheme (Ferro et al., 2003), which follows the representation of dates and times in the ISO 8601 standard. Whether this Is desirable also for annotating period lengths, quantifications, frequencies and repetitions deserves further study.

Most importantly, the outcome of this study is a version of the ISO-TimeML abstract syntax and the decoding function from concrete to abstract syntax that promise to provide a solid basis for the further interlinking of ISO-TimeML annotations and QuantML annotations.

References

J. Allen and M. Core. 1997. *DAMSL: Dialogue Act Markup in Several Layers (Draft 2.1). Technical Report*. University of Rochester, Rochester, NY.

H. Bunt. 2009. The DIT++ taxonomy for functional dialogue markup. In *Proceedings of AAMAS 2009 Workshop 'Towards a Standard Markup Language for Embodied Dialogue Acts'*, pages 13–24, Budapest.

H. Bunt. 2019. An annotation scheme for quantification. In *Proceedings 14th International Conference on Computational Semantics (IWCS 2019)*, pages 31–42, Gothenburg, Sweden.

H. Bunt. 2020. The annotation of quantification: The current state of ISO 24617-12. In *Proceedings of the 16th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-16)*, Marseille.

H. Bunt. 2024. Combining Annotation schemes through Iinterlinking. In *Proceedings of the 20th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-20)*, pages 83 – 95, Turin, Italy.

H. Bunt and J. Pustejovsky. 2010. Annotating temporal and event quantification. In *Proceedings ISA-5, Fifth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 15–22. City University of Hong Kong.

H. Bunt, J. Pustejovsky, and K. Lee. 2018. Towards an ISO Standard for the Annotation of Quantification. In *Proceedings 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Myazaki, Japan. ELRA.

Harry Bunt. 2018. Downward Compatible Revision of Dialogue Annotation. In *Proceedings 14th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation (ISA-9)*, page 241, Santa Fe, New Mesico, USA.

R. Cooper. 1983. *Quantification and syntactic theory*. Reidel, Dordrecht.

ISO. 2012. *ISO 24617-1: 2012, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and events*. International Organisation for Standardisation ISO, Geneva.

ISO. 2014. *ISO 24617-4: 2014, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 4: Semantic roles*. Geneva: International Organisation for Standardisation ISO.

ISO. 2015. *ISO 24617-6:2015, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 6: Principles of semantic annotation*. International Organisation for Standardisation ISO, Geneva.

ISO. 2024. *Preliminary Work Item proposal PWI 24617-17, Interlinking*. International Organisation for Standardisation ISO, Geneva.

ISO. 2025. *ISO24617-12:2025, Language Resource Management: Semantic Annotation Framework (SemAF) - Part 12: Quantification*. International Standard. International Organisation for Standardisation ISO, Geneva.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

K. Lee, H. Bunt, J. Pustejovsky, A. Fang, and C. Park. 2025. Representing ISO Annotated Dynamic Information in UMR. In *Proceedings 6th International Workshop on designing meaning representations (DMR 2025)*, Prague.

I. Mani, C. Doran, D. Harris, J. Hitzeman, R. Quimby, J. Richer, B. Wellner, S. Mardis, and S. Clancy. 2010. SpatialML: Annotatopn, Resources, and Evaluation. *Language Resources and Evaluation*, 44 (3):263–280.

R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings 6^{th} International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

J. Pustejovsky, J. Castano, R. Ingria, R. Gaizauskas, G. Katz, R. Saurí, and A. Setzer. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, pages 337–353, Tilburg, Netherlands.

The representation of QuantML annotations in UMR - an exploration

Harry Bunt
Tilburg University / Tilburg, The Netherlands
harry.bunt@tilburguniversity.edu

Kiyong Lee
Korea University / Seoul, Korea
ikiyong@gmail.com

Abstract

This paper explores the possibilities and the problems in using Unified Meaning Representations (UMRs) for representing annotations of quantification phenomena, according to the ISO standard scheme QuantML (ISO 24617-12:2025). We show that the semantic information in QuantML annotations can we expressed in UMR, provided that some powerful semantic concepts are introduced and a slightly more general approach is adopted for the representation of multiple scope relations. Conversion functions are defined that transform the XML-based representations of QuantML into UMR structures and vice versa. The consequences are discussed that can be drawn from this regarding the possible role of UMR and the semantics of UMR representations of quantification.

1 Introduction

Quantification is one of the most studied topics in semantics. Its complexity gives rise to a plethora of questions, conceptual, linguistic, logical, and computational. Montague (1971) used a higher-order intensional logic with categorial grammar to treat quantification in natural language. In contrast, Abstract Meaning Representation (AMR) and its extension, Uniform Meaning Representation (UMR) provide a first-order Neo-Davidsonian semantics which does not address many aspects of quantification, but which is attractive for its conceptual simplicity, especially when representing meaning in the form of a rooted graph. Intuitively, the nodes and edges of such a graph are similar to the entity and link structures of QuantML. For these reasons, and further motivated by the increasing popularity of AMR and UMR in natural language processing (see Lee et al., 2025), this paper explores the possibility of representing the rich QuanML annotations of quantification in the form of UMRs.

The organization of this paper is as follows. In Section 2 we consider some of the characteristic features of QuantML and UMR, in particular the 3-level architecture of QuantML with (a) a concrete syntax, which defines an XML-based representation format, (b) an abstract syntax which defines annotation structures in a format-independent way, using set-theoretic constructs, and (c) a semantics, which specifies semantic interpretations of the annotation structures of the abstract syntax. Section 3 discusses some of the fundamental concepts in the annotation of quantification. In Section 4 we examine the potential application of UMR expressions in the semantic annotation of quantification. We do this in two steps. First, we compare the annotation of a variety of quantification phenomena using (a) the XML-based reference format of QuantML, which we will refer to as QuantML/XML, and (b) a UMR-based representation format, which we will refer to as QuantML/UMR. This comparison provides insight into those aspects where the two forms seem little more than notational variants and those aspects for which the relation between the two is more complex. Second, we investigate the possibility of converting representations in QuantML to QuantML/UMR and vice versa. In the concluding Section 5 we discuss how QuanML/UMR could fit into the architecture of ISO SemAF annotation standards and we consider the semantics of Quant/UMR annotations in this context.

2 Background

2.1 QuantML

QuantML (ISO 24617-12:2025) is part 12 of the multi-part ISO standard *Semantic Annotation Framework (SemAF)* for semantic annotation. Following the Principles of semantic annotation (ISO 24617-6:2015), the parts of SemAF have the same 3-level architecture (see also Pustejovsky et al. 2017), consisting of:

1. An abstract syntax, which specifies the class of well-defined *annotation structures* as pairs,

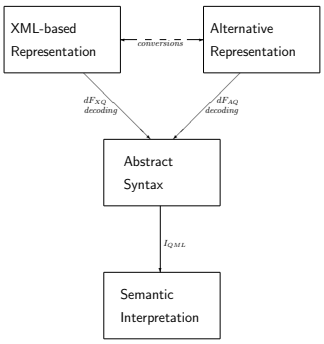


Figure 1: Architecture of SemAF parts.

triples, and other set-theoretical constructs containing quantification-related concepts. Annotation structures consist of *entity structures*, which contain information about a stretch of primary data, and *link structures*, which contain information relating two (or more) entity structures.

2. semantics, which specifies the meaning of the annotation structures defined by the abstract syntax. QuantML has an interpretation-by-translation semantics, which translates annotation structures to discourse representation structures (DRSs, (Kamp and Reyle, 1993)). The use of DRSs is mainly motivated by the fact that this formalism is also used in other SemAF parts.

3. A concrete syntax, which specifies a representation format for annotation structures. The QuantML definition includes an XML-based reference format, primarily motivated by the widespread use of XML in other standards.

The three levels are interrelated by encoding (eF), and interpretation functions; see Figure 1. Since the semantics is defined at the level of the abstract syntax, alternative representation formats may be used that share the same abstract syntax, as indicated in Figure 1, and are thus semantically equivalent. This adds to the interoperability of the annotation schema.

QuantML is semantically rooted in Davidsonian event semantics (Davidson (1967), Parsons (1990)) and in the theory of generalised quantifiers (GQT, Barwise and Cooper (1981), Cooper, 1983). Generalised quantifiers are not the logical counterparts of determiners, such as *"every"* and *"some"*, but of NPs like *"More than fifty students"* and *Three of the five men"*. In Davidsonian semantics, verbs are viewed as denoting events and their NP arguments as denoting participants in the events in certain semantic roles. Combining the two approaches, a sentence with quantified participants like *"More than fifty students protested"* has at least two readings, depending on whether it is taken to describe a single protest event with a set of more than fifty students as agents, or a set of protest events with individual students or smaller numbers of students as agents, the total number of participants involved in the events adding up to more than 50.

QuantML supports the annotation of quantified participation in events by taking into account the following categories of information:

(1)
1. quantification domain
2. determinacy (determine/indeterminate)
3. distributivity (individual/collective/ unspecific)
4. individuation (count/mass)
5. involvement (absolute and proportional)
6. semantic role
7. exhaustivity
8. polarity
9. participant scope
10. event scope
11. repetitiveness
12. size of reference domain
13. restrictiveness of modifiers
14. linking of modifiers (inverse or linear)
15. modality (e.g. epistemic)
16. genericity (generic or specific).

The categories 1 - 14 correspond to attributes of XML elements in the concrete syntax of QuantML. Some of these items are optional, in the sense of having a default value: polarity is by default 'positive', exhaustivity is 'negative', event scope is 'narrow', and repetiveness is 'at least once'. The attributes

14, 15 and 16 are exceptional in that they exist only in the concrete syntax; they do not correspond to anything in the abstract syntax or the semantics. They have been added purely to support searches in corpora where generic or modal quantification is marked up. For explanations and discussion of all the categories see (Bunt, 2024).

2.2 UMR Formalism

Three equivalent alternative formats are used in AMR-UMR: logical, graph, and PENMAN formats. This paper focuses on the last. form. Example (2) illustrates how meaning is represented in the PENMAN format at the two levels of UMR: the sentence and document levels.[1] Representation 2 is understood as saying that the sentence 's1' contains an instance 's' of the event of two women sharing a pizza 'p' and this occurred yesterday, the date of which depends on the document creation time ('DCT') (see Van Gysel et al. (2022), part 1).

(2) Sentence 1: Two women shared a pizza.

```
Snt1: Two women shared a pizza
      today.
%% Sentence (predicate) level
(s / share
   :agent (w / woman
             :quantity 2)
   :patient (p / pizza
               :quantity 1)
   :temporal (t / yesterday))
%% Document (discourse) level
(s1 / sentence
    :temporal ((DCT :depends-on s1t)
               (s1t :contained s1s)))
```

Some of the variables that are introduced in the PENMAN representation ('s','w', 'p', and 't') at the predicate-structure level. may recur at the document (discourse-structure) level, being prefixed with their root variable 's1' in this level, as seen in (2).

This format makes use of two operators: the slash / for concepts and the colon : for semantic relations. The slash form like (s / share) represents the variable *s* as an *instance* of the semantic concept *share*, not as a word in the text, and it is logically represented as *instance(s, share)*.[2] The concept *share* might be realized as a verb "shared", a noun "share", or a participle "sharing".

The colon `':'` indicates a binary relation between two concept variables or between a variable and some logical element like the negation, a numeral, or a name like "John". For example, the relation :agent in example (2) relates the concept variable 'w' for woman to the root concept variable 's' for the 'share' predicate. It is logically represented as *agent(s, w)*, stating that *w* is the agent in the share event *s*.

The current version of the UMR guidelines (Van Gysel et al., 2022) treats coreference, temporal, and modal relations at the document (discourse-structure) level representation, with representations different from those at the sentence level. To be consistent, we propose that they should be represented as follows:

(3) Proposed discourse-level representation

```
(s1 / sentence
    :temporal (d / depends-on
                 :arg1 s1t
                 :arg2 DCT)
    :anchoring (c / contained
                  :arg1 s1s
                  :arg2 s1t))
```

For using UMRin the annotation of quantification, we have to consider the representation of scope. The original AMR format is too limited in this respect, and this is one of the motivations for extending it to UMR. According to the UMR Guidelines (Van Gysel et al., 2022) quantification scope is represented with an 'inverse relation', such as `pred-of`, as illustrated by example (4):

(4) Someone answered all the questions.

```
(a / answer_01
  :ARG0 (p / person)
  :ARG1 (q / question
           :quant all)
  :pred-of (s / scope
              :arg1 p
              :arg2 q))
```

Here, it is understood that the first argument scopes over the second. This would be clearer if instead of the term *scope* a term like *scopes-over* would be used. On this approach the event predicate remains the root of the structure.

Second, Pustejovsky et al. (2019) have proposed a treatment in which a *scope* concept is used as the root of the representation, rather than the event predicate. On this approach, the sentence in (4) is represented as follows:

(5) Representation in (Pustejovsky et al. (2019)

[1] Lee et al. (2025) prefer to call the two levels of UMR *predicate-structure level* and *discourse-structure* level, respectively. We follow their practice in this paper. Note also that the discourse-structure level may deal with a single sentence.

[2] By introducing the notion of *instance*, events and properties are treated not as functional types like $(t \rightarrow e)$, but as first-class objects.

```
(s / scope
    :pred (a / answer_01
            :ARG0 (p / person)
            :ARG1 (q / question
                     :quant all)
    :arg1 p
    :arg2 q))
```

Again, the intended interpretation could be made more explicit by using *scopes-over* as the concept name. QuantML distinguishes other scope relations besides *wider*, viz. *equal* and *dual*. In this paper we therefore propose a slightly different representation, in which the scope concept has besides the two scope-linked arguments also an explicit representation of the scope relation. This leads to the following representation:

(6) Representation eith explicit scope relation

```
(s / scope
   :pred (a / answer_01
            :ARG0 (p / person)
            :ARG1 (q / question
                     :quant all)
  :arg1 p
  :arg2 q
  :scopeRel wider)
```

Both the approach proposed by Pustejovsky et al. and the one suggested in (6) are limited to representation of a single scope relation. For a sentence with multiple scope relations, like (7), a more general approach is needed.

(7) Some teachers gave every pupil a present.

One possibility is to extend the approach proposed by Pusttejovsky et al. (2019) by allowing a list of scope relations, all of which refer to the same predicate-level substructure.

Alternatively, the scope concept used in (6) can be generalized to a more complex 'scope structure' concept that has a list of scope relations. For a three-argument sentence like (7) this would look schematically as follows.

(8) Snt1 Some teachers gave every pupil a present.

```
(s / scope
   :scoping (sc / scopeLinks
                :op1 (sl1 / scopeLink
                   :arg1 x1
                   :arg2 x2
                   :scopeRel wider)
                :op2 sl2 / scopeLink
                   :arg1 x1
                   :arg2 x3
                   :scopeRel wider)
   :pred (g / give
            :  ...))
```

Even when enriched with a representation of scope, the UMR format is still quite limited in its representation of quantification, being restricted to the use of a `quant` relation in expressions of the form `:quant all`, which corresponds to the representation of involvement, including numerical involvement, as expressed in QuantML/XML by the value of the @involvement attribute. As noted in Section 2.1, QuantML supports the annotation of 16 aspects of quantification, of which involvement is one. In the next section we consider how of these aspects could be represented in UMR.

3 Annotations of Quantification

3.1 Participation in Events

Some of the aspects of quantification listed in (1) can be represented as properties of events or participants, but some aspects cannot, since they express properties of *the way* certain participants are involved in an event. This is the case for the distributivity of a quantification, as illustrated by the example *"The three men had a beer after moving the pianos"*. Here we see the same three men involved individually in drinking beer and collectively in moving pianos.

Another case is the exhaustivity of a quantification. Consider the difference between *"Two women smiled"* and *"TWO women smiled"*. In the latter case, no more (and no less) than two women smiled, and there is an implication that certain other women did not smile. This distinction can hardly be captured by a property of the women who did smile.

A third case is the *event scope*. This sometimes overlooked aspect concerns the relative scoping of events and participants. For example, the sentence *"Ninety-six passengers survived a crash"* describes in one interpretation a crash in which 96 participants survived, and in another interpretation the total number of passengers who survived in a number of crashes. In the former interpretation, an event scopes over set of passengers, in the latter interpretation it's the other way round.

In QuantML, *participation* and *predication* structures are used for indicating how events and participants are related.[3] Such structures contain the specification of semantic role, distributivity, and event scope (and optionally exhaustivity and polarity). These aspects of quantification can be represented in UMR by introducing the participation and prediction concepts. The use of these concepts in

[3] Predication structures are used for copular verbs and verbs with adjectival complements

QuantML/XML and QuantML/UMR is illustrated in (13) and (14) and discussed in the next section.

3.2 Domains of Quantification

Quantifications in natural language have a certain domain. The domain defined by the head noun of a quantifying NP, such as the domain of women in *"two women"* and the domain of students in *"all the students"* is called the *source domain*. Occurrences of NPs are nearly always intended to refer to a subset of the source domain - in *"two women smiled"* the quantification most likely does not refer to the set of all women, and in *"all the students protested"* the quantification does not range over the set of all students in the world, but they refer to certain contextually determined subsets of women and students, respectively. This more restricted domain is called the *reference domain*[4]. For NPs with bare head nouns the reference domain and the source domain may coincide, but in general the specification of a reference domain combines nouns, modifiers and conjunctions, as illustrated in (9)

(9) Twenty valuable ((Chinese vases) and (Japanese drawings))

Modifiers are interpreted in QuantML using a concept of *modification* that has some similarities with *participation*. A modification structure captures properties of the way a modifier relates to its arguments. These properties include distributivity, restrictiveness, and form of linking (linear or inverse). The following examples illustrate the distributivity and linking of modifiers, respectively.

(10) a. I'm carrying these heavy books to the new library building.

b. Two students from every Dutch university participated in the talks.

For describing such properties QuantML makes use of <entity> elements with a @domain attribute whose value refers to a reference domain, represented by a <refDomain> element. Such an element may contain references to multiple subdomains and modification structures. The recursion in such representations ends when a subdomain consists of a single component without modifiers. Such a domain is represented by a <sourceDomain> element. For example, the quantification domain of "older (men and women)" is represented in (11).

[4]Also known as 'context set' (Westerståhl, 1985)

(11) Older men and women.

Markables:
m1 = "older men and women",
m2 = "older", m3 = "men and women",
m4 = "men", m5 = "women".

<entity xml:id="x1" target="#m1" domain="#x2"
individuation="count" involvement="some"/>
<refDomain xml:id="x2" target="#m3"
:subdomains="#y1 #y2" determinacy="indet"
restrictions=#r1"/>
<adjMod xml:id="r1" target="#m2"
pred="older"/>
<sourceDomain xml:id="y1" target="#m4"
pred="man"/>
<sourceDomain xml:id="y2" target="#m5"
pred="woman"/>

Complex quantification domains can be represented in UMR by introducing the concepts of reference domain, source domain, and modification structure, leading to the following representation.

(12) Older men and women.

```
(x1 / entity
   :individuation count
   :involvement some
   :domain (x2 / refDomain
    :subdomains (c / conjunction
       :op1 (y1 / sourceDomain
          :pred man)
       :op2 (y2 / sourceDomain
          :pred woman))
    :determinacy / indet
    :restr (m / modification
        :distributivity individual
        :restrictiveness restrictive
        :linking linear))
```

When we compare the XML and the UMR representations of this sentence we can see some interesting similarities and differences. These are analysed in the next section.

4 Comparing and Converting Representations

The unit of annotation in QuantML is a clause, i.e. a grammatical unit describing an event (or set of events) and the participants involved. The QuantML/XML annotation of the clause in example (2) is represented as follows.

(13) Two women shared a pizza.

Markables: m1 = "Two women", m2 = "women",
m3 = "shared", m4 = "a pizza", m5 = "pizza".

<event xml:id="e1" target="#m2" pred="share"/>
<entity xml:id="x1" target="#m1" domain="#x2"
individuation="count" involvement="2"/>
<refDomain xml:id="x2" target="#m2"
subdomains="#x3" determinacy="indet"/>
<sourceDomain xml:id="x3" target="#m2"
pred="woman"/>
<entity xml:id="x4" target="#m4" domain="#x5"

individuation="count" involvement="some"/>
<refDomain xml:id="x5" target="#m5"
subdomains="#x3" determinacy="indet/>
<sourceDomain xml:id="x3" target="#m5"
pred="pizza"/>
<participation event="#e1" participant="#x1"
semRole="agent" distributivity="individual"/>
<participation event="#e1" participant="#x2"
semRole="patient" distributivity="individual"/>
<scope arg1="#x1" arg2="#x4" scopeRel="wider"/>

Using the representation of scope shown in (6), the corresponding QuantML/UMR representation is as follows.

(14) Two women shared a pizza.

```
%% Sentence (predicate) level
    Snt1: Two women shared a pizza.
    (sc / scope
       :scoping (sLs / scopeLinks
         :op1 (sL1 / scopeLink
             :arg1 x1
             :arg2 x4
             :scopeRel wider)
       :pred (sh / share
         :arguments (a / argStructure
          :op1 (p1 / participation
             :semRole agent
             :participant (x1 / entity
                :individuation count
                :involvement 2
                :domain (x2 / refDomain
                  :determinacy indet
                  :subdomains
                     (x3 / sourceDomain
                          :pred woman))))
              :distributivity collective
              :eventScope narrow)
           :op2 (p2 / participation
             :semRole patient
             :participant (x4 / entity
                :individuation count
                :involvement 1
                :domain (x5 / refDomain
                  :determinacy indet
                  :subdomains
                     (x6 / sourceDomain
                          :pred pizza))))


                :pred pizza
                :individuation count
                :involvement 1)
             :distributivity individual
             :eventScope narrow)))
```

Inspecting the clause representations in (13) and (14) and the NP representations in (11) and (12) we can observe the following correspondences,

(15) a. For all entity structures[5], i.e. for the XML elements <event>, <entity>, <cardinality>, <amount> and the elements representing modifiers (<adjMod>, <relClause>, <ppMod> and <nnMod>) there is a simple correspondence:[6]

<E xml:id="x" a1="v1" ... ak="vk" />
⇐⇒
(x / E
:a1 v1
...
:ak vk)

b. For participation and predication link structures, the corresponding UMR structure is the same, except that in this case the variable that is introduced does not have a counterpart in the XML representation.

```
<L a1="v1" ...  ak="vk"/>
⇐⇒
(z / L
    :a1 v1
     ...
    :ak vk)
```

c. For XML attributes with value type 'IDREF" (i.e. a pointer to another substructure of the representation):

aj="#xj"
⇐⇒
the UMR structure that corresponds to the XML element with xml:id="xj".

For example, the following correspondence holds because of clause (15a) and the present clause:

<refDomain xml:id="x2" subDomains=
"#x3" determinacy="indet"/>
<sourceDomain xml:id="x3"
pred="woman">
⇐⇒

```
(x2 / refDomain
     :determinacy indet
     :subDomain
             (x3/sourceDomain
                :pred woman))
```

d. For QuantML/XML attributes with value type 'IDREFS" (i.e. a list of pointers to other substructures):

```
aj="#xj #x2...  #xk"
⇐⇒
(op1 U1
:op2 U2
```

[5]In the sense of the distinction between entity structures and link structures.

[6]ISO annotation schemes include a fine-grained representation of the anchoring of annotation components in the primary data using maarkables, following the TEI standard. Since UMR does not represent the anchoring of annotation components, we suppress for comparison the markables in QuantML/XML representations in the rest of this paper.

```
...
:opk Uk)
```
where Uj is the UMR structure that corresponds to the XML element with identifier 'xj', i.e. the entity structure with xml:id="xj".

Together, the items in (15) define correspondences between NP representations, as exemplified in (16) for the phrase *"Seven black kittens"*.

(16)
```
<entity xml:id="x1" domain="#x2"
    individuation="count" involvement="#c1"/>
<cardinality xml:id="c1" numRel="greater-
    or-equal" number="7">
<refDomain xml:id="x2" subDomains="#x3"
     restrs"#r1" determinacy="indet" />
<source xml:id="x3" pred="kitten"/>)
<mod xml:id="#m1" pred="black">
    distributivity="individual">
⇐⇒
(x1 / entity
  :domain (x2 / refDomain
     :subdomains (x3 / sourceDom
                      :pred kitten)
     :restrs (m1 / mod
          :distributiviy individual
          :pred black
     :determinacy indet
  :individuation count
  :involvement (c1 / cardinality
    :numRel greater-or-equal
    :number 7))
```

The correspondences between parts of XML- and UMR-representations suggest that it should not be too difficult to define conversion functions $F_{X\to U}$ and $F_{U\to X}$ that transform XML-representations to UMR-representations and vice versa. A complication, however, is formed by a structural difference between the two.

Where a QuantML/XML representation is a list of XML elements, describing events and participants, connected by pointers, a UMR representation is a rooted nested structure with an event node (also called 'pred' node in UMR) as the root and with the participation in the event as properties of the event. Moreover, for a sentence with scope relations, which is almost every sentence, the proposals for scope representation discussed in Section 2.2 add a scope node as the root directly 'above' the event node, By contrast, scope relations in QuantML/XML are simply items in the list of XML elements, arbitrarily positioned.

The conversion form XML to UMR places the converted scope representation immediately 'above' the converted event representation.

To define the XML - UMR conversion function we note hat a QuantML/XML structure represents a clause. In the list of elements of such a representations three parts can be distinguished: (1) an <event> element, (2) a list of <scope> elements, and (3) elements describing participants and participation. We use the notation in (17) to designate such a list:

(17) $L = [L_e, L_s, L_p]$

For the elements in L_e and L_p the XML-UMR conversion function is defined by the correspondences in (15).

For converting the scope relations $l_{s1}, l_{s2}, ...l_{sn}$ forming L_s it should be noted that their representation in QuantML/XML depends on the fact that they occur in a clause representation with a single <event> elemen. Their conversion to UMR is therefore part of the conversion of the clause representation as a whole, specified in (18).

(18) $F_{X\to U}(L) =$
```
(s / scope
    :scoping (sLs / scopeLinks
       :op1 (sL1 / scopeLink
                   :arg1 x1
                   :arg2 x2
                   :scopeRel sr1)
       ...
       :opk (sLk / scopeLink
                   :arg1 xj
                   :arg2 xk
                   :scopeRel srk)
     :pred F_{X→U}(L_e, L_p))
```
where
$F_{X\to U}(L_e, L_p) = \bigcup^3(F_{X\to U}(L_e), F_{X\to U}(L_p),$ "arguments")
and
$F_{X\to U}(L_p) =$
```
( a / argumentStructure
    :op1 F_{X→U}(L_p1)
    ...
    :op2 F_{X→U}(L_p2)
    :opk F_{X→U}(L_pk))
```

The operation $\bigcup^3$ is defined as follows. Given two UMR representations, the operation adds the second argument to the first argument through the relation expressed by the third argument. Schematically:

(19) $\bigcup^3$
```
( (a / A,      (b / B,     R )
    :a1 v1        :b1 w1
    :  ...         ...
    :ak vk)        :bn wn)
= (a / A
    :a1 v1
    :  ...
    :ak vk
    :R (b / B
          :b1 w1
          :  ...
          :bn wn))
```

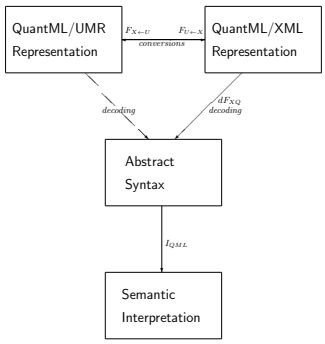


Figure 2: SemAF Architecture with UMR .

Applied to the QuantML/XML representation in (13) this conversion function produces the QuantML/UMR representation in (14).

Looking in the other direction, the correspondence relations in (15) are equally useful for converting subexpressions of UMR representations to XML, due to their symmetry. For example, based on the correspondence (15a), the conversion of entity structure representations is defined in (20).

(20) $F_{U\to X}$((x / E
:a1 v1
:...
:ak vk)) =
= <E xml:id="x" a1="v1"...ak="vk" />

The structural differences and the different treatments of scope are easier to convert from UMR to XML than in the other direction. The list of scope relations in the (sLs / scopeLinks) list is converted to the list of their converted representations, with all the elements linked to the <event> element corresponding to the value of :pred relation in the (s / scope) representation. This is illustrated in (21).

Since QuantML/UMR representations can be converted to QuantML/XML, the existence of a decoding function from annotations in the latter format have to expressions of the QuantML abstract syntax shows that the QuantML/UMR representations have the same semantics as those of QuantML/XML.

To complete the reasoning about the two formats, we consider the decoding of QuantML/UMR representations directly by a decoding function to expressions of the QuantML abstract syntax.

5 Conclusions: UMR in the SemAF architecture

The inter-convertibility of the XML-based and the UMR-based representations of QuantML annotations means that QuantML/UMR fits perfectly into the three-level architecture of ISO SemAF annotation schemes, as visualized in Figure 2. It does involve the addition to UMR of some of the fundamental concepts of QuantML, notable those of participation (in events), of predication for dealing with copular verbs, and of adnominal modification, and it requires a more general treatment of scope.

As noted in Section 2.1, annotations made by a SemAF annotation schema have a semantics, defined by an interpretation function which is applicable to the structures of the abstract syntax. In the case of QuantML, the semantics combines event semantics and generalized quantifier theory with Discourse Representation Theory. This semantics applies to the representation structures defined by a concrete syntax via 'decoding' functions that express the semantic information in these representations in the set-theoretic structures of the abstract syntax.

From the inter-convertibility of QuantML/XML and QuantML/UMR plus the existence of a decoding function for QuantML/XML, defined in ISO 24617-12:2025 and Bunt (2023), it follows that QuantML/UMR representations can also be decoded in the QuantML abstract syntax, and thus share the semantics of QuantML/XML.

UMR and its predecessor AMR are commonly viewed as a representation format for first-order logic. Indeed, Bos (2016) presented a first-order semantics for AMR. The enrichment of AMR to UMR with scope links, inspired by QuantML and the extension to generalized quantifiers, as well as with other than individual participation, brings the expressiveness of UMR to a higher level. Since the semantics of the QuantML abstract syntax is second-order (viz. second-order DRT), needed for dealing with generalized quantifiers, it follows that QuantML/UMR representations also have a second-order semantic interpretation.

(21) $F_{U\rightarrow X}$(s / scope
:scoping (sLs / scopeLinks
:op1 (sL1 / scopeLink
:arg1 x1
:arg2 x2
;scopeRel R1)
...
:opk (sLk / scopeLink
:arg1 x1
:arg2 x2
;scopeRel Rk))
::pred (e / event
:name 'eventName'
:arguments (a / argumentStructure
:participant (x1 / entity
:domain d1
:individuation i1
:involvement q1
:semRole sR1
:distributivity di1
:opk (pk / participation
:participant (xk / entity
:domain dk
:individuation ik
:involvement qk
:semRole sRk
:distributivity dik

:individuation ik
: ...)))

= $F_{U\rightarrow X}$((sL1 / scopeLink
:arg1 x1
:arg2 x2
;scopeRel R1))
...
$F_{U\rightarrow X}$((sLk / scopeLink
:arg1 xi
:arg2 xj
;scopeRel Rk))
$F_{U\rightarrow X}$((e / event
:name 'eventName'
:arguments (a / argumentStructure
:op1 (p1 / participation
:participant (x1 / entity
:individuation i1
: ...))
:...
:opk (pk / participation
:participant (xk / entity
:individuation ik
: ...))

= <scope arg1="#x1 arg2="x2" socpeRel="R1">
...
<scope argi="#xi arg2="xj" socpeRel="Rk"/>
<event xml:id="e" pred="eventName"/>
<participation event="#e" participant="#x1" distributivity="d1" semRole="sR1"/>
< entity xxml:id=x1 individuation="i1".../>
...
<participation event="#e" participant="#xk" distributivity="dk" semRole="sRk"/>
< entity xxml:id=x1 individuation="ik".../>
...
etc.

References

J. Barwise and R. Cooper. 1981. Generalized Quantifiers and Natural Language. *Linguistics and Philosophy*, 4:159–219.

J. Bos. 2016. Expressive Power of Abstract Meaning Representations. *Computational Linguistics*, 42:3:527–535.

J. Bos. 2020. Separating Argument Structure from Logical Structure in AMR. In *Proceedings Second International Workshop on Designing Meaning Representations (DMR 2020)*, pages 13–20. Association for Computational Linguistics ACL.

H. Bunt. 2019. An annotation scheme for quantification. In *Proceedings 14th International Conference on Computational Semantics (IWCS 2019)*, pages 31–42, Gothenburg, Sweden.

H. Bunt. 2023. The Compositional Semantics of QuantML. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-19)*, pages 56 – 65, Nancy, France.

H. Bunt. 2024. Combining annotation schemes through interlinking. In *Proceedings ISA-20, Twentieth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 84–94, Turin, Italy.

R. Cooper. 1983. *Quantification and syntactic theory*. Reidel, Dordrecht.

D. Davidson. 1967. The Logical Form of Action Sentences. In N. Resher, editor, *The Logic of Decision and Action*, pages 81–95. The University of Pittsburgh Press, Pittsburgh.

ISO. 2012. *ISO 24617-1: 2012, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 1: Time and events*. International Organisation for Standardisation ISO, Geneva.

ISO. 2014. *ISO 24617-4: 2014, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 4: Semantic roles*. Geneva: International Organisation for Standardisation ISO.

ISO. 2015. *ISO 24617-6:2015, Language Resource Management - Semantic Annotation Framework (SemAF) - Part 6: Principles of semantic annotation*. International Organisation for Standardisation ISO, Geneva.

ISO. 2025. *ISO24617-12:2025, Language Resource Management: Semantic Annotation Framework (SemAF) - Part 12: Quantification*. International Standard. International Organisation for Standardisation ISO, Geneva.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer Academic Publishers, Dordrecht.

K. Lee, H. Bunt, J. Pustejovsky, A. Fang, and Chongwon Park. 2025. Representing ISO-annotated dynamic information in UMR. In *Proceedings of the Sixth International Workshop on Designing Meaning Representations (DMR 2025)*, pages 49–58.

R. Montague. 1971. The proper treatment of quantification in ordinary language. In R.Thomason, editor, *Formal Philosophy*. Yale University Press, New Haven.

T. Parsons. 1990. *Events in the Semantics of English: A Study in Subatomic Semantics*. MIT Press, Cambridge, MA.

V. Petukhova, H. Bunt, et al. 2008. LIRICS Semantic Role Annotation: Design and Evaluation of a Set of Data Categories. In *LREC*.

J. Pustejovsky, H. Bunt, and K. Lee. 2010. ISO-TimeML. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta. ELDA, Paris.

J. Pustejovsky, H. Bunt, and A. Zaenen. 2017. Designing annotation schemes: From theory to model. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 21–72. Springer, Berlin.

J. Pustejovsky, K. Lai, and N. Xue. 2019. Quantification and Scope in Abstract Meaning Representations. In *Proceedings of the First International Workshop on Designing Meaning Representations (DMR 2019), Florence, Italy*, pages 28–33. Association for Computational Linguistics.

J. Van Gysel, M. Vigus, J. Zhao, and N. Xue. 2022. *Uniform Meaning Representation 0.9 Specification*. Unpublished document.

D. Westerståhl. 1985. Determiners and context sets. In Johan van Benthem and Alice ter Meulen, editors, *Generalized Quantifiers in Natural Language*, pages 45–71. Foris, Dordrecht.

Cococorpus: a corpus of copredication

Long Chen **Deniz Ekin Yavaş** **Laura Kallmeyer** **Rainer Osswald**
Heinrich Heine University Düsseldorf
{chen.long,deniz.yavas,laura.kallmeyer,rainer.osswald}@hhu.de

Abstract

While copredication has been widely investigated as a linguistic phenomenon, there is a notable lack of systematically annotated data to support empirical and quantitative research. This paper gives an overview of the ongoing construction of Cococorpus, a corpus of copredication, describes the annotation methodology and guidelines, and presents preliminary findings from the annotated data. Currently, the corpus contains 1500 gold-standard manual annotations including about 200 sentences with copredications. The annotated data not only supports the empirical validation for existing theories of copredication, but also reveals regularities that may inform theoretical development.

1 Introduction

Inherently polysemous nouns have multiple meaning facets. For example, the noun 'breakfast' has an object facet referring to its food reading, and an event facet referring to its dining reading. It is common to analyze its semantic type as a "dot-type" *food • event* (Pustejovsky, 1995; Cruse, 1995). Different meaning facets of a dot-type noun can be predicated by multiple predicates at the same time. This phenomenon is referred to as *copredication* (Pustejovsky, 1995; Asher, 2011). In (1), the verb 'bring' targets the object facet while 'late' targets the event facet.[1]

(1) Go *bring* your new sister some *late* **breakfast**.

Coercion, by contrast, refers to the phenomenon where the predicate targets a facet which is not available in the noun. (2) is an example of coercion. The verb 'resist' targets an event facet, but 'novel' is an *object • info* noun without event facets.

(2) I *resisted* a second **novel** for 14 years until Jack became a way out of a trap I got myself into with a multi-book contract.

Copredication has been studied extensively in linguistics. For example, a number of studies focus on the restrictions and the orders of copredication. Asher (2011) observed an asymmetry in the copredication of the polysemous noun 'city'. Retoré (2014) noticed that *football team* reading of the polysemous noun 'Liverpool' cannot copredicate with other readings. Chatzikyriakidis and Luo (2015) concluded that the copredication of 'newspaper' related to the *organization* reading is relatively restricted compared to its other facets. Ortega-Andrés and Vicente (2019) proposed the concept of 'activation package' to explain the possibility of the copredication over 'school'. Sutton (2022) discovered that the *physical entity* and *eventuality* reading of 'statement' cannot cooccur in a copredication construction, but either reading can copredicate with the *informational content* reading. Murphy (2024) claimed that complexity and coherence are the decisive factors for the predicate order within a copredication. Michel and Löhr (2024) further suggested that context is a more fundamental factor and explained the order of copredication with the notion of "expectation". Chen et al. (2025) proposed a distinction between 'primary facets' and 'secondary facets' to explain the asymmetry in the copredication of *food • event* nouns. However, most of these previous work is based on a small number of introspectively constructed sentences or cherry-picked typical cases, and there is little quantitative research on copredication to prove or disprove the proposed theories. Also, previous work mainly focused on prototypical copredication

[1] The example sentences in the paper are all taken from our annotated data, which come from BookCorpus (Zhu et al., 2015), accessed via https://huggingface.co/datasets/bookcorpus/.

instances, while a lot of borderline cases are not well-represented.

These limitations show the need of an annotated corpus of sentences with copredication. Currently there are few corpora related to copredication. Hanks and Pustejovsky (2005) created a lexicon with corpus usage patterns of words with semantic types information. However, the focus is on verbs instead of polysemous nouns and copredication constructions are not specifically addressed. Alonso et al. (2013) annotated in total 4500 sentences from three languages containing regular polysemous nouns, but most of the sentences contain single predications instead of copredications, and copredication is only treated as a kind of underspecification. Another valuable resource focusing on semantic types targeted by predicates is T-PAS (Typed Predicate Argument Structures; Jezek et al. 2014). T-PAS offers argument structure patterns for Italian verbs, annotated with semantic types. It also provides corpus instances for each verb pattern. However, although T-PAS includes a range of semantic types and is not limited to simple type nouns, instances involving dot-types are relatively infrequent, and the focus is on single-type predications. To our knowledge, no corpus to date has a specific focus on copredication.

In this paper we describe the construction of Cococorpus, an ongoing project that aims at a corpus with copredication and coercion annotations. The current version is restricted to English. Up to now, we have mainly been targeting the annotation of copredication, covering three kinds of dot-types, and we have manually annotated about 1500 gold-standard sentences from BookCorpus (Zhu et al., 2015) where a polysemous noun is predicated by multiple predicates, including around 200 copredication sentences. Cococorpus also contains around 18000 silver-standard sentences acquired through automatic annotation.

2 Annotation overview

For our copredication annotation, we focused on three common dot-types in language: *food • event*, *info • event*, and *object • info*, which are related to the three major ontological categories *phys(ical)-obj(ect)*, *information*, and *event*.[2] These dot-types are selected because they are relatively better-studied and their facets are rather easy to distinguish from each other.

[2] *food* is a subtype of *phys-obj*

For each dot-type, five nouns with relatively high frequency and little ambiguity are selected. The selected nouns are listed in Table 1. At the moment, we focus on the copredication construction V+Adj+N for annotation. The source of our data is BookCorpus (Zhu et al., 2015), not only because it is free and easily accessible with a considerable number of sentences, but also because of the diversity of the genres of the texts and the contemporary, naturalistic language of the texts.

The annotation pipeline includes the following stages:

1. Automatic extraction of target constructions containing the selected nouns
2. Preliminary annotation using the classifier from Yavas et al. (2023)
3. Manual annotation by two trained annotators
4. Disagreement resolution through discussion

More specifically, the annotation focuses on the relationship between each selected noun and the adjective or verb in the sentence that predicates the noun. For example, in (3), the relation between 'ate' and 'breakfast' would be annotated as ARTIFACT,[3] and the relation between 'quick' and 'breakfast' would be annotated as EVENT.

(3) They packed their bags, *ate* a *quick* **breakfast** of dry cereal and headed south.

Our labeling scheme primarily follows the original labels of the classifier. The basic facet selection labels consists of ARTIFACT for the predication over the object facet, INFORMATION for the predication over the info(rmation) facet, and EVENT for the predication over the event facet. To account for cases where a predicate simultaneously targets multiple facets,[4] we added four composite labels: ARTIFACT_INFO, ARTIFACT_EVENT, EVENT_INFO, and ARTIFACT_EVENT_INFO. Furthermore, coercion is distinguished from facet selection during human annotation, so four additional labels are employed: COERCION_ARTIFACT, COERCION_EVENT, COERCION_INFO and COERCION_OTHER, which indicate

[3] This label stands for the object facet. It comes from the T-PAS taxonomy, from which our classifier has been trained on. In this paper the annotation labels are presented in small capitals, and semantic types are presented in italics.

[4] e.g. in 'read a book', 'read' targets both the object facet and the info facet of 'book'.

Dot types	Selected nouns
food • event	breakfast, buffet, dinner, feast, meal
info • event	conversation, lecture, response, speech, submission
object • info	brochure, diary, novel, summary, textbook

Table 1: The selected nouns of each dot-type

coercion to the object facet, event facet, info facet, and other facets, respectively. Although theoretically, facet selection and coercion can both happen in one single predication, and coercion can involve multiple facets, our current annotated corpus does not contain such instances, and therefore the corresponding combined labels are not implemented yet. Additionally, it is usually unclear which facets light verbs target in copredication constructions, so we specifically annotate light verbs with the label LVB and exclude them in our current research, such as the relation between 'have' and 'dinner' in the light verb construction 'have an early dinner'. And for technical reasons, deleting sentences directly during annotation is not possible, so we also employ a label called DELETE to mark the exclusion of a sentence.

3 The construction of the corpus

3.1 Automatic extraction and preliminary classification

We extract sentences containing the candidate nouns from the BookCorpus and parse them using spaCy's transformer-based pipeline for English.[5] Our goal is to identify sentences where candidate nouns are the direct object of a verb (`dobj`) and modified by an adjective (`amod`) simultaneously. Sentences meeting both criteria are classified in the next step using the classifiers developed by Yavas et al. (2023) for copredication detection.

In their study, Yavas et al. (2023) develop multilingual classifiers to identify semantic argument types in both verbal and adjectival predications. They train separate binary support vector machine classifiers for several semantic types. These classifiers employ contextualized word embeddings generated by pre-trained language models, specifically the multilingual RoBERTa model (Conneau et al., 2020). Given their contextualized word embeddings in the sentences as input, the classifiers classify the relation between the predicate and its argument based on the targeted semantic type. Fig. 1 illustrates the classification process. Yavas et al. (2023) demonstrate that these classifiers effectively detect verb-adjective copredications across multiple languages, making them well-suited for our study. Specifically, we employ six classifiers corresponding to three semantic types for both verbal and adjectival predications: *information*, *event*, and *artifact*. An example of copredication detection using the classifiers is illustrated in Fig. 2.

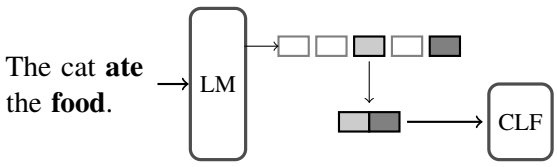


Figure 1: The figure illustrates the working principles of the classifiers developed by Yavas et al. (2023). The classifiers are trained to classify the relation between a predicate and its argument in a sentence using their contextualized word embeddings from a pre-trained language model as input.

3.2 Manual annotation

3.2.1 Annotation platform and format

The manual annotation and adjudication were conducted using the INCEpTION annotation platform (Klie et al., 2018).[6] As illustrated in Fig. 3, the annotation interface displays the automatic annotation, with each relation between a predicate and a noun represented by a labeled directional arrow. Annotators could modify the relations by selecting the corresponding arrows and adjusting the assigned labels through a drop-down menu on the right side of the page.

The adjudication interface (Fig. 4) employs a color-coded system to indicate the annotation status of the sentences. On the left side, the green cells and the white cells mark sentences that require no further modification: Green indicates that the annotators did not change the automatic annotation, and white indicates that they changed it in the same way. Red cells indicate unresolved disagreements.

[5] spacy.io/models/en#en_core_web_trf

[6] inception-project.github.io

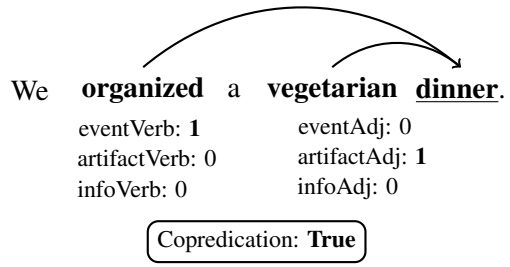


Figure 2: Classification of verbal and adjectival predications in a sentence. Copredication is detected when different types of classifiers assign positive labels to different predications.

The adjudication panel presents a comparative view of both annotators' decisions in the central display area. The adjudicator could either select one of the existing annotations or create a new annotation in the upper panel to establish the final decision.

3.2.2 Annotation guidelines

Our annotation distinguishes facet selection from argument coercion. For example, in (4-a), the predicate 'finish' targets the type event, which is not a meaning facet of the noun 'novel', so the relation between 'finish' and 'novel' is annotated as COERCION_EVENT;[7] in (4-b), although the noun 'novel' does have an info facet, coercion still occurs because according to the context, the content the person 'posted' is the metadata of the novel (e.g. description or advertisement) instead of the actual content, so the relation should be annotated as COERCION_INFORMATION instead of INFORMATION.

(4) a. I really have to *finish* this current **novel** before researching another one.
b. I have *posted* the new **novel** Rome's Evolution on Amazon, B & N, Kobo and Smashwords.

Regarding facet selection, the nouns in the current annotation task are preselected, so it is clear which facets these nouns have. The major annotation task thus reduces to identifying which facets are targeted by given predicates. Operationally, we distinguish selected facets by substitution. Taking the object facet as an example, replace the dot-type noun with physical objects such as 'stone'. The phrase 'throw the stone' is felicitous but '#memorize the stone' is semantically anomalous, which proves that 'throw' targets object facets but 'memorize' cannot.

The facet selection of certain predicates is context-dependent, requiring annotators to determine the most possible targeted facets based on the context from the sentence. As a representative case, the verb 'remember' can target only the object facet of 'breakfast' (as in (5-a)) or target both the event and the object facets of 'meal' simultaneously (as in (5-b)).

(5) a. Her pleased expression tells me she likes that I *remember* her favorite **breakfast**.
b. Once passed the initial security checkpoint it occurred to Jane that she could not *remember* her last **meal**, but there were lines outside all the food stalls.

Some predicates exhibit a high flexibility in meaning facet selection or their predication involve other mechanisms instead of standard facet selection. For example, in 'love the book', any part of the book can be targeted by 'love'; in 'his own book', the adjective 'own' is more focused on the possession relation instead of the facets of 'book'. In such cases, determining the targeted facets becomes both methodologically challenging and theoretically uninformative. Therefore, these predicates were systematically excluded from our annotation. These predicates include:

- 'like' verbs: like, love, hate, prefer
- 'equal' verbs: be, equal, mean
- quoting verbs: say, mention
- adjectives describing type/token: certain, particular, single, same, such
- adjectives describing number: many, much, more, enough, some, few, extra
- adjectives describing entirety/identity: own, whole, entire, complete, actual, real, other, another
- adjectives describing quality: good, nice, fine, best, great, fantastic, wonderful, amazing, bad, awful

[7] While events such as reading and writing are often analyzed as being part of the qualia structure of 'novel', they do not count as formal quale but as telic quale (cf. Pustejovsky 1995, Sect. 6.2); that is, they do not qualify as event facets of the noun.

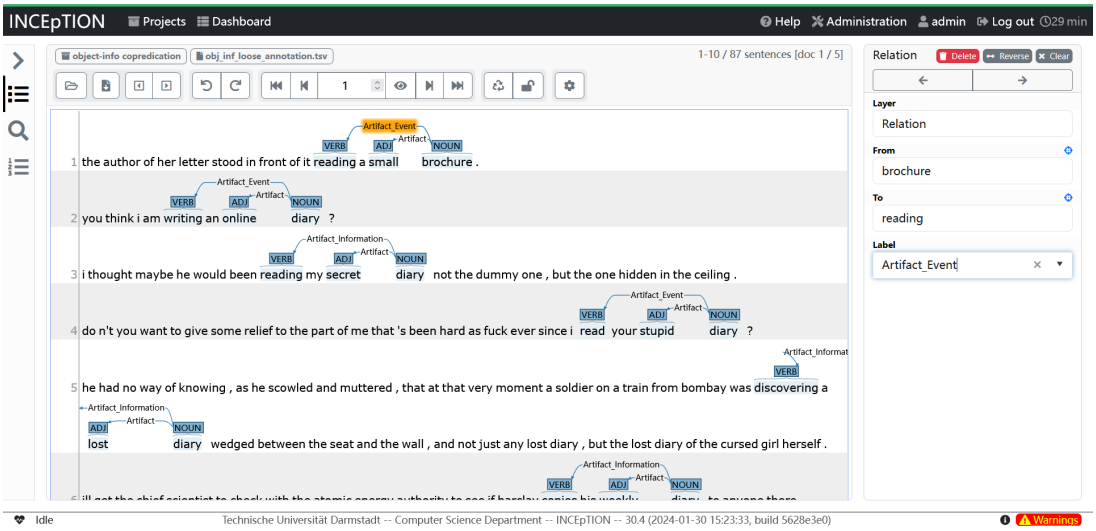


Figure 3: Annotation interface

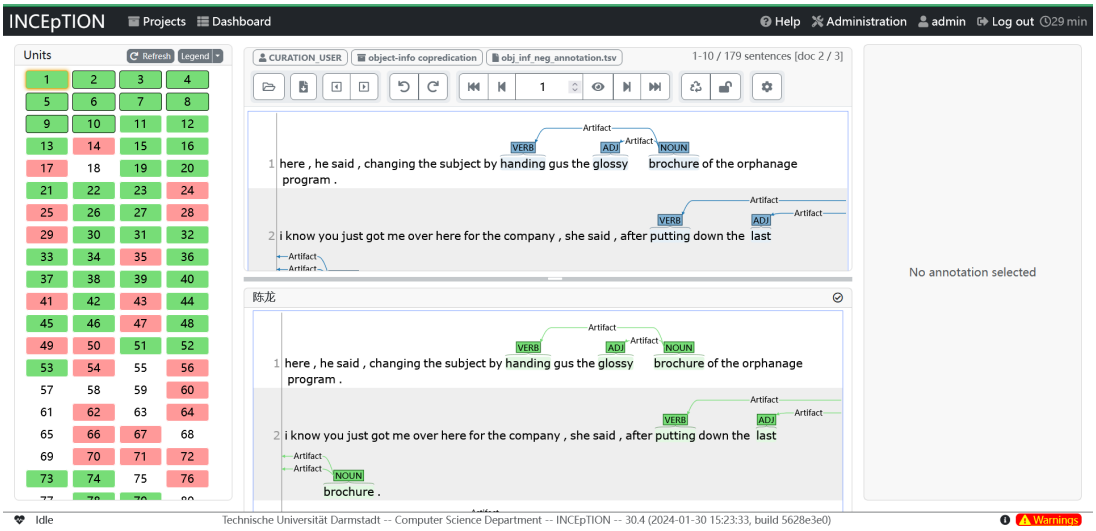


Figure 4: Adjudication interface

4 The analysis of the annotation

4.1 Statistics of the annotated data

The extraction process yielded varying numbers of candidate sentences across different dot-types. For the selected *food • event* nouns, we extracted 6838 sentences with multiple predication from BookCorpus. From this subcorpus, we randomly selected 300 'positive candidates' where the classifier detected copredication (the verb-noun relation and the adjective-noun relation are different) and 300 'negative candidates' (both predicates target the same facet). For *info • event* nouns, we got 9129 sentences with multiple predication, from which we similarly selected 300 positive and 300 negative cases for annotation. *object • info* nouns provided only 832 sentences in total due to the relatively low frequency of the five chosen nouns. Consequently, we selected only 150 'positive' and 150 'negative' candidates for annotation.

	V_{obj}	V_{ev}	V_{both}	$V_{coercion}$	Σ
A_{obj}	303	16	12	1	332
A_{ev}	108	42	17	2	169
A_{both}	2	4			6
Σ	413	62	29	3	507

Table 2: Statistics of the annotation of *food • event* nouns

4.1.1 *food • event*

Among the 600 chosen sentences of the five *food • event* nouns, 507 valid annotations were obtained. Around 100 sentences were excluded,

mainly because they involve light verb constructions or parsing errors. Table 2 reveals the frequency of the predicate tokens that target different facets of the nouns.[8] The majority of verbs (413 out of 507) target the object facet, and approximately two-thirds of adjectives (332 out of 507) target the object facet. Additionally, around 30 sentences contain predicates that simultaneously target both facets. Furthermore, three sentences were identified as involving a coercion.

(6) a. "I *promised* you guys a hot **meal**", said Ellen, sighing.
b. Her expression changes to that of a lioness *stalking* her next **meal**.

(6) presents two annotated instances of coercion. In (6-a), the verb 'promise' typically targets an event facet and the sentence means 'I promised to get you guys a hot meal'. Although the noun 'meal' has an event facet, the type of the event facet is *eating*, which is not the giving event 'promise' targets. Thus, the relation between 'promise' and 'meal' is annotated as COERCION_EVENT. In (6-b), the object of the verb 'stalk' usually needs to be an animate object rather than food, so the predication is annotated as COERCION_OTHER.

While our annotation can provide empirical evidence for some theoretical analyses on copredication, we are not yet able, due to the limited number of dot types and copredication instances, to fully validate or falsify theories related to dot objects with multiple facets (as in Asher, 2011, Ortega-Andrés and Vicente, 2019, Sutton, 2022, etc.) or specific to word items and context (as in Michel and Löhr, 2024). Murphy (2024) proposed a principle called Incremental Semantic Complexity (ISC) and concluded that in copredication, the concrete readings would come earlier than abstract readings (in linear order). It is assumed that physical objects are more concrete than information, and information is more concrete than events. This assumption is supported to a large extent by our annotation statistics on copredication over *food • event* nouns. As shown in Table 2, in most copredication instances, the verb, which precedes the adjective, is targeting the object facet. However, there are still 16 cases where the verb targets the event facet, indicating that the ISC is rather a tendency than a strict principle.

As observed in Chen et al. (2025), for *food • event* nouns in the copredication pattern V+Adj+N, the verb targeting the event facet and the adjective targeting the object facet is a preferred order. When the adjective instead targets the event facet, copredication is only possible when the adjective is *facet-addressing*. Facet-addressing adjectives, contrary to *facet-picking* adjectives, are adjectives that do not affect the availability of a facet of a dot-type noun. For example, 'quick' is a facet-addressing adjective, because 'quick lunch' is still a dot-type and can be copredicated by object-targeting verbs like 'cook' and 'order'. By contrast, 'slow' is a facet-picking adjective since the object facet is not available anymore in 'slow lunch', and '#cook a slow lunch' or '#order a slow lunch' are not acceptable. However, according to Table 2, only 16 cases align with the assumption that object facet is targeted first, whereas there are 108 cases where the copredication works in the less preferred direction. This discrepancy can be partially explained by the dominance of object-targeting verbs in the corpus.

Further analysis of the 108 copredication cases with the event-targeting adjectives reveals that there are only 23 adjective types in the 108 cases. These adjectives fall into the following four semantic categories:

- order related: *first, last, next, fourth, new*
- time related: *quick, slow, early, late, occasional*
- specialness related: *special, customary, regular, unexpected, obligatory, worthy, easy*
- other: *romantic, solitary, civilized, corporate*

Notably, all of them except 'slow' are facet-addressing adjectives. In all the copredication instances involving 'slow', the verb was always 'eat' in our dataset. We discovered in previous studies that 'eat' can also take *food • event* nouns modified by any adjectives as objects, probably due to its relatively light meaning in the context of meals. Moreover, the phrase 'eat a slow meal' could be understood as eating a meal slowly, which makes the phrase felicitous.

[8] V_{obj} stands for the cases where the verb targets the object facet; V_{both} means that the verb targets both facets; $V_{coercion}$ means the verb is annotated as having coercion. Similar interpretations apply to the other symbols of the table.

	V_{inf}	V_{ev}	V_{both}	$V_{coercion}$	Σ
A_{inf}	53	28	1	1	83
A_{ev}	14	189	2		205
Σ	67	117	3	1	288

Table 3: Statistics of the annotation of *info • event* nouns

4.1.2 *info • event*

From the initial set of 600 chosen sentences containing *info • event* nouns, only 288 yielded valid annotations. The low proportion of valid annotations is due to two factors: (1) the selected nouns are deverbal nouns and frequently occur in light-verb constructions, and (2) 'response', 'submission' and 'speech' also have other meanings unrelated to info facets. The distribution of facet selection for these *info • event* nouns is presented in Table 3.

In the 288 instances of multiple predication, there are only 42 instances (14.6%) of copredication. In 14 of the instances, the verb targets the info facet and the adjective targets the event facet; in the other 28 instances copredication works in the other order. This implies either the ISC from Murphy (2024) might be too strict or *information* and *event* are actually close to each other in terms of complexity.

The proportion of copredication over *info • event* nouns is significantly lower than that observed with *food • event* nouns. This is consistent with the observation in Chen et al. (2025) that for *info • event* nouns, both facets tend to be secondary facets, that might be inaccessible if the other facets are targeted first, and copredication in the construction V+Adj+N can only happen when the adjective is facet-addressing.

The 42 copredication instances only include 10 different adjective types. These adjectives can be classified into the following four semantic categories and they are all facet-addressing predicates:

- order related: *last*
- time related: *rapid, earlier, lengthy, little*
- atmosphere related: *bickering, heated*
- speaker related: *private, unwilling, hasty*

4.1.3 *object • info*

The distribution of predications over *object • info* nouns is presented in Table 4, derived from 253 valid annotations out of 300 candidate sentences.

	V_{obj}	V_{inf}	V_{both}	$V_{coercion}$	Σ
A_{obj}	56	2	12		70
A_{inf}	30	31	96	23	180
$A_{coercion}$	1	1	1		3
Σ	87	34	109	23	253

Table 4: Statistics of the annotation of *object • info* nouns

	V_{obj}	V_{inf}/V_{both}	$V_{coercion}$	Σ
A_{obj}	56	14		70
A_{inf}	30	127	23	180
$A_{coercion}$	1	2		3
Σ	87	143	23	253

Table 5: Updated statistics of the annotation of *object • info* nouns

This dot-type exhibits a notably higher frequency of coercion, with the coercion to the event facet being particularly predominant. In 22 of the 23 coercion instances, the verb targets an event facet, suggesting a possible tendency of the direction of coercion.

Regarding copredication, the 32 instances revealed a significant asymmetry, which is consistent with the ISC suggested by Murphy (2024) but seems contradictory to the observation in Chen et al. (2025). According to Chen et al. (2025), there is little restriction on the copredication of *object • info* nouns and copredication can happen in both orders. However, in only two cases in our annotated data, the verb targets the event facet and the adjective targets the object facet. This may be attributed to two factors. First, the frequency of info-targeting adjectives is relatively high (180 out of 253 instances). Secondly, high-frequency verbs like 'read', 'write', and 'publish' are annotated as targeting both facets, resulting in a high number in the third column of the table. Interestingly, in 96 of the 109 instances, the adjective targets the info facet, suggesting that verbs like 'read' probably "mainly" targets the info facet. If we take this into account and combine the cases where the verb targets the info facet and the verb targets both facets, the updated statistics (as in Table 5) reveals a more balanced distribution of copredications.

4.2 Disagreement analysis

The inter-annotator agreement is listed in Table 6. The primary reasons of inter-annotator disagree-

Dot types	Agreement
food • event	0.64
info • event	0.43
obj • info	0.67

Table 6: The inter-annotator agreement in Cohen's Kappa

ment can be summarized as follows.

4.2.1 The exclusion of the sentences

As is in (7), the adjective 'complete' refers to part-whole relations and can target any facets; so one of the annotator followed the guideline and labeled DELETE while the other annotator decided by context, which suggests that 'complete' targets the info facet, and annotated INFORMATION.

(7) To buy the *complete* **novel**, Trail of Storms, click here.

4.2.2 Difficult contexts

In some sentences, the predicate can target both facets of the noun, but the context provided by the sentence is insufficient or complicated, which also results in the disagreement between annotators.

(8) a. Do you *remember* that first **dinner**?
b. He was cooking a *special* **dinner** for her and he had finally found the perfect ring to put on her finger, a heart shaped diamond surrounded by smaller stones and set in platinum.

The annotation of (8-a) presented a challenge due to lack of context. The annotators labeled ARTIFACT and EVENT respectively. The adjudication process selected EVENT as the final decision, because conceptually, it is more plausible to recall a dining event while forgetting specific culinary details than remembering only the food while forgetting the associating eating event. Thus, the event facet is established as the default selection for such contextually underspecified cases in terms of remembering a meal.

In (8-b), the context provides competing clues. The verb 'cook' implies a special food preparation, while the sentential context subsequently indicates the dining experience being 'special', resulting in the divergent annotation of the relation between 'special' and 'dinner'. The final decision is that 'special' targets the event facet, as the contextual evidence provided no substantive indication of unusual food characteristics that warrants an object facet selection. This annotation example reflects the difficulty in the identification of copredication regarding predicates with a wider choice of facets.

4.2.3 Borderline light verbs

There is little consensus regarding the definition of light verb constructions, which is reflected in our annotation of high-frequency *info • event* noun patterns including 'give a speech/lecture', 'make a response/speech/conversation' and 'deliver a speech'. One annotator label them as LVB and the other treat them as regular verb phrases, contributing substantially to the relatively low agreement in the annotation of *info • event* nouns.

To resolve the disagreement, we implemented the diagnostics proposed by Fleischhauer and Neisani (2020), such as replacing the verb with its synonyms and examining the acceptability and the meaning of the new phrase. Application of these diagnostics reveal that the verbs 'deliver', 'give' and 'make' cannot be easily substituted ('#send a speech/lecture', '#produce a conversation' are ungrammatical). Consequently, the verbs mentioned above are regarded as light verbs during the final adjudication.

4.2.4 Unclear distinctions between facets

Some of the disagreement arises from the unclear distinctions between facets, especially the object facet and the info facet of *object • info* nouns.

(9) a. He tried to talk a lot about theories and make funny stories at times to let students feel like they were not drones *downloading* the latest **textbook** that the publishing company decided could be revised for the twelfth time in a row for twelve years straight.
b. During this time, she published a *short* **novel**.

The ontological status of digital texts presents a significant challenge to our annotation, as illustrated in example (9-a), which involves the predication over electronic versions of textbooks rather than traditional physical books. It is arguable whether the PDF file and the strings in the computers are a kind of physical object or more about the information.

A similar puzzle also exists with traditional paper books, as demonstrated in example (9-b). The

adjective 'short' unambiguously targets the info facet of the 'novel', but at the same time, the length of the printed characters in the physical book is also short. The printed symbols are ontologically different from the object facet of the 'novel', which is usually made of paper and consists of covers, and also different from the info facet of the 'novel', which does not have a physical form. Frequent verbs including 'read' and 'write' also have the same problem. The entity a person 'read' in a novel is not the paper material but rather the printed matter on the paper, which is not exactly the object facet of 'novel'. The statistics shown in Fig. 4 also suggests that the facet these verbs target is probably closer to info facet than object facet. Currently, the label INFORMATION is decided for both controversial cases, but these disagreements highlight a need for a comprehensive revisit of the analysis of *object • info* nouns and an investigation of the possible existence of a third meaning facet.

4.2.5 Borderline coercions

The unclear distinction between coercion and facet selection is also a reason for the disagreement between annotators.

(10) a. Anticipating an *angry* **conversation** he would not want to overhear, Mark hurried to the shower.
b. It will handle all your daily chores, provide *intelligent* **conversation** and need absolutely no maintenance.

The examples (10-a) and (10-b) exemplify a systematic pattern of human-targeting adjectives modifying *info • event* nouns, relating to the behavioral manner of the participant during the event. On the one hand, the constructions display typical features of coercion. For example, the syntactic transformation of these phrases are restricted. Transformations such as '?The conversation is angry', '#It provides a conversation that is intelligent' are marginally acceptable or unacceptable. Furthermore, these adjectives cannot easily modify the event facet or info facet of the nouns of other types, e.g. '?an angry book' requires some context to be acceptable, and '#an intelligent meeting' is infelicitous.

On the other hand, the treatment of these cases as facet selection can also be justified. First, the interpretation of the phrase is specific, unlike the typical coercion examples such as 'finish the book', where the event that is 'finished' is implicit and needs to be specified by context. Second, the usage of human-targeting adjectives for *info • event* nouns is productive. Other adjectives of this kind including 'cheerful', 'friendly', 'polite', and 'honest' modifying *info • event* nouns also exist in our annotated data.

Currently, these instances are annotated as facet selection. A more comprehensive analysis and annotation on coercion will be left for future research.

5 Conclusion and future work

The construction of Cococorpus is an ongoing project. Currently, we achieved an annotation of more than 1000 sentences with multiple predications over inherently polysemous nouns, among which 198 sentences exhibit copredication. The annotated data can serve as an empirical evidence for some linguistic analyses on copredication, such as a tendency of ISC from Murphy (2024) and the distinction between primary and secondary facets of polysemous nouns from Chen et al. (2025). The annotated data and annotation guideline are published at the project Github page (https://github.com/CoCoCo-Project).

Many aspects of copredication remain to be addressed in future annotation efforts. The current annotation framework is limited to the construction V+Adj+N, while other typical copredication constructions, such as (reduced) relative clauses and multiple adjectives, are yet to be incorporated. Furthermore, the existing coverage of dot-types and nouns is also relatively limited. We plan to expand our annotation scope to include: (1) additional dot-types and lexical items, particularly nouns exhibiting multiple facets (e.g., *school*, *city*) and those with debatable facet classifications (e.g., *annotation*); (2) a broader range of predicate types; (3) cross-linguistic investigations to examine potential variations in copredication phenomena across different languages.

Acknowledgements

This work is part of the project "Coercion and Copredication as Flexible Frame Composition" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number 440934416. We would like to thank the anonymous reviewers for their valuable comments.

References

Héctor Martínez Alonso, Bolette Sandford Pedersen, and Núria Bel. 2013. Annotation of regular polysemy and underspecification. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 725–730.

Nicholas Asher. 2011. *Lexical meaning in context: A web of words*. Cambridge University Press.

Stergios Chatzikyriakidis and Zhaohui Luo. 2015. Individuation criteria, dot-types and copredication: A view from modern type theories. In *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 2015)*, pages 39–50.

Long Chen, Laura Kallmeyer, and Rainer Osswald. 2025. Primary vs. secondary meaning facets of polysemous nouns. *Empirical issues in syntax and semantics: Selected papers from CSSP 2023*, page 27.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

D. Alan Cruse. 1995. Polysemy and related phenomena from a cognitive linguistic viewpoint. In Patrick Saint-Dizier and Evelyne Viegas, editors, *Computational lexical semantics*, pages 33–49. Cambridge University Press.

Jens Fleischhauer and Mozhgan Neisani. 2020. Adverbial and attributive modification of persian separable light verb constructions. *Journal of Linguistics*, 56(1):45–85.

Patrick Hanks and James Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de linguistique appliquée*, 10(2):63–82.

Elisabetta Jezek, Bernardo Magnini, Anna Feltracco, Alessia Bianchini, and Octavian Popescu. 2014. T-PAS; a resource of typed predicate argument structures for linguistic analysis and semantic processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 890–895, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna 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.

Christian Michel and Guido Löhr. 2024. A cognitive psychological model of linguistic intuitions: Polysemy and predicate order effects in copredication sentences. *Lingua*, 301:103694.

Elliot Murphy. 2024. Predicate order and coherence in copredication. *Inquiry*, 67(6):1744–1780.

Marina Ortega-Andrés and Agustín Vicente. 2019. Polysemy and co-predication. *Glossa: a journal of general linguistics*, 4(1).

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Christian Retoré. 2014. The Montagovian Generative Lexicon ΛTy_n: a type theoretical framework for natural language semantics. In *TYPES: International Workshop on Types and Proofs for Programs, April 2013, Toulouse, France*, pages 202–229.

Peter Roger Sutton. 2022. Restrictions on copredication: a situation theoretic approach. In *Proceedings of the 32nd Semantics and Linguistic Theory Conference*, pages 335–355.

Deniz Ekin Yavas, Laura Kallmeyer, Rainer Osswald, Elisabetta Ježek, Marta Ricchiardi, and Long Chen. 2023. Identifying semantic argument types in predication and copredication contexts: A zero-shot cross-lingual approach. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 310–320.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

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

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)

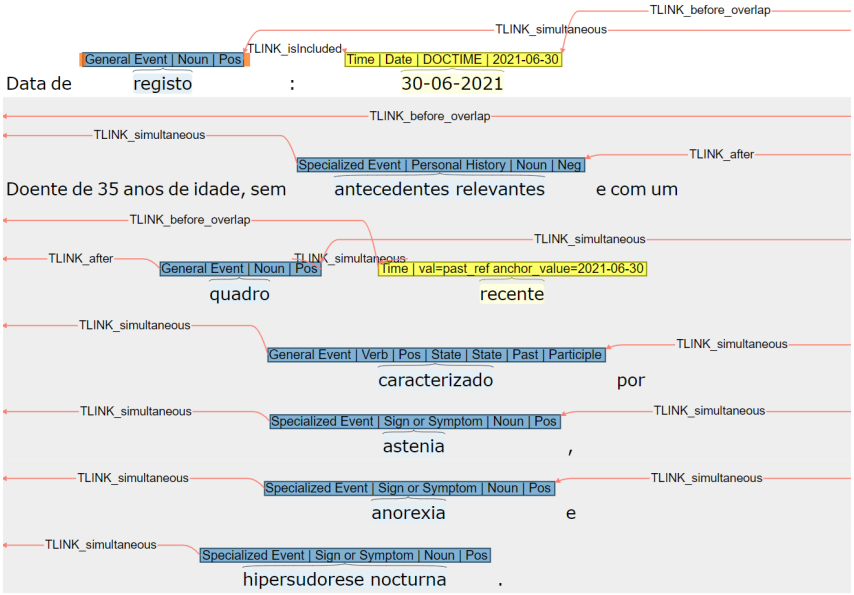


Figure 1: Annotation of an excerpt from a medical report using the Med2Story scheme. Events are marked in blue and temporal expressions in yellow. The annotated excerpt illustrates the identification of various attributes associated with both events and temporal expressions, as well as the temporal relations 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.

Acknowledgments

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the support UID/50014/2023

(https://doi.org/10.54499/UID/50014/2023). The authors also acknowledge the support of the StorySense project (DOI 10.54499/2022.09312.PTDC).

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

Enhancing ISO 24617-2: Formalizing Apology and Thanking Acts for Spoken Russian Dialogue Annotation

Ksenia Klokova
MIPT
Moscow, Russia
klokova.ks@mipt.ru

Anton Bankov
Independent researcher
Moscow, Russia
ant.s.bankov@gmail.com

Nikolay Ignatiev
HSE, MIPT
Moscow, Russia
ignatievnickolay@gmail.com

Abstract

This paper refines ISO 24617-2's Social Obligations Management dimension by formalizing apology and thanking acts for Russian dialogue annotation. Addressing gaps in formal definitions and limited response strategies, we propose culture-neutral semantic cores using Wierzbicka's universal primes and update semantics. We introduce three response functions: address (minimal acknowledgment), downplay (mitigation), and decline (reinforcement). Validated through qualitative analysis, this framework captures empirical strategies—including non-response, formulaic minimization, and strategic obligation maintenance—unaddressed in the current standard. Our approach maintains ISO compatibility while eliminating unsubstantiated elements like obligatory response pressure, enhancing annotation accuracy for Russian dialogue.

1 Introduction

Natural dialogue involves nuanced negotiation of social obligations that extends beyond binary frameworks. While ISO 24617-2:2020 provides a comprehensive framework for Social Obligations Management (SOM), its treatment of core functions like *apology* and *thanking* reveals opportunities for refinement. The standard currently lacks formal definitions for these functions and offers limited coverage of response strategies beyond acceptance — a gap particularly evident when applied to casual spoken dialogues.

Building on Wierzbicka's universal primes and empirical observations, we propose a formalization of the *apology/thanking* and the hierarchy of respective response functions. This approach accommodates cross-linguistic variation while addressing empirical observations from Russian dialogue data, where conventional response taxonomy fail to capture strategies like non-committal addressing, downplaying, or explicit declination. By providing formal definitions and expanded response taxonomies for apology and thanking functions, this work offers a framework that can be utilized in implementing more nuanced social reasoning in automated systems.

Our primary contributions are:

- Formal definitions for refined *apology* and *thanking* using update semantics compatible with ISO 24617-2.
- Extended response taxonomy introducing *addressApology/Thanking*, *downplayApology/Thanking*, and *declineApology/Thanking* functions.
- Validation framework demonstrating applicability to Russian through qualitative analysis of movie dialogues.

This paper is structured as follows: Section 2 reviews ISO 24617-2's SOM dimension and prior work; Section 3 establishes our theoretical foundation in politeness and semantic primitives; Section 4 describes the Russian Multimedia Politeness Corpus and annotation methodology; Sections 5 and 6 present formalizations and case studies for apologies and thankings, respectively; Section 7 discusses implications for dialogue annotation standards.

2 Related Work

The ISO 24617-2 standard (Bunt et al., 2012, 2020; ISO Central Secretary, 2020) is a multidimensional dialogue act annotation scheme based on the Dynamic Interpretation Theory (DIT) (Bunt, 2000b) and the DIT++ taxonomy (Bunt, 2009). It provides the semantic framework for the analysis of the communicative behaviour of dialogue participants and

has been sucessfully applied to the dialogue act annotation in a number of languages (Petukhova et al., 2014; Yoshino et al., 2018; Ngo et al., 2018; Roccabruna et al., 2021; Oleksy et al., 2022; Hwaszcz et al., 2023).

Communicative functions in the Standard are mapped to 10 different dimensions across 5 contexts (Bunt, 2000a). These dimensions are responsible for the information about the task or activity which motivates the dialogue, including dealing with social obligations. The corresponding Social Obligations Management dimension includes functions related to greeting, introduction, apology, thanking, leave-taking, etc. The extended taxonomy suggested by Gilmartin et al. (2017, 2018) includes functions to account for the common social intentions such as politeness questions.

Communicative functions are defined by their update semantics: updates they impose onto addressee's context (Bunt, 2011; Petukhova, 2011). These updates consist of basic semantic concepts called semantic primitives. There are both general-purpose semantic primitives, which can be used in forming updates for communicative functions in any of the dimensions, and dimension-specific primitives. Although the ISO Standard does not provide formal definitions of communicative functions in terms of these updates, they were used during the process of creating the Standard (Bunt et al., 2010). To our knowledge, the specified formal definitions of the communicative functions (Bunt, 2012, 2014) did not include the definitions for *apology*, *thanking* and their response functions.

3 Pragmatic Perspectives on Apology and Thanking

While apologies universally function as remedial acts addressing normative violations through expressions of regret and responsibility acceptance, their specific linguistic realizations and contextual applications are culture-bound. While the politeness autonomy-based framework (Brown and Levinson, 1987) and cross-linguistic preconditions (speaker involvement, recognized breach, perceived harm) (Blum-Kulka and Olshtain, 1984) offer valuable analytical tools, they do not provide a framework for cross-cultural generalization.

Similarly, thanking acts manifest culture-specific expressions of benefit acknowledgment and debt management (Coulmas, 1981), operating as positive politeness strategies within the universal politeness framework (Brown and Levinson, 1987). Both speech acts fundamentally negotiate social valence—apologies repairing negative equilibrium, thankings reinforcing positive bonds — yet their concrete realizations vary cross-culturally.

Wierzbicka (1991) proposed that universal primes provide the most suitable framework for capturing this variation, reducing apology/thanking to:

> I did/didn't do something (to/for you).
> I feel something bad because of this.
> (Wierzbicka, 1991, p. 126)

These culture-neutral explications avoid ethnocentric presuppositions inherent in other models.

However, any operationalization grounded in real language necessitates culture-linked correlates. When formalizing definitions and annotating examples, naming associated feelings is unavoidable. In the proposed formalization (Sections 5.2 and 6.2):

- Apology operationalized through the semantic primitive *regret* (Petukhova, 2011, p. 158), which encapsulates Wierzbicka's definition. A corresponding emotional correlate that dominates Russian interactions concerning breaches of social norms is *guilt*.
- Similarly, in thanking we operate through the primitive that describes the state of being *grateful* (Petukhova, 2011, p. 158) and its associated correlate *indebtedness*.

While *regret*/*guilt* and *grateful*/*indebtedness* may serve as operational correlates in other languages, we acknowledge that these specific emotional mappings reflect Russian cultural patterns and should not be assumed universal. Our methodological contribution lies in proposing that annotation of apology- and thanking-related acts in any language should identify and employ the specific core emotions that culturally drive these speech act realizations.

4 Data and Annotation

This study leverages dialogues from the Russian Multimedia Politeness Corpus (Klokova et al., 2023), a resource designed to model politeness phenomena in contemporary spoken Russian. The RMPC comprises manually transcribed excerpts from modern Russian films, capturing five core

politeness scenarios: greetings, acquaintance rituals, apologies, thankings, and leave-takings. Each transcript includes punctuation annotation and is enriched with paralinguistic features — specifically non-verbal markers (e.g., gestures, facial expressions) and sociopragmatic variables grounded in foundational sociolinguistic frameworks (Brown and Levinson, 1987; Helfrich, 1979; Holmes, 1995; Mills, 2003).

For the purpose of this study, we initially applied the ISO 24617-2:2020 annotation framework to apology and thanking sequences within the corpus. The discovered limitations in the current primary (*apology, thanking*) and response (*acceptApology/Thanking*) functions necessitated an iterative extension of the taxonomy. The extended scheme was then tested through targeted annotation of salient sequences — rather than full dialogue application — to validate its descriptive adequacy.

4.1 Annotation Procedure

We selected 110 dialogues containing apology-relevant frames and 105 dialogues containing thanking-relevant frames from the corpus. Each dialogue was independently annotated by two annotators, with a third annotator performing final reconciliation of disagreements.

Our annotation approach focused exclusively on segments that instantiated the core emotional and pragmatic functions of apologizing and thanking, conceptualized through the emotional correlates of 'regret' (guilt) and 'gratitude' (indebtedness), respectively. These states were primarily inferred through linguistic cues (formulae and explicit admissions of such feelings). The corresponding (un)resolved states were inferred through subsequent conversational context, the speaker's intonation, and non-verbal signals.

This framework required excluding supportive strategies that, while co-occurring within apology or thanking sequences, serve different pragmatic functions. For apologies, we excluded strategies such as offers of repair or accounts that lack essential expressions of regret or guilt acknowledgment. Similarly, for thankings, we excluded strategies like exclamations of surprise or compliments that do not signal indebtedness. Although such strategies might substitute for nuclear expressions, their reliable annotation presents considerable challenges due to dependence on multiple contextual variables, ranging from the scale of the wrong- or right-doing to the speaker's actual psychological state.

This principled exclusion maintains theoretical coherence by distinguishing core speech acts from accompanying strategies. Conflating these would obscure the distinct pragmatic mechanisms underlying different politeness phenomena and prevent accurate identification of linguistic realizations of regret and gratitude.

Following this filtering, our final dataset comprised 103 dialogues with apology acts and 92 with thanking acts. Inter-annotator agreement (Krippendorff's alpha) reached 0.92 for all segments and 0.98 for matching segments, indicating high reliability. Representative examples are analyzed in Sections 5.3 and 6.3.

5 Apology

5.1 Current ISO 24617-2 Definition

Current ISO 24617-2 offers two communicative functions for annotating apology interactions: *apology* and *acceptApology*, which refers to the downplay response strategy. Their definitions are given as follows:

> **/apology/:** Communicative function of a dialogue act performed by the sender, S, in order to signal that he/she wants the addressee, A, to know that S regrets something; S puts pressure on A to acknowledge this.
>
> **/acceptApology/:** Communicative function of a dialogue act performed by the sender, S, in order to mitigate the feelings of regret that the addressee, A, has expressed.

Empirical evidence from our data reveals frequent absence of responses to apologies (Section 5.3). This demonstrates that apologies do not universally impose response pressure on the addressee, contrary to the current specification. We therefore refine the definition of the *apology* function by eliminating the pressure-to-respond component.

5.2 Formalization

The exact formalized preconditions of the apology-related communicative functions are provided in the Table 1.

Apology consists of the elementary update functions which inform the addressee of the *regret* (Section 3), which the speaker experiences, and their

Comm. Function	Update Semantics	Explanation
Apology	$\mathbf{Regret}(S, \mu)$ $\mathbf{Want}(S, \mathbf{Bel}(A, \mathbf{Regret}(S, \mu)))$	Sender, S, informs the addressee, A, that S regrets some action or information, μ
Address Apology	$\mathbf{Bel}(S, \mathbf{Regret}(A, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Regret}(A, \mu))))$	Sender, S, acknowledges addressee's, A, regret for some action or information, μ
Downplay Apology	$\neg\mathbf{Want}(S, \mathbf{Regret}(A, \mu))$ $\mathbf{Bel}(S, \mathbf{Regret}(A, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Regret}(A, \mu))))$	Sender, S, acknowledges addressee's, A, regret for some action or information, μ, and wants to mitigate the A's feelings
Decline Apology	$\mathbf{Want}(S, \mathbf{Regret}(A, \mu))$ $\mathbf{Bel}(S, \mathbf{Regret}(A, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Regret}(A, \mu))))$	Sender, S, acknowledges addressee's, A, regret for some action or information, μ, and wants to reinforce the A's feelings

Table 1: Formalized preconditions for the proposed communicative functions of apology and response to apology (S = sender, A = addressee, μ = some action or information)

desire to communicate this *regret* to the addressee. Given that an apology does not imply pressure to respond, the elementary update function responsible for the pressure was omitted.

In order to account for possible response strategies, we propose the following hierarchy of *apology*-related response functions: *addressApology* is the broader concept and *downplayApology* and *declineApology* are its conceptual domain (see Figure 1).

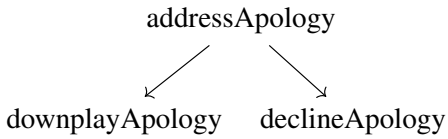


Figure 1: Apology responses

The *addressApology* function is responsible for the more general and non-committal responses. Thus, its formalized preconditions consist only of the elementary update functions corresponding to the acknowledgement of the *regret*, expressed by the addressee with the *apology* function previously in the dialogue. The *downplayApology* function corresponds to the *acceptApology* in ISO 24617-2 and *Apology-downplay* in DIT++. In this function, the sender wants to mitigate the addressee's feeling of *regret*, i.e. the sender communicates that they do not want the addressee to experience *regret*; thus, $\neg\mathbf{Want}(S, \mathbf{Regret}(A, \mu))$ elementary update function is present. On the contrary, in *declineApology* the sender wants to reinforce the addressee's feeling of *regret*; thus, $\mathbf{Want}(S, \mathbf{Regret}(A, \mu))$ elementary update function is used.

5.3 Response Strategies to Apology

Our annotation identified 140 segments as apology acts. The majority of apologies (99 instances, 71%) received no explicit response from the recipient. Among the responses that did occur, downplaying emerged as the most frequent strategy, accounting for 17% (24 instances) of all apologies, followed by declining at 8% (11 instances) and addressing at 6% (9 instances). The remaining three cases involved reciprocal apologies, where recipients responded with their own apology acts.

5.3.1 Apology without Response

Our annotation reveals contexts where apologies elicit no response, particularly when functioning as discourse organizers rather than debt-negotiation acts (1). Other possible contexts involve apologizing for misspeaking, at the end of an interaction (2), and preemptive apologies.

(1) [Rus] Дмитрий: *Извините*... Мне бы Галаганову Нину Сергеевну найти? Девушка: В зале она, молодой человек.

[Translation] Dmitry: *Excuse me*... I'm looking for Nina Sergeevna Galaganova. [SOM:apology]

Young woman: She's in the hall, young man.

(2) [Rus] Олег: Всё, ухожу, ухожу. *Извини, пожалуйста*. *уходит*

[Translation] Oleg: Alright, I'm going, I'm going. *Forgive me, please*. *leaves* [SOM:apology]

Example (2) demonstrates that pressure within the *apology* function is not merely ignored by the addressee but is absent from the speaker's communicative intent. In this instance, the speaker concludes a conflictual exchange for which they bear responsibility. While the speaker's utterance stems from guilt and constitutes an apology act, the speaker simultaneously terminates the dialogue — a move that would be incompatible with exerting pressure on the addressee to elicit a response to the apology.

5.3.2 Addressing Apologies

Addressing constitutes non-committal acknowledgment without guilt mitigation, typically realized through minimal tokens (3) or clarification requests (4) that maintain rather than resolve guilt conditions or avoid evaluating the offense's validity.

(3) [Rus] Маша: *Извини*, я не могла раньше, честное слово.
Костя: *Угу*.

[Translation] Masha: *Sorry*, I couldn't make it earlier, I swear. [SOM:apology]
Kostya: *Uh-huh*. [SOM:addressApology]

(4) [Rus] Маша: Виктор Сергеевич, *простите меня*.
Виктор: *За что*?

[Translation] Masha: Viktor Sergeyevich, *forgive me*. [SOM:apology]
Viktor: *For what*? [SOM:addressApology]

5.3.3 Downplaying Apologies

Downplaying responses actively mitigate the apologizer's guilt burden, characterized by an intention to reduce perceived offense severity. Common formulaic apology minimizers could include ничего страшного (*no worries*) or все в порядке (*it's ok*) as in Example (5).

(5) [Rus] Фима: *Извините, что поздно звоню*. Вам говорить удобно?
Вадим: Да, Фима. *Ничего, ничего*.

[Translation] Fima: *Sorry for calling so late*. Can you talk? [SOM:apology]
Vadim: Yes, Fima. *It's alright, no problem*. [SOM:apologyDownplay]

Two other strategies are shown in Example (6): the speaker employs apology minimization through jocular disbelief followed by imperative termination, systematically dismantling the addressee's guilt assertion.

(6) [Rus] Иван: *Прости нас с матерью*, сынок.
Дмитрий: *Да ты чё*, бать? (...)
Иван: *Всё, что с тобой случилось, это наша вина*.
Дмитрий: *Перестань*, бать.

[Translation] Ivan: *Forgive me and your mother*, son. [SOM:apology]
Dmitry: *What're you on about*, Dad? [SOM:apologyDownplay] (...)
Ivan: *All that happened to you it's our fault*. [SOM:apology]
Dmitry: *Stop it*, Dad. [SOM:apologyDownplay]

5.3.4 Declining Apologies

The *decline* response type maintains or reinforces guilt rather than alleviates it. The most straightforward examples of such utterances would be я тебя не прощаю (*I don't forgive you*) or мне не нужны твои извинения (*I don't need your sorry*). Other possible strategies could involve highlighting negative consequences or costs incurred, and establishing avoidable fault by specifying an alternative action (as in (7)). In Example (8) the speaker declines an apology by postponing resolution.

(7) [Rus] Маша: Максим, *извините*, вы, наверное, меня не дождались?
Максим: *Я час ждал*.
Маша: Ой, а меня на работе задержали.
Максим: *Могли бы позвонить*.

[Translation] Masha: Maxim, *I'm sorry*, you probably didn't wait for me, did you? [SOM:apology]
Maxim: *I waited for an hour*. [SOM:declineApology]

Masha: Oh, they kept me late at work.
Maxim: *You could have called* [SOM:declineApology].

(8) [Rus] Олег: Кать, *прости меня*.
Катя: Олег, *давай всё потом, пожалуйста*? Я тебя очень прошу, ну я очень хочу спать.

[Translation] Oleg: Katya, *forgive me*. [SOM:apology]
Katya: Oleg, *let's talk later, please*? I'm begging you, come on... I really want to sleep. [SOM:declineApology]

6 Thanking

6.1 Current ISO 24617-2 Definition

Similar to apology, ISO 24617-2 offers two communicative functions for annotating thanking interactions: *thanking* and *acceptThanking*, which refers to the downplay response strategy. Their definitions are given as follows:

> **/thanking/:** Communicative function of a dialogue act performed by the sender, S, in order to inform the addressee, A, that S is grateful for some action performed by A; S puts pressure on A to acknowledge this.
>
> **/acceptThanking/:** Communicative function of a dialogue act performed by the sender, S, in order to mitigate the feelings of gratitude which the addressee, A, has expressed.

However, once again, our data shows that non-committal and even declining response strategies to thanking are possible and that thanking does not necessarily pressure the addressee to respond. For illustrative examples, refer to Section 6.3.

6.2 Formalization

The exact formalized conditions of the thanking-related communicative functions are provided in the Table 2.

Similar to the *apology* function, *thanking* function consists of the elementary update functions, informing the addressee of the *gratitude*, which the sender experiences, and their desire to communicate this *gratitude*. It does not include the elementary update function responsible for the pressure to acknowledge addressee's *gratitude* in response.

The hierarchy of response communicative functions to *thanking* is similar to that of *apology*: *addressThanking* is the broader concept, while *downplayThanking* and *declineThanking* are its conceptual domain (see Figure 2).

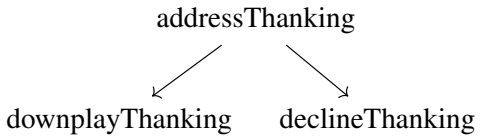


Figure 2: Thanking responses

The *addressThanking* function occurs in more general and non-committal responses and its formalized conditions consist only of the elementary update functions corresponding to the acknowledgement of the *gratitude*, expressed by the addressee. In *downplayThanking* (*acceptThanking* in ISO 24617-2 and *Thanking-downplay* in DIT++), the sender wants to mitigate the addressee's feeling of *gratitude*; thus, $\neg\mathbf{Want}(S, \mathbf{Grateful}(A, T, \mu))$ elementary update function is present. Contrary to downplaying, in *declineThanking* the sender wants to reinforce the addressee's feeling of *gratitude* (for example, expecting favor in the future); thus, $\mathbf{Want}(S, \mathbf{Grateful}(A, T, \mu))$ elementary update function is used.

6.3 Response Strategies to Thanking

Our annotation resulted in 118 segments classified as thanking acts. Most expressions of gratitude (82 instances, 70%) went without explicit acknowledgment. When responses were present, addressing was the predominant strategy at 16% (19 instances) of all thankings, while downplaying occurred in 5% (6 instances) and declining in 3% (3 instances). Similar to apologies, the remaining eight cases featured reciprocal thanking.

6.3.1 Thanking Without Reply

Similar to apologies, our annotation of thanking sequences reveals contexts where no reply is pragmatically required. Particularly, this is the case in casual interactions or when the gratitude expression terminates the conversational exchange (e.g., service encounters or farewells). This absence of response may be conditioned by: 1) conversational position – terminal thanking acts often lack replies;

Comm. Function	Update Semantics	Explanation
Thanking	$\mathbf{Grateful}(S, T, \mu)$ $\mathbf{Want}(S, \mathbf{Bel}(A, \mathbf{Grateful}(S, T, \mu)))$	Sender, S, informs the addressee, A, that S is grateful to some person(s), T, for some action or information, μ (most often T coincides with A)
Address Thanking	$\mathbf{Bel}(S, \mathbf{Grateful}(A, T, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Grateful}(A, T, \mu))))$	Sender, S, acknowledges addressee's, A, gratitude to some person(s), T, for some action or information, μ
Downplay Thanking	$\neg\mathbf{Want}(S, \mathbf{Grateful}(A, T, \mu))$ $\mathbf{Bel}(S, \mathbf{Grateful}(A, T, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Grateful}(A, T, \mu))))$	Sender, S, acknowledges addressee's, A, gratitude to some person(s), T, for some action or information, μ, and wants to mitigate the A's feelings
Decline Thanking	$\mathbf{Want}(S, \mathbf{Grateful}(A, T, \mu))$ $\mathbf{Bel}(S, \mathbf{Grateful}(A, T, \mu))$ $\mathbf{Bel}(S, \mathbf{Want}(A, \mathbf{Bel}(S,$ $\mathbf{Grateful}(A, T, \mu))))$	Sender, S, acknowledges addressee's, A, gratitude to some person(s), T, for some action or information, μ, and wants to reinforce the A's feelings

Table 2: Formalized conditions for the proposed communicative functions of thanking and response to thanking (S = sender, A = addressee, T = some person(s), μ = some action or information)

2) socio-relational factors – familiarity (short social distance), age and hierarchical disparity may override the expectation of a response. Example (9) demonstrates this pattern in a child-adult interaction with a clear status asymmetry.

(9) [Rus] Девочка: Дядя, а можно, пожалуйста, мячик?
Игорь: Конечно.
Девочка: *Спасибо*.

[Translation] Little girl: Mister, can I have the ball please?
Igor: Of course.
Little girl: *Thank you*. [SOM:thanking]

6.3.2 Addressing Thanking

Addressing responses function as a minimal non-committal response (*gratitude* is recognized without an attempt to alleviate or reinforce the speaker's expressed obligation). They often involve formulaic markers пожалуйста (*welcome*), interjections угу (*uh-huh*) (10).

(10) [Rus] Егор: Я понимаю, что достал. В общем, *спасибо, что помогаешь мне*.
Катя: *Угу*. Давай, пока.

[Translation] Yegor: I know I'm being annoying. Anyway, *thanks for helping me out*. [SOM:thanking]
Katya: *Uh-huh*. Alright, see ya. [SOM:addressThanking]

Example (11) illustrates strategic addressing in conflict discourse. Here, the response *I see* acknowledges the expression of *gratitude* and avoids mitigation.

(11) [Rus] Алина: *Я очень благодарна твоей маме*.
Борис: *Я вижу*.

[Translation] Alina: *I'm really grateful to your mom*. [SOM:thanking]
Boris: *I see*. [SOM:addressThanking]

6.3.3 Downplaying Thanking

We operationalize *thankingDownplay* as a response only when the speaker actively minimizes the addressee's debt acknowledgment. This could be achieved with intentional mitigation via explicit verbal cues (e.g., rhetorical questions, minimizers) that negate the need for *gratitude* (12). Common minimizers could include phrases like без проблем (*no problem*) or забей (*forget it*).

(12) [Rus] Отец Кати: (...) *Спасибо, что доехали до меня*.
Катя: Пап, *ну ты шутишь что ли*? *Ну как же бы мы не доехали*?

[Translation] Katya's father: And I love you too. *Thank you for coming to see me*. [SOM:thanking]
Katya: Dad, *are you kidding*? [SOM:thankingDownplay]

How could we not come? [SOM:thankingDownplay]

In contrast to formulaic expressions, a downplay can also be characterized by pragmatic markedness. In (13) there is a three-part thanking sequence which emphasizes the debt acknowledgment, followed by a blatant rejection of the debt frame.

(13) [Rus] Алина: *Спасибо*, дорогой мой. Золотце моё, *ты меня очень сильно выручил*. *Я твоя должница*.
Юра: *Да иди ты*, *что ты говоришь*? Всё, давай, целую, до понедельника.

[Translation] Alina: *Thank you* [SOM:thanking], my dear. My darling one, *you really helped me out big time* [SOM:thanking]. *I owe you one*. [SOM:thanking]
Yura: *Screw you* [SOM:thankingDownplay], *what are you talking about* [SOM:thankingDownplay]? Alright then, kisses, see you Monday.

6.3.4 Declining Thanking

Declining responses actively maintain or intensify debt obligations. Possible strategies include the affirmation of the outstanding obligation (и правильно (*rightly so*)) or explicit debt reminders (не забудь, кто тебе помог – *don't forget who helped you*). Example (14) illustrates the strategic avoidance of debt closure. The speaker's topic shift *Shall we go?* occurs as a response to preemptive thanking before request fulfillment. By doing so, she acknowledges the gratitude and creates conditional obligation (debt remains pending).

(14) [Rus] Борис: Всё-таки, пожалуйста, не кричи на неё. Мы давно живём, я хорошо знаю, когда ты громко говоришь, когда кричишь.
Алина: *молчит*
Борис: *Спасибо*.
Алина: *Поехали*?

[Translation] Boris: Still, please don't yell at her. We've lived together long enough - I know well when you're just speaking loudly and when you're actually shouting.
Alina: *remains silent*
Boris: *Thank you*. [SOM:thanking]
Alina: *Shall we go*? [SOM:declineThanking]

7 Conclusion

In this paper we proposed the refinement to the ISO 24617-2's Social Obligations Management dimension through formal extensions for apology- and thanking-related functions. By grounding definitions in Wierzbicka's universal primes, we established culture-neutral semantic cores for apology and thanking, while accommodating Russian-specific realizations through correlates like *guilt* and *indebtedness*.

Our key contributions — the *address*, *downplay*, and *decline* response functions — resolve empirical gaps in the current Standard, enabling precise annotation of strategies observed in Russian dialogues: non-response (e.g. in the opening positions), formulaic downplaying, and strategic declination. The proposed formalizations maintain ISO compatibility through update semantics while omitting unsubstantiated elements like automatic "pressure to respond".

Future work involves extending our research to other Social Obligations Management functions, such as leave-taking and greetings, using the same methodology of culture-neutral formalization grounded in empirical data. The second objective lies in complete annotation of the communicative functions in the dialogues from the Russian Multimedia Politeness Corpus. Ultimately, this research aims to contribute to the development of evaluation frameworks that can assess how well conversational agents, large language models, and other dialogue systems understand and deploy appropriate politeness strategies within specific cultural and contextual parameters.

References

Shoshana Blum-Kulka and Elite Olshtain. 1984. Requests and apologies: A cross-cultural study of speech act realization patterns (ccsarp). *Applied Linguistics*, 5(3):196–213.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Harry Bunt. 2000a. Dialogue pragmatics and context specification. In Harry Bunt and William Black, editors, *Abduction, Belief and Context in Dialogue: Studies in computational pragmatics*, pages 81–149. John Benjamins Publishing Company, Amsterdam.

Harry Bunt. 2000b. Dynamic interpretation and dialogue theory. In M. Martin Taylor, Françoise Néel, and Don Bouwhuis, editors, *The Structure of Multimodal Dialogue*, pages 139–188. John Benjamins Publishing Company, Amsterdam.

Harry Bunt. 2011. The semantics of dialogue acts. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Harry Bunt. 2012. The semantics of feedback. In *Proceedings of the 16th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2012), Paris, France*, pages 118–127. University Paris-Diderot, Paris Sorbonne-Cite.

Harry Bunt. 2014. A context-change semantics for dialogue acts. In Harry Bunt, Johan Bos, and Stephen Pulman, editors, *Computing Meaning: Volume 4*, pages 177–201. Springer Netherlands, Dordrecht.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. 2010. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Harry Bunt, Volha Petukhova, Emer Gilmartin, Catherine Pelachaud, Alex Fang, Simon Keizer, and Laurent Prévot. 2020. The ISO standard for dialogue act annotation, second edition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 549–558, Marseille, France. European Language Resources Association.

H.C. Bunt. 2009. The dit++ taxanomy for functional dialogue markup. In *Proceedings of EDAML@AAMAS, Workshop Towards a Standard Markup Language for Embodied Dialogue Acts*, pages 13–24. International Foundation for Autonomous Agents and Multi-agent Systems.

Florian Coulmas. 1981. *"Poison to Your Soul" Thanks and Apologies Contrastively Viewed*, pages 69–92. De Gruyter Mouton, Berlin, New York.

Emer Gilmartin, Christian Saam, Brendan Spillane, Maria O'Reilly, Ketong Su, Arturo Calvo, Loredana Cerrato, Killian Levacher, Nick Campbell, and Vincent Wade. 2018. The ADELE corpus of dyadic social text conversations:dialog act annotation with ISO 24617-2. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Emer Gilmartin, Brendan Spillane, Maria O'Reilly, Christian Saam, Ketong Su, Benjamin R. Cowan, Killian Levacher, Arturo Calvo Devesa, Lodana Cerrato, Nick Campbell, and Vincent Wade. 2017. Annotation of greeting, introduction, and leavetaking in dialogues. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Hede Helfrich. 1979. Age markers in speech. In Klaus R. Scherer and Howard Giles, editors, *Social markers in speech*, chapter 3, page 63–107. Cambridge University Press, Cambridge, UK.

Janet Holmes. 1995. *Women, Men and Politeness*. Routledge, London, UK.

Krzysztof Hwaszcz, Marcin Oleksy, Aleksandra Domogała, and Jan Wieczorek. 2023. ISO 24617-2 on a cusp of languages. In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 40–46, Nancy, France. Association for Computational Linguistics.

ISO Central Secretary. 2020. Language resource management — semantic annotation framework (semaf) part 2: Dialogue acts. Standard ISO 24617-2:2020, International Organization for Standardization, Geneva, Switzerland.

Ksenia Klokova, Maxim Krongauz, Valery Shulginov, and Tatiana Yudina. 2023. Towards a russian multimedia politeness corpus. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2023"*, pages 233–244.

Sara Mills. 2003. *Gender and politeness*. Cambridge University Press, Cambridge, UK.

Thi-Lan Ngo, Pham Khac Linh, and Hideaki Takeda. 2018. A Vietnamese dialog act corpus based on ISO 24617-2 standard. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Marcin Oleksy, Jan Wieczorek, Dorota Drużyłowska, Julia Klyus, Aleksandra Domogała, Krzysztof Hwaszcz, Hanna Kędzierska, Daria Mikoś, and Anita Wróż. 2022. DiaBiz.Kom - towards a Polish dialogue act corpus based on ISO 24617-2 standard. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3631–3638, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Volha Petukhova. 2011. *Multidimensional dialogue modelling*. Phd thesis, Tilburg University, Tilburg, Netherlands.

Volha Petukhova, Martin Gropp, Dietrich Klakow, Gregor Eigner, Mario Topf, Stefan Srb, Petr Motlicek, Blaise Potard, John Dines, Olivier Deroo, Ronny Egeler, Uwe Meinz, Steffen Liersch, and Anna Schmidt. 2014. The DBOX corpus collection of spoken human-human and human-machine dialogues. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 252–258, Reykjavik, Iceland. European Language Resources Association (ELRA).

Gabriel Roccabruna, Alessandra Cervone, and Giuseppe Riccardi. 2021. Multifunctional iso standard dialogue act tagging in italian. In *Proceedings of the Seventh Italian Conference on Computational Linguistics CLiC-it (2020)*, pages 365–372, Bologna, Italy. Accademia University Press.

Anna Wierzbicka. 1991. *Cross-cultural pragmatics: The semantics of human interaction.* Mouton de Gruyter, Berlin.

Koichiro Yoshino, Hiroki Tanaka, Kyoshiro Sugiyama, Makoto Kondo, and Satoshi Nakamura. 2018. Japanese dialogue corpus of information navigation and attentive listening annotated with extended ISO-24617-2 dialogue act tags. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

An annotation scheme for financial news in Portuguese

António Leal
University of Macau
University of Porto
CLUP
antonioleal@um.edu.mo

Purificação Silvano
University of Porto
CLUP
INESC TEC
msilvano@letras.up.pt

Zuo Qinren
University of Porto
up202202310@edu.letras.up.pt

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

Alípio Jorge
University of Porto
INESC TEC
amjorge@fcc.up.pt

Abstract

We present an annotation scheme designed to capture information related to the maintenance or change in the price of some goods (fuels, water, and vehicles) in news articles in Portuguese. The methodology we used involved adapting an existing annotation scheme, the Text2Story scheme (Silvano et al., 2021; Leal et al., 2022), which is based on different parts of ISO 24617 to capture the essential information for this project. Adaptations were needed to accommodate specific information, namely, information related to quantitative data and comparative relations that are abundant in this type of news. In this paper, we provide an overview of the annotation scheme, highlighting attributes and values of the entity and link structures specifically designed to capture financial information, as well as some problems we had to overcome in the process of building it and the rationale of some decisions behind its overall architecture.

1 Introduction

Corpora annotation is fundamental for theoretical linguistics research and for faster progress in improving Natural Language Processing and Information Extraction tasks by providing training material and gold standards for model evaluation (e.g., Levi and Shenhav, 2022). In recent years, projects have been carried out to build annotation schemes and annotated corpora to capture the content of texts with more or less generic themes (e.g., Groningen Meaning Bank (GMB) (Basile et al., 2012; Bos et al., 2017); Manually Annotated Sub-Corpus (MASC) (Ide et al., 2008); Georgetown University Multilayer Corpus (GUM) (Zeldes and Simonson, 2016; Zeldes, 2017); for Portuguese, Nunes et al., 2024).

However, several studies have also emerged that seek to capture the content of texts from specific domains, such as the medical domain (cf., e.g., Sun et al., 2013 and Campillos et al., 2018). These studies seek to overcome the particular difficulties of these domains, such as the specialized lexicon, which requires a more detailed ontological categorization of entities or specific links between these entities. The financial domain is one of these domains where efforts to build resources are scarce (cf., e.g., Lee et al., 2022). In fact, there is a significant lack of annotated corpora to train models in this domain, which is not unrelated to the lack of explicitly designed annotation schemes to capture financial information.

The project we present in this article aims to contribute to filling this gap. Our general objective was to create an annotation scheme that captured information related to the maintenance or change in the price of some goods (fuels, water, and vehicles) in European Portuguese (EP) news.

The methodology we used involved adapting an existing annotation scheme, the Text2Story (T2S) scheme, to capture the essential information for this project. The reasons for choosing this methodology are related to the characteristics of the T2S scheme and the objectives of this annotation project, which are to capture the maintenance and variation of prices of some goods. On the one hand, the T2S scheme is a scheme created based on the ISO 24617 standard (Silvano et al., 2021; Leal et al., 2022), which guarantees interoperability, particularly if you want to add in the future distinct semantic layers to the annotation scheme, such as, for example, discourse relations. On the other hand, the T2S scheme was created to capture the narrative structure of news of generalist themes in EP and has already proven suitable for this purpose (Silvano et al., 2023, 2024). Thus, it was predictable that, despite having to make a certain number of changes to the T2S scheme to adapt it to the objectives of this specific project, the T2S scheme should be

suitable, given that the type of text (news) and the language (EP) would be the same.

In the following section, we present some previous work on annotating financial information in texts and, in particular, on annotating quantities. In the third section, we present the annotation scheme we built for annotating financial news and describe some problems we have encountered during the process. We also present some justifications for specific choices we have made while constructing this annotation scheme. We conclude with some final remarks and future work.

2 Related work

Reasoning based on numerical information is common in all human domains (Thawani et al., 2021), particularly in scientific domains, in which one seeks to deduce scientific facts from premises that, in empirical sciences, are constructed based on quantitative data. For these reasons, in NLP, research on measurement expressions has focused on specific domains (Göpfert et al., 2022). For example, the medical domain is particularly interesting, as electronic health records contain much quantitative information (e.g., blood test results) relevant, for instance, to including patients in clinical trials. However, there are other interesting domains, such as the financial domain. In this field, research has focused more on the value of company shares and crude oil and, to a lesser extent, on the factors underlying the price variation of certain commodities (Lee et al., 2022).

Extracting these quantification expressions from texts in natural languages faces several challenges, starting with the format of these expressions (Göpfert et al., 2022). They are typically formed with a numerical value and a unit (which may or may not correspond to a standardized value). However, the numerical value can be presented in different ways (with letters or numbers); it can correspond to some vague form of quantification (e.g., 'a couple') or denote a numerical range (e.g., '3-5 g.'). The numerical value can correspond, for example, to a percentage relative to another quantification expression, which may be implicit in the text (e.g., 'The price of gasoline increased by 4%').

Quantification expressions may also contain some modification, which alters their overall meaning (cf. '3 g.' vs. 'approx. 3 g.'). One problem found in some works related to the annotation of numerical expressions (e.g., Ning et al., 2022) is the fact that they only annotate cardinal quantifiers, leaving aside modifiers of these quantifiers: for example, in "at most 3%", they only annotate "3%" leaving aside the modifier "at most"; the same happens in "at least 2%". Leaving aside the modifiers, they lose part of the nominal expressions' semantics. In this specific case, the distinction between right monotone increasing determiners ('at least') and right monotone decreasing determiners ('at most')[1] is lost (cf., e.g., Barwise and Cooper, 1981; Partee et al., 2012).

It is not just the identification of quantification expressions that is problematic. Indeed, identifying how these quantification expressions measure entities or properties of entities (explicit or implicit in the texts) is also a problem in Information Extraction. For example, in 'gasoline costs 1.5 euros', the quantification expression is locating a characteristic of the entity denoted by "gasoline", its price per liter, on a numerical scale associated with a particular monetary unit. Another problem is identifying price changes over time, as in 'Gasoline costs 4 cents more than last week'. All these problems are further aggravated by a lack of uniformity in the terms used: in different projects, the same concept may correspond to different terms; conversely, the same term may have non-equivalent definitions in other projects (cf. Göpfert et al., 2022). In this regard, the work carried out under ISO 24617, parts 11 and 12, seeks to answer some of these questions. ISO 24617-12 (ISO-24617-12, 2024) is a proposal to deal with quantification issues mainly related to the quantificational information of nominal expressions but also of eventualities expressions and the scope relations in which these expressions are involved. As for ISO 24617-11 (ISO-24617-11, 2021), it deals with aspects related to measurable quantitative information, that is, with the standardization of expressions involving a quantity n and a unit u, as in 'two meters'. Also noteworthy is the work of Abzianidze and Bos (2017), who propose a semantic tagset for use in a new NLP task to encode information that better characterizes lexical semantics than POS tags[2]. These tags describe the

[1] The following examples illustrate the distinct inferences of the two semantic types:
(i) right monotone increasing determiners (e.g., 'at least n')
If at least two men walk fast, then at least two men walk
(ii) right monotone decreasing determiners (e.g., 'at most n')
If at most two men walk, then at most two men walk fast.

[2] We thank one of the anonymous reviewers for bringing

semantic contribution of particular words or punctuation marks concerning the meaning of the whole expression and are grouped into 13 meta-tags. For example, the ATT (attribute) meta-tag includes the QUC (quantity of concrete) tag, which applies to words such as 'two', and the NAM (named entity) meta-tag includes the UOM (unit of measurement) tag, which applies to words such as 'euro' or 'percent'.

Datasets available to train models specifically with quantitative information are not abundant, particularly full-text datasets. Quantitative information is particularly relevant in texts from specialized domains, so these domains, such as medicine or finance, are necessary for this type of annotation. However, the question of finding suitable annotators for this type of task is also pertinent: annotators must have grammatical knowledge and, in some cases, knowledge of the specific domain (for instance, in cases of quantities given relative to a standard; cf. Göpfert et al., 2022). Projects have sought to use LLMs as annotators to overcome this difficulty, given that these models have proven effective in annotating general domain datasets. Aguda et al. (2024) show that LLMs are a possible alternative to non-expert crowd workers for domain-specific tasks. However, they do not surpass domain-specific human experts.

Another problem with datasets in the financial domain is that they often consist only of news headlines, and the annotation aims to perform sentiment analysis on stock prices. There are a few publicly available datasets with information about commodity news. Sinha and Khandait (2021) present one of these datasets, which is a dataset consisting of 11,412 news headlines about gold. This dataset was manually annotated with various information, namely whether the price is attributed to the past or the future ("Past/Future Price Information") and whether this value corresponds to an increasing, decreasing, or maintenance trend ("Price Up/Constant/Down").

Regarding news annotation work on commodities, Lee et al. (2022) report that datasets of this type are scarce, and the authors assume that their dataset may be the only annotated corpus. This is a dataset of 425 news articles about crude oil in English, manually annotated by undergraduate students from a School of Business. This dataset has some similarities to the one we developed in our project. The dataset from Lee et al. (2022) contains information about Entities (divided into 21 types) and Events (triggers, argument roles, and properties such as polarity, modality, and intensity). However, this annotation scheme has several problems. For example, in the list of Entity types, we can find ontologically very distinct entities, such as "commodities" (oil), "dates" (1998), "locations" (Europe), "Money" (USD 50), "percent" (25%), "Price unit" (USD 58 per barrel), "Quantity" (18 million tonnes) and "Production Unit" (29 million barrels per day). Furthermore, there is no unified treatment in the annotation of quantification expressions. Regarding Events, these authors identify 18 types, ranging from events that describe the change in price (e.g., CAUSED-MOVEMENT DOWN-LOSS; MOVEMENT-DOWN-LOSS; MOVEMENT-FLAT) to events that indicate the cause for the price change (e.g., CRISIS; EMBARGO). However, the strictly linguistic annotation is significantly reduced: polarity (POSITIVE and NEGATIVE); modality (ASSERTED and OTHER); and intensity (NEUTRAL, INTENSIFIED, and EASED). All in all, this annotation scheme is similar to our scheme in that it contains events and participants. However, the scheme we have developed is richer from the point of view of the grammatical information provided and the information contained in the annotation of quantification expressions, in addition to including different types of links, namely, semantic function links.

To sum up, identifying quantification expressions is a complex task, but one that is fundamental in several areas of research (cf. Göpfert et al., 2022 for a survey on measurement extraction in NLP tasks). In our project, we are particularly interested in extracting the value attributed to specific products over time: fuels (gasoline, diesel), motor vehicles, and essential goods (water, electricity). We will present a corpus of newspaper articles about finances, given that these texts often present numeric information representing different aspects of the situations they describe: in the case that interests us, the price of commodities (Roy et al., 2015).

3 Description of the scheme for annotating commodities price information

The scheme we created includes the following:

(i) a series of entity structures with information

this proposal to our attention.

about events, the participants in these events, and temporal and measurement expressions;

(ii) a series of link structures: between entities of a temporal nature, such as events and temporal expressions (TLinks); between participants, quantification expressions, and events (semantic roles); between participants (OLinks); and between quantification expressions and participants. All (entity and link) structures are composed of several attributes and values. The scheme also includes a function associating each document with its publication time. See the Appendix A for an overview of the annotation scheme.

Most of the content of this annotation scheme is based on parts of the ISO standard ISO24617, Language Resource Management - Semantic annotation framework: Part 1: Time and events (ISO-24617-1, 2012); Part 4: Semantic roles (ISO-24617-4, 2014); Part 7: Spatial information (ISO-24617-7, 2020); and Part 9: Reference annotation framework (ISO-24617-9, 2019). However, it was necessary to create some attributes and links to capture the information relevant to this annotation project, which will be detailed below.

Although a part of ISO deals with quantification, ISO 24617 - Part 12: Quantification (ISO-24617-12, 2024), we decided not to include it in this project. There are several reasons for this. For example, the expressive power of ISO 24617 - 12, notably in the annotation of quantificational information of nominal expressions and in the scope relations between nominal expressions and between these expressions and eventualities or other operators, such as negation, would not be used because this kind of information does not appear in the news that constitute our corpus. So, although ISO 24671-12 is quite expressive, our project did not require the level of detail proposed in this part of the ISO. We argue that the simplicity of the annotation scheme is crucial if one wants to recruit annotators without specialized linguistic knowledge.

Using ISO 24617-12 also poses some problems. One is that this part of the standard does not cover some of the expressions we had to annotate, namely, the quantification expressions of the "non-exact" type. Another problem is that many of the eventualities in the annotated news are expressed by nominal expressions with adjectives expressing quantification. Neither ISO 24617-12, for quantification, nor ISO 24617-9, for referential annotation, suggests ways to deal with these adjectival expressions that express some form of quantification over eventualities expressed by nouns. Examples (1) and (2) illustrate these cases. These sentences contain nominal expressions that, semantically, are equivalent to the sentences in (3) and (4), with explicit quantification within the verb phrase.

(1) *A gasolina registou um aumento expressivo.*
'Gasoline registered a significant increase.'

(2) *A gasolina registou uma descida ligeira.*
'Gasoline registered a slight drop.'

(3) *A gasolina aumentou expressivamente.*
'Gasoline prices have increased significantly.'

(4) *A gasolina desceu ligeiramente.*
'Gasoline prices have fallen slightly.'

Although ISO 24617-12 proposes a scheme for annotating quantified nominal expressions that could potentially apply to the annotation we intend to make (cf., for example, the @involvement attribute of entities, i.e., the information about the number of entities or part of the entity involved in a particular eventuality[3]), this proposal is not effectively applicable directly to the data we had to annotate, as in (5), but to constructions of another type, as in (6), which do not exist in the news that makes up our corpus.

(5) *A gasolina custa 1,5 euros.*
'Gasoline costs 1.5 euros.'

(6) *Comprei 3 euros/ dois litros de gasolina.*
'I bought 3 euros/two liters of gasoline.'

Finally, ISO 24617-11 (ISO-24617-11, 2021) was also assessed in this annotation project. However, we considered that its use could be dispensable, which is why we chose not to include it. For example, ISO 24617-11 proposes using the measure link and the comparison link. As for the measure link, which connects an entity to a measure, its use would unnecessarily complicate our scheme, given that it overlaps with the semantic roles. The comparison link, which is similar to the ARG1 and

[3]The value of the @involvement attribute indicates how many/much or which fraction of the reference domain is contained in the participant set (ISO-24617-12, 2024).

ARG2 that we propose, is a link that is established between two measures (according to ISO 24617-11). This type of link is irrelevant to our annotation, given that the comparison relations in the news are not between measures but between entities.

The news annotations were performed in BRAT (Stenetorp et al., 2012). The manual was built incrementally, following the MAMA cycle (Pustejovsky et al., 2017), taking the Text2Story scheme as a starting point. A PhD student in linguistics annotated a set of news items and identified problems in the annotation process. In a meeting with two senior linguistics researchers, these problems were discussed, and solutions were proposed. The student revised the annotation according to these solutions, and in the following meeting, the new results were analyzed. If the problems persisted, new solutions were proposed. If the problems were solved, the student annotated a new set of news items. This process was repeated cyclically until the entire set of news items was annotated, and no problems were left in the annotation. Finally, the news annotations (N=98) were reviewed by both senior linguistics investigators to identify and correct any lapses or inconsistencies that might persist.

3.1 Overview of the annotation scheme

1. Events

The markables of events follow the same rules defined in ISO 24617-1, the basis for the T2S annotation scheme. Regarding the attributes, it was decided to simplify them. Thus, the @class attribute was eliminated, as it was not relevant to the objectives of this annotation project. The @aspect, @vform, and @mood attributes were included in the @tense attribute. This attribute was adapted to capture the verb forms in Portuguese texts, thus making the annotation task less complex and easier to implement. Table 2 in the appendix presents the values used in the @tense attribute. The @pos attribute was limited to just two values (noun and verb). The @polarity attribute was maintained. Finally, in the case of the @movement attribute, it was decided to simplify the annotation to two values, "upward" and "downward," to capture the rise or fall of prices in the case of events or progressive states. Example (7) illustrates the case of price increase (upward movement).

(7) *A gasolina subiu esta semana. / A gasolina está a subir esta semana.*

'Gasoline rose this week. / Gasoline is rising this week.'

2. Time expressions

The time expressions in this annotation project follow the annotation proposed in T2S, which in turn abides by the rules of ISO-24617-1 (2012) proposal. The tag spans are the same, as are the @type attributes (with the values "date," "time," "duration," and "set") and the "publication time".

3. Participants

Regarding participants, this annotation project follows the T2S proposal (based on ISO-24617-9, 2019), with few adaptations, only those necessary to capture some specificities of news about price variations (recall that the T2S scheme was created to capture the structure of news about narratives of generalist themes, which is why it lacks specific tags for entities in the financial domain). Thus, the tag spans and the annotation of @lexical-head, @individuation, and @involvement remained the same, as did the objectal relations between participants. In the case of the @type attribute, some values were added to capture specific information. Therefore, the value "relation-price/unit" was included for referents corresponding to an abstract numerical relation between a measuring unit and a value in some monetary system. This value is subdivided into "average-value" (when the relation corresponds to an average) and "precise-value" (when the relation corresponds to an exact value). The relevant measuring unit is indicated in the text box (in Brat). Examples (8) and (9) illustrate these two possibilities.

(8) *O preço de referência do litro de gasolina em Portugal é actualmente de 1,741 euros enquanto o do gasóleo vale 1,521 euros.*

'The reference price of a liter of gasoline in Portugal is currently 1.741 euros, while that of diesel is 1.521 euros.'

(relation-price/unit-precise-value; unit=liter)

(9) *O preço médio do gasóleo na quarta-feira era de 1.410 euros.*

'The average price of diesel on Wednesday was 1,410 euros.'

(relation-price/unit-average-value; unit=liter)

The value "TARIFA" (TAR) is also included for cases where the referent is a table (or a range of that table) of rates charged for a given service. Example (10) is one of those cases.

(10) *O primeiro escalão dos consumidores de água teve uma descida de 3%. (TAR)*

'The first tier of water consumers saw a 3% drop.'

Table 3 in the appendix summarizes all of the @type values used in the annotation.

4. Semantic roles

The semantic roles used in this annotation project are a subset of those proposed in ISO-24617-4 (2014), with definitions explicitly oriented to the financial domain (cf. Table 4 in the appendix). Examples (11) and (12) are instances of this annotation layer.

(11) *Ao mesmo tempo, o crude apreciava 1,31% para 92,87 euros.*

'At the same time, crude oil appreciated by 1.31% to 92.87 euros.'

(92,87 euros — Goal — apreciava; O crude – Theme — apreciava; 1,31% — Amount — apreciava)

(12) *O gasóleo permanece em 1,22 euros por litro.*

'Diesel remains at R$1.22 per liter.'

(O gasóleo — Pivot — permanece; 1,22 euros — Attribute — permanece)

5. Quantification expressions

One of the objectives of this project is to capture information related to changes in the prices of goods. Therefore, annotating each expression that conveys quantitative information related to prices is essential (see Figure 2 for an overview of the occurrence of entity structures in the annotated corpus). In this annotation scheme, these expressions are annotated in entity structures called "quantification expressions". These expressions appear in the news in very different formats. However, they can be grouped into two groups: those of the "exact" type, providing quantitative information that is expressed, in some way, by a numerical value, and those of the "non-exact" type, whose quantitative information corresponds to a non-numerical value and is typically expressed by an adjective or an adverb (see Table 1 for the distribution of quantification expressions in the annotated corpus).

From a linguistic point of view, expressions of the type "exact", often referred to as "measurement phrases", correspond, in most cases, to argumental nominal expressions in the oblique case (cf. Gonçalves and Raposo, 2013). The verbs with which these expressions occur are varied: (i) verbs such as *custar* 'to cost' (which, like *durar* 'to last', *medir* 'to measure' or *pesar* 'to weigh', express the value of physical or abstract entities on a quantitative scale); (ii) verbs of movement (which, in this case, have fictive readings; cf. Talmy, 1996), such as *subir* 'to go up' and *descer* 'to go down' and their synonyms; (iii) verbs that denote variation of properties in general, such as *aumentar* 'to increase' or *reduzir* 'to decrease'; (iv) verbs lexically specialized in price variation, such as *encarecer* 'to make more expensive' or *custar* 'to cost'; (v) eventive verbs that are not lexically specified as to the nature of the variation, such as *evoluir* 'to evolve' or *encerrar* 'to close', which lexically encode only some type of change. These measurement phrases can also occur as predicatives with copula verbs or verbs that proceed to some static location in cases where no variation in price is expressed, but an indication is given of the value associated with the good in a given time interval (e.g., *ser* 'to be,' *manter-se* 'to maintain,' and *situar-se* 'to be located'). Since these "exact" quantification expressions are included in a more general group of expressions associated with measurement verbs, such as *medir* 'to measure,' we annotated them according to the ISO-24617-7 (2020) proposal, using the "measure" tag. Thus, quantification expressions have the attributes @value, @unit, and @modifier. The @value attribute provides the indication of the quantity of the entity, denoted by cardinal numeral quantifiers or other quantifiers; the @unit corresponds to the monetary unit used in the quantification expression; the @modifier corresponds to an expression that modifies the quantification in terms of quantitative, circumstantial or temporal information. Example (13) illustrates this part of the annotation.

(13) *Os combustíveis já aumentaram, entre 1 de janeiro e 1 de março, cerca de nove cêntimos.*

'Fuel prices have already increased by around nine cents between January 1st and March 1st.'

(@modifier - cerca de; @value - 0,09; @unit – euro)

Moving on to "non-exact" quantification expressions, they correspond to seven subtypes, which are syntactically very diverse and distinct from the

type described previously. Therefore, creating specific attributes for these expressions was necessary, which are not provided in any part of ISO 24617. In what follows, we describe those attributes.

- *value-minimum*: when it corresponds to a minimum price value in the time interval considered in the sentence. Typically, it corresponds to a nominal expression with an adjective in the superlative degree.

  (14) *O preço do barril de Brent, o petróleo que serve de referência ao mercado português, desvalorizou 4,57% para 37,93 dólares, o que representa* ***o valor mais baixo*** *desde dezembro de 2008.*
  'The price of a barrel of Brent, the oil used as a reference for the Portuguese market, fell 4.57% to 37.93 dollars, representing the lowest value since December 2008.'

- *value-maximum*: when it corresponds to a maximum price value in the time interval considered in the sentence. Typically, it corresponds to a nominal expression with an adjective in the superlative degree.

  (15) *Em maio, a matéria-prima superou a fasquia dos 80 dólares por barril,* ***o valor mais alto*** *desde novembro de 2014.*
  'In May, the raw material surpassed the US$ 80 per barrel mark, the highest value since November 2014.'

- *value-much*: when it corresponds to a high differential price value in the time interval considered in the sentence. Typically, it corresponds to an adjective that denotes an interval in a scale whose extension is greater than a reference value (which, in most cases, is implicit).

  (16) *A gasolina registou um aumento* ***expressivo***.
  'Gasoline registered a significant increase.'

- *value-low*: when it corresponds to a low differential price value in the time interval considered in the sentence. Typically, it corresponds to an adjective that denotes an interval in a scale whose extension is smaller than a reference value (which, in most cases, is implicit).

  (17) *A gasolina registou uma descida* ***ligeira***.
  'Gasoline registered a slight drop.'

- *value-comparative-superiority*: when a differential value of superiority of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (18) *A gasolina é* ***mais cara do que*** *o gasóleo.*
  'Gasoline is more expensive than diesel.'

- *value-comparative-inferiority*: when a differential value of inferiority of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (19) *A gasolina é* ***menos cara do que*** *o gasóleo.*
  'Gasoline is less expensive than diesel.'

- *value-comparative-equality*: when a value of equality of the price is expressed relative to the price of another entity referred to in the sentence. It corresponds to an adjective in the comparative degree.

  (20) *A gasolina é* ***tão cara como*** *o gasóleo.*
  'Gasoline is as expensive as diesel.'

As with "exact" quantification expressions, "non-exact" quantification expressions may be subject to some modification. For this reason, these expressions can also be annotated with the @modifier attribute, as exemplified in (21).

(21) *Este combustível continuará a ter um preço de venda média muito perto de máximos de agosto de 2015.*

'This fuel will continue to have an average selling price very close to the highs of August 2015.'

(máximos – quantification expression; non-exact; value-maximum; @modifier - muito perto de)

Quantification expressions that express non-exact values of the Value-Comparative type (superiority/inferiority/equality) establish a relationship between two entities that share the same scalar property, in this case, the price. Therefore, these three quantification expressions trigger comparison relationships. Since these relations are not provided

for in ISO 24617, they were created specifically for this work, seeking to follow the same general principles of this standard. Thus, these comparison relations link quantification expressions of Value-comparative and the respective compared entities (identified through the links "argument 1" (Arg1) and "argument 2" (Arg2), which correspond to the first and second terms of comparison, respectively). The relation is established from the quantification expression to the entity that plays the role of Arg1 or Arg2. Example (22) shows how this link is used.

(22) *A gasolina é mais cara do que o gasóleo.*

'Gasoline is as expensive as diesel.'

mais cara do que – quantification expression; value-comparative-superiority

a gasolina – Argument 1 – mais cara do que

o gasóleo – Argument 2 – mais cara do que

Example (23) illustrates how this annotation scheme was implemented (see also Figure 1).

(23) *Segundo o presidente da Anarec, Augusto Cymbron, o preço do gasóleo passará dos actuais 1,339 euros para os 1,369 euros.*

'According to the president of Anarec, Augusto Cymbron, the price of diesel will increase from the current R$1.339 to R$1.369'

In this example, the change in the price of diesel is captured. The participant of the "relation price-unit" type is expressed by the noun phrase *o preço* (the price), which denotes a count entity. The substance is represented by the noun phrase *o diesel*, a participant that is a mass-type "object." The relationship between these two participants is captured by the objectal link "partOf." The price change event is annotated in the markable *passará* (will pass): it is an event expressed by a verb in the future tense. It encodes a price increase (an upward movement on the numerical scale associated with the price). The expressions *1.339 euros* and *1.369 euros* correspond to participants of the "quantification expression" type: both correspond to exact information, and their unit of measurement is "euro," varying only in their respective measurement values. Semantic roles link these entity structures. Thus, *o preço* (the price) is connected to *passará* (will pass) by the SR Theme, indicating that it is the entity that suffers a change of state during the event (in this case, it is the entity whose position on the numerical scale associated with the price changes). In turn, the quantification expressions *1.339 euros* and *1.369 euros* are linked to *passará* by the SR Source and Goal, respectively: the first expression denotes the starting point of the price change, while the second denotes the end point of the price change.

Category	Count	Perc. (%)
Exact	422	78.00
Non-Exact	119	22.00
Max	40	7.39
Much	32	5.91
Min	19	3.51
Low	10	1.85
Comparative-superiority	8	1.48
Comparative-equality	6	1.11
Comparative-inferiority	3	0.55
Total	541	100.00

Table 1: Distribution of Quantification Categories

3.2 Building the annotation scheme: some problems

Throughout the process of building this annotation scheme, we encountered several problems that had to be overcome, from the simplest problems found in any annotation project (for example, the correct identification of the @type of a nominal expression, which is referentially ambiguous), to the more complex problems this particular type of text poses. In fact, the variation in expressions that indicate numerical values and the variation in expressions of eventualities that denote price variation or price maintenance are challenging if one wants to develop a comprehensive but straightforward annotation scheme. We will discuss some of those cases in this section.

The expression denoting a fuel's "price/liter" ratio, such as gasoline, may appear in the following formats (non-exhaustive list).

(24) *O preço do litro da gasolina / O preço da gasolina / O litro da gasolina / O preço do litro no caso da gasolina / A gasolina / O valor da unidade de gasolina está a 1,60 euros.*

'The price of a liter of gasoline / The price of gasoline / The liter of gasoline / The price of a liter in the case of gasoline / Gasoline / The value of the gasoline unit is R$1.60.'

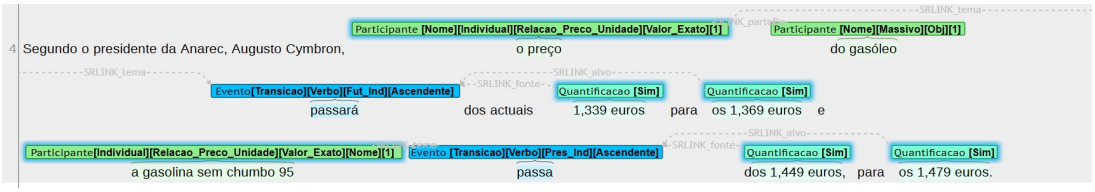


Figure 1: (23): an example of annotation in BRAT (fuel, news 22)

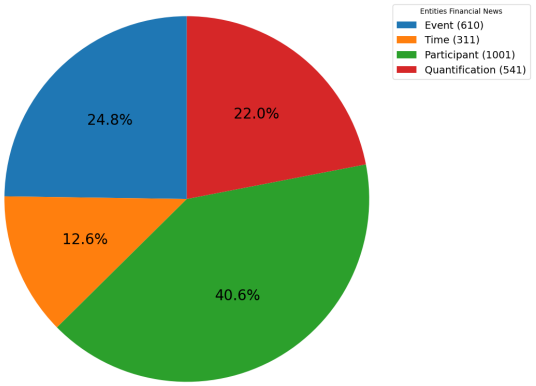


Figure 2: Proportion of Entities in Dataset

(25) *A inflação colocou o custo do litro da gasolina nos 1,60 euros/ colocou o custo do litro nos 1,60 euros no caso da gasolina.*

'Inflation put the cost of a liter of gasoline at R$1.60/ put the cost of a liter at R$1.60 in the case of gasoline.'

Another problem related to the different manners in which the price variation can be expressed. In all of the examples (26)-(29), the event corresponds to a decrease in price, but in example (26), the markable is of the nominal type (*descida*), while in the others ((27)-(29)), the markables are of the verbal type. In this second set of examples, the decrease can be encoded lexically by the verb (such as *cair* 'to fall,' in (27), in a case of fictive motion), by a semi-lexicalized expression (such as *evoluir em baixa* 'to evolve downwards', in (28)), or compositionally, as in (29), through the combination of the verb *atingir* 'to reach' and the direct object that contains a word that corresponds to a quantification expression of the type "non-exact-value-minimum" (*mínimos deste ano* 'minimums of this year').

(26) *A variação do euro face ao dólar deverá determinar uma descida de 1,5 cêntimos no preço de venda da gasolina e de 2 cêntimos no gasóleo.*

'The variation of the euro against the dollar should determine a drop of 1.5 cents in the sale price of gasoline and 2 cents in diesel.'

(27) *O preço do gasóleo deverá cair 2 cêntimos, colocando o custo do litro nos 1,399 euros.*

'The diesel price is expected to fall by 2 cents, putting the cost per liter at 1.399 euros.

(28) *O brent evoluiu hoje em baixa, depois de ter estado a subir durante várias sessões.*

'Brent fell today after rising for several sessions.'

(29) *O brent atingiu mínimos deste ano.*

'Brent reached its lowest price this year.'

4 Concluding remarks

The annotation of information about quantities is a topic that still requires further study, both in terms of annotation schemes that adequately capture the various nuances that quantification expressions can have in natural languages and in terms of annotated corpora that can be used to train models or as gold standards in evaluation tasks. In this article, we seek to contribute to the discussion on annotation schemes and present a new scheme to capture information related to the rise, fall, or maintenance of commodities prices in news articles. To that end, we propose an extension of the Text2Story scheme, which, in turn, was built by combining four parts of ISO 24617 (parts 1, 4, 7, and 9). The new scheme contains new attributes and values specially designed to capture the quantificational information typical of news in the financial domain.

This new scheme can also be extended to capture other types of information. In the future, we plan to include other ISO resources, such as the entire semantic role array of ISO 24617-4, and other parts, such as ISO 24617–8 (ISO, 2016), using the discourse relations framework to analyze the underlying reasons for price changes. Another line of research is this scheme's application (with the necessary adaptations) to other domains with

abundant quantificational information, such as electronic health records. Finally, formalizing the semantics of the proposed annotation scheme is still missing.

Acknowledgments

This work was financed in part 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 would also like to acknowledge project StorySense, with reference 2022.09312.PTDC (DOI10.54499/2022.09312.PTDC)..

References

Abzianidze, L. and Bos, J. (2017). Towards universal semantic tagging. In Gardent, C. and Retoré, C., editors, *Proceedings of the 12th International Conference on Computational Semantics (IWCS) — Short papers*.

Aguda, T., Siddagangappa, S., Kochkina, E., Kaur, S., Wang, D., Smiley, C., and Shah, S. (2024). Large language models as financial data annotators: A study on effectiveness and efficiency. *arXiv preprint arXiv:2403.18152*.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. In *Philosophy, language, and artificial intelligence: Resources for processing natural language*, pages 241–301. Springer.

Basile, V., Bos, J., Evang, K., and Venhuizen, N. (2012). Developing a large semantically annotated corpus. In *LREC 2012, Eighth International Conference on Language Resources and Evaluation*.

Bos, J., Basile, V., Evang, K., Venhuizen, N. J., and Bjerva, J. (2017). The groningen meaning bank. *Handbook of linguistic annotation*, pages 463–496.

Campillos, L., Deléger, L., Grouin, C., Hamon, T., Ligozat, A.-L., and Névéol, A. (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.

Gonçalves, A. and Raposo, E. P. (2013). Verbo e sintagma verbal. In Raposo, E. P., Mota, M. A., Segura, L., and Mendes, A., editors, *Gramática do Português*, pages 1155–1220. FCG, Lisboa.

Göpfert, J., Kuckertz, P., Weinand, J., Kotzur, L., and Stolten, D. (2022). Measurement extraction with natural language processing: a review. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2191–2215.

Ide, N., Baker, C., Fellbaum, C., Fillmore, C., and Passonneau, R. (2008). Masc: The manually annotated sub-corpus of american english. In *6th international conference on language resources and evaluation, LREC 2008*, pages 2455–2460. European Language Resources Association (ELRA).

ISO (2016). ISO 24617-8. 2016. Language resource management, part 8: Semantic relations in discourse (DR-Core). Standard, International Organization for Standardization, Geneva, CH.

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

ISO-24617-11 (2021). Language resource management-semantic annotation framework (semaf) - part 11: Measurable quantitative information. Standard, Geneva, CH.

ISO-24617-12 (2024). Language resource management-semantic annotation framework (semaf) - part 12: Quantification. Standard, Geneva, CH.

ISO-24617-4 (2014). Language resource management-semantic annotation framework (semaf) - part 4: Semantic roles (semaf-sr). Standard, Geneva, CH.

ISO-24617-7 (2020). Language resource management-semantic annotation framework (semaf) - part 7: Spatial information. Standard, Geneva, CH.

ISO-24617-9 (2019). Language resource management-semantic annotation framework (semaf) - part 9: Reference annotation framework (raf). Standard, Geneva, CH.

Leal, A., Silvano, P., Amorim, E., Cantante, I., Silva, F., Jorge, A. M., and Campos, R. (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.

Lee, M., Soon, L.-K., Siew, E.-G., and Sugianto, L. F. (2022). Crudeoilnews: An annotated crude oil news corpus for event extraction. *arXiv preprint arXiv:2204.03871*.

Levi, E. and Shenhav, S. R. (2022). A decomposition-based approach for evaluating inter-annotator disagreement in narrative analysis. *arXiv preprint arXiv:2206.05446*.

Ning, Q., Zhou, B., Wu, H., Peng, H., Fan, C., and Gardner, M. (2022). A meta-framework for spatiotemporal quantity extraction from text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2736–2749.

Nunes, S., Jorge, A. M., Amorim, E., Sousa, H., Leal, A., Silvano, P. M., Cantante, I., and Campos, R.

(2024). Text2story lusa: A dataset for narrative analysis in european portuguese news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15773–15782.

Partee, B. B., Ter Meulen, A. G., and Wall, R. (2012). *Mathematical methods in linguistics*, volume 30. Springer Science & Business Media.

Pustejovsky, J., Bunt, H., and Zaenen, A. (2017). Designing annotation schemes: From theory to model. In Ide, N. and Pustejovsky, J., editors, *Handbook of Linguistic Annotation*, pages 31–63. Springer, Dordrecht.

Roy, S., Vieira, T., and Roth, D. (2015). Reasoning about quantities in natural language. *Transactions of the Association for Computational Linguistics*, 3:1–13.

Silvano, M. d. P., Amorim, E., Leal, A., Cantante, I., Jorge, A., Campos, R., and Yu, N. (2024). Untangling a web of temporal relations in news articles. In *Proceedings of Text2Story 2024-seventh workshop on narrative extraction from texts*.

Silvano, M. d. P., Amorim, E., Leal, A., Cantante, I., Silva, M. d. F. H. d., Jorge, A., Campos, R., and Nunes, S. S. (2023). Annotation and visualisation of reporting events in textual narratives. In *Proceedings of Text2Story 2023: Sixth Workshop on Narrative Extraction From Texts*.

Silvano, P., Leal, A., Silva, F., Cantante, I., Oliveira, F., and Jorge, A. M. (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.

Sinha, A. and Khandait, T. (2021). Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

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

Talmy, L. (1996). Fictive motion in language and ception.

Thawani, A., Pujara, J., Szekely, P. A., and Ilievski, F. (2021). Representing numbers in nlp: a survey and a vision. *arXiv preprint arXiv:2103.13136*.

Zeldes, A. (2017). The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Zeldes, A. and Simonson, D. (2016). Different flavors of gum: Evaluating genre and sentence type effects on multilayer corpus annotation quality. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 68–78.

A Annotation Scheme and Descriptions

Overview of the annotation scheme

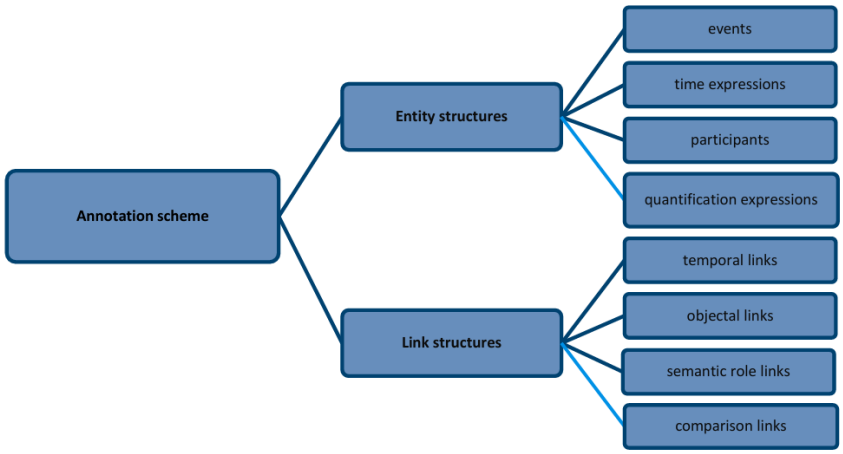


Figure 3: Overview of the Annotation Scheme for Finance news

Indicativo	Conjuntivo	Progressivo	Gerúndio	Infinitivo	IR + Infinitivo
Presente do Indicativo – Pres-Ind	Presente do Conjuntivo – Pres-Conj	Presente progressivo (está a + inf) – PresPro	Gerúndio Simples – GS	Infinitivo Simples – INF-S	IR (presente) + infinitivo (futuro) – ir(Pres)+INF-S
Pretérito Perfeito Simples – PP-Ind	Pretérito Imperfeito – PIMP-Conj	Passado progressivo (esteve/estava a + inf) – PstPro	Gerúndio Composto – GC	Infinitivo Composto – INF-C	IR (futuro) + infinitivo (futuro) – ir(Fut)+INF-S
Pretérito Imperfeito – PIMP-Ind	Pretérito Perfeito – PPC-Conj				
Pretérito Perfeito Composto – PPC-Ind	Pretérito mais-que-perfeito – PMP-Conj				
Pretérito mais-que-perfeito – PMP-Ind	Futuro Simples – Fut-Conj				
Futuro Simples – Fut-Ind	Futuro Composto – Fut-C-Conj				
Futuro Composto – Fut-C-Ind					

Table 2: Verb tense attributes

Type	Definition
PES	The referent is a person.
ORG	The referent is an organization.
OBJ	The referent is a tangible object, whether or not made by a human being.
relation_price/unit	The referent is an abstract numerical relationship between a unit of measurement and a value in some monetary system.
LOC	The referent is a concrete or abstract location.
TAR	The referent is a table (or a range of that table) of rates that are charged for a specific service performed.

Table 3: Type values and definitions

Semantic Role	Definition
Agent	Entity that intentionally causes a price change.
Cause	Entity that causes a price change unintentionally.
Theme	Entity that changes price, in an event.
Pivot	Entity that maintains the price, in a state.
Quantity	Quantification expression that indicates the amount of change operated in the price, in events.
Attribute	Quantification expression that indicates the quantity at which the price is maintained, in states.
Goal	Quantification expression indicating the end point of the price change.
Source	Quantification expression that indicates the starting point of the price change.
Locative	Expression that represents the place (concrete or abstract) where an entity is located or the event is held.

Table 4: Semantic roles and definitions

Enhanced Evaluative Language Annotation through Refined Theoretical Framework and Workflow

Jiamei Zeng[1*], Haitao Wang[2*], Harry Bunt[3], Xinyu Cao[2], Sylviane Cardey[4], Min Dong[5], Tianyong Hao[6], Yangli Jia[7], Kiyong Lee[8], Shengqing Liao[9], James Pustejovsky[10], François Claude Rey[4], Laurent Romary[11], Jianfang Zong[2], and Alex C. Fang[1**]

[1]City University of Hong Kong, PR China (jiamezeng3-c@my.cityu.edu.hk, acfang@cityu.edu.hk)
[2]China National Institute for Standardization, PR China ({wanght,caoxy,zongjf}@cnis.edu.cn)
[3]University of Tilburg, The Netherlands (Harry.Bunt@tilburguniversity.edu)
[4]University of Franche-Comté, France (sylviane.cardey@univ-fcomte.fr, francois_claude.rey@edu.univ-fcomte.fr)
[5]Beihang University, PR China (mdong@buaa.edu.cn)
[6]South China Normal University, PR China (haoty@m.scnu.edu.cn)
[7]Liaocheng University, PR China (jiayangli@lcu.edu.cn)
[8]Korea University, Korea (ikiyong@gmail.com)
[9]Fudan University, PR China (sqliao@fudan.edu.cn)
[10]Brandeis University, USA (jamesp@cs.brandeis.edu)
[11]National Institute for Research in Digital Science and Technology, France (laurent.romary@inria.fr)

Abstract

As precursor work in preparation for an international standard *ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16: Evaluative language*, we aim to test and enhance the reliability of the annotation of subjective evaluation based on Appraisal Theory. We describe a comprehensive three-phase workflow tested on COVID-19 media reports to achieve reliable agreement through progressive training and quality control. Our methodology addresses some of the key challenges through the refinement of targeted guideline refinements and the development of interactive clarification tools, alongside a custom platform that enables the pre-classification of six evaluative categories, systematic annotation review, and organized documentation. We report empirical results that demonstrate substantial improvements from the initial moderate agreement to a strong final consensus. Our research offers both theoretical refinements addressing persistent classification challenges in evaluation and practical solutions for the implementation of the annotation workflow, proposing a replicable methodology for the achievement of reliable annotation consistency in the annotation of evaluative language.

1 Introduction

Annotating evaluative language has become increasingly important in the present era of artificial intelligence, particularly given the need for high-quality language resources for training large language models and understanding human emotions and personal stance. In response to this need, we are working on an international standard (*ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16: Evaluative language*) for the annotation of evaluative language, to be adopted by the International Organization for Standardization. To ensure a sound practical application of the standard, we apply Appraisal Theory (AT) as a foundational framework of analysis, fully described in Martin and White (2005). While AT has proven influential across diverse research contexts, its practical implementation through manual annotation reveals significant methodological challenges. Research addressing annotation methodology remains scarce, with studies rarely reporting inter-coder agreement measures or addressing reliability issues (Fuoli, 2018). Similarly, the development of annotation workflow has received limited attention. While Fuoli (2018) proposed a stepwise approach to Appraisal annotation, the methodological guidance remains insufficient for addressing the complex practical challenges encountered in the annotation of evaluative language, which is fundamentally subjective. The limited focus on annotation methodologies has created a significant gap between theoretical significance and operational reliability, affecting the advancement of scientific interpretative approaches to the understanding of expressions of human stance and attitude.

We aim to address these methodological gaps through the development of annotation workflows, refinement of guideline procedures, and tool-facilitated quality control mechanisms. We describe a comprehensive approach that combines refined theoretical grounding with practical solutions for the achievement of reliable inter-annotator agreement based on the annotation of a corpus of authentic texts sampled from media reports. In what follows, we describe a methodological framework that encompasses a multi-stage training protocol, problem identification through pilot annotations, and targeted intervention strategies that reduce annotator disagreements.

* Equal contribution
** Corresponding author

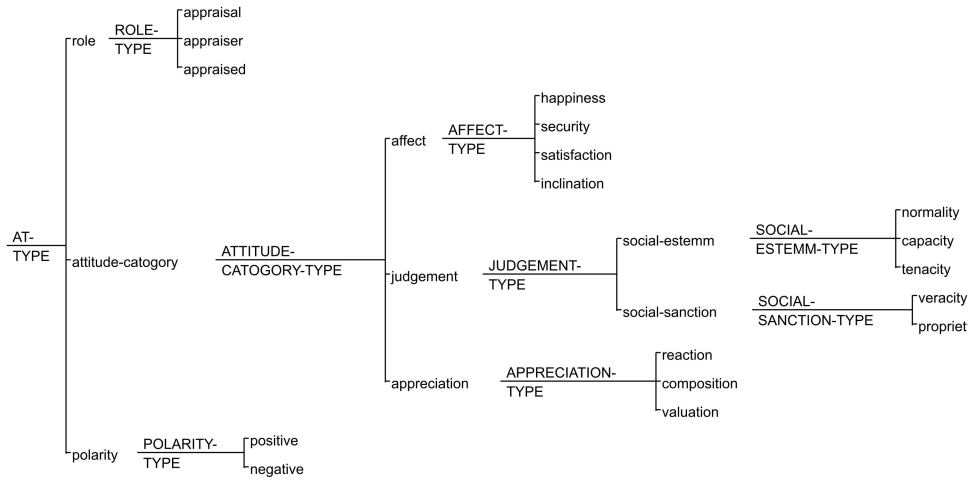


Figure 1: Annotation framework for Appraisal theory

2 Annotation Framework and Methodology

2.1 Annotation Framework

Our annotation framework is fundamentally grounded in AT fully described and published in Martin and White (2005) and specifically focuses on the Attitude system. In our approach, we explicitly define three core roles for each identifiable (markable) textual segment expressing an evaluation, namely, Appraiser (the entity making the evaluation), Appraised (the target being evaluated), and Appraisal (the appraisal element itself). The annotation process followed a hierarchical framework as shown in Figure 1.

2.2 Core Annotation Principles

Our annotation methodology is governed by two fundamental principles: (1) Minimality Principle, according to which annotators mark only the core evaluative lexical items and their essential modifiers, avoiding the unnecessary expansion of the textual segment, and (2) Completeness Principle, according to which the annotation segment must preserve the semantic integrity of evaluative expressions. Critical elements that influence evaluative meaning, particularly negation markers, must be included within the segment to maintain accurate interpretations and polarity determination. Consider

[1] *Research from earlier in the pandemic* ***does not yield definitive clues***. (USN0122)

In [1], while "definitive" represents the core evaluative term, the negation "does not" fundamentally alters the evaluative stance and must be included in the annotation span as "does not yield definitive clues" to preserve semantic completeness.

These principles establish baseline standards rather than rigid constraints. Recognizing the inherent complexity of the segment boundaries of evaluative language, we adopt a flexible approach to boundary matching for inter-annotator agreement assessment. Following Wiebe et al. (2005) and Read and Carroll (2012), overly strict boundary matching can be counter-productive. Therefore, when segments marked by different annotators overlap and demonstrate complete agreement on Main Category, Subcategory, and Polarity, we consider these as valid matches, specifically categorized as "Match with Overlap". This methodology maintains analytical rigour in categorical agreement while accommodating the subjective nature of evaluative language boundaries.

2.3 Corpus and Annotators

The texts used in our study are drawn from a corpus of COVID-19 news reports. The corpus comprises a total of 144 news articles related to COVID-19 balanced across four media outlets – China Daily, South China Morning Post, The Guardian, and The New York Times – with 36 articles from each. The articles were sampled from Factiva, covering the period from January 2020 to December 2022, with one article selected per month to ensure temporal balance. The descriptive statistics are presented in Table 1.

Articles were selected from this corpus to test the annotation of evaluative language, with each round

using articles from the same time periods across the four outlets, ensuring temporal and cross-media representation while maintaining manageable annotation loads.

	CD	SCMP	TG	NYT	Overall
Articles	36	36	36	36	144
Tokens	20536	23863	31432	42704	118535
Types	3710	4327	5062	6045	10701
TTR (%)	18.07	18.13	16.10	14.16	9.03
Mean size	570	663	873	1186	823

Table 1: Corpus composition and descriptive statistics across four media outlets

Two annotators were selected based on three criteria: (1) proficiency in English comprehension, (2) enrolment in a graduate-level programme in linguistics or English studies, and (3) foundational linguistic knowledge, including familiarity with Systemic Functional Linguistics (SFL). We intentionally selected annotators without prior independent experience in the annotation of evaluative language. The selection of novice annotators allows us to assess whether our framework can achieve satisfactory inter-annotator agreement through systematic training protocols, rather than relying on prior experience.

2.4 Annotation Software Tool

Many software tools facilitate basic annotation tasks but generally lack sophisticated comparison capabilities beyond simple agreement-disagreement identification. This limitation particularly affects the crucial intermediate review process, where detailed comparative analysis is essential for quality control and guideline refinement by providing organized documentation of annotator classifications and problematic and classic cases. Given these limitations, we adopt a hybrid approach combining existing and custom-developed tools. The initial annotation was conducted using the UAM corpus tool (O'Donnell, 2018). A web-based review tool has been designed and implemented that supports simultaneous upload of processed annotation results and enables detailed comparison of all annotation units within their original textual contexts, across the pair of annotators.

The tool also categorizes comparison results into six distinct types to provide fine-grained characterization of annotation consistency:

1. Annotator 1 Only: Annotations identified solely by Annotator 1
2. Annotator 2 Only: Annotations identified solely by Annotator 2
3. Match: Complete agreement in annotation content
4. Match with Overlap: Overlapping appraisal elements with identical classifications
5. Conflict: Identical text spans with different classification or polarity
6. Conflict with Overlap: Overlapping appraisal elements with different classifications or polarity

The interface enables comparative viewing of both annotators' work through individual or category-based review, with functions to flag typical or problematic cases and add comments. It identifies specific points of disagreement, supports targeted feedback, and offers traceable evidence for iterative guideline refinement. The tool also exports processed data for statistical analysis, significantly reducing manual comparison workload with minimal need for post-processing.

2.5 Statistical Analysis Methods

Our statistical approach is based on Read and Carroll (2012)'s evaluation methodology with a key modification. Read and Carroll (2012) note that "the 'number correct' (COR) will differ for each annotator in the pair under evaluation" due to multiple matching scenarios in text span annotation. We address this challenge through a different approach by implementing an explicit six-category classification system as a pre-processing step: Match, Match with Overlap, Conflict, Conflict with Overlap, Annotator 1 Only, and Annotator 2 Only, which represents a key functionality of our software tool, as detailed in the previous section. This approach simplifies statistical calculations by providing structured input for subsequent metrics while offering annotators and reviewers clear insight into agreement and disagreement patterns.

For all subsequent agreement calculations, including text anchor agreement, appraisal type agreement, and chance-corrected measures such as Cohen (1960)'s Kappa, we treat both Match and Match with Overlap categories as agreement instances. This approach recognizes that boundary variations do not necessarily indicate substantial disagreement in evaluation identification or classification, consistent with our flexible annotation principles outlined in Section 2.2.

3 Annotation Workflow Development

3.1 Training and Calibration Workflow

Our procedure for annotation training is designed as a systematic workflow to achieve reliable inter-annotator agreement through progressive skill development and quality control. The workflow encompasses three distinct phases, each serving specific methodological purposes in building up annotation competency.

The Foundation and Initial Practice phase establishes both theoretical knowledge and basic operational skills. Annotators begin with comprehensive study of the fundamentals of AT, followed by structured tutorials that bridge theoretical concepts and practical application. The initial practice involves supervised annotation of four articles with iterative feedback, ensuring that the annotators have developed proper software proficiency and terminological accuracy before proceeding to independent work in subsequent stages.

The Pilot Study and Problem Identification phase serves as both a competency test and a diagnostic tool. Eight articles were annotated independently by both annotators, generating comprehensive data for multi-dimensional statistical analysis. This phase is designed to reveal recurrent inconsistencies and conceptual confusions that require targeted intervention. The pilot study functions as a critical checkpoint to reveal specific areas where annotation guidelines need refinement and where annotators require additional training.

The Iterative Training and Quality Assurance phase addresses identified problems through targeted training cycles that alternate between collaborative learning and independent practice. Joint annotation sessions provide real-time guidance on problematic cases while independent practice allows for skill consolidation and performance monitoring. This iterative approach continues until the final assessment demonstrates that annotators can consistently achieve the targeted reliability threshold of 0.8 inter-annotator agreement.

3.2 Pilot Annotation and Performance Analysis

3.2.1 Initial Agreement Analysis

The pilot study produced 810 total annotation instances across both annotators, with substantial variations in annotated segments. Annotator 1 identified 520 segments of evaluation while Annotator 2 identified 631, indicating different thresholds for marking up evaluative language. Only 341 instances (42.1%) were commonly annotated, suggesting significant differences between the two. The distribution in Figure 2 reveals the extent of disagreement across the six-category system: Match (113 instances), Match with Overlap (21), Conflict (160), Conflict with Overlap (47), Annotator 1 Only (179), and Annotator 2 Only (290). The substantial proportions of unique annotations observed here highlight the scope of challenges in the annotation or, rather, the subjective interpretation of evaluative language.

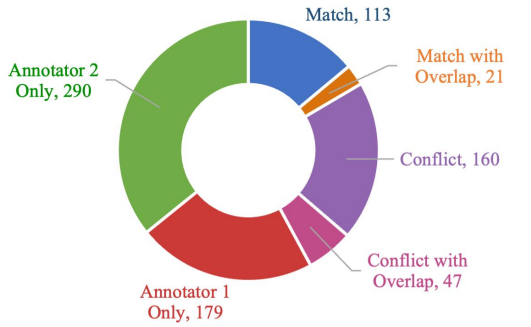


Figure 2: Distribution of annotation pairs (pilot study)

An agreement analysis using our extended classification framework revealed moderate consistency levels. The AGR values show asymmetric agreement patterns: AGR(1‖ 2) = 65.6% and AGR(2‖ 1) = 54.0%. These values are lower than comparable studies. For instance, Read and Carroll (2012) reported 70.6% and 68.6% respectively. This reflects significant disagreement between the two annotators regarding the identification of evaluative segments, reinforcing our initial suspicion that the interpretation of evaluative language is subjective and hence fundamentally controversial, a primary concern that triggered the present study in the first place.

	F_1	REC	PRE	ERR	UND	OVG
1 w.r.t. 2	0.233	0.212	0.258	0.835	0.460	0.344
2 w.r.t. 1	0.233	0.258	0.212	0.835	0.344	0.460
Mean	**0.233**	**0.235**	**0.235**	**0.835**	**0.402**	**0.402**

Table 2: MUC-7 test scores applied to all annotation instances (pilot study)

MUC-7 metrics (Chinchor, 1998) were applied to provide the detailed assessment in Table 2. Results indicate an overall F_1 score of 0.233 and a high error rate of 83.5%, revealing substantial room for improvement.

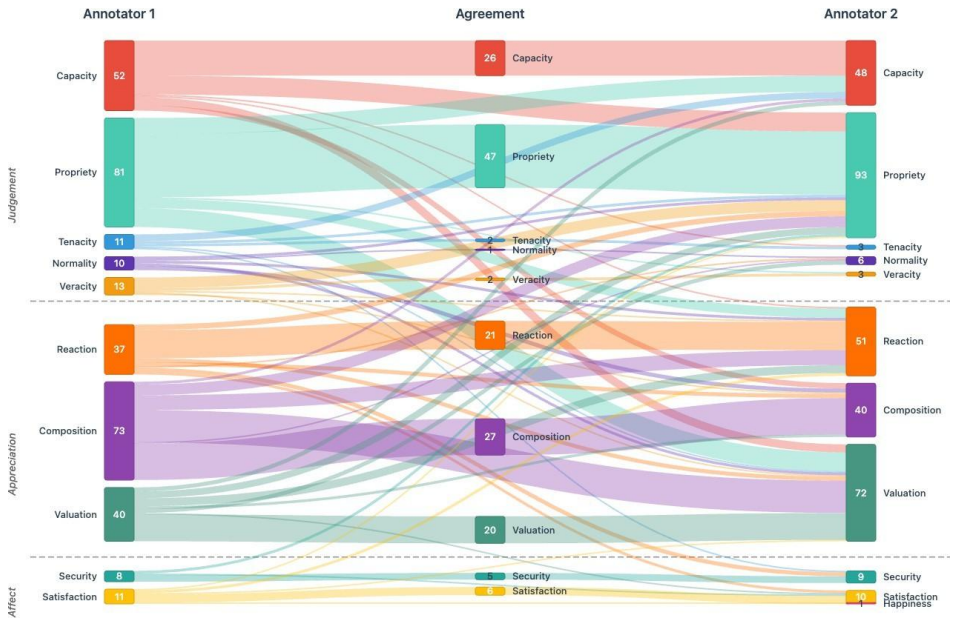


Figure 3: Annotation framework for Appraisal theory

3.2.2 Hierarchical Performance Analysis

To assess annotator agreement at different levels of complexity, we conducted a hierarchical analysis examining four progressive levels of classification requirements:

Level 0: Role identification (appraisal[1])
Level 1: Role + Main Category (Affect, Judgement, Appreciation)
Level 2: Role + Main Category + Subcategory
Level 3: Complete classification (Role + Main Category + Subcategory + Polarity)

Level	Overall	Affect	Judgement	Appreciation
1	0.551	–	–	–
2	0.360	0.478	0.429	0.506
3	0.354	0.400	0.380	0.420

Table 3: Cohen's Kappa values at different levels (pilot study)

Coohen's Kappa values, as presented in Table 3, show strong agreement at the main category level (κ=0.551) but moderate agreement for subcategories (κ=0.360) and complete classification (κ=0.354). Category-specific analysis reveals that Judgement subcategories were most problematic (κ=0.429), suggesting the need for targeted intervention in this area.

[1]Since Level 0 showed complete consistency (all tested instances were appraisal elements), our data analysis begins from Level 1.

3.2.3 Annotation Pattern Analysis

A Sankey diagram in Figure 3 visualizes the annotation flow patterns and annotators' classification strategies. Annotator 1 demonstrates a conservative approach (267 annotations), preferring Propriety over Capacity (81 vs 52) within Judgement, Composition over Valuation (73 vs 40) within Appreciation. Annotator 2 exhibits a liberal approach (405 annotations), identifying more instances across categories, particularly in Reaction (51 vs 37) and Valuation (72 vs 40).

The diagram helps to identify persistent confusion areas. Within Judgement, significant cross-flows between Capacity and Propriety highlight the difficulties distinguishing the nuances of evaluative language. Appreciation subcategories reveal a significant confusion between Composition and Valuation. Furthermore, cross-category flows between Judgement and Appreciation evidence the theoretical challenges in the determination and, indeed, the ambiguation in relation to the two primary categories of Attitude. These empirical revelations and findings jointly serve to lay an informed foundation for the need of a targeted guideline refinement to address the operational disagreements in annotation scopes, category classes, and subcategory classes.

3.3 Guideline Refinement and Problem Resolution

Based on the findings arising from the pilot study, three major problematic areas were found to require targeted intervention: annotation scope clarification, Judgement versus Appreciation distinction, and subcategory classification refinement.[2] Each area received targeted treatment through specific guideline modifications and focused training exercises.

3.3.1 Annotation Scope Clarification

The pilot study revealed significant disagreements about what constitutes an appraisal expression. The differences were reviewed with a primary focus on annotations marked as 'Annotator 1 Only' and 'Annotator 2 Only', which helped to identify two major areas of disagreement:

Area 1: Factual and administrative reporting
This category includes routine procedural language that one annotator mistakenly identified as evaluative. Consider

[2] *On the one hand, Chinese state media have* ***reported*** *test kit shortages and processing bottlenecks, which could produce an undercount.* (USN0220)

In Example [2], one annotator considered "reported" as indicating a capacity, but we determined that expressions like "confirmed," "reported," and "declared" should not be annotated as they represent a procedural documentation rather than evaluation.

Area 2: Scientific terminology
Similarly, research reporting language was frequently misidentified as appraisal. Consider Example [3]:

[3] *The researchers also* ***found*** *that one deer with Omicron already had a high level of antibodies.* (USN0222)

Here, scientific reporting verbs including "reveal," "found," and "decide" in research contexts function as neutral technical descriptors without evaluative implications.

The two areas jointly constituted a disproportionately large portion of our news corpus and addressing them resulted in significant improvement.

[2]It should be noted that throughout the multiple rounds of training, the annotators encountered several other minor issues, however, this study primarily focuses on addressing the three most significant problems identified during the pilot phase.

3.3.2 Judgement versus Appreciation Distinction

The second major problematic area involved confusion between Judgement and Appreciation, as evidenced by the cross-category flows in Figure **??**. This theoretical challenge has been widely recognized with various proposed solutions. Unlike Bednarek (2009)'s non-prioritized dual-criteria approach, Taboada and Carretero (2012)'s lexical-priority ethics-aesthetics distinction, and Starfield et al. (2015)'s subcategory relocation strategy that over-broadens Valuation, our refined framework builds upon Martin and White (2005) and Thompson (2014) but goes further by establishing systematic target-based classification through entity subdivision and dual-test verification for complex cases.

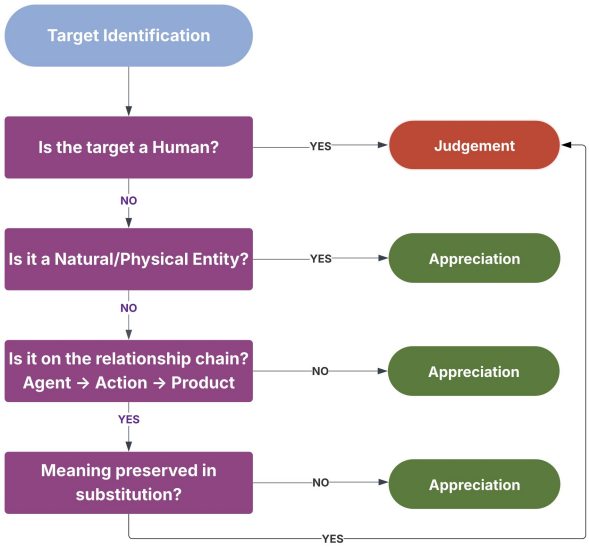


Figure 4: Decision tree for Judgement-vs-Appreciation classification

As illustrated in Figure 4, our framework begins with target identification. By establishing what entity is being evaluated as the foundational step, it ensures that classification decisions are grounded in the evaluative relationship rather than solely in lexis, thereby changing the traditionally lexical orientation of evaluative analysis. We first distinguish between human and non-human targets as the most fundamental step. Human targets receive Judgement classification, consistent with Martin and White (2005). In practice, our approach extends this to include organizations, institutions, and government entities, recognizing that these entities function as collective human agents, as illustrated in the following example:

[4] *Lipkin said he knew of one lab running 5,000*

samples a day, which might produce some false-positive results, [inflating]$_{Judgement}$ *the count.* (USN0220)

In Example [4], the appraised entity "lab" is treated as a collective human agent, and "inflating" is then classified as Judgement.

For non-human targets, rather than directly assigning all non-human targets to Appreciation, an approach that appears to simplify Judgement-vs-Appreciation distinction but creates complications in subsequent subcategorization, we subdivide non-human targets into Natural/Physical Entities and Human-Derived Entities. Natural/Physical Entities receive direct Appreciation classification, consider this example:

[5] *Omicron immediately [caused concern]*$_{Appreciation}$ *in the scientific community because it had 50 mutations compared with the original virus, many of which were known to produce.* (UKG0122)

In [5], the evaluative segment "caused concern" has "Omicron" as its appraised entity, which is a natural/physical entity and therefore classified as Appreciation.

Human-Derived Entities undergo two tests for Judgement classification:

Test 1: Agent → Action → Product relationship: Does the entity represent a human-action product traceable to human agents?

Test 2: Substitution test for meaning preservation: Does evaluative meaning remain consistent when transferred from product to agent?

We demonstrate these criteria through the following examples.

[6] *That development threatens what had been one of the most [important]*$_{Appreciation}$ *defenses against Covid: monoclonal antibodies.* (USN1122)

In [6], while "defenses" derives from human action, transferring the evaluation from "important defenses" to "the developers are important" changes the semantic meaning, failing the substitution test. Consequently, it receives Appreciation classification.

In contrast, consider an example that passes both tests:

[7] *Some information was [fabricated]*$_{Judgement}$ *to spread panic on purpose.* (CNC0620)

Example [7] passes both tests. In this example, "fabricated information" can be traced to human agents (those who fabricated it) and the evaluation transfers meaningfully to the agents: "the agents fabricated information" preserves the evaluative meaning, warranting Judgement classification.

This approach aligns with Martin and White (2005)'s and Thompson (2014)'s emphasis on target-based classification while maintaining practical clarity. By establishing systematic criteria for complex cases, our framework retains theoretical consistency while providing clear decision-making procedures for the category ambiguities that generated substantial disagreements in the pilot study.

3.3.3 Subcategory Classification Refinement

The third major problematic area emerged from fine-grained confusions within the main categories, particularly among the subcategories of Judgement, with subcategory agreement declining to F_1 = 0.460 and Cohen's Kappa values showing moderate agreement (κ = 0.360 overall, κ = 0.429 for Judgement subcategories). The Sankey diagram in Figure 3 demonstrated substantial cross-flows between related subcategories, indicating recurring confusion in fine-grained distinctions.

To address these issues, we developed an interactive clarification tool for annotators encountering classification difficulties during the annotation. This tool provides detailed distinctions for confusable pairs of subcategories, emphasizing functional rather than purely lexical distinctions, with actual annotation examples.

This reference tool allows annotators to quickly resolve subcategory uncertainties during the annotation process, directly targeting the confusion patterns revealed in our analysis of the pilot study. It offers a comprehensive facility for the fine-grained classification decisions that prove most challenging in the initial pilot phase.

4 Results and Discussion

4.1 Progressive Training and Assessment Results

Following targeted guideline refinement, we implemented iterative training to address identified problems. Training began with Round 1 (initial pilot study) that revealed three major problems. After implementing guideline refinements addressing annotation scope, Judgement-vs-Appreciation boundaries, and subcategory distinctions, we conducted successive rounds with targeted feedback and problem-specific interventions. Rounds 2-3

focused on applying refined guidelines, with Round 4 providing final assessment.

As shown in Figure 5, progressive training markedly increased annotation consistency across rounds. Subcategory-level agreement rose from moderate (κ = 0.360) to strong (κ = 0.794), while match rates improved consistently, indicating annotators achieved consistent boundary identification and scope determination. Final assessment reached 85.5% for Main Category agreement and 79.4% for subcategory classification, demonstrating that reliable inter-annotator agreement for fine-grained evaluative annotation can be achieved through progressive training and refined guidelines. These empirical results have provided a convincing validation of our approach and methodology, establishing a replicable framework for achieving the consistent annotation of evaluative language while maintaining a theoretical consistency with the established principles of AT.

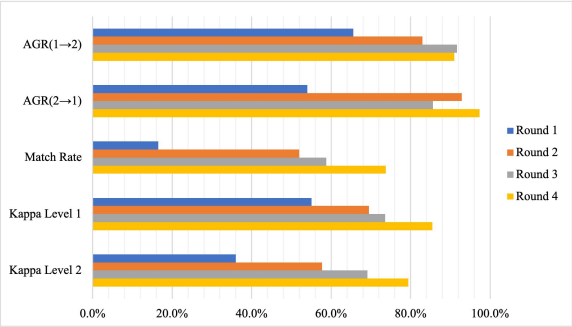


Figure 5: Consistency metrics across four progressive training rounds

4.2 Theoretical and Practical Contributions

This study has significantly contributed to the annotation of evaluative language and related methodologies in several key areas. Firstly, it has established a replicable progressive training workflow capable of transforming novice annotators through structured theoretical instruction and iterative practical feedback. Secondly, our hybrid annotation and comparison tool has proved to be effective in addressing the critical gap in existing annotation software systems, particularly for its detailed intermediate comparison and review process. Thirdly, we introduce a six-category pre-classification system that enhances accuracy and clarity in annotation comparisons, enabling precise targeting of quality control interventions. Finally, the theoretical refinements proposed in this study, particularly regarding Judgement-vs-Appreciation boundary determinations and treatment of human-derived entities, have offered practical and effective operational guidelines while maintaining an alignment with Martin and White (2005)'s foundational principles.

5 Conclusion

This article has described a precursor study in preparation for an international standard (*ISO/PWI 24617-16 Language resource management – Semantic annotation – Part 16*) for the annotation of evaluative language, aiming at providing an empirical basis for the setting of the standard. The study addressed a critical gap in AT research by developing a systematic annotation workflow that achieved reliable inter-annotator agreement for fine-grained analysis of evaluative language that is inherently subjective and ambiguous. Through progressive training protocols and targeted guideline refinements, the study yielded substantial improvements in reliability and consistency, progressing from a moderate agreement (κ = 0.360) in an initial pilot study to a strong consistency (κ = 0.794) in the final assessment. The research reported in this article has contributed to methodological issues relating to the application of AT in three key areas: a replicable workflow to ensure the reliable annotation of evaluative language, the refined theoretical guideline to address persistent classification challenges found particularly in Judgement-vs-Appreciation distinction, and a tool-facilitated approach to ensure targeted corrections and quality control.

Several limitations warrant further research. The study focused on news discourse within the COVID-19 domain, potentially limiting generalizability to other text types and evaluative contexts. The two-annotator design, while sufficient for establishing methodological principles, would benefit from an expansion to multiple annotators for broader validations. Additionally, the theoretical refinements require further testing across diverse discourse domains to confirm their general applicability. Future research should also explore human-AI collaborative annotation approaches, where our refined guidelines and training protocols could inform AI model development, while AI-assisted pre-annotation could potentially enhance human annotator efficiency and consistency in evaluative language annotation.

Acknowledgement

Research described in this article was partially supported by grants received from China National Social Science Fund (Project No 24&ZD28), City University of Hong Kong (Project Nos 9361013, 7020036 and 9360115) and Beijing Social Sciences Foundation (Project Nos 18JDYYA005 and 19YYA001).

References

Monika Bednarek. 2009. Language patterns and ATTITUDE. *Functions of Language*, 16(2):165–192.

Nancy Chinchor. 1998. MUC-7 test scores introduction. In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Matteo Fuoli. 2018. A stepwise method for annotating APPRAISAL. *Functions of Language*, 25(2):229–258.

J. R. Martin and P. R. White. 2005. *The language of evaluation: Appraisal in English*. Palgrave Macmillan, Basingstoke.

Mick O'Donnell. 2018. UAM corpus tool. Computer software, version 3.3.

Jonathon Read and John Carroll. 2012. Annotating expressions of appraisal in English. *Language Resources and Evaluation*, 46(3):421–447.

Sue Starfield, Bridget Paltridge, Robert McMurtrie, Anthony Holbrook, Allyson Bourke, Susan Fairbairn, and Toni Lovat. 2015. Understanding the language of evaluation in examiners' reports on doctoral theses. *Linguistics and Education*, 31:130–144.

Maite Taboada and Mar´ıa Carretero. 2012. Contrastive analyses of evaluation in text: Appraisal in English, German, Spanish and French. *Linguistics and the Human Sciences*, 6(1-3):275–295.

Geoff Thompson. 2014. AFFECT and emotion, target-value mismatches, and Russian dolls: Refining the APPRAISAL model. In Geoff Thompson and Laura Alba-Juez, editors, *Evaluation in Context*, pages 47–66. John Benjamins, Amsterdam.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Multimodal Common Ground Annotation for Partial Information Collaborative Problem Solving

Yifan Zhu[1], **Changsoo Jung**[2], **Kenneth Lai**[1], **Videep Venkatesha**[2], **Mariah Bradford**[2]
Jack Fitzgerald[2], **Huma Jamil**[2], **Carine Graff**[2], **Sai Kiran Ganesh Kumar**[2]
Bruce Draper[2], **Nathaniel Blanchard**[2], **James Pustejovsky**[1], **Nikhil Krishnaswamy**[2]
[1]Brandeis University, Waltham, MA USA
[2]Colorado State University, Fort Collins, CO USA
{zhuyifan,jamesp}@brandeis.edu, nkrishna@colostate.edu

Abstract

This project note describes challenges and procedures undertaken in annotating an audio-visual dataset capturing a multimodal situated collaborative construction task. In the task, all participants begin with different partial information, and must collaborate using speech, gesture, and action to arrive a solution that satisfies all individual pieces of private information. This rich data poses a number of annotation challenges, from small objects in a close space, to the implicit and multimodal fashion in which participants express agreement, disagreement, and beliefs. We discuss the data collection procedure, annotation schemas and tools, and future use cases.

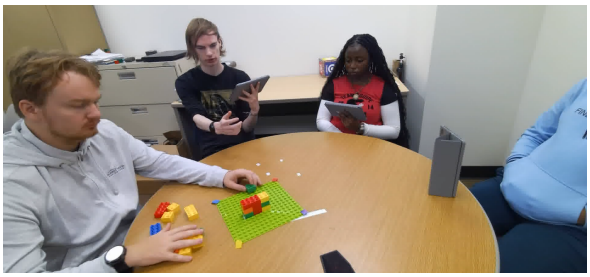


Figure 1: Left to right: a *builder* and 3 *directors* participating in the collaborative construction task with a partially-completed structure on the board. Director 1 (second from left) is indicating the position of a block using a combination of language and gesture with the accompanying utterance "Coming towards me then it's the red long block."

1 Introduction

In collaborative tasks, participants may convey their beliefs, desires, and intentions (BDI) through language, gesture, gaze, and action. These modalities communicate explicit beliefs, disambiguate references, and signal implicit attitudes, enabling participants with different backgrounds or knowledge to build a shared *common ground*—the set of task-relevant facts and evidence jointly accepted by the group. The Edinburgh Map Task (Anderson et al., 1991) is a well-known example of multimodal, conversational, collaborative task annotation and has long served as a benchmark for studying dialogue, spatial reference, and grounding. Our work builds on this tradition with a more complex, co-situated construction task with multiple *instruction givers* and integrates gesture, speech, and action while supporting the study of common ground under structural and spatial ambiguity.

In this project note, we briefly describe the collection and annotation of a novel collaborative problem solving dataset centered around this task. The data is being annotated with multiple modal channels, and the process implicates a number of interesting challenges toward creating semantic annotations that are interoperable across modalities.

The problem of *common ground tracking* (CGT) has been addressed in previous work such as Clark and Brennan (1991); Traum (1994); Ginzburg et al. (1996); Stalnaker (2002); Asher and Gillies (2003); Traum and Larsson (2003), and Hadley et al. (2022). Multimodal approaches to common ground tracking include Khebour et al. (2024b) and Vanderhoeven et al. (2025). However, the tasks addressed in these and similar approaches (Khebour et al., 2024a) suffer from a number of drawbacks, including problems with 1) **agreement/disagreement**: there are few opportunities for disagreement as the task is well-structured with clear solutions at each step; 2) **complexity**: cognitive and interpretive complexity is low as disagreements typically center questions of single-step procedures or computations; 3) **reusability**: once a group has completed the task, they know the answer and cannot organically perform the task again. Our task has been designed to mitigate these shortcomings to to enable the robust study of common ground tracking in multimodal dialogue.

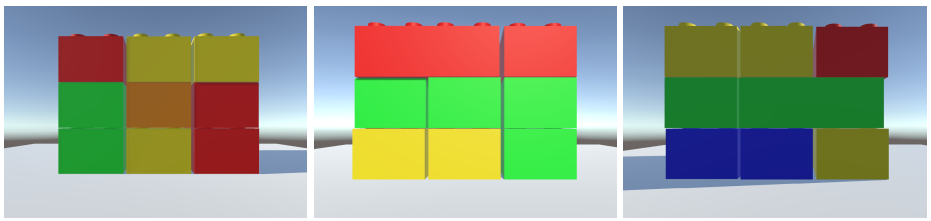


Figure 2: 3 individual side views of a complete structure, each given to a director.

2 Task Description

The task we focus on is a group collaborative construction task structured to satisfy the three conditions enumerated in Sec. 1, that we previously identified as being shortcomings in existing tasks used in the study of multimodal CGT. Namely, we designed the task to create meaningful disagreements within the group about the right course of action, be sufficiently complex such that there are multiple likely solutions toward the goal, and allow participants to do the task multiple times by creating a novel goal each session.

The task is designed for 4 people: 3 *directors* and 1 *builder* (see Fig. 1). Each builder receives a different side view of a 3D structure made of large blocks (see Fig. 2; the directors receive their images on a personal tablet). There are an assortment of blocks on the table before the group, but only the builder (identified in Fig. 1 as the only person without a tablet) is allowed to touch the blocks. The directors are not allowed to show their private images to each other or to the builder. The group must then collaborate to instruct the builder to build a single coherent structure that is consistent with the images given to all the directors. Dialogue is free form and there are no restrictions on what the participants may say, do, or ask each other, as long as the directors do not touch the blocks or show their private images to anyone else. Since there are four sides to the structure but only three images provided, there may be multiple valid solutions. A novel test pattern is generated at the beginning of each session. Thus this task satisfies the 3 desiderata listed above by distributing *partial* information throughout the group, and also simulates a scenario in which a group of people with different background knowledge, expertise, and skills must collaborate to solve a problem.

3 Data Collection

Data was collected at 2 sites, both universities in the United States. The task takes place on a tabletop and is captured using 3 Microsoft Kinect Azure cameras to capture different angles of the task space. Audio is recorded on a single conference-style tabletop microphone. The study was approved by university Institutional Review Boards (IRBs) and participants received USD 15.00 each.

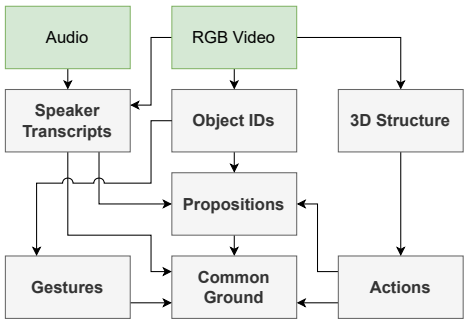


Figure 3: Dependencies between annotations of different modalities. Arrows represent required inputs to the target annotation. Boxes with bolded text are each described in subsections in Sec. 4.

Novel test structures were generated for each group, either manually by the researchers, or procedurally using a script written in the Unity game engine. Test structures consist of blocks arranged in a 3D grid configuration, and screenshots of 3 side views are taken and distributed to the 3 directors. Each session consists of two phases. In the first phase, the test structure contains strictly square or rectangular blocks arranged in a $3 \times 3 \times 3$ grid (wdh – see Fig. 2), with no gaps permitted in the structure. In phase 2, the footprint of the structure is expanded to $4 \times 4 \times 3$, the blocks involved may have curved or angled components, and gaps in the structure are permitted.

In total, after removing recordings with technical or procedural errors, 38 usable group recordings were retained. Most were 20-40 minutes in length.

4 Annotation Schemas

The technical challenges in annotating this data are manifold. Speech overlaps, the objects are close together, actions and gestures may have multiple physical manifestations and interpretations. Additionally, complete annotation of one modality frequently depends on information from another modality, creating dependencies in the annotation

pipeline (Fig. 3). Finally, the sheer amount of data makes purely manual annotation an infeasible task. Therefore, for most modalities, we adopt a semi-automated machine annotation with human validation and post-correction strategy. Specific challenges and methods for each individual modality are given in their respective sections below.

4.1 Speech Transcriptions

Spoken dialogue is transcribed via automatic transcription with the Whisper ASR model (Radford et al., 2023), combined with PyAnnote (Plaquet and Bredin, 2023) for speaker diarization. Annotators review the ASR transcriptions while watching the relevant video, and correct errors in segmentation, transcription, or speaker attribution. We save both the manually-corrected and automatic transcriptions, as research has shown that automated segmentation and transcription errors can have an impact on downstream task performance (Terpstra et al., 2023; Ibarra et al., 2025; VanderHoeven et al., 2025; Venkatesha et al., 2025a), but training models against noisy transcriptions can mitigate this effect (Nath et al., 2025). A zero-shot LLM (Llama-3.1-8B-Instruct) is then used to extract relations among blocks referenced in dialogue, which constitute the project's primary task-relevant signal. All outputs were subsequently manually reviewed and corrected by a human annotator, with omitted instances added, yielding a curated annotation set.

4.2 Gestures

Participant gestures are annotated using Gesture AMR (GAMR; Brutti et al. (2022)), an abstract meaning representation format designed to capture gesture semantics. Gestures may be deictic, iconic, or emblematic, indicating structural descriptions (e.g., *side by side*), block attributes (e.g., *square*, *curved*), or actions (e.g., *bring forward/backward*, *rotate*). These annotations are time-stamped against the video, enabling alignment with utterances and actions. Annotation is performed while watching the video, with access to object IDs so that gestures referring to specific objects can be concretely recorded. For example, a GAMR annotation:

```
(d / deixis-GA
  :ARG0 (d1 / director-1)
  :ARG1 (bs1 / blue-square-1)
  :ARG2 (g / group))
```

indicates that Director 1 (`ARG0`) is pointing not just at a generic blue square, but at a specific block (`ARG1`), with the intended recipient of this reference being the group (`ARG2`). In practice, we treat deictic references as pointing to objects or locations. More fine-grained distinctions—such as path or manner— are considerably harder to interpret. While GAMR could in principle encode these with roles like `:manner` or `:mode`, we do not capture them in our current annotations.

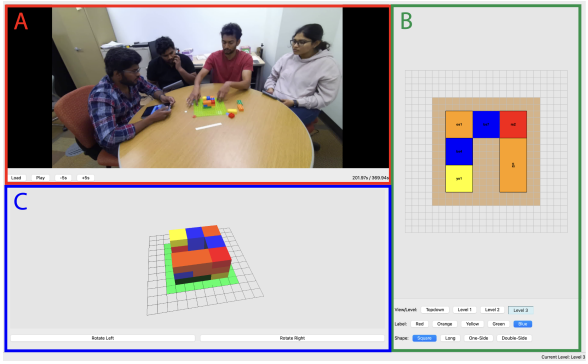


Figure 4: Structure Annotation Tool. **A**: Video Player, **B**: Interactive Area, **C**: 3D View.

Gesture signals are inherently context-dependent and do not provide direct evidence for belief states alone. To address this, we adopt a two-pass contextualization procedure. In the first pass, emblematic attitudinal gestures (e.g., nods, head-shakes) are identified, as these directly license belief updates from gesture evidence. In the second pass, gestures are aligned with the discourse and action layers to assess co-occurrence and temporal contingency (e.g., accompanying an utterance or preceding an action). When such alignment holds, we annotate the corresponding gesture-derived belief state.

4.3 3D Structure

3D structure annotations are intended to capture the state of the board after each time the builder places, moves, or removes a block. To capture these we created a Structure Annotation Tool (SAT), whose interface is shown in Fig. 4. The SAT interface consists of a Video Player (A), an Interactive Area (B), and a 3D View (C). The Video Player displays the video being annotated and the annotator can scrub back and forth as needed. The Interactive Area is a drag-and-drop tool where, if a block is placed on the board in the video, the annotator chooses the color (red, orange, yellow, green, blue) and shape (square, long, single curve, and double curve) of the block and places it on a grid representing the placement board. Blocks can be placed on the bottom level or on top of other blocks in the top-down view, and can then be selected, moved,

rotated, or deleted. The 3D view shows the current structure in 3D, which is rotatable for better visibility. Any actions taken in the Interactive Area are instantly reflected in the 3D View. The construction history is autosaved to JSON as timestamped data that contains all actions along with object IDs and coordinate information on the grid.

4.4 Actions

Annotator actions in the Structure Annotation Tool reflect block placement, movement, or removal actions in the task video. Therefore, actions are automatically extracted from the saved structure annotations. If a block appears at a location where there was none previously, a put action is registered. If a block disappears from a location, a $remove$ action is registered. $move$ can be considered a combination of $remove$ and put such that $move(b, \ell_1, \ell_2)$ can be reified to $remove(b, \ell_1)$ followed by $put(b, \ell_2)$. Locations may be absolute coordinates or relational predicates extracted over coordinates. Relational predicates are restricted to a fixed set such as on and $left$ to avoid creating ambiguous annotations, e.g., where $left(a, b)$ and $right(b, a)$ refer to the same configuration.

4.5 Object Identification

Since the small size of individual objects and the dense configurations of stuctures created during the task pose a tractability challenge for manual annotation, we use a semi-automated approach. However, since automated object trackers struggle to reidentify objects that have disappeared from view, some level of human annotation, validation, and correction is required. We adopt a pipeline as shown in Fig. 5 to maximize accuracy while minimizing human labor cost. The original video is split into 30 second segments and in the first frame of each segment, an annotator manually labels points on distinct objects in the frame. These labeled video segments are then fed into the Segment Anything 2 (SAM-2) model (Ravi et al., 2024) which makes an initial prediction of bounding boxes. The same segments are also fed to a fine-tuned instance of YOLOv11x (Khanam and Hussain, 2024) and the YOLO and SAM-2 bounding boxes are compared for validation. Where SAM-2 missed detections, the failed frames are extracted and returned to the manual keypoints annotation stage and the process repeats. We find that compared to strictly manual bounding box annotation of objects, this pipeline results in up to a 240× speed-up in processing time. Object detection was intended to automatically identify (a) the targets of deictic gestures, (b) the targets of actions, and (c) the positions of blocks within or outside of the structure. The broader goal was to use automatically detected objects to generate complete 3D structures. However, this proved challenging in practice, so to ensure usable data for downstream annotation and analysis, we provided teams with the 3D structure annotations directly.

4.6 Propositions

In Khebour et al. (2024b), common ground is computed in part by extracting expressed task-relevant propositions. Relatedly, Venkatesha et al. (2024, 2025b) develop propositional extraction methods in multiple tasks that realize task relevant propositions as relations between task items or between items and properties. Similarly, in this task, each proposition is indexed by participant ID, timestamp, and the relative relation among blocks, enabling participant-specific retrieval and temporal alignment in the general form "`<timestamps> <person> <block> <relation> <block>`". Propositions must also capture the perspective from which the annotated relation is seen, and the layer of the structure.

The structured propositions are extracted from three modalities—speech, action, and gesture—and integrate them into a unified representation format. Annotations from each modality are first collected independently and then merged into a single CSV file. The entries are sorted chronologically and converted into a standardized belief-annotation schema to support common-ground computation. Speech-based propositions are derived using off-the-shelf large language models (LLMs). For each target mention, the model receives a ten-utterance window (five preceding, five following) to capture discourse cues and resolve coreference. Generic block descriptors are then replaced with specific block identifiers from the action annotations, resulting in propositions such as $on(o1, y2, D3_{side}, layer_1)$. Action-based propositions are converted from absolute spatial coordinates to relative positions between blocks, while side information is supplemented using cues from the speech modality. As gesture annotations contain only the gesturer's identity and gesture content, we incorporate contextual information from both speech and action to complete the final set of gesture-related propositions.

4.7 Common Ground

We track participants' epistemic state updates and define common ground as propositions mutually

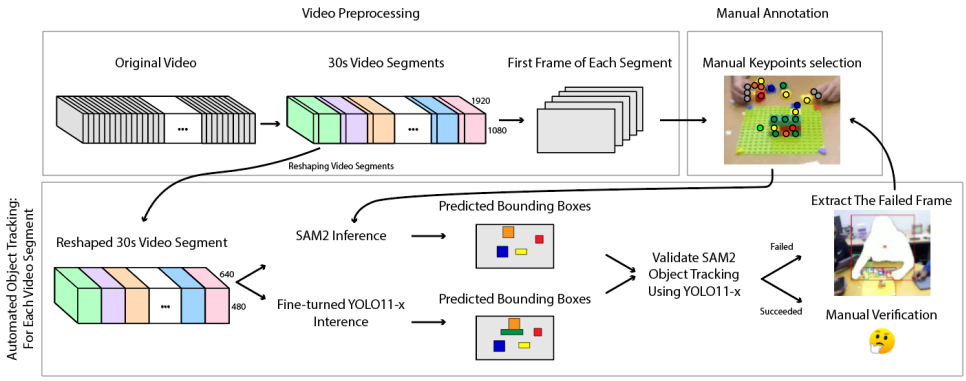


Figure 5: Semi-automated pipeline for capturing object bounding boxes and IDs.

agreed upon. Propositions from different modalities are normalized into a belief-annotation format, $B_x\phi$, where participant x believes (B) propositional content ϕ. Belief formation is licensed by three axioms—Seeing is Believing (Bolander, 2014), Saying is Believing, and Acting is Believing—aligning with the three modalities from which we extract evidence: speech, gesture, and action (gaze is treated as implicit). For example, if participant x asserts φ in dialogue at time t, we record $B_x\varphi$ under the axiom Saying is Believing.

Annotations across different modalities are merged into a single chronologically ordered file and converted into the standardized belief-annotation schema for time-indexed computation. A proposition is considered common ground when the same normalized content is attributed to two or more participants, represented as $CG_{a,b,\ldots}\phi$, where $a, b, \ldots$ denotes the set of participants jointly committed to proposition ϕ. In our dataset, the maximal common-ground set involves four participants. Because beliefs are dynamic, B_x may be revised over time as implicit intentions become explicit in speech or as actions provide evidence that licenses new belief updates.

To capture commitment to or rejection of others' propositions, our scheme further includes $ACCEPT$ and $DOUBT$ labels. When a participant accepts another's proposition, the corresponding common ground annotation is updated to reflect shared understanding; when a participant expresses doubt, a disagreement annotation is recorded to mark epistemic conflict.

5 Conclusion

We have outlined the desiderata, processes, and challenges involved in annotating common ground in a co-situated, multimodal, partially observable collaborative problem-solving task. This type of annotation requires integrating multiple communicative channels with converging dependencies and raises a range of technical, design, and interpretive challenges, for which we have described our approaches and techniques. More broadly, annotation of data of this kind presents challenges familiar to the annotation community, and we hope that our experiences can serve as a useful reference point. Although a gold-standard annotation set and corresponding inter-annotator agreement (IAA) analysis are not yet available, developing them remains a priority for future work. We plan to obtain human annotations, quantify agreement using standard measures (e.g., Cohen's κ, Krippendorff's α), and evaluate the computed annotations against this gold standard. The annotation remains ongoing and a fully-annotated dataset will be released at a future date.

The task's multi-party, partial information setting represents a novel contribution in the age of LLMs. The resulting corpus captures the conversational and information dynamics of a collaboration that is not fully transparent to any of the participants, including any AI system observing the interaction. Therefore, in the context of LLM-driven agents for problem-solving support or human-AI collaboration, our data captures how each participant expresses their implicit "theory of mind" of the other participants' beliefs and goals. The ability to infer such belief states has been shown to be challenging for modern LLMs (Ullman, 2023; Hu et al., 2025), and this challenge is amplified by how even granular task-relevant propositions may be expressed multimodally in this task, including through speech, gesture, and actions. Thus, to fine-tune or assess LLMs for this and similar tasks, or to provide a modern LLM or VLM with sufficient information to interpret participant behaviors in context, an interoperable annotation scheme that captures the semantic relations across modalities

and across time, is required. The efforts described represent a step toward a corpus that would be suitable for fine-tuning, constructing scenarios suitable for assessment of zero-shot prompting, or for benchmarking the recoverability of information in modalities of interest from other modalities.

Acknowledgments

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and by award W911NF-25-1-0096 from the U.S. Army Research Office (ARO). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Nicholas Asher and Anthony Gillies. 2003. Common ground, corrections, and coordination. *Argumentation*, 17:481–512.

Thomas Bolander. 2014. Seeing is believing: Formalising false-belief tasks in dynamic epistemic logic. In *European conference on social intelligence (ECSI 2014)*, pages 87–107.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract meaning representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

Jonathan Ginzburg et al. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*, 1:221–237.

Lauren V Hadley, Graham Naylor, and Antonia F de C Hamilton. 2022. A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1(1):42–54.

Jennifer Hu, Felix Sosa, and Tomer Ullman. 2025. Re-evaluating theory of mind evaluation in large language models. *Philosophical Transactions B*, 380(1932):20230499.

Benjamin Ibarra, Brett Wisniewski, Corbyn Terpstra, Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard. 2025. Investigating automated transcriptions for multimodal cps detection in groupwork. In *International Conference on Human-Computer Interaction*, pages 214–224. Springer.

Rahima Khanam and Muhammad Hussain. 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne M Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, Nikhil Krishnasamy, and James Pustejovsky. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*, 10(1).

Ibrahim Khalil Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A Brutti, Christopher Tam, Jingxuan Tu, Benjamin A Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602.

Abhijnan Nath, Carine Graff, Andrei Bachinin, and Nikhil Krishnaswamy. 2025. Frictional Agent Alignment Framework: Slow Down and Don't Break Things. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Corbyn Terpstra, Ibrahim Khebour, Mariah Bradford, Brett Wisniewski, Nikhil Krishnaswamy, and Nathaniel Blanchard. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, pages 75–81.

David Traum. 1994. A computational theory of grounding in natural language conversation.

David R Traum and Staffan Larsson. 2003. The information state approach to dialogue management. *Current and new directions in discourse and dialogue*, pages 325–353.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Hannah VanderHoeven, Brady Bhalla, Ibrahim Khebour, Austin C Youngren, Videep Venkatesha, Mariah Bradford, Jack Fitzgerald, Carlos Mabrey, Jingxuan Tu, Yifan Zhu, and James Krishnaswamy, Nikhil ane Pustejovsky. 2025. Trace: Real-time multimodal common ground tracking in situated collaborative dialogues. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 40–50.

Videep Venkatesha, Mariah Bradford, and Nathaniel Blanchard. 2025a. Dude, where's my utterance? evaluating the effects of automatic segmentation and transcription on cps detection. In *International Conference on Artificial Intelligence in Education*, pages 144–151. Springer.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 169–180.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, Hannah VanderHoeven, Brady Bhalla, Austin Youngren, James Pustejovsky, and Nikhil Krishnaswamy. 2025b. Propositional extraction from collaborative naturalistic dialogues. *Journal of educational data mining*, 17(1):183–216.

Second Workshop on Multimodal Semantic Representations (MMSR II) at IWCS 2025, Düsseldorf, Germany, merged with the ISA-21 workshop, September 24.

MMSR II Organising Committee

Richard Brutti
Lucia Donatelli
Nihil Krishnaswamy
Kenneth Lai Brandeis
James Pustejovsky

MMSR II Programme Committee

Selene Baez Santamaria	Universidad Nacional Autónoma de Mexico
Abhidip Bhattarachayya	University of Massachusetts at Amherst
Claire Bonial	Army Research Lab
Johan Bos	University of Groningen
Harry Bunt	Tilburg University
Stergios Chatzykiriakidis	University of Crete
Jayeoi Chun	Brandeis Unversity
Robin Cooper	Gothenburg University
Maria Esipova	Bar-Ilan University
Annamaria Friedrich	University of Augsburg
Jonathan Ginzburg	Université Paris Cité
Nikolai Ilinykh	University of Gothenburg
Elisabetta Jezek	Università degli Studi di Pavia
Casey Kennington	Boise State University
Marketa Lopatkova	Charles University Prague
Andy Lücking	Université de Paris Cité
Massimo Moneglia	University of Florence
Larry Moss	Indiana University Bloomington
Ledtitia Paralabescu	Aleph Apha IPAI
Djamé Seddah	INRIA, Paris
Rosella Varvara	University of Turin
Xiao Zhang	University of Groningen

Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content

Maucha Andrade Gamonal[1], **Tiago Timponi Torrent**[1,2], **Ely Edison Matos**[1],
Adriana S. Pagano[2,3], Frederico Belcavello[1], Flávia Affonso Mayer[3,4],
Arthur Lorenzi[1], Natália S. Sigiliano[1], Helen de Andrade Abreu[1],
Lívia Vicente Dutra[1,5], Marcelo Viridiano[1] , André Coneglian[3],
Victor A. S. Herbst[1] , Franciany O. Campos[1],
Kenneth Brown[1], Lívia Pádua Ruiz[1], Lisandra Carvalho Bonoto[1],
Luiz Fernando Pereira[1] and Yulla Liquer Navarro[1]
[1]*FrameNet Brasil, Federal University of Juiz de Fora*
[2]*Brazilian National Council for Scientific and Technological Development – CNPq*
[3]*Observatory for Language and Inclusion, Federal University of Minas Gerais*
[4]*Federal University of Paraíba*
[5]*Department of Multilingualism, Gothenburg University*
maucha.andrade@visitante.ufjf.br; tiago.torrent@ufjf.br

Abstract

This paper presents a multimodal semantic analysis of accessible Brazilian short films using a frame-based annotation approach. We introduce a subset of the *Audition* dataset, comprising six short films from the animation and documentary genres. We analysed three communicative modes: original audio, audio description, and visual content. Trained annotators semantically annotated each mode following the FrameNet Brazil multimodal methodology. To compare meaning across modalities, we used cosine similarity over frame-semantic representations. Results show that audio description aligns more closely with video content than original audio, reflecting its role in translating visual meaning into language. Our findings demonstrate the effectiveness of frame semantics in modelling meaning across modalities and provide quantitative evidence of audio description as a bridge between visual and verbal communication. The dataset and annotation strategies are a valuable resource for research on multimodal representation, semantic similarity, and accessible media.

1 Introduction

Large Language Models (LLMs) have significantly propelled research in Natural Language Processing (NLP). However, these models are still predominantly trained on textual data, while human communication is inherently multimodal, construing meaning through spoken language, sound, gestures, and visual content (Li et al., 2022; Cánovas et al., 2020; Radford et al., 2021). Modelling multimodality is essential not only for advancing NLP in general, but also for developing accessible technologies.

A critical domain for multimodal understanding is production of accessible audiovisual content (Ma et al., 2024; Ye et al., 2024; Lee et al., 2024; Hendricks et al., 2017; Han et al., 2023). Creating inclusive media requires systems capable of interpreting and generating meaning across modalities — particularly between audio and visual content — to support users with visual impairment.

This paper reports an experiment on cross-modal semantic similarity using *Audition*, a frame-annotated multimodal dataset of accessible Brazilian Portuguese short films. We compute similarity across three communicative modes — original audio, audio description and video content — using a hybrid metric that combines frame-based spread activation and cosine similarity.

Our work aims to advance semantically grounded methods for multimodal alignment, contributing both to assistive technologies and to multimodal NLP.

2 Background

2.1 Theoretical foundation: Frame Semantics and FrameNet Brazil

Frame Semantics is a theory of meaning representation developed by (Fillmore, 1982), which posits that understanding linguistic expressions requires access to schematic representations of situations, known as *frames*. Each frame models a conceptual scene that involves participants, entities and relevant objects, referred to as frame elements (FEs).

Words evoke frames and their meanings are interpreted in relation to these structured conceptualizations. A central tenet is that meaning emerges from

the relation between lexical items and the frames they evoke, as well as from contextual factors such as world knowledge and the communicative situation (Fillmore, 1985). For example, in the sentence "*He put the keys in the drawer*," the verb *put* evokes the `Placing`[1] frame, which presupposes the presence of frame elements such as AGENT (*He*),THEME (*the keys*), and GOAL (*the drawer*).

Frame Semantics has been computationally implemented through FrameNet, a large lexical database of English based on the annotation of frames, lexical units and their syntactic and semantic valences (Baker et al., 1998). Lexical Units (LU) are words or expressions that evoke frames. The annotation process begins with a LU linked to a frame annotated in real linguistic contexts and captures semantic information through frame elements, and their syntactic realizations and grammatical functions.

FrameNet includes a set of frame-to-frame relations that ensure conceptual interconnection — such as *Inheritance*, *Subframe*, and *Perspective_on*. These relations demonstrate how lexical meaning is structured within an articulated conceptual system. For instance, the *Cause_motion* frame inherits from *Transitive_action*, decomposes into `Placing` and `Removing` as subevents, and serves as background for frames such as `Bringing`, `Excreting`, `Gathering_up`, and `Ingestion`.

FrameNet Brazil (henceforth FN-BR) builds on this structure, adapting and expanding it for Brazilian Portuguese (Torrent and Ellsworth, 2013). FN-BR introduces additional dimensions to the FrameNet architecture: (i) *frame elements-to-frame* relations, which connect frame elements to other frames and enable recursive modelling of complex scenes; (ii) a mechanism for *metonymy modelling* to represent semantic shifts in which a frame elements evokes a related entity; and (iii) *ternary qualia relations* inspired by the Generative Lexicon (Pustejovsky, 1995), which formalize inferential links among concepts. These extensions significantly expand the model's semantic representational capacity (Torrent et al., 2022).

Additionally, FN-BR extends frame-based modelling to include **multimodal semantic representation**. Drawing on the premise that not only verbal expressions, but also images, gestures, and visual scenes may evoke frames, frame activation operates across multiple communicative modalities, with their elements instantiated by linguistic expressions, visual cues, or bodily signals, thereby enhancing the interpretative scope of the semantic network. This is central to the *ReINVenTA* network described in the following section.

2.2 ReINVenTA and related datasets

Building on the theoretical and architectural advances of FN-BR, ReINVenTA — Research and Innovation Network for Vision and Text Analysis — integrates NLP and accessibility research to advance computational semantic modelling of multimodal meaning. Grounded in Frame Semantics, the network develops gold-standard annotated[2] datasets and methods for multimodal meaning representation.

Two major datasets have already been developed: Frame^2 (Belcavello et al., 2024) and Framed Multi30k (Viridiano et al., 2024).

Frame^2 is a frame-annotated multimodal dataset based on 230 minutes of a Brazilian travel television program. It includes 11,796 semantic annotations for transcribed speech and subtitles, and 6,841 semantic annotations for video segments. Through bounding boxes, each video annotation is associated with a frame and one or more frame elements, linked to a lexical unit.

A recent study based on Frame^2 (Samagaio et al., 2024) investigates the notion of semantic permanence in interlingual subtitling. By comparing frame-semantic annotations in original audio transcriptions and their translated subtitles, it demonstrates how cosine similarity can detect semantic shifts due to translation strategies, as well as to the temporal and spatial constraints in the subtitling process. Ongoing work with Frame^2 includes the development of a multimodal model for turn organization in conversation in audiovisual discourse (Abreu and Matos, 2025).

Framed Multi30k expands the widely used Multi30k dataset (Elliott et al., 2016) with 158,915 image captions in Brazilian Portuguese — both originally created in Portuguese and translated into this language — and more than 4.5 million frame and frame element annotations for English and Portuguese. It also enriches the Flickr30k Entities dataset (Plummer et al., 2015) by aligning phrase-

[1] Following established conventions, frame names are set in `Courier`, and FEs in SMALL CAPS.

[2] For details on the semantic annotation process see sections 3.2 and 4.

to-region correlations with semantic frames.

Building on Framed Multi30k, a new multimodal corpus is currently being constructed for the journalistic domain. Based on image-text pairs extracted from online news portals, it is designed to be automatically processed by NLP and computer vision systems to identify visual entities, events, and semantic relations in real-world discourse-situated contexts.

2.3 Motivation for Audition

Accessible Audiovisual Translation encompasses a set of practices, such as audio description, subtitling, and closed captioning, designed to make audiovisual content accessible to audiences with hearing and visual impairment. This is an inherently multimodal task, requiring the translation of meaning across spoken and visual modalities in an accurate and accessible manner, which poses a significant challenge to current NLP systems, primarily trained on textual data with little regard for accessibility.

The Audition dataset was created to address this gap, by providing a semantically annotated multimodal resource covering original audio, audio description, and video content from accessible Brazilian Portuguese short films. *Audition*'s frame-based structure supports the development and evaluation of NLP models for assistive technologies grounded in multimodal semantic understanding.

3 Design of the Audition dataset

3.1 Corpus composition and communicative modes covered

Audition[3] is a multimodal dataset comprising Brazilian short films with accessibility features spanning a variety of cinematic genres[4], including animation, fiction, autobiography, performance and documentary. Its full version, currently under finalization, comprises more than 240 minutes (over four hours) of audiovisual material.

The dataset includes semantic annotation across verbal and non-verbal communicative modes: original audio (dialogue and narration), audio description, subtitles, closed captions, overlaid on-screen text, and video content. These modes are defined as follows:

- **Original audio (OA)**: spoken language of the film's original soundtrack, including dialogue and narration.
- **Audio description (AD)**: scripted additional narration verbalizing visual information such as actions, gestures, body language, emotional states, settings, and costumes.
- **Subtitles**: written translations of the original dialogue into another language, usually synchronized with spoken content and displayed on screen.
- **Closed captions (CC)**: on-screen textual representations of spoken language and relevant sound cues (e.g., music, sound effects).
- **Overlaid on-screen text (text-overlays)**: written language integrated into the visual composition of the film for artistic or informational purposes, such as titles, narrative inserts, or stylized captions.
- **Video content (VC)**: moving image stream segmented into meaningful visual scenes or shots conveying narrative, spatial, and emotional information.

This work focuses on a subset of the dataset, comprising 71 minutes of semantically annotated audiovisual content extracted from six Brazilian short films belonging to the animation and documentary genres.

Our analysis centres on three communicative modes: **Original Audio**, **Audio Description** and **Video Content**, selected because of their strong semantic interplay: each provides a distinct but interdependent representation of a given narrative sequence. While original audio and audio description operate in the verbal channel, dynamic visuals convey non-verbal semantic content. Their alignment enables a robust investigation of cross-modal semantic similarity and frame-level coherence.

3.2 Methodology and annotation tool

Semantic annotation in Audition follows FrameNet's full text methodology, complemented by the multimodal guidelines established by FN-BR. This framework ensures systematic labelling of frames, lexical units, frame elements, and visual entities in all communicative modalities of the dataset.

[3]The first version of the dataset is available at: https://huggingface.co/datasets/FrameNetBrasil/Audition.

[4]we classify animation as a cinematographic genre distinct from fiction based on(Gordeef, 2023)

We used Webtool [5], a web-based multimodal platform developed within FN-BR to support the integrated annotation of text, audio, images, gestures, and video. This tool allows annotators to work simultaneously with verbal and visual data, grounding their decisions in the conceptual structure defined by FN-BR.

Its interface supports: (i) synchronized visualization of aligned modalities (e.g., audio, audio description, and video segments); (ii) selection of lexical units or visual entities that evoke frames; (iii) assignment of frame elements to verbal spans or visual regions.

Annotation was performed by seven undergraduate junior researchers[6], who were trained on both the annotation tool and FN-BR multimodal guidelines. Expert linguists supervised the process, reviewed annotations when necessary, and solved ambiguous cases collaboratively. Tasks were assigned and completed within comparable timeframes, annotation time per instance not being measured.

We adopted a perspectivist approach to semantic annotation (Basile et al., 2021), which acknowledges that frame selection may vary depending on annotators' perspectives, background knowledge, and contextual interpretation.

4 Semantic Annotation across Communicative Modes

The semantic annotation process covered the three communicative modes under analysis: audio, audio description, and video. The dataset was segmented into aligned analytical units: sentences for verbal modalities and bounding boxes representing events, states and other coherent visual segments for the video. This alignment enables a cross-modal comparison of the semantic structures instantiated in each communicative mode.

Annotation comprised identifying the frames evoked in each modality and labelling their associated frame elements. For verbal content, both in the original audio and in the audio description, frames were assigned based on lexical units identified in the spoken language. For the visual modality, frames were annotated through delimitation of bounding boxes that mark and categorize the salient visual elements, including participants, actions, spatial configurations, and perceptually relevant events.

Whereas in verbal text annotation, frames are directly evoked by lexical units, in video annotation frames are instantiated based on their frame elements within the scene. In addition, visual entities are also annotated and linked to semantic frames, refining object recognition through alignment with conceptual structures.

Semantic annotation was performed at three levels: (i) identification of the evoked frame; (ii) labelling of the associated frame elements; (iii) linking to a lexical unit (in verbal modalities) or to a visual entity (in the video) responsible for frame evocation.

This structure was applied to segments from audio transcriptions and audio descriptions, as well as to video segments involving characters, actions, and objects. Although each modality was independently annotated, using a shared conceptual structure based on FN-BR enabled semantic alignment between communicative modes. Alignment was foundational for the cross-modal similarity analyses conducted in our experiment.

For video annotation, we used bounding boxes that identify scene entities associated with frame evocation. YOLOv3 was initially tested to assist this process, but its generic object detection categories were not adequate for the situational entities required in frame-based annotation. Bounding boxes had to be predominantly created manually by trained annotators, ensuring consistency with the semantic framework.

First, we collected and preprocessed the audiovisual materials, transcribing the audio and aligning transcriptions with the corresponding video segments. Next, we conducted semantic annotation, assigning frames, frame elements, lexical units, and visual entities to both textual and video content.

4.1 Audio annotation

Verbal content derived from audio transcriptions is annotated in two communicative modes - original audio (OA) and audio description (AD) - using the same method, *full text annotation* (Ruppenhofer et al., 2016), in which all semantic frames evoked within the scope of the sentence are identified and labelled.

Figure 1 illustrates audio semantic annotation

[5] Both the Webtool annotation software and the annotated data for FN-BR are available at: https://webtool.frame.net.br/. The FN-BR repository on GitHub can be accessed at: https://github.com/FrameNetBrasil.

[6] All student annotators were funded by Brazilian research agencies.

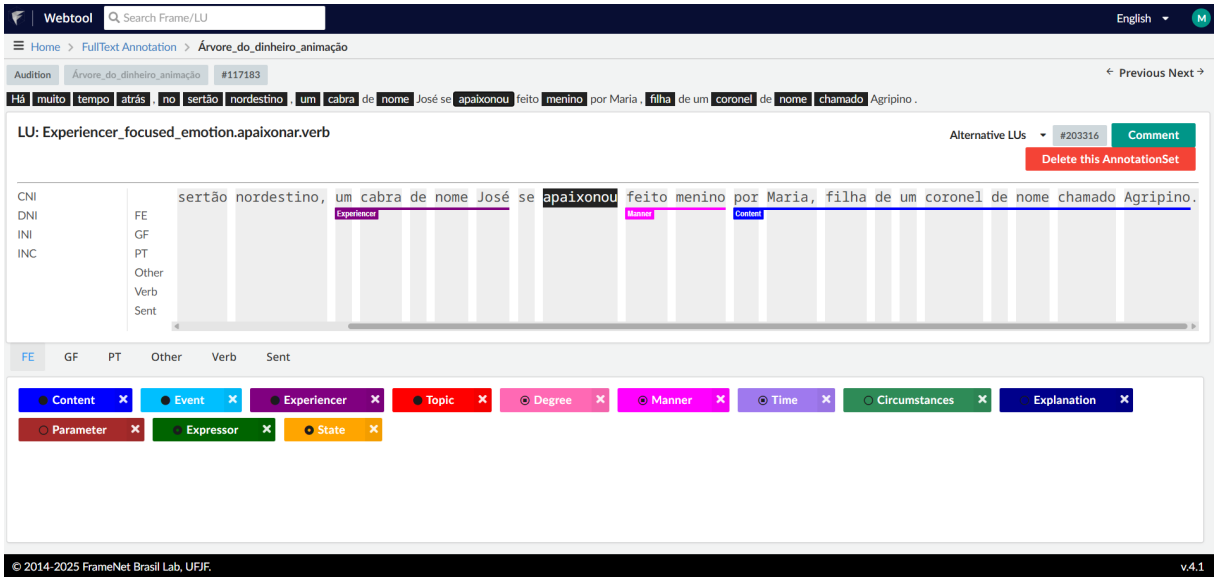


Figure 1: Multimodal annotation of audio: full text annotation

using the sentence[7]:

Há muito tempo atrás, no sertão nordestino, um cabra de nome José se apaixonou feito menino por Maria, filha de um coronel de nome chamado Agripino.

(*A long time ago, in the backlands of the Northeast, a guy named José fell in love like a boy with Maria, the daughter of a colonel named Agripino.*)

Several Lexical Units (LUs) were annotated in this sentence. Figure1̃ highlights the LU *apaixonar-se.v* (*fall in love.v*), which evokes the frame `Experiencer_focused_emotion`. The Frame Elements annotated in the sentence are: EXPERIENCER (*José*), MANNER (*feito menino*), and CONTENT (*por Maria, filha de um coronel de nome chamado Agripino*).

4.2 Video annotation

Video content (VC) is annotated through the creation of bounding boxes across the entire duration of the audiovisual material. These boxes are linked to the Frame Elements of semantically relevant frames. The procedure follows a *text-oriented* approach (Belcavello et al., 2024), semantic annotation being guided by transcribed audio. Our study annotated the visual representation of states, events, processes, and relations that may contribute to the audience cinematic experience.

Since the annotation task is text-oriented, annotators have access to the previous completed textual annotation as well as to the full video. Figure 2 shows an example of this annotation process.

The video sequence corresponds to the narration presented in Figure 1. The annotation consists in marking the visual entity that refers to the EXPERIENCER in the frame `Experiencer_focused_emotion`. In addition to this annotation, the visual entity classified as a Framed Entity is also marked and linked to the appropriate semantic frame, in this case, *homem.n* the frame `Person`.

Bounding boxes semantically anchor visual elements that contribute to narrative construction in both original audio and audio description. Depending on the situations perceived throughout the integration of audio, audio description, and video segments, the annotator can choose to duplicate the same bounding box and associate it with more than one frame by assigning different frame elements.

This is the case in Figure 3. The bounding box delimiting the person is annotated twice: first with the FE SUPPLIER in the frame `Service_client_supplier_interaction`, and second with the FE AGENT in the frame `Manipulation`, since the worker is holding a razor, the instrument used to perform the service. The audio description sequence[8] is as follows:

[7]Transcribed audio retrieved from the original audio of the short film *Árvore do Dinheiro* (Genre: Animation) Available at: https://cinematecapernambucana.com.br/filme/?id=2551. Access on: August 13, 2025.

[8]This transcribed audio is derived from audio description of the short film *Cinema Glória* (Genre: Documentary). Available at: https://cinematecapernambucana.com.br/filme/?id=3267. Acessed on: August 13, 2025.

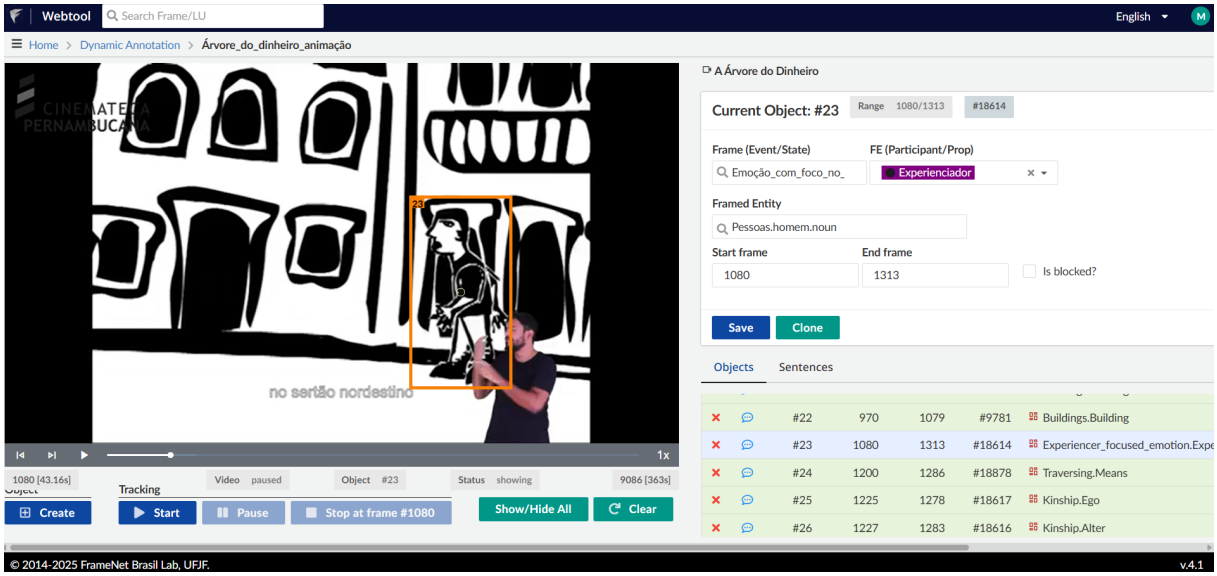


Figure 2: Multimodal annotation of video: bounding boxes anchored to frame element

Embaixo da sombra da árvore, um barbeiro atende um cliente./ O homem está com a cabeça recostada na cadeira de barbeiro./ Um babador branco sobre a camisa e o rosto com espuma./O barbeiro usa uma navalha.

(Under the shade of the tree, a barber attends to a client. / The man has his head on the barber's chair. / A white cape over his shirt and foam on his face. / The barber uses a razor.")

5 Dataset Metrics

5.1 Annotation Totals

Table 1 shows the total number of sentences, annotation sets, frame elements and semantic boxes used in this experiment, based on the selected subset of the Audition dataset, split by film genre. The dataset comprises 894 sentences in the verbal modality (464 from OA and 430 from AD), with 3,979 corresponding semantic annotation sets. Each **annotation set** includes a Lexical Unit (LU), an evoked frame, and the corresponding Frame Elements (FEs) identified in the sentence, totalling 8,189 FEs across both OA and AD. The documentary genre accounts for the majority of the data across all modalities, contributing over 69 percent of the total annotations[9]

The video modality includes 1,103 semantic boxes, representing visual entities anchored to the frame elements of the annotated frames. Each visual entity is associated both with a frame that aligns with the auditory narrative and with a generic entity type and its corresponding frame, supporting automatic visual entity recognition with semantic refinement.

	anim.	doc.	total
AO			
Sentences	142	322	464
Annotation sets	605	1597	2202
FEs	1172	3120	4292
AD			
Sentences	139	291	430
Annotation sets	501	1276	1777
FEs	1604	2293	3897
VC			
Semantic boxes	304	799	1103

Table 1: Data overview of animation, documentary, and total counts for AO, AD, and VC.

5.2 Semantic Similarity Across Modalities

The frame-based semantic similarity metric used in Viridiano et al. (2024) is particularly suitable for this study, as it quantifies semantic alignment across heterogeneous and multimodal data while being grounded in the cognitive principles of Frame Semantics. In this work, FrameNet categories are used to semantically annotate the communicative modes present in the dataset, namely, verbal language and visual representations. The metric evaluates how original audio, audio description and

[9]Dataset balancing will be addressed in future work, after all semantic annotations are completed. Issues related to genre diversity and modality distribution will be explored in subsequent studies, including the validation of the results reported here.

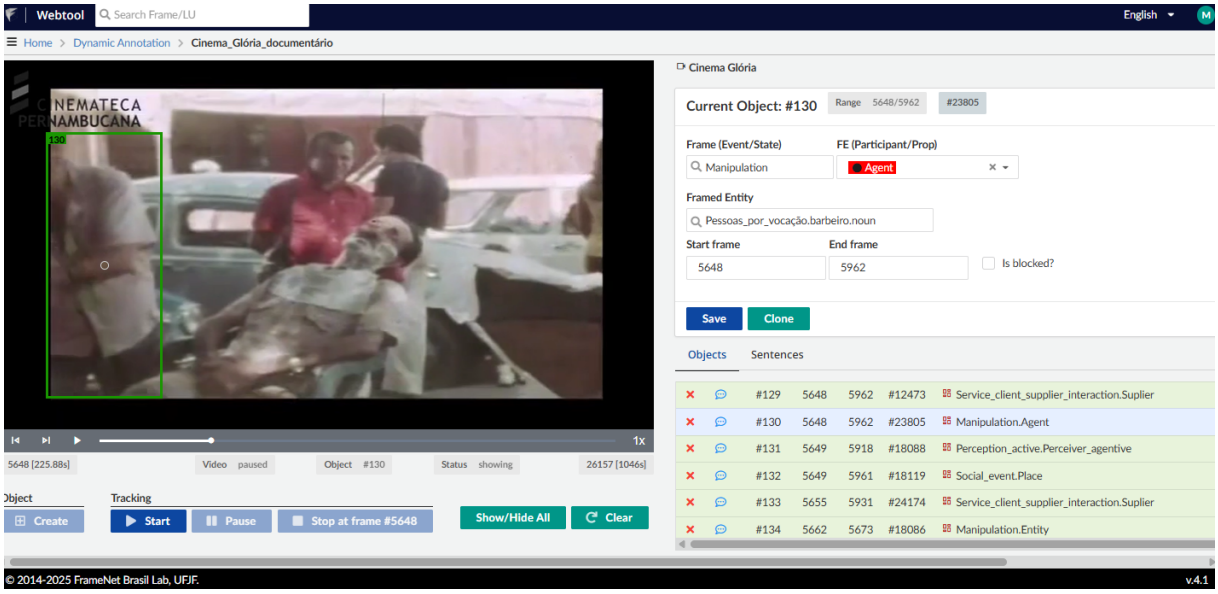


Figure 3: Multimodal annotation of video: duplicated bounding box anchored in more than one semantic frame

video content converge or diverge in terms of frame activation.

This method is crucial for multimodal analysis, where meaning is distributed between semiotic modes with distinct representational properties. The spread activation mechanism enables graded comparisons of semantic similarity, accounting not only for direct overlap but also for conceptual proximity mediated by FrameNet's relational structure. Specifically, the metric quantifies the degree of semantic adherence between video content and original audio and audio description. In this way, it provides an objective measure of how these modalities jointly contribute to filmic meaning construction.

The metric operates in three steps. First, we construct an **association matrix** that encodes frame-to-frame relations defined in FrameNet, including Inheritance, Subframe, Perspective_on, Causative_of, Inchoative_of, and Using. Each relation is weighted according to its semantic proximity, with stronger semantic links receiving higher weights (Gouws et al., 2010). Second, a **spread activation algorithm** propagates activation from the frames directly evoked in the annotations throughout the FrameNet graph.

Activation strength decays exponentially with graph distance, giving higher weights to semantically closer frames and lower weights to distant frames. Third, the spread activation process generates a **vector representation** for each annotated instance, where each vector dimension corresponds to a frame and its cumulative activation value. The semantic similarity between two instances is then computed as **cosine similarity** between their respective activation vectors, producing a normalized score between 0 (no similarity) and 1 (identical semantic content).

As described in 4, the verbal modes selected for comparison are annotated for frames and frame elements derived from the lexical units identified in the transcriptions. The visual mode, in turn, is annotated through bounding boxes, where each box is linked to one or more frame element, depending on the frame evoked. In addition, a framed entity is labelled to designate the entity delimited by the bounding box. This labelling adds semantic refinement to the model by associating a frame from the entity category with its corresponding lexical unit, as illustrated in Figure 2.

Therefore, besides comparing modes with one another, possible variation in the annotations can be compared as seen in Table 2. The comparison can target exclusively the lexical units in the annotations (AD LU / OA LU), which would indicate the extent to which the same entities are present in the verbal modes and in the video. It can also target the frames and FEs used in the annotations (AD Frame and FE / OA Frame and FE), which focuses on measuring the extent to which similar events, states, processes and relations are mentioned across modes. Finally, the comparison can target all annotations simultaneously (AD Full / OA Full).

Cosine similarities obtained for each comparison type are shown in Table 2, which indicates

number of pairs of semantic representations compared (pairs), average normalized cosine similarity (avg), variance (var) and standard deviation (stdev) obtained[10].

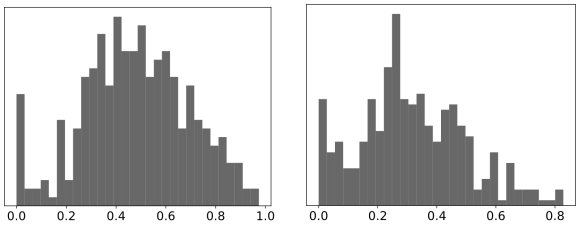
	Video Content			
	pairs	avg	var	stdev
AD LU	323	0.42	0.05	0.22
AD Frame and FE	323	0.38	0.05	0.22
AD Full	323	0.49	0.05	0.21
OA LU	357	0.23	0.03	0.17
OA Frame and FE	357	0.29	0.04	0.19
OA Full	357	0.32	0.03	0.18

Table 2: Cosine similarity between frame-based semantic representations of verbal language modes and video content.

From Table 2, we observe that comparisons including both events, states, processes, relations, and entities mentioned in verbal language with those shown in the video segments yield the highest average cosine similarities. This result is statistically significant across all comparisons, except for the comparison between OA Frame and Frame Element and OA Full (with test statistic $t(317828) = -1.82$, $p < 0.0067$).

These results indicate that audio description is more similar to visual content than to original audio. This result was expected, given that the purpose of audio descriptions is precisely that of providing some sort of access to the content shown in video through audio. By contrast, original audio comes from the film script, which, for most cases, assumes that the video content will be accessed by the person experiencing the film.

The distributions of the cosine similarity values shown in Figure 4 corroborate this claim.



(a) AD Full×VC (b) OA Full×VC

Figure 4: Distribution of similarity values between Audio Description Full, Original Audio Full and Visual Content.

6 Conclusion and Future Work

This paper introduced the first release of the *Audition* dataset, a multimodal corpus of accessible Brazilian short films annotated within a frame-based semantic approach. This initial subset established a methodological basis for cross-modal comparisons and provided quantitative evidence of how accessibility resources such as AD mediate between non-verbal and verbal communication.

We applied FN-BR's multimodal semantic annotation methodology, labelling Lexical Units, frames, frame elements, and visual entities. Cosine similarity analyses show that audio description aligns more closely with video content than with original audio, confirming its role as a mediating modality. These results point to the effectiveness of Frame Semantics for modelling meaning across communicative modes.

A central contribution of our work is the frame-based similarity metric for multimodal data. This methodological innovation has direct implications for accessibility studies, audio description evaluation, and semantic modelling. Its architecture draws on cognitive models of semantic activation spreading, maintaining theoretical alignment with Frame Semantics, which reinforces the interpretability and the conceptual consistency of our analyses.

Future developments will extend the analysis to the full *Audition* dataset, including additional film genres, as well as subtitles, closed captions, and overlaid on-screen text to investigate their roles in multimodal meaning construction. From an NLP perspective, we plan to train and evaluate models for the automatic identification of frames and frame elements across modalities, and explore applications such as the automatic generation and evaluation of audio descriptions grounded in visual content.

7 Acknowledgments

The development of the Audition dataset is one of the initiatives of ReINVenTA—Research and Innovation Network for Vision and Text Analy-

[10]The reference point for cosine similarity is the audio time span as determined by the annotated transcriptions. Thus, all visual entities annotated within the time span of the sentences are considered in the similarity measurement. The alignment between sentences and video annotations is not strictly one-to-one.

sis of Multimodal Objects—, funded by the Minas Gerais State Agency for Research and Development (FAPEMIG – grant RED-00106-21) and the Brazilian National Council for Scientific and Technological Development (CNPq – grant 420945/2022-9). M. A. Gamonal is a postdoctoral fellow supported by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES – grant 88887.015648/2024-00). The construction of the dataset was also funded through grants 88887.936139/2024-00 (CAPES) and 151361/2023-1 (CNPq). T. T. Torrent has a grant from CNPq (311241/2025-5). A. S. Pagano has grants from CNPq (404722/2024-5; 313103/2021-6) and FAPEMIG's program for internationalization of scientific, technological and innovation institutions of Minas Gerais. F. Belcavello was supported by CNPq (200270/2023-0). The authors acknowledge the reviewers of the *Beyond Language: Multimodal Semantic Representations II* Workshop for their valuable feedback and suggestions.

References

Helen de Andrade Abreu and Ely Edison da Silva Matos. 2025. A framenet brasil approach to annotation of pragmatic frames evoked by turn organization gestures. *Caligrama: Revista de Estudos Românicos*, 30(1):94–109.

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.

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.

Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Maucha Gamonal, Natalia Sigiliano, Livia Vicente Dutra, Helen de Andrade Abreu, Mairon Samagaio, Mariane Carvalho, Franciany Campos, Gabrielly Azalim, Bruna Mazzei, Mateus Fonseca de Oliveira, Ana Carolina Loçasso Luz, Lívia Pádua Ruiz, Júlia Bellei, Amanda Pestana, Josiane Costa, Iasmin Rabelo, Anna Beatriz Silva, Raquel Roza, Mariana Souza, and Igor Oliveira. 2024. Frame2: A FrameNet-based multimodal dataset for tackling text-image interactions in video. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7429–7437, Torino, Italia. ELRA and ICCL.

Cristóbal Pagán Cánovas, Javier Valenzuela, Daniel Alcaraz Carrión, Inés Olza, and Michael Ramscar. 2020. Quantifying the speech-gesture relation with massive multimodal datasets: Informativity in time expressions. *PLoS ONE*, 15.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Charles Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

Charles J. Fillmore. 1982. Frame Semantics. In Linguistics Society of Korea, editor, *Linguitics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea. Pages:111–138.

Eliane M. Gordeef. 2023. Avaliação sobre animação e cinema de vida real: semelhanças e diferenças. *Diálogo com a Economia Criativa*, 8(24):50–63.

Stephan Gouws, G-J van Rooyen, and Herman A. Engelbrecht. 2010. Measuring conceptual similarity by spreading activation over Wikipedia's hyperlink structure. In *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, pages 46–54, Beijing, China. Coling 2010 Organizing Committee.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. Autoad ii: The sequel – who, when, and what in movie audio description.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language.

Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. 2024. Nowyousee me: Context-aware automatic audio description.

Zhenhao Li, Marek Rei, and Lucia Specia. 2022. Multimodal conversation modelling for topic derailment detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5115–5127, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jian Ma, Wenguan Wang, Yi Yang, and Feng Zheng. 2024. MS2SL: Multimodal spoken data-driven continuous sign language production. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7241–7254, Bangkok, Thailand. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Mairon Samagaio, Tiago Torrent, Ely Matos, and Arthur Almeida. 2024. Semantic permanence in audiovisual translation: a FrameNet approach to subtitling. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 168–176, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Tiago Timponi Torrent and Michael Ellsworth. 2013. Behind the labels: Criteria for defining analytical categories in framenet brasil. *Veredas-Revista de Estudos Linguisticos*, 17(1):44–66.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Frederico Belcavello, Marcelo Viridiano, Maucha Andrade Gamonal, Alexandre Diniz da Costa, and Mateus Coutinho Marim. 2022. Representing context in framenet: A multidimensional, multimodal approach. *Frontiers in Psychology*, Volume 13.

Marcelo Viridiano, Arthur Lorenzi, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Natália Sathler Sigiliano, Maucha Gamonal, Helen de Andrade Abreu, Lívia Vicente Dutra, Mairon Samagaio, Mariane Carvalho, Franciany Campos, Gabrielly Azalim, Bruna Mazzei, Mateus Fonseca de Oliveira, Ana Carolina Luz, Livia Padua Ruiz, Júlia Bellei, Amanda Pestana, Josiane Costa, Iasmin Rabelo, Anna Beatriz Silva, Raquel Roza, Mariana Souza Mota, Igor Oliveira, and Márcio Henrique Pelegrino de Freitas. 2024. Framed Multi30K: A frame-based multimodal-multilingual dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7438–7449, Torino, Italia. ELRA and ICCL.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD:multimodal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, Torino, Italia. ELRA and ICCL.