22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22)@LREC 2026

12 May 2026

Proceedings of the Workshop

Harry Bunt, editor

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation

ISBN 978-2-493814-64-7

Preface

Welcome to the proceedings of the 2026 edition of the annual ISA workshops on interoperable semantic annotation and representation. The twenty-second edition of this year, hosted by the LREC conference, had only half a day available, where on previous occasions the ISA workshops usually took a full day. The restricted duration of the workshop implied (a) a severe limitation of the number of papers that could be accepted for presentation, and (b) that, following the recommendations of the Program Committee, half of the accepted papers had a full oral presentation and half a short oral presentation and an accompanying poster.

Altogether, a compact, very attractive program has resulted, accompanied by the rich collection of papers in these proceedings.

We thank the members of the Program Committee for their thorough and timely review work, without which this workshop would not have been possible.

The organizers,

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

Organizing Committee

Harry Bunt (chair)
Nancy Ide
Kiyong Lee
Volha Petukhova
James Pustejovsky
Laurent Romary

Program Committee

Jan Alexandersson
Maxime Amblard
Claire Bonial
Johan Bos
Harry Bunt (chair)
Stergios Chatzykiriakidis
Jae-Woong Choe
Robin Cooper
Ludivine Crible
Rodolfo Delmonte
David DeVault
Jens Edlund
Alex Fang
Robert Gaizauskas
Koiti Hasida
Nancy Ide
Elisabetta Jezek
Nikhil Krishnaswamy
Kiyong Lee
Philippe Muller
Rainer Osswald
Catherine Pelachaud
Volha Petukhova
Massimo Poesio
Laurent Prevot
James Pustejovsky
Laurent Romary
Merel Scholman
Purificação Silvano
Manfred Stede
Thorsten Trippel
Carl Vogel
Menno van Zaanen
Annie Zaenen
Heike Zinsmeister

Table of Contents

Workshop Program

09:00 **Workshop opening**

09:00 *Med2Story Referential: A Domain-Specific Extension of ISO 24617-9 for Clinical Narratives Annotation*
Ana Luisa Fernandes, Purificação Silvano, Nuno Guimarães, Luís Filipe Cunha, Rita Rb-Silva and Alípio Mario Jorge

09:20 **Short presentations**

Polysemy and Ambiguity: The Case of the French Modal Verb Devoir
Anna Colli and Delphine Battistelli

A Frame and Canvas-Based Perspective-Encoding Methodology for Multimodal Semantic Annotation of Classroom Settings
Claudia Ferraz, Ely E. Matos, Frederico Belcavello, Julia Gasparetto, Juliana de Oliveira, Janina Wildfeuer and Tiago Timponi Torrent

Leveraging LLMs for Semantic Type Annotation of Verb Arguments
Elisabetta Jezek and Gabriele Errico

Korean Quantification in Abstract Meaning Representation
Kiyong Lee, Chongwon Park, Younggyun Hahm, Harry Bunt and Byongrae Ryu

Towards Corpus-Based Population and Visualization of ISO 24617-8 Ontology (Short Paper)
Maciej Ogrodniczuk and Dariusz Czerski

Tracing Consensus Formation in Meetings: Annotation and Incremental Decision Modelling in the MEET Corpus
Ghazaleh Esfandiari-Baiat and Jens Edlund

GeoAffect: A Multi-Layer Annotation Schema and Few-Shot LLM Evaluation for Geoaffective Analysis of Literary Texts
Fotini Koidaki and Stergios Chatzykiriakidis

10:20 **Poster visit**
10:30 **Coffee break and poster visit (cont'd)**

11:00 *Evaluating the Impact of LLM-Assisted Annotation in a Perspectivized Setting: The Case of FrameNet Annotation*
Frederico Belcavello, Ely E. Matos, Arthur Lorenzi, Lisandra Bonoto, Livia Pádua Ruiz, Luiz Fernando Pereira, Victor Herbst, Yulla Liquer Navarro, Helen de Andrade Abreu, Lívia Vicente Dutra and Tiago Timponi Torrent

11:20 *Annotating Word Meanings over Time: The Trade-off between Scalability, Reliability and Expressivity Power*
Pierluigi Cassotti and Nina Tahmasebi

11:40 *Gaze Behaviour & Conversation Unfolding in the HCRC Map Task Corpus*
Anaïs Claire Murat and Carl Vogel

12:00 Short break

12:05 *CATS: An Annotation Scheme of Causality and Temporal Structure*
Nana Yu, Purificação Silvano, Luís Filipe Cunha and Alípio Jorge

12:25 *ISO-TimeML Semantics for Interlinking Annotations*
Harry Bunt, Alex Chengyu Fang, Kiyong Lee, Volha Petukhova, James Pustejovsky and Purificação Silvano

12:40 *From Categories to Decisions: A Framework for Attitudinal Analysis of Evaluative Language*
Jiamei Zeng, Haitao Wang, Harry Bunt, Xinyu Cao, Min Dong, Tianyong Hao, Kiyong Lee, James Pustejovsky, Laurent Romary, Jianfang Zong, François Claude Rey, Sylviane Cardey, Yangli Jia, Shengqing Liao and Alex Chengyu Fang

12:55 Workshop closing.

Med2Story Referential: A Domain-Specific Extension of ISO 24617-9 for Clinical Narratives Annotation

Ana Luísa Fernandes[*‡‖], **Purificação Silvano**[*‡‖]
Nuno Guimarães[*†], **Luís Filipe Cunha**[*†] **Rita Rb-Silva**[§¶], **Alípio Jorge**[*†]
[*] Institute for Systems and Computer Engineering, Technology and Science (INESC TEC)
R. Dr. Roberto Frias, Porto, Portugal
{ana.l.fernandes, purificacao.silvano, nuno.r.guimaraes, luis.f.cunha, alipio.jorge}@inesctec.pt
[†] Faculty of Sciences, University of Porto
Rua do Campo Alegre, Porto, Portugal
[‡]Faculty of Arts and Humanities, University of Porto
Via Panorâmica, Porto, Portugal
[‖]Centre for Linguistics of the University of Porto (CLUP)
Via Panorâmica, Porto, Portugal
[§] CI-IPOP
Rua Dr. António Bernardino de Almeida, Porto, Portugal
[¶] MedTechnologist
Rua Dr. Manuel de Arriaga, Valongo, Portugal
rrsilva@med.up.pt

Abstract

The semantic annotation of clinical narratives is particularly challenging due to the complexity of medical discourse and the need to integrate linguistic, semantic, and domain-specific information within a unified framework. Existing schemes tend to fall into two categories: general-purpose frameworks, which offer robust linguistic modelling but lack specialised medical representation, and domain-specific schemes, which capture clinical content yet often fail to distinguish fundamental semantic types, especially eventive expressions and referential entities. To address this gap, this study proposes Med2Story Referential, a new extension of the Text2Story annotation scheme (Silvano et al., 2021; Leal et al., 2022) (based on ISO 24617-9: 2019) dedicated to referential entities in clinical narratives. Building on previous work that introduced a specialised branch for eventive entities (Fernandes et al., 2025a), and informed by the UMLS Metathesaurus and expert validation from a consultant haematologist, the extension introduces eight referential categories that refine the representation of clinical actors, substances, biological entities, instruments, and documentation. The results show that ISO 24617-9: 2019 can be applied to this type of text; however, several adaptations are required, particularly with regard to the grammatical domain and the inclusion of specialised domain labels. Nonetheless, the annotation experiment conducted to validate our proposal showed that the annotation scheme and its accompanying guidelines enable a comprehensive and detailed representation of both grammatical and medical aspects. Moreover, the results indicate that the scheme can be applied effectively by annotators without medical expertise.

Keywords: clinical narratives, Electronic Health Records, annotation scheme, referential annotation, health data

1. Introduction

Electronic Health Records (EHRs) constitute a rich source of clinical information and may play a decisive role in addressing issues related to healthcare quality, clinical decision support, and the reliable flow of information among individuals and departments involved in patient care (Kundeti et al., 2016). In this context, Natural Language Processing (NLP) techniques play a fundamental role by transforming unstructured data into structured data, thereby enabling systematic exploration. Information extraction may be carried out through methods such as Named Entity Recognition (NER) and Relation Extraction (RE), which are central NLP components applied to the clinical domain (Durango et al., 2023). These tasks allow for the identification, extraction, and classification of clinical terms, such as diseases, symptoms, treatments, tests, pharmaceuticals, procedures, and body parts, thereby enabling the automatic recognition of a wide range of clinical concepts (Durango et al., 2023). However, these concepts differ substantially from both a linguistic and ontological perspective, which poses significant challenges to their annotation and computational modelling. For instance, diseases and symptoms correspond semantically to states; treatments, tests, and procedures correspond to events; and pharmaceuticals and body parts to entities with referential existence. Capturing these distinctions in a consistent and operational way requires annotation frameworks capable of representing the

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 1–12
12 May 2026.

semantic and referential properties of clinical entities in a structured manner.

Manual annotation constitutes one of the fundamental pillars of linguistic research and NLP (cf. Snow et al., 2008, Flickinger et al., 2017). In addition to enabling the empirical study of linguistic phenomena, it provides essential reference labels for the training and evaluation of language models across diverse tasks (cf.Pustejovsky and Stubbs, 2012, Pustejovsky et al., 2017, Levi and Shenhav, 2022). It also contributes to the formalisation of linguistic theories by offering a structured framework for their empirical validation (Hovy and Lavid, 2010). The development of robust and coherent annotation schemes is therefore of particular relevance: a well-designed scheme should ensure systematicity, consistency, interoperability, and comprehensiveness, thereby allowing complex linguistic phenomena to be represented in a theoretically grounded and operationally applicable manner by annotators (Beck et al., 2020).

In the clinical domain, existing annotation schemes tend to treat uniformly entities that differ significantly in their linguistic and ontological nature (cf. Campillos-Llanos et al., 2018, Sun et al., 2013, Kundeti et al., 2016). For instance, the i2b2 temporal annotation scheme classifies drugs (e.g., *Diltiazem*) as events, even though they correspond to participants rather than events in linguistic terms. This uniform treatment of ontologically heterogeneous entities hinders the rigorous establishment of referential and anaphoric relations, as well as the construction of a coherent temporal line of clinical events and states. It is therefore essential to adopt consistent semantic principles and linguistic conventions that facilitate the automatic extraction of information and its holistic interpretation since studies have shown significant improvements in clinical information processing resulting from the use of appropriate annotation frameworks (Deléger et al., 2017).

Among existing schemes, Text2Story (Silvano et al., 2021, Leal et al., 2022) stands out as a generalist annotation scheme based on ISO 24617 and adapted to European Portuguese, proposing a layered annotation approach encompassing temporal, referential, spatial, and semantic role layers. The scheme has been successfully applied to the annotation of journalistic narratives in Portuguese (Silvano et al., 2023), literary narratives in Portuguese, English, and Spanish (Cantante et al., 2024), and financial narratives in Portuguese (Leal et al., 2025). Moreover, the temporal layer of the scheme has already been adapted to clinical narratives, resulting in the development of Med2Story Temporal (Fernandes, 2025). In Text2Story, events, i.e., eventive entities that denote something that occurs (e.g., “the patient underwent a full blood count.), are annotated in the temporal layer, whereas participants, i.e., entities with referential status that play an important role in the narrative (e.g., “the full blood count was added to the patient’s electronic record”), are annotated in the referential layer. Each layer includes different labels and attributes that enable a general semantic and morphosyntactic characterisation of the represented entities and the relations between them. While these tags may be suitable for representing general narrative structures, clinical narratives pose additional challenges due to the density and specificity of medical information they contain. In particular, medical texts frequently refer to clinical procedures, examinations, conditions, and treatments whose interpretation requires a more fine-grained representation of their semantic and grammatical properties. The current referential layer of Text2Story does not provide sufficiently detailed labels and attributes to capture this type of domain-specific information in a consistent and structured way. As a consequence, the representation of medical referential entities remains limited, which may affect downstream tasks such as clinical information extraction, normalization, and interoperability with medical knowledge resources.

In this work, we address this limitation by investigating how the referential representation can be extended to better accommodate the specific characteristics of clinical narratives. More specifically, we propose to: (i) assess the extent to which the ISO 24617-9: 2019 standard is suitable for representing the referential structure of clinical narratives in European Portuguese; and (ii) define and integrate an extension of the referential layer of the Text2Story annotation scheme (based on ISO 24617-9: 2019) that enables the consistent representation of medical and grammatical information (Med2Story Referential).

Our main contributions are the following: (i) a referential annotation framework for clinical data based on ISO 24617-9: 2019; (ii) the validation of the proposed annotation framework through an annotation experiment.

The remainder of the article is structured as follows. Section 2 reviews the related work. Section 3 presents the framework used for referential annotation: in Section 3.1, we describe the methodology adopted to develop the extension of the annotation scheme, and in Section 3.2, we introduce the proposed Med2Story Referential framework. Section 4 presents the experimental study: Section 4.1 outlines the research design, while Section 4.2 reports the results of the annotation experiment. Finally, Section 5 presents the conclusions and future work, and Section 6 discusses the limitations of the study.

2. Related work

In the clinical domain, annotation guidelines assume particular importance for the preparation of structured data and for the representation of textual content for human interpretation and information retrieval. They also support the training of supervised learning models and enable the comparative evaluation of language technologies through reference datasets (gold standards). Clinical annotation tasks are, however, especially demanding due to the heterogeneity of clinical texts and the breadth of knowledge they encompass, spanning multiple dimensions such as diseases, signs, symptoms, clinical findings, and procedures, as well as contextual factors including temporality, factuality, and other relevant interpretive parameters (Schulz et al., 2023). Despite this complexity, many existing annotation guidelines have been developed primarily within the context of shared tasks aimed at addressing specific NLP, such as NER and RE (cf. Harnoune et al. (2021); Yao et al. (2015); Miranda-Escalada et al. (2020)). This orientation has often resulted in relatively shallow annotation policies focused on immediate application-driven objectives rather than on comprehensive linguistic and semantic modelling (Schulz et al., 2023).

Several annotation frameworks adopt a layered architecture that formally distinguishes between entities such as participants and events. Examples include the ISO 24617 standards (ISO 24617-1: 2012; ISO 24617-9: 2019) and the Text2Story annotation scheme (Silvano et al., 2021; Leal et al., 2022). While such general-purpose frameworks enable fine-grained morphosyntactic and semantic analysis, they remain largely domain-independent and may therefore fail to capture specialised information required in highly technical fields such as medicine.

With regard to domain-specific annotation, a number of frameworks have been developed, particularly for the temporal dimension of clinical texts. Notable examples include THYME-TimeML (Styler IV et al., 2014), an adaptation of TimeML (Pustejovsky et al., 2003) tailored to clinical narratives. Other TimeML-inspired initiatives include the i2b2 project for English clinical records (Sun et al., 2013), the MERLOT corpus for French (Campillos-Llanos et al., 2018), and MEDSPANER for Spanish (Campillos-Llanos et al., 2025). In addition, recent shared tasks, such as the 2024 Chemotherapy Treatment Timeline Extraction challenge (Yao et al., 2024), aim to advance the automatic extraction of clinical event timelines from EHRs. González-Moreno et al. (2025) proposed an annotation scheme for Spanish clinical narratives concerning nut allergy; however, they do not distinguish between eventive entities (e.g. comorbidities such as "asthma") and purely referential entities (e.g. allergies such as "cat or dog"). More broadly, existing schemes frequently lack clearly differentiated annotation layers and tend to conflate distinct semantic categories, including eventive expressions and referential entities. This limitation reduces analytical precision and constrains interoperability across annotated resources.

General-purpose medical annotation tools have increasingly incorporated functionalities for the recognition and classification of medical entities, as exemplified by MedLEE (Friedman, 2000), MetaMap (Aronson and Lang, 2010), and MetaMap Lite (Demner-Fushman et al., 2017), as well as mechanisms for negation and assertion detection, such as those implemented in cTAKES (Savova et al., 2010) and CLAMP (Soysal et al., 2018). With respect to NER strategies, these tools frequently rely on dictionary-based approaches, which enable the identification of entities in unsupervised contexts and effectively handle infrequent mentions (cf. Campillos-Llanos et al., 2025). Such approaches have been combined with rule-based and machine learning methods and, more recently, with neural models adapted or pre-trained on large volumes of biomedical and clinical data (Campillos-Llanos et al., 2025). However, they remain insufficient for capturing specialised information concerning discourse-relevant referential entities.

To the best of our knowledge, no existing annotation scheme simultaneously satisfies three key requirements: (i) an explicit semantic distinction between eventualities (events and states) and participants (discourse-relevant referring expressions); (ii) a rigorously layered architecture integrating multiple linguistic levels; and (iii) the systematic representation of grammatical, semantic, and domain-specific medical information within a unified framework. This gap highlights the need for a comprehensive annotation model capable of capturing the full complexity of clinical narratives while remaining interoperable, linguistically grounded, and computationally exploitable.

3. Framework for Referential Annotation

3.1. Methodology for Building the Referential Layer

This study aims to assess the applicability of the ISO 24617-9: 2019 framework to the referential annotation of clinical narratives in Portuguese and to develop an extension of the referential component of the Text2Story annotation scheme specifically designed for clinical narratives, hereafter *Med2Story Referential*. To this end, we adopted the four-phases methodological framework proposed by Fer-

nandes et al. (2025b), originally used in the development of the temporal-layer extension of Text2Story for clinical narratives (*Med2Story Temporal*).

Phase 1 – Foundational Preparation: We began with a literature review of annotation schemes capable of supporting referential annotation in clinical narratives. This review showed that most existing schemes do not systematically distinguish between eventive nouns and nouns with referential existence (cf. MERLOT (Campillos-Llanos et al., 2018) and i2b2 (Sun et al., 2013)), and, as previously observed by Fernandes (2025), clinical-domain labels tend to be overly broad. In light of these findings, the referential layer of the Text2Story annotation scheme was selected as the base framework due to its comprehensiveness, interoperability, language independence, and ability to integrate annotations across multiple semantic layers within a harmonised architecture. Furthermore, Text2Story derives from an adaptation of the ISO 24617 framework, incorporating adjustments tailored to European Portuguese, making it particularly suitable for the linguistic characteristics of the dataset analysed. Its interoperability has already been demonstrated through successful adaptations to multiple domains, providing further evidence of its flexibility and robustness while preserving a coherent annotation structure and supporting domain-specific extensions and specialised analytical requirements. Having selected the base framework, we conducted an experiment to identify strengths and weaknesses of the scheme. This involved manual annotation of five pseudonymised medical reports derived from group consultations with five patients diagnosed with Acute Myeloid Leukemia (AML) and followed at IPO-Porto, Portugal. These reports were annotated according to the referential layer of Text2Story. The annotation focused on participants, i.e., discourse-relevant referential expressions.

Phase 2 – Design and Specification: The resulting dataset was analysed to characterise referential expressions and to identify the adaptations required for the Text2Story scheme to adequately represent clinical narratives. To ensure sufficient domain coverage and determine the full set of required medical-domain labels, a broader exploratory analysis was conducted on a larger corpus consisting of 38 pseudonymised real medical reports from nine AML patients followed at the same institution. This extended analysis enabled the identification of additional domain-specific categories and ensured that the proposed extension captured all relevant referential phenomena present in clinical discourse. The proposed extension is presented in 3.2.

Phase 3 – Empirical Validation: A pilot annotation study was conducted with two annotators and one curator. Annotation results were analysed with particular attention to inter-annotator agreement (IAA), sources of disagreement between annotators and the curator, and qualitative feedback. Annotators also evaluated the scheme and its guidelines using Likert-scale questionnaires. Based on these results, refinements were introduced to both the annotation scheme and the guidelines. Findings from this phase are reported in 4.

Phase 4 – Consolidation and Refinement: Finally, the annotation scheme and its guidelines were consolidated and further improved in light of the empirical results obtained in Phase 3.

3.2. Med2Story Referential Proposal

The referential layer of Text2Story, grounded in ISO 24617-9: 2019, focuses on the identification of referring expressions in the text, the determination of the discourse entities they denote, and the specification of relations among those entities (Silvano et al., 2021). With regard to Text2Story and specifically to discourse-entity structure, the attribute *Lexical_Head* was considered, originally comprising the values *Noun* and *Pronoun*. For referential-expression structure, the attributes *Domain* and *Involvement* were incorporated. The former encodes *Individuation*, with the values *Set*, *Individual*, or *Mass*, and includes the attribute *Type*, whose values derive from named-entity classification typologies, namely *PER* (person), *ORG* (organisation), *LOC* (location), *OBJ* (object), *NAT* (nature), and *OTHER*. The *Involvement* attribute specifies the degree of participant engagement in an event, with the values 0, 1, >1, all, or undefined, as defined in the scheme (Silvano et al., 2021).

To represent nominal anaphora mechanisms, Text2Story also employs the objectal links defined in ISO 24617-9, enabling the identification of relations such as objectal identity, part–whole (part_of), membership (member_of), and referential disjunction (Silvano et al., 2021).

Following the experimental annotation carried out in Phase 1, it became evident that certain modifications to the Text2Story referential layer were required in order to enhance its informativeness for capturing both grammatical and domain-specific medical information in clinical narratives. Concerning the *Lexical_Head* attribute, two additional values were introduced: *adverb*, to capture instances such as *Foi transferida para o IPO, onde foi confirmado o diagnóstico.* [“She was then transferred to the IPO, where the diagnosis was confirmed.”], and *adjective*, to account for expressions such as *infeção urinária* ["urinary infection]. Although the head of *infeção urinária* is the noun *infeção*, analysing the adjective as part of the event markable would result in the loss of essential medical information, namely the location of the infection. A similar approach was adopted by Campillos-Llanos et al. (2018), who annotate adjectives such as “bilat-

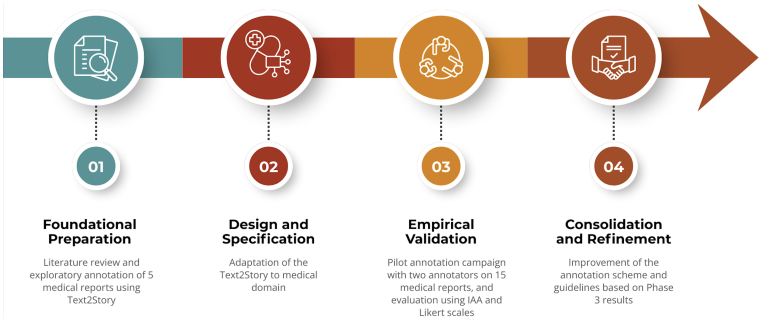

Figure 1: Methodology used for the development and validation of the annotation scheme, based on Fernandes et al. (2025b).

eral" separately from the noun phrase to which they belong. With regard to the *Involvement attribute*, the original set of values defined in the scheme was retained. In contrast, most of the values originally provided for the *Type* attribute proved unsuitable for the medical domain, resulting in many participants remaining unclassified. This observation indicated the need for a comprehensive revision and domain-specific adaptation of the attribute set. To this end, a larger corpus comprising 38 medical reports of different types was re-examined. In order to determine the medical categories required to capture the relevant information contained in the reports, new attributes were proposed inductively as participants emerged in the corpus, using the UMLS Metathesaurus (Bodenreider, 2004) as a conceptual reference framework.

The resulting inventory was organised into eight major categories, each comprising top-level attributes (Figure 2) and second and third-level attributes (Figure 3).

Because numerous nominal expressions with eventive semantics may also receive referential interpretations depending on context, the study additionally involved the development of a set of diagnostic tests designed to distinguish eventive from referential entities. These tests were developed based on the works of Grimshaw (1990), Brito and Oliveira (1997), Borer (2003), and Roy and Soare (2013). They were incorporated into the annotation guidelines to support annotators in consistently differentiating between the two categories and thereby improve annotation reliability.

Regarding the OLINKs, we noticed that they also required significant modifications, which will be addressed in future work. In this study, the focus will be on the participants and their attributes.

In Figure 4, we present an example of a medical report annotated according to the *Med2Story Referential* annotation scheme.

The development of Med2Story Referential was undertaken by a linguist with a background in Pharmaceutical Sciences. As argued by Roberts et al. (2009), the integration of linguistic expertise with

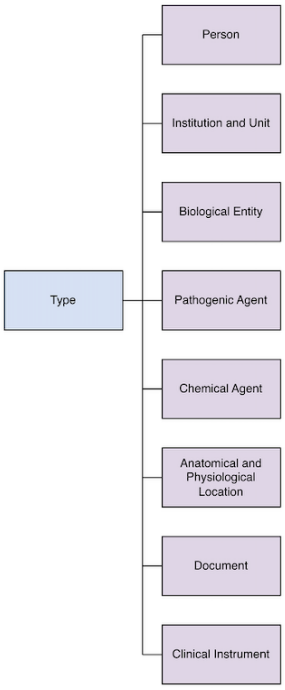

Figure 2: Top-level attributes of *Type*

clinical knowledge in the annotation process tends to enhance annotation quality and reliability. In line with this perspective, the linguistic labels were validated by two specialists from distinct areas of Semantics, while the medical-domain labels were validated by a physician specialising in Haematology. The Med2Story Referential guidelines used for the pilot-study annotation of Phase 3, can be consulted in the the project's GitHub Repository[1].

[1] https://github.com/analuisacardosofernandes/Med2Story-Referential/

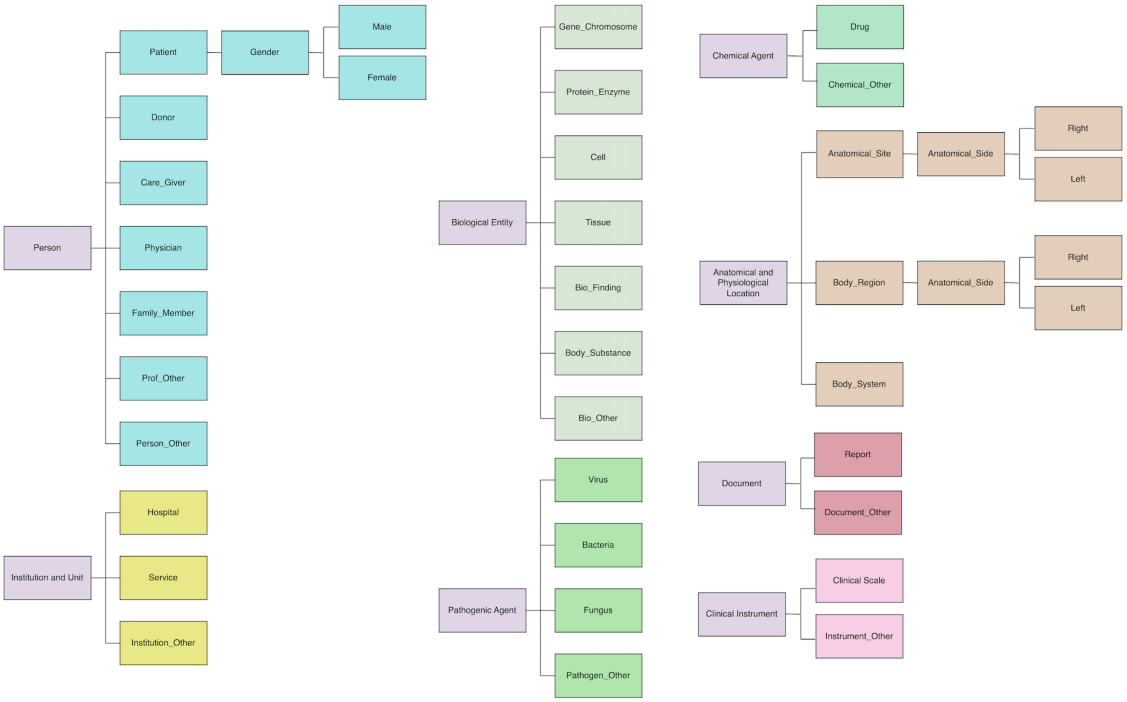

Figure 3: Second and third-level attributes of *Type*

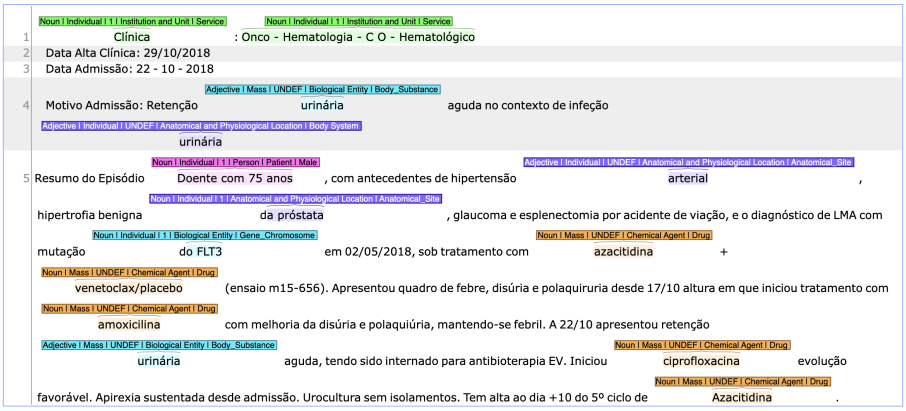

Figure 4: Excerpt from a medical report annotated according to the *Med2Story Referential* annotation scheme. *Episode Summary: A 75-year-old patient with a medical history of arterial hypertension, benign prostatic hyperplasia, glaucoma, and splenectomy following a road traffic accident, and a diagnosis of AML with an FLT3 mutation on 02/05/2018, undergoing treatment with azacitidine + venetoclax/placebo (trial M15-656). The patient presented with fever, dysuria, and pollakiuria since 17/10, when treatment with amoxicillin was initiated, resulting in improvement of dysuria and pollakiuria, although the fever persisted. On 22/10 the patient developed acute urinary retention and was admitted for intravenous antibiotic therapy. Ciprofloxacin was initiated, with favourable clinical evolution. Sustained apyrexia since admission. Urine culture showed no bacterial growth. The patient was discharged on day +10 of the 5th cycle of azacitidine.*

4. Experimental Study

4.1. Research Design

In order to validate the annotation scheme and its guidelines, with particular emphasis on assessing consistency, reliability, and interpretability, we proceeded to Phase 3 of the study. This phase consisted of a small-scale pilot annotation conducted by two linguistics students familiar with the Text2Story scheme, who had extensive experience in referential annotation but no medical expertise. The annotation was subsequently curated by a linguist with a background in Pharmaceutical Sciences, who was also responsible for developing the Med2Story Referential scheme. The corpus selected for this phase comprised 15 expert-generated synthetic medical reports concerning

patients diagnosed with AML. The reports were authored by a Haematologist from IPO-Porto and corresponded to three fictitious patients, each represented by five reports: one group consultation report, three discharge reports, and one general report. The corpus contains a total of 1623 words; detailed corpus statistics are presented in Table 1. Annotation was performed using the INCEpTION platform (Klie et al., 2018).

Table 1: Summary statistics of report length across patient cases, including total word count (TWC), mean report length in words (MRL), and standard deviation (SD).

Patients	TWC	MRL	SD
Patient 1	611	122,8	13
Patient 2	494	98,8	24
Patient 3	518	103,6	16

Prior to the annotation task, a joint meeting was held involving the annotators and the curator. During this session, the guidelines were discussed, doubts were clarified, and one sample medical report was annotated collaboratively to ensure a shared understanding of the scheme. Upon completion of the annotation process, the corpus underwent curation, and the principal sources of disagreement were analysed, both between the annotators and between the annotators and the curator.

Subsequently, inter-annotator agreement (IAA) was calculated. High IAA scores typically indicate clear and operational guidelines, whereas low agreement may arise from a range of factors (Artstein, 2017; Basile et al., 2021; Bayerl and Paul, 2011), such as ambiguities or conceptual difficulties that warrant further refinement of the scheme.

After completing the annotation task, the annotators were asked to complete a questionnaire using five-point Likert scales (1–5). The questionnaire assessed several dimensions: the clarity of the definition of markables to be annotated; the effectiveness of the adaptation of Text2Story to the medical domain; the adequacy of the scheme for identifying and classifying grammatical and medical-domain information; the clarity of definitions and illustrative examples; the robustness of the proposal in resolving ambiguities; the extent to which the scheme provides tools to clearly distinguish events from participants; and the overall coherence and clarity of the framework. The items evaluated through the Likert scales are presented in Table 2.

4.2. Annotation Results

The IAA results for the annotation done in Phase 3 are presented in Table 3. IAA was evaluated using Krippendorff's alpha. In addition, agreement with the curator, who developed the annotation guidelines, was also calculated. The curator's annotations were used as a gold standard, enabling the assessment of how closely the annotators' decisions align with the gold annotation.

Annotator 1 (ANN1) and Annotator 2 (ANN2) agreed with each other on the identification of 135 participants, as well as on the assignment of the corresponding top-level type (e.g. Person, Institution and Unit, Biological Entity, etc.). The gold standard, that is, the curator's annotation, identified 174 participants. ANN1 showed a higher level of agreement with the Curator (CUR), identifying 160 of the 174 participants marked by the curator, whereas ANN2 identified 142 of these 174.

Cases involving discrepancies in the identification (or non-identification) of markables corresponded to two categories: missed markables and non-aligned markables (cf. Table 3). The former refers to cases in which annotators failed to identify a markable present in the gold standard. The latter includes instances in which a markable was identified by the annotator but did not fully match the gold standard span, i.e., the begin and/or end boundaries did not correspond exactly. One example concerns the annotation of *gengivorragias* ["gingival bleeding"] as *Anatomical and Physiological Location*, although this nominalisation denotes an event and therefore should not be annotated in this layer. Non-aligned markables involved cases such as *o aspirado medular mostrou 55% de blastos* ["the bone marrow aspirate showed 55% blasts"]. In such instances, the annotators marked *o aspirado medular* ["the bone marrow"] as a *Biological Entity* with the sub-attribute *Bio_Finding*, whereas only *medular* should have been annotated as *Biological Entity* with the second-level attribute *Tissue*. In this context, the noun phrase *aspirado medular* refers to a diagnostic procedure and therefore corresponds to an event. Only *medular* provides referential information, corresponding to "of the marrow", which denotes a tissue.

Among the markables and top-level types on which ANN1 and ANN2 agreed (135 cases), they also agreed on the second-level attributes (e.g. Patient, Cell, Bio_Finding, etc.) in 108 cases. Agreement on second-level attributes was higher between ANN1 and CUR (146/160) and between ANN2 and CUR (129/142).

Krippendorff's alpha for IAA measuring agreement on markable identification and top-level type assignment was 0.644 for ANN1/ANN2, 0.848 for ANN1/CUR, and 0.701 for ANN2/CUR. With regard to agreement on second-level attributes, the values were 0.771 (ANN1/ANN2), 0.901 (ANN1/CUR), and 0.894 (ANN2/CUR). Overall, these results may be considered robust, indicating that the guidelines were generally clear and could be reliably applied,

Table 2: Items included in the annotator questionnaire, rated on five-point Likert scales.

Parameters	Questions
Markables	How clear and consistent are the guidelines for defining markables?
Grammatical Domain	How effective are the guidelines for identifying and classifying participants at the grammatical level?
Medical Domain	How well-defined and comprehensive are the tags for the medical domain?
Definitions	How clear and detailed are the definitions of each tag?
Ambiguities	How effective is the approach to resolving ambiguities?
Coherence and Clarity	How coherent and clear are the provided guidelines?
Examples	How relevant and illustrative of the tags are the examples?
Text2Story	How effective is the adaptation of the Text2Story scheme for medical reports?
Events vs. Participants	The guidelines provide clear tests that allow events to be distinguished from participants?

Table 3: IAA between ANN1 and ANN2, between ANN1 and the CUR, and between ANN2 and the CUR.

Metrics	ANN1/ANN2	ANN1/CUR	ANN2/CUR
complete_markables	135	160	142
missed_markables	1	3	1
not_aligned_markables	56	20	46
complete_attributes	108	146	129
missed_class_attributes	27	14	13
krip_alpha_markable_total	0.644	0.848	0.701
krip_alpha_attributes	0.771	0.901	0.894

even by annotators without formal training in the medical domain.

One of the main sources of disagreement, which may account for the lower IAA values observed at the second-level attribute, concerns the identification of patient gender. Within the top-level type *Person*, attributes include *Patient*, *Donor*, *Physician*, among others. Within the attribute *Patient*, the sub-attribute *Male*/*Female* is specified. In several cases, gender was not explicitly stated but could be inferred from contextual information, as in *Doente de 74 anos com hiperplasia benigna da próstata* ["74-year-old patient with benign prostatic hyperplasia"] or *Doente de 47 anos, com antecedentes de fibromiomas uterinos* ["47-year-old patient with a history of uterine fibroids"]. Reference to the prostate allows one to infer that the patient is male, whereas reference to uterine fibroids implies a female patient. In some cases, this information only appeared in the patient's third report, making it impossible to infer the patient's sex from the previous reports. The annotators handled such cases differently: ANN1 assigned the sub-attribute female (or male) to subsequent occurrences of the patient once gender had been inferred, whereas ANN2 refrained from doing so.

One way to address this issue would be to resort to cross-document annotation, incorporating the *Objectal_LINK* (OLINK) relation of the *Identity* type in order to capture these cases of coreference. Thus, once the annotator identifies that the patient is female or male, they could mark the nearest markable corresponding to *Patient* and establish, through an *Identity* link, the required coreferential relation, thereby resolving the issue without the need to annotate all preceding markables.

Regarding the analysis of other IAA values for second-level attributes, high agreement was achieved for most attributes (e.g., ANN1/ANN2, ANN1/CUR, and ANN2/CUR all reached 1.000 for *Gene_Chromosome*). However, a low IAA value was observed for the attribute *Cell* between ANN1 and ANN2 (-0.023). The IAA between ANN1 and CUR was similarly low (-0.019), whereas the agreement between ANN2 and CUR was high (1.000). This pattern suggests that the disagreement may be attributable to a lack of expertise on the part of ANN1, while ANN2 encountered no comparable difficulties. Detailed IAA results for all attributes are available in the project's GitHub Repository.

With regard to the analysis of the Likert scale results, ANN1 assigned the maximum score to all parameters (5/5). ANN2 assigned the maximum score to the parameters related to the *grammatical domain*, *coherence and clarity*, *examples*, the adaptation of *Text2Story*, and the tests provided to facilitate the distinction between events and participants. ANN2 assigned a score of 4/5 to the parameters related to the definition of *markables*, noting that the guidelines were clear, with some areas for improvement; to the *medical domain*, stating that the tags are clear and well contextualised, with minor adjustments needed; and to the definitions and ambiguity resolution, remarking that good strategies are provided to handle ambiguities, with a few exceptions. ANN2 reported that the main difficulty concerned the annotation and classification of adjectival participants (e.g., *urinary retention*). To address this issue, we refined the definition of *markable* in the annotation manual and included additional examples. The Likert scale results for each annotator are available in the GitHub Repository.

The results of the Phase 3 curation revealed the identification of a total of 174 participants. With regard to the *Lexical_Head* category, the majority of participants correspond to nouns (152 occurrences) (*paciente com 75 anos* ["75-year-old patient"]), followed by adjectives (21 occurrences) (*hipertensão arterial* ["arterial hypertension"]) and a single occurrence of a pronoun (*iniciou vancomicina 1 g EV a 10/08/2018, que cumpriu durante 7 dias* ["vancomycin 1 g IV was initiated on 10/08/2018, which was administered for 7 days"]).

Table 4: Number of occurrences (Count) of top-level attributes of *Type*.

Top-level attributes of *Type*	**Count**
Biological Entity	51
Chemical Agent	36
Person	28
Anatomical and Physiological Location	24
Institution and Unit	24
Pathogenic Agent	7
Document	4

Regarding the *individuation* attribute, 121 instances received the value *individual* (*o FT3* ["the FLT3"]), 48 the value *mass* (*aspirado medular* ["the bone marrow aspirate"]), and 5 the value *set* (*o cariótipo* ["the karyotype"]).

Concerning the *involvement* attribute, the majority of occurrences (105) were classified with the value *1* (*a próstata* ["the prostate"]), 57 as *UND* (*tratamento com azacitina* ["treatment with azacitidine"], 10 as *>1* (*blastos* ["blasts'"]), and 2 as *all* (*hipertensão arterial* ["arterial hypertension"]).

With respect to the medical domain tags, the results for top-level attributes are summarised in Table 4. The most frequent top-level attribute was *Biological Entity* (51 occurrences), with *Gene_Chromosome* being the most represented *second-level attribute* within this class (33 occurrences) (*o FT3* ["the FLT3"]). By contrast, the least frequent attribute was *Document* (4 occurrences) (*passo o presente relatório a pedido do doente* ["I issue this report at the patient's request"]). The results for the second- and third-level attributes can be found in the GitHub Repository.

5. Conclusions and Future Work

In this work, our objectives were to assess the applicability of ISO 24617-9: 2019 to the annotation of medical reports in European Portuguese and to define and integrate an extension adapted to the medical domain of the referential layer of the Text2Story annotation scheme (based on ISO 24617-9: 2019, with adaptations for European Portuguese).

Our results show that ISO 24617-9: 2019 can be applied to medical reports; however, several adaptations are required, particularly with regard to the inclusion of domain-specific labels such as *Biological Entity*, *Pathogenic Agent*, and *Anatomical and Physiological Location*, among others.

With respect to OLINKs, these are not sufficient to capture the relevant information in this type of text, making it necessary to propose additional links, which we intend to address in future work. Future research will also involve applying the proposed scheme to a larger dataset of real medical reports.

6. Limitations

Regarding the limitations of the annotation scheme, the main issue we would like to highlight is the current absence of links, which would facilitate the capture of relevant information. The links proposed in ISO 24617-9: 2019 are not sufficient to represent the referential information present in medical reports, as they are too generic and not adapted to the clinical context. Therefore, future work will involve proposing a set of links specifically designed for this annotation layer.

Another aspect that may be viewed either as a limitation or as a strength concerns the presence of several second- and third-level attributes within the Type category. Although these attributes make the annotation process denser and more demanding for the annotators, they also enable a more fine-grained and informative representation of the data.

Finally, another limitation concerns the use of synthetic reports and the size of the dataset. Although these reports were produced by a medical specialist, they lack the variability and complexity typically found in real clinical narratives. Nevertheless, this approach was adopted in order to address the ethical constraints associated with the sharing and use of authentic medical reports. In addition, the dataset is relatively small, which may limit the generalisability of the findings. However, the purpose of this annotation was to validate the proposed annotation scheme, which will be implemented in a larger dataset in future work.

7. Acknowledgements

This work was supported by national funds through the Fundação para a Ciência e a Tecnologia (FCT), under PhD grant 2025.00440.BD. The authors also acknowledge support from the StorySense project (DOI: 10.54499/2022.09312.PTDC). We thank Cláudia Couto and Cecília Ortiz for their contribution to the annotation process.

8. Bibliographical References

Alan R. Aronson and Francois M. Lang. 2010. An overview of metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

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

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo

Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

Petra Saskia Bayerl and Karsten Ingmar Paul. 2011. What determines inter-coder agreement in manual annotations? a meta-analytic investigation. *Computational Linguistics*, 37(4):699–725.

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

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

Hagit Borer. 2003. Exo-skeletal vs. endo-skeletal explanations. In John Moore and Maria Polinsky, editors, *The Nature of Explanation in Linguistic Theory*. CSLI and University of Chicago Press.

Ana Maria Brito and Fátima Oliveira. 1997. Nominalization, aspect and argument structure. In Isabel Faria, Gabriela Matos, Maria de Miguel, and Inês Duarte, editors, *Interfaces in Linguistic Theory*, pages 57–80. Porto, Portugal.

Leonardo Campillos-Llanos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie 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.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2025. Hybrid natural language processing tool for semantic annotation of medical texts in spanish. *BMC Bioinformatics*, 26(1):7.

Inês Cantante, Rute Rebouças, Purificação Silvano, António Leal, and Evelin Amorim. 2024. "el niño who sobreviveu": Anotação semântica do *Harry Potter* em inglês, português europeu e espanhol. In *Book of Abstracts of the 40th National Meeting of the Portuguese Association of Linguistics*, pages 78–79. University of the Azores.

Louise Deléger, Leonardo Campillos-Llanos, Anne-Laure Ligozat, and Aurélie Névéol. 2017. Design of an extensive information representation scheme for clinical narratives. *Journal of Biomedical Semantics*, 8:37.

Dina Demner-Fushman, Willie J. Rogers, and Alan R. Aronson. 2017. Metamap lite: An evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

María C. Durango, Ever A. Torres-Silva, and Andrés Orozco-Duque. 2023. Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research*, 29(4):286–300.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Evelin Amorim, and Alípio Jorge. 2025a. Can ISO 24617-1 go clinical? extending a general-domain scheme to medical narratives. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 41–52, Düsseldorf, Germany. Association for Computational Linguistics.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Rb-Silva, Luís Filipe Cunha, and Alípio Jorge. 2025b. The incremental process of building an annotation scheme for clinical narratives in Portuguese: the contribution of human variation analysis. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 332–343, Vienna, Austria. Association for Computational Linguistics.

Ana Luísa Fernandes. 2025. Temporal structure of clinical narratives in portuguese. Master's thesis, University of Porto.

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

Carol Friedman. 2000. A broad-coverage natural language processing system. In *Proceedings of the AMIA Symposium*, page 270. American Medical Informatics Association.

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

Jane Grimshaw. 1990. *Argument Structure*. MIT Press.

Ayoub Harnoune, Maryem Rhanoui, Mounia Mikram, Siham Yousfi, Zineb Elkaimbillah, and

Bouchra El Asri. 2021. Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*, 1:100042.

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

ISO. 2019. *ISO 24617-9: 2019, Language resource management – Semantic annotation framework – Part 9: Reference annotation framework (RAF)*. The International Organization for Standardization, Geneva. Project leaders: Laurent Romary and Kiyong Lee (SC 4/ WG 2 convenor).

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, Santa Fe, New Mexico. Association for Computational Linguistics.

Srinivasa Rao Kundeti, J Vijayananda, Srikanth Mujjiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1937–1945.

António Leal, Purificação Silvano, Zuo Qinren, Evelin Amorim, and Alípio Jorge. 2025. An annotation scheme for financial news in Portuguese. In *Proceedings of the 21st Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 63–75, Düsseldorf, Germany. Association for Computational Linguistics.

António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Jorge, and Ricardo Campos. 2022. The place of iso-space in text2story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL–ISO Workshop on Interoperable Semantic Annotation within LREC 2022*, pages 61–70. European Language Resources Association.

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

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 2020. Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, CEUR Workshop Proceedings, pages 303–323.

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

James Pustejovsky, José M. Castaño, Robert Ingria, Roser Saurí, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

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

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966.

Isabelle Roy and Elena Soare. 2013. Event Related Nominals. In Gianina Iordachioaia, Isabelle Roy, and Kaori Takamine, editors, *Categorization and Category Change*, pages 123–152. Cambridge Scholars Publishing, Newcastle upon Tyne.

Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Stefan Schulz, Warren Del-Pinto, Lifeng Han, Markus Kreuzthaler, Sareh Aghaei, and Goran Nenadic. 2023. Towards principles of ontology-based annotation of clinical narratives. In *Proceedings of the 14th International Conference on Biomedical Ontologies (ICBO 2023)*, pages 36–47, Sweden. CEUR Workshop Proceedings. 14th International Conference on Biomedical Ontologies (ICBO 2023), 28 Aug–1 Sep 2023.

Purificação Silvano, Alípio Mário Jorge, António Leal, Evelin Amorim, Hugo Sousa, Inês Cantante, Ricardo Campos, and Sérgio Nunes. 2023. Text2story lusa annotated corpus. Dataset.

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.

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

Ergin Soysal, Jin Wang, Min Jiang, Yanshan Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. Clamp – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

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

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Suppl 0):S5–S12.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.

Lin Yao, Hong Liu, Yi Liu, Xinxin Li, and Muhammad Waqas Anwar. 2015. Biomedical named entity recognition based on deep neural network. *International Journal of Hybrid Information Technology*, 8(8):279–288.

Polysemy and Ambiguity: the case of the French modal verb Devoir

Anna Colli, Delphine Battistelli
Laboratoire MoDyCo (Université Paris Nanterre, CNRS)
{acolli, dbattist}@parisnanterre.fr

Abstract

This article focus on a methodology for representing the semantics of polysemous markers whose meanings cannot (or do not have to) be disambiguated, even in context. We name this task (multi-)sense representation and present here the French modal verb *devoir* as a case study. Specifically, we reframe this task — traditionally treated as a multi-class problem — as a multi-label classification problem to account for instances that remain ambiguous due to contextual and intentional factors. In order to fine-tune our model (CamemBERT), we implement an active learning loop to enhance the annotation process and we demonstrate that combining global and local features yields the best results (F1-micro = 0.83; F1-macro = 0.79). The model is then applied on two distinct corpora, showing that the automatic analysis of *devoir*'s modal senses provides deeper insights into modal verb usage and facilitates comparisons across corpora differing in *medium* (spoken vs. written) or genre (e.g. legal discourse). Furthermore, our multi-label approach enables the detection and analysis of double-labeled instances, offering valuable applications, as for example legal discourse interpretation and second language acquisition.

Keywords: modality, polysemy, amibiguity, multi-label classification

1. Introduction

Modal verbs, like other modal cues, give information about the speaker's attitude toward a propositional content (Lyons, 1977; Quirk et al., 1985; Gosselin, 2010). For example, in the English translation of the French sentence (1), the modal verb *must* can signify that Paul is probably at home (in this case, it is an epistemic reading); it can also signify that Paul has the obligation to be at home (in this case, it is a deontic reading). The ambiguity between the two modal readings cannot be resolved without other information. The same ambiguity is already present in the French version, with the modal verb *doit*, and persists in its translation.

(1) Paul doit être à la maison
(en) Paul must be at home

While additional context can sometimes clarify an intended modal sense, corpora often leave such ambiguities unresolved—either because the speaker does not complete the sentence or (more or less) consciously maintains the ambiguity. Let's take the examples (2) to (4), extracted from two spoken corpora [1], which use the French modal verb *devoir*.

(2) je pense qu'elle elle devait se dire que ça craignait rien quoi. (13_11)
(en) I think she must have been telling herself it wasn't risky or anything. (13_11)

(3) pourtant il y avait, il devait y avoir des médias, avant le match, mais, mais ça a été beaucoup moins relayé que le reste, quoi (13_11)
(en) Still, there must have been some media attention before the match, but it didn't get nearly as much coverage as everything else. (13_11)

(4) La recherche de la puissance euh financière doit être un un moyen d'assurer le contrôle des autres et on y est prédestiné. (CFPP)
(en) The pursuit of financial power um must be a way of ensuring control over others and we're destined for it. (CFPP)

In (2), the speaker adds the modal verb *devait* to mark that he is not certain, but there is a high probability that his friend (elle) thought the situation was not scary (epistemic meaning). In (3), the role of *devait* is ambiguous : it could either express the speaker's assumption that there was likely media coverage for the match (epistemic meaning) or that media presence is required by norms in these situations (deontic meaning). Similarly, in (4), the use of *doit* is ambiguous : the speaker might be presenting his declaration as highly probable (epistemic meaning) or as an objective truth based on natural necessity (alethic meaning).

Modal verbs appear very frequently in corpora, whatever the type of corpus (either when regarding its *medium* (spoken vs. written) or its genre (journalistic, encyclopedic, etc.)) are. The fact remains that both their overall frequency and the distribution of their modal senses appear to vary depending on the corpus. Disambiguating their occurrences, including identifying instances where multiple modal senses coexist (ambiguous instances), is then essential when pursuing a diverse type corpus-based analysis of modality. The present study presents a step toward this issue. It presents a model for the French modal verb *devoir* that aims at (multi-)sense representation rather than simple disambiguation,

[1] More details will be give about those corpora in Section 3.1)

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 13–21
12 May 2026.

in both written and spoken corpora. This approach enables the model to disambiguate only when possible and to represent persistent ambiguity when disambiguation is not feasible, whether due to limited context or the possible speaker's/writer's deliberate choice. This article is organized as follows. Section 2 presents different kinds of needs for corpus-based analysis of modality and the state of the art for automatic analysis of modality. Section 3 introduces the corpora we use and the linguistic model employed to describe the modal values of the modal verb *devoir*. Section 4 presents the annotation procedure and the integration of an active learning loop to improve the efficiency and quality of the annotation process. Section 5 presents our BERT-based model and its performance. Finally, in Section 6 we test our model on two corpora (a spoken one and a written one) and we discuss the findings in Section 7.

2. State of the art

Research in various domains examines modal verbs frequency and sense distribution through corpus-based approaches by manually disambiguating modal verbs instances. In second-language acquisition, modal verbs are used to compare learner and native corpora to improve textbook design. In fact, modals are essential for communication but challenging to teach due to their polysemy and lack of direct equivalents in learners' native languages as stated by Li (2024). For example, Bouhlal et al. (2018) compared English spoken corpora Nation (2012) with Québécois learner corpora Martini (2012), finding discrepancies in modal verb use, particularly in the underuse of the deontic *might* and epistemic *should* and *must*. Legal discourse studies likewise analyse modal verbs to highlight their role in shaping legal meaning and authority. For instance, Wu et al. (2025) examined epistemic *may*, must, and related markers in war crimes tribunal trials. Other studies on legal language, for example Jaskot and Wiltos (2017), focus on the translation of modal verbs, given the complexity of their senses and the lack of direct equivalence across languages. For example, the English modals *must*, *may*, *might*, and *should* are all translated into French with the verb *devoir*. As these kinds of studies use manual procedures for disambiguating modal verbs instances, it remains a challenge to do this automatically. Our study aims to bridge this gap by proposing an automatic approach that, on one hand, disambiguates the French verb *devoir* when possible, and on the other, represents ambiguity when it persists.

The earliest studies on modal verbs in NLP relied on rule-based systems or feature-based Support Vector Machines (SVM) aiming to disambiguate modal senses via multi-class classification. The aims of these studies in NLP were varied; for example, enhance modal verbs translation (Baker et al., 2012), analyze their epistemic usage as hedges in the biomedical domain (Light et al., 2004), extract rules from legal regulations (Wyner and Peters, 2011). From a methodological point of view, we can distinguish diverse studies which have followed the evolution of NLP, for example: Ruppenhofer and Rehbein (2012) which proposed an annotation scheme for each modal verb in English, an annotated news domain corpus and logistic regression models based on an ensemble of hand-crafted features, Marasović et al. (2016) which extended this original feature set and applied it to a CNN architecture. Finally, more recent studies which have attempted to solve the problem as a classical modal sense classification task by probing BERT architecture Devlin et al. (2019). As examples for this last case, we can mention Wagner and Zarrieß (2023) which showed that BERT, given the same semantic value, encodes it differently for each modal verb. For this reason, individual classifiers for each verb perform better than a classifier for each modal sense. Finally, Dehouck and Denis (2023) performed classification on BERT's last hidden layer representations of the English modal verbs and their context showing that BERT-based models outperform the frequency baseline and previous models. However, as Owan et al. (2022) noted, this remains a non-trivial task, even for expert annotators. They compare annotations based on two different frameworks and highlight proximity in the interpretation of some labels, even when grounded on distinct theoretical frameworks, and the ambiguity that often leads to multiple acceptable interpretations. Regarding French, little research has focused on the disambiguation of modal verbs using a machine learning or deep learning approach. Nissim et al. (2013) proposed an annotation scheme for French and Italian for modality markers at large and use it to annotate on a spoken corpus. (Colli et al. (2024)) propose a fine-tuned CamemBERT for the modal sense disambiguation on the verb pouvoir. Nissim et al. (2013) and (Colli et al. (2024)) highlight the challenges of disambiguating modal verbs in spoken contexts. Limited contextual cues and intentional speaker ambiguity often make interpretation difficult also for human annotators, increasing task complexity and leading to a scarcity of annotated data. Moreover, although not all instances can be linguistically disambiguated, all methods treat the task as a multi-class classification problem.

In this context, Active Learning provides an effective strategy to overcome data scarcity in annotation. Active Learning is a method for reducing annotation costs and effort by selecting the most informative instances to label for the training pro-

cess (Settles, 2009). Recent studies demonstrate its effectiveness in boosting BERT's performance, especially in low-resource and class-imbalanced settings (Ein-Dor et al., 2020). Active Learning has proven particularly effective, even outperforming larger language models, in domain-specific tasks (Lu et al., 2023), or multi-label classification task, for example, user's intent classification (Zhang and Zhang, 2019) where annotation is expensive and time-consuming.
This study develops a French-specific model that focuses on the (multi-)sense representation of the modal verb *devoir* across written and spoken corpora. By representing modal senses, our approach both disambiguates *devoir* when the context allows only one interpretation and represents ambiguity when it persists. The model's task is framed as a multi-label classification problem to capture cases of contextual or intentional ambiguity, using a BERT-based architecture combined with active learning to enhance annotation efficiency and quality.

3. Corpus and linguistic framework

In this section we present our corpus (3.1) and the linguistic model (3.2) on which the annotation scheme is based.

3.1. Corpus

We train our model on two French corpora, one spoken - named here ES_CP - and one written - named here T2K.

- ES_CP (approximately 250.000 tokens) is composed of 112 semi-structured interviews and short monologues extracted from two different corpora. In the first corpus, named Eslo (Université d'Orléans and CNRS - Laboratoire Ligérien de Linguistique (LLL), 2015), we selected 20 interviews featuring questions to the citizens of Orléans about their habits and feelings regarding their city as well as some conversations. In the second one, named CFPP (CLESTHIA, 2024), we selected 6 interviews containing similar questions but focusing on the city of Paris.

- T2K (approximately 100.000 tokens) is composed of 100 written texts of two different genres : newspaper (60 texts) and encyclopedic (40 texts). The corpus is part of Battistelli et al. (2022).

Lately, we applied our model on two other corpora, 13_11 and legal_CODE in order to show real life applications of our model.

- 13_11 is a corpus of 1750 transcribed interviews (approximately 20 million tokens) about the November 13, 2015 terrorist attacks in Paris and in the Île-de-France region. These interviews are collected as part of the interdisciplinary study Étude 1000 within the Programme 13-Novembre.[2]

- legal_CODE is a corpus of 70 French legal codes (approximately 3 million tokens) extracted from a site that offers an open-source updated, enhanced version of the main French legal codes.[3]

3.2. Linguistic framework

In French, several studies have focused on elucidating the various meanings of the modal verb *devoir*, e.g. (Kronning, 1996; Gosselin, 2010; Barbet, 2012; Vetters and Barbet, 2006). We choose to rely on the model of Gosselin (2010) that retains three possible modal categories for *devoir*. Table 1 presents those three semantic values of our annotation scheme with some examples.

Our task is defined as multi-label annotation. Each instance of the verb *devoir* is assigned one or more labels. Instances annotated with multiple labels reflect either insufficient contextual information for disambiguation or inherent semantic ambiguity, which persists even when the surrounding context is available. For example, in (5) ambiguity persists between deontic ("At that moment, I had to follow my daughters") and epistemic sense ("At that moment, I probably followed my daughters") for *dû*.

> (5) À ce moment là, j'ai dû suivre une des mes filles (CFPP).

4. Corpus annotation

In order to annotate our corpus, we conducted manual annotation in two steps. In the first step (4.1), we performed multi-label annotation on a portion of the corpus. In the second step (4.2), after assessing annotator agreement, we enhanced annotation quality by incorporating an active learning framework into the process.

4.1. Annotation procedure

Firstly, we annotated 207 instances from our training corpus described in Section 3.1. The annotation was carried out by three expert annotators using Glozz (Widlöcher and Mathet, 2012). We then calculated Inter-Annotator Agreement (IAA) using Krippendorff alpha obtaining a score of 0.8.

[2] https://www.memoire13novembre.fr/
[3] https://codes.droit.org

Label	Definition and Examples
deontic	**Definition:** *devoir* with a sense of obligation whose source is a human being, norms or conventions, or external circumstances. **Example:** "Je crois qu'après la troisième euh les enfants doivent passer un un examen obtenir un diplôme" — (en) *I think that after middle school uh children must take an exam to obtain a diploma.* (ESLO_ENT_089)
alethic	**Definition:** *devoir* with a sense of necessity grounded in a law of nature. **Example:** "Pour moi la voiture est un un outil c'est un véhicule rien de plus ça doit être pratique" — (en) *For me a car is a tool it's a vehicle nothing more it has to be practical.* (CFPP)
epistemic	**Definition:** *devoir* with a value of strong probability. **Example:** "Il devait être 22 heures, entre, ouais à peu près, 22h ou 10 heures et quart." — (en) *It must have been around 10 p.m., yeah, roughly 10 or a quarter past ten.* (13_11)

Table 1: Annotation scheme of *devoir* based on Gosselin (2010)'s approach.

4.2. Active Learning loop

After the first manual annotation step, we observed that many instances of *devoir*, particularly in the ES_CP corpus, required extensive reflection by the annotators. In fact, as noted by Owan et al. (2022), the modal verb disambiguation annotation task is challenging even for expert annotators, as it requires more attention than many other linguistic features. To address this, we integrated an active learning loop into the annotation procedure to prioritize the most informative examples, to improve both annotation speed and quality. We chose to implement an uncertainty-based active learning loop as this strategy shows better performances over random sampling for a classification task for BERT models (Jacobs et al., 2021) and LLMs (Lu et al., 2023). Uncertainty-based Active Learning aims to identify the most uncertain instances for subsequent training iterations. Within each iteration, the model operates on a randomly sampled subset of the training data locating instances with the highest entropy in model predictions. We provide a pool of 150 unannotated texts from our corpus. At each iteration, the model (described in Section 5) —initially trained on the dataset annotated in Step 1 (4.1)—is used to infer probabilities over the pool. The *n* instances for which the model is least confident are then selected for annotation. Annotators label these examples and can stop the loop at any point. After each iteration, the model is retrained on the newly annotated batch and is ready for the next iteration. To further validate the effectiveness of this approach over a random sampling baseline, we plan an additional experiment in which annotators will label a comparable set of instances selected at random, allowing us to directly compare model performance improvements across the two sampling strategies. During this active learning step we annotate 214 devoir instances reaching a total of 418 instances with the following distribution (see Figure 1).

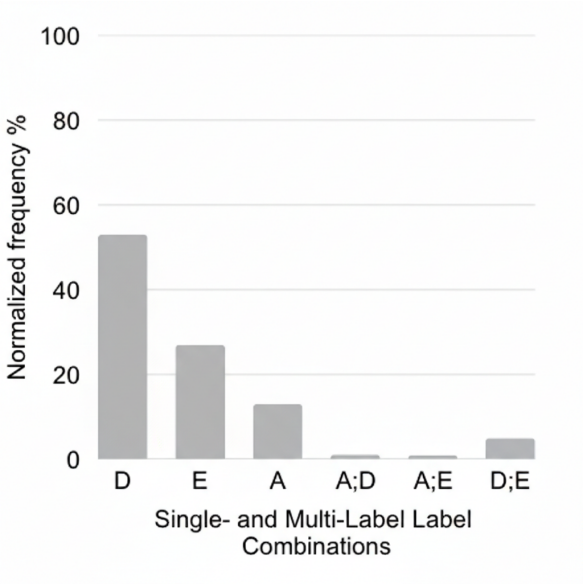

Figure 1: Distribution of annotated instances.

5. Experiments and results

In our experiments, the primary objective was to provide a multi-label classification of *devoir* in order to take into account instances that cannot be disambiguated due to lack of context or intentional ambiguity by the speaker.

5.1. Model architecture

Before fine-tuning the model, the dataset was segmented into individual sentences, and instances of *devoir* were automatically identified using Stanza syntactic analyzer (Qi et al., 2020). The positions of these instances within each sentence were recorded and later used to extract their corresponding embeddings for classification. In sentences containing multiple *devoir* instances, the positions of all instances were considered separately, as each occurrence may convey a distinct modal meaning.

This phenomenon is particularly common in the spoken corpus, where sentences tend to be longer. In example (6), the same sentence contains an epistemic devoir (*dû*) and a double-labeled epistemic-deontic one (*devais*).

> (6)Avec le mode de garde de cette époque-là je devais l'avoir sûrement en milieu de semaine à ces moments-là donc j'ai dû passer du temps avec ma fille et aussi passer à autre chose (ESLO_ENT)
> (en) *With the custody arrangement at that time, I must have had her probably in the middle of the week back then so I must have spent time with my daughter and also moved on to other things.* (ESLO_ENT)

We fine-tuned the CamemBERT model [4] (Martin et al., 2019) for multi-label classification performing 5-fold cross-validation using 80% of the data for training, 10% for validation and 10% for testing. The model was fine-tuned using a learning rate of 2e-5 and a batch size of 8. Training was conducted for a maximum of 10 epochs, with early stopping. To address class imbalance and well-calibrate the model, we incorporated class-specific weights in the BCEWithLogitLoss function. We experimented two embeddings architectures :

- *local*: in this setup, each sentence was encoded using CamemBERT, and the last hidden layer representation (12th layer) corresponding to the target *devoir* token was extracted. This 768-dimensional token embedding was then fed directly into a classification head.
- *global_local*: in this setup, after encoding each sentence, we extracted the embedding of the target *devoir* token and the sentence embedding [CLS]. These embeddings are then concatenated and passed through the classifier.

local embeddings follow the standard contextual embedding architecture. They are designed to capture local information about the target token (*devoir*), while inherently encoding some contextual information. In contrast, *global_local* embeddings enrich the local embeddings by incorporating global sentence-level information via the [CLS] token. As demonstrated by Miaschi and Dell'Orletta (2020), although many sentence-level properties are implicitly encoded within individual word embeddings, sentence representations are generally more effective at capturing syntactic and structural features, whereas token-level embeddings better encode raw text and morphosyntactic properties. By combining both, the *global_local* leverages the complementary strengths of local and sentence-level representations, improving the model's ability to capture both token-specific meaning and sentence context. Our approach was inspired by Zhang et al. (2022), who demonstrated that a BERT-based model that combines the [CLS] embedding (representing global features) with selected token embeddings (representing local features) outperforms a model that uses only the [CLS] token for multi-label text classification.

[4] https://huggingface.co/almanach/camembert-base

Model	Micro F1	Macro F1	D	E	A
majority baseline	0.56	0.24			
camemBERT	0.66	0.60	0.76	0.65	0.39
local before AL	0.74	0.73	0.75	0.80	0.64
global_local before AL	0.78	0.76	0.80	0.81	0.67
***global_local* after AL**	**0.83**	**0.79**	**0.83**	**0.86**	**0.67**

Table 2: Results for automatic multi-label classification

5.2. Model performances

Results are presented in Table 2, comparing *local* and *global_local* models against two baselines: a majority baseline based on the most frequent labels combination and CamemBERT in its original setup. Overall, the *global_local* configuration yields the best results across all evaluation metrics. For this reason, we selected the *global_local* configuration and used it to implement the active learning loop described in Section 4.2. Table 2 reports the model's performance before (*global_local* BEFORE AL) and after (*global_local* AFTER AL) the active learning step.

6. Application on 13_11 and legal_CODE

We applied our model to the 13_11 and legal_CODE corpora (described in Section 3.1) to automatically classify instances of *devoir*. We expected different distributions due to differences in the *medium* (spoken versus written) of these corpora and their genre (testimonies related to terrorist attacks in 13_11 versus legal rules in legal_CODE). Specifically, we expected a higher prevalence of epistemic *devoir*, which marks uncertainty, in 13_11, and a higher prevalence of deontic *devoir*, which marks obligation, in legal_CODE. In 13_11, we observe that half of instances of *devoir* express an epistemic meaning. It shows a higher frequency compared to our annotated corpus (see Section 4.2) where epistemic instances were 27% and to the legal_CODE corpus where the number of epistemic values is minimal (0.02%). However, frequency of the double label epistemic-deontic (5%) remains stable between the annotated corpus and 13_11. On the other hand, in the legal_CODE corpus, we observe that almost all instances of *devoir* (85%) convey a deontic meaning, followed by the alethic–deontic double label, which indicates

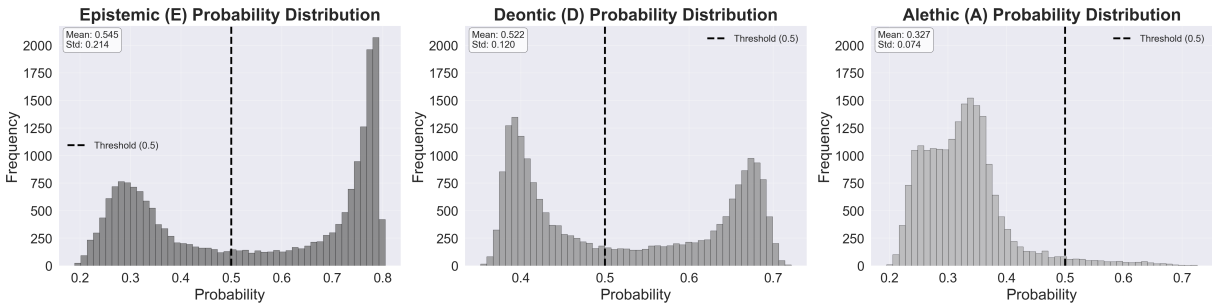

Figure 2: Probabilities distributions for epistemic (E), deontic (D), and alethic (A) labels in 13_11

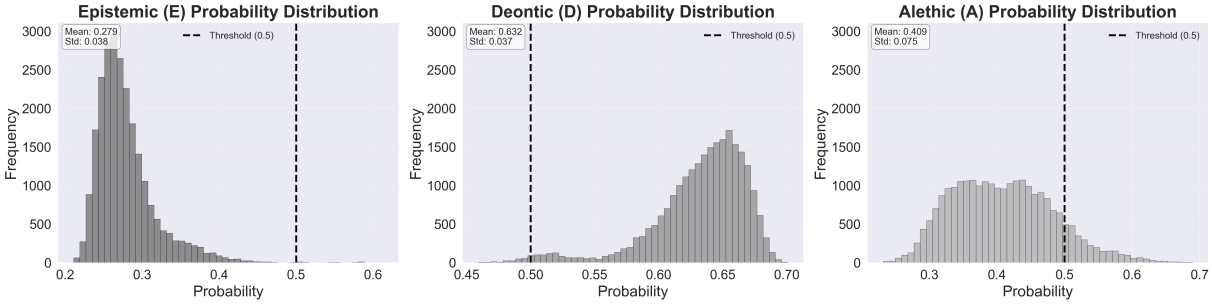

Figure 3: Probabilities distributions for epistemic (E), deontic (D), and alethic (A) labels in legal_CODE

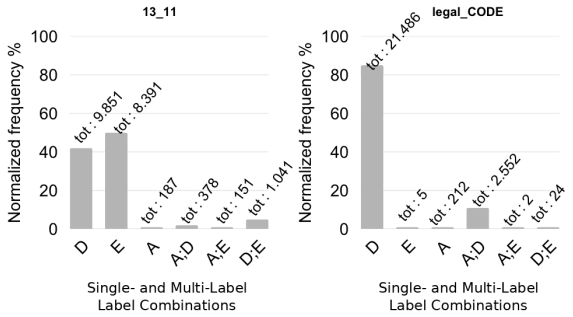

Figure 4: Distribution of (automatically) annotated instances in 13_11 (left) and legal_CODE (right).

ambiguity between the two senses. Comparing this distribution with that of our manually annotated corpus, we find that - as expected due to the legal nature of these texts - deontic *devoir* is much more frequent in legal_CODE, dominating the other categories. Ambiguity between deontic and alethic persists, with a higher frequency in the legal_CODE corpus (11%) compared to the manually annotated corpus (1,1%). This highlights the proximity between the two labels, less evident in the manual annotated corpus, and the difficulty of distinguishing between them.

We also examined, for each corpus, the extent to which the model's label assignments were decisive. Figures 2 and 3 show the distribution of predicted probabilities for the E, D, and A labels in 13_11 (figure 2) and legal_CODE (figure 3). In 13_11, the model appears highly decisive for epistemic classifications, often assigning either very high or very low probabilities, with relatively few instances receiving intermediate values. In contrast, for deontic classifications, the model is less decisive, as a larger number of instances fall within the middle probability range. In the legal_CODE corpus, the model is decisive for both deontic and epistemic classifications— assigning very high probabilities to deontic instances and very low probabilities to epistemic ones—but less decisive for the alethic label, where a greater proportion of instances receive intermediate probabilities.

7. Discussion

In this article, we fine-tuned a BERT-based model to represent the modal (multi-)senses of *devoir* instances. This methodology corresponds to a modal sense disambiguation or, when disambiguation is not possible, to an ambiguity representation. In first place, we performed manual annotation on a dataset of written and spoken French. After obtaining a stable IAA (0.8) we integrated an active learning loop based on uncertainty sampling in our annotation procedure to improve annotation speed and quality. Concerning the BERT-based model, we tested two embedding architectures for classification: a single token embedding of *devoir* (*local*) and a concatenation of the sentence-level [CLS] embedding with the *devoir* token embedding (*global_local*). The *global_local* architecture

yielded better performance reaching an F1-micro of 0.83 and F1-macro of 0.79. We did not attempt classification relying solely on the [CLS] embedding, because some sentences contain multiple *devoir* instances with different modal values (e.g., Example 6) that needed to be treated individually. Finally, we applied our model to 13_11 corpus and legal_CODE corpus. The results of the automatic (multi)-sense representation of *devoir* demonstrate how modal sense classification can capture corpus-specific characteristics by contributing to the definition of each corpus's modal profile. As introduced in (*anonymous*), a modal profile characterizes a corpus according to how modality is expressed, including which modal markers are used and the frequency of different modality categories. On the one hand, 13_11 is characterized by a high frequency of epistemic *devoir*, a marker of uncertainty; on the other hand, legal_CODE is dominated by deontic *devoir* and marked by a persistent ambiguity between deontic and alethic uses. Automatic (multi-)sense representation of all instances also allows us to perform some linguistic analysis in order to find, for example, correlations between modal senses and moods. In 13_11, 99% of epistemic *devoir* instances are indicative and only 0,2% are in a conditional mood. On the contrary, conditional mood covers 18% of deontic meaning. This result shows that, although the conditional mood is usually considered a marker of uncertainty, in this context it appears to be linked to deontic *devoir*, modulating obligation to express a suggestion (7) or what is considered appropriate by the speaker (8).

(7) Mais vous ne devriez pas y aller (13_11)
(en) *But you shouldn't go there.* (13_11)

(8) On devrait pas, on devrait pas mourir parce que quelqu'un a décidé qu'on devrait mourir. (13_11)
(en) *We shouldn't — we shouldn't die because someone decided that we must die.* (13_11)

On the contrary, in the legal_CODE corpus, deontic *devoir* appears predominantly in the indicative mood (94%), with only 1% of instances in the conditional mood. This is expected, as normative texts rarely express modulated obligations. These results suggest that the relationship between a specific modal value (deontic) and a particular mood (conditional) is corpus-specific.

8. Conclusion

This study represents a first step in the automatic classification of the French verb *devoir* in both spoken and written corporain order to compare how *medium* or genre are related or not to specific semantic values of this modal verb. Traditionally, the task has been framed as a multi-class classification problem, but we decided to reframe it as a multi-label classification problem. This allows us to account for instances in which *devoir* cannot be disambiguated due to insufficient context or intentional ambiguity. Furthermore, we show that a combination of global and local features—achieved by concatenating the *devoir* token embedding with the [CLS] sentence embedding—yields better performance than classification relying solely on the *devoir* embedding. In future work, we will explore additional representation strategies, particularly classification based on the [CLS] embedding by introducing special tokens around the targeted *devoir* instance. In addition, to better identify the appropriate context for *devoir* instances and to reduce the number of multiple occurrences within the same sentence—common in spoken corpora—we will experiment with segmenting the data into discourse units rather than relying on punctuation-based sentence boundaries, as proposed in (Prevot and Muller, 2025). However, we argue that (multi-)sense representations in spoken corpora may not benefit from a shift from sentence-level to discourse-level analysis, due to the rapid nature of conversation. The application of our model to two corpora, 13_11 and legal_CODE, which differ in *medium* (spoken vs. written) and genre (testimonies about a terrorist attack vs. legal code), demonstrates how correctly representing the modal senses of *devoir* instances can help enrich the modal profile of a corpus, thereby enabling comparisons with other corpora that differ in *medium* or genre. This model will be useful for studies in second language acquisition and legal discourse that focus on the use of modal verbs and and currently depends on manual annotation for their analysis. In addition, as in the NLP domain, research in these two fields generally frames the task as a multi-class classification problem, where each *devoir* instance is assigned to only one modal category. Our multi-label classification approach provides richer insights into how modal verbs are used in corpora. Moreover, in the context of learner corpora, our model can help identify and focus on double-labeled instances—that is, cases where ambiguity persists—in order to reformulate their expression for pedagogical purposes. Reframing the disambiguation task as a (multi-)sense representation problem may be more appropriate for cases traditionally treated as disambiguation, but where ambiguity can still persist. This perspective, which is closer to linguistic theory, could be relevant for general sense-labeling tasks.

9. Limitations

There are certain limitations to our work. First, for active learning, we only tested one sampling

method. In future work we aim to explore others methods and compare their performances with the current approach and random sampling to further validate the effectiveness of the active learning approach. Second, we aim to enhance data diversity and include additional written texts in order to balance the distribution of *devoir* instances between written and spoken corpora. Finally, we would like to experiment other methods to concatenate embeddings to improve our model's performances.

10. Bibliographical References

K. Baker, M. Bloodgood, B. J. Dorr, C. Callison-Burch, N. W. Filardo, C. Piatko, L. Levin, and S. Miller. 2012. Modality and negation in SIMT: Use of modality and negation in semantically-informed syntactic MT. *Computational Linguistics*, 38(2):411–438.

C. Barbet. 2012. Devoir et pouvoir, des marqueurs modaux ou évidentiels? *Langue française*, 173(1):49–63.

D. Battistelli, A. Etienne, R. Rahman, C. Teissèdre, and G. Lecorvé. 2022. Une chaîne de traitement pour prédire et appréhender la complexité des textes pour enfants d'un point de vue linguistique (a processing chain to explain the complexity of texts for children from a linguistic and psycho-linguistic point of view). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 236–246, Avignon, France. ATALA.

F. Bouhlal, M. Horst, and J. Martini. 2018. Modality in ESL textbooks: Insights from a contrastive corpus-based analysis. *The Canadian Modern Language Review*, 74(2):227–252.

A. Colli, D. Rossini, and D. Battistelli. 2024. A modal sense classifier for the French modal verb pouvoir. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 233–243, Pisa, Italy. CEUR Workshop Proceedings.

M. Dehouck and P. Denis. 2023. Revisiting modal sense classification with contextual word embeddings. In *Models of Modals: From Pragmatics and Corpus Linguistics to Machine Learning*, chapter 8, pages 225–253. De Gruyter Mouton, Berlin, Boston.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

L. Ein-Dor, A. Halfon, A.n Gera, E. Shnarch, L. Dankin, L. Choshen, M. Danilevsky, R. Aharonov, Y. Katz, and N. Slonim. 2020. Active learning for BERT: An empirical study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Laurent Gosselin. 2010. *Les modalités en français: La validation des représentations*. Rodopi.

P. F. Jacobs, G. Maillette de Buy Wenniger, M. Wiering, and L. Schomaker. 2021. Active learning for reducing labeling effort in text classification tasks. In *Benelux Conference on Artificial Intelligence*, pages 3–29. Springer International Publishing, Cham.

M. P. Jaskot and A. Wiltos. 2017. An approach to the translation of deontic modality in legal texts: The case of the polish and english versions of the "charter of fundamental rights of the european union". *Cognitive Studies| Études cognitives*, (17).

Hans Kronning. 1996. *Modalité, cognition et polysémie : sémantique du verbe modal 'devoir'*.

L. X. Li. 2024. Developing strategies to improve textbook design using synergy of native and learner corpora. *Journal of Psycholinguistic Research*, 53(6):76.

M. Light, X. Y. Qiu, and P. Srinivasan. 2004. The language of bioscience: Facts, speculations, and statements in between. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pages 17–24.

Y. Lu, B. Yao, S. Zhang, Y. Wang, Pe. Zhang, T. Lu, T. J.-J. Li, and D. Wang. 2023. Human still wins over LLM: An empirical study of active learning on domain-specific annotation tasks.

John Lyons. 1977. *Semantics: Volume 1*. Cambridge University Press.

A. Marasović, M. Zhou, A. Palmer, and A. Frank. 2016. Modal sense classification at large: Paraphrase-driven sense projection, semantically enriched classification models and cross-genre evaluations. *Linguistic Issues in Language Technology*, 14.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, E. V. de La Clergerie, B. Sagot, et al.

2019. Camembert: A tasty french language model.

Juliane Oliveira Martini. 2012. High frequency vocabulary in a secondary quebec esl textbook corpus. Unpublished thesis.

A. Miaschi and F. Dell'Orletta. 2020. Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 110–119, Online. Association for Computational Linguistics.

M. Nissim, P. Pietrandrea, A. Sansò, and C. Mauri. 2013. Cross-linguistic annotation of modality: A data-driven hierarchical model. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 7–14. Association for Computational Linguistics.

R. Owan, M. Gini, and D. Kang. 2022. Quirk or palmer: A comparative study of modal verb frameworks with annotated datasets. *arXiv preprint arXiv:2212.10152*.

L. Prevot and P. Muller. 2025. A few shades of supervision for discourse segmentation: Experiments on a french conversational corpus. *Dialogue & Discourse*, 16(2):35–73.

P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the Association for Computational Linguistics (ACL) System Demonstrations*. Association for Computational Linguistics.

R. Quirk, S. Greenbaum, G. Leech, and J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman, London.

J. Ruppenhofer and I. Rehbein. 2012. Yes we can!? annotating the senses of english modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1538–1545.

Burr Settles. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin-Madison.

C. Vetters and C. Barbet. 2006. Les emplois temporels des verbes modaux en français: Le cas de devoir. *Cahiers de Praxématique*, 47:187–210.

J. Wagner and S. Zarrieß. 2023. Probing BERT's ability to encode sentence modality and modal verb sense across varieties of English. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 28–38, Nancy, France. Association for Computational Linguistics.

A. Widlöcher and Y. Mathet. 2012. The glozz platform: a corpus annotation and mining tool. In *Proceedings of the 2012 ACM Symposium on Document Engineering*, DocEng '12, page 171–180, New York, NY, USA. Association for Computing Machinery.

S. Wu, M. Amini, and O. H. A. Mahfoodh. 2025. Unveiling certainty and doubt: A systemic functional exploration of epistemic modality in courtroom discourse. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*.

A. Wyner and W. Peters. 2011. On rule extraction from regulations. In *Legal Knowledge and Information Systems*, pages 113–122. IOS Press.

L. Zhang and L. Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 107–111. Association for Computing Machinery (ACM).

R. Zhang, Y.-S. Wang, Y. Yang, T. Vu, and L. Lei. 2022. Exploiting local and global features in transformer-based extreme multi-label text classification.

11. Language Resource References

CLESTHIA. 2024. Cfpp2000. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.

Nation, Paul. 2012. *The BNC/COCA word family lists v.2*. ILC-CNR for CLARIN-IT repository hosted at Institute for Computational Linguistics "A. Zampolli", National Research Council, in Pisa, ISLRN 532-206-426-067-2. PID https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/paul-nations-publications/publications/documents/Information-on-the-BNC_COCA-word-family-lists.pdf.

Université d'Orléans and CNRS - Laboratoire Ligérien de Linguistique (LLL). 2015. *Corpus ESLO: Enquête Sociolinguistique à Orléans*. LLL / CNRS / Université d'Orléans. ORTOLANG (Open Resources and TOols for LANGuage), ESLO resources, 2.0.

A Frame and Canvas-Based Perspective-Encoding Methodology for Multimodal Semantic Annotation of Classroom Settings

Cláudia Ferraz[1,2], Ely Matos[1], Frederico Belcavello[1], Júlia Gasparetto[1], Juliana de Oliveira[1], Janina Wildfeuer[2], Tiago Timponi Torrent[1,3]

[1] Federal University of Juiz de Fora | FrameNet Brasil, [2] University of Groningen,
[3] Brazilian National Council for Scientific and Technological Development – CNPq
{claudia.ferraz, julia.gasparetto, juliana.oliveira}@estudante.ufjf.br, {ely.matos, fred.belcavello, tiago.torrent}@.ufjf.br, j.wildfeuer@rug.nl

Abstract

We propose a methodology for the multimodal semantic annotation of classroom interactions that takes interactional canvases and semantic frames as its core analytical categories. The approach enables the systematic recording of semantic correlations among interactants, communicative modes, and material supports involved in situated meaning-making processes. The methodology encodes participant perspective by relying on the temporal alignment of multiple video recordings of the same instructional event captured from different viewpoints, allowing for the representation of how meaning construction unfolds across perceptual and interactional positions. To operationalize the proposal, we introduce an annotation tool that implements the scheme and supports the integration of multimodal data streams within a unified semantic representation framework. We conclude by discussing the limitations of the current proposal and the possibilities for extending it to other interactional settings.

Keywords: Multimodal Semantic Annotation, FrameNet, Canvas, Perspective

1. Introduction

In recent years, there has been a growing demand for more comprehensive approaches to multimodal data annotation, capable of offering systematic, consistent, and interconnectable annotation schemes that capture the rich diversity of communicative situations and all media and modes involved (Bateman et al., 2017a; Pflaeging et al., 2021). Given the increasing complexity of multimodal data analyzed in different domains, this demand is driven by academic and social motivations.

As a particular challenge, the immense diversity and complexity of multimodal communicative situations makes it necessary to develop tools and schemes that are tailored to the different types of data at question: page-based data, such as documents, diagrams, or comics, are analyzed with regard to their spatial arrangement, while linear data, such as films or face-to-face interactions, request a temporal and dynamic annotation. At the same time, there is a shortage of annotation methodologies that explicitly consider the real contexts of data production and interpretation, particularly those related to education. In such environments, where multiple participants engage in joint meaning-making processes, the perspective of each participant towards the communicative event is key.

Existing FrameNet-based annotation schemes allow for a myriad of multimodal analyses (Belcavello et al., 2024; Viridiano et al., 2024; Abreu and Matos, 2025; Sigiliano, 2025). They are based on an adaptation of the FN-Br WebTool (Torrent et al., 2024),

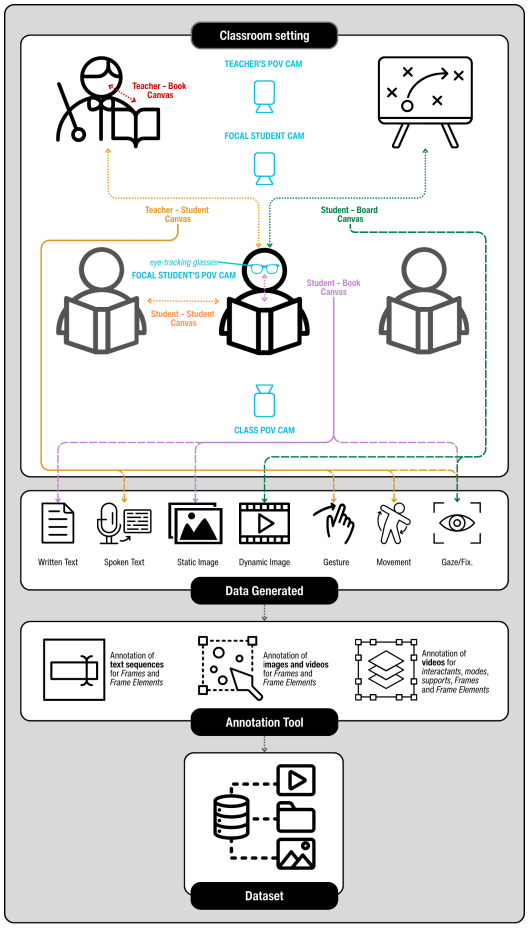

Figure 1: Annotation scheme and resulting dataset

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 22–32
12 May 2026.

which enables fine-grained semantic annotation of multimodal corpora, particularly combinations of text and image (Belcavello et al., 2022). The tool allows annotators to choose frames and locate Frame Elements (FEs) in both text and images. When annotating videos, it allows tracking the time interval in which these elements are active in the image and in the text transcript of spoken audio (Belcavello et al., 2020).

Such annotation schemes were aimed at analyzing multimodal corpora built from mostly two-dimensional media, but this foundation allows for the development of new annotation frameworks that focus on even more complex communicative situations, such as those in educational settings. In-classroom language teaching, for example, involves multiple modes of expression, such as speech, gesture, image, and writing, while at the same time using material (books, blackboard) and digital (tablets or computers) supports, which again feature a diversity of semiotic modes. This multidimensional environment combining both static and dynamic materialities that make space for direct, indirect, or subordinate modes (Bateman et al., 2017a) has brought three new challenges:

1. How to annotate the simultaneous and dynamically unfolding use of multiple semiotic modes (such as speech, gesture, image, etc.) and their interrelationships between interactants and objects in order to capture and represent information from frame evocation processes?

2. How to delimit the analysis considering the co-occurrence and interdependence of these modes, enabling semantic analysis of granularity resulting from the combination of different modalities?

3. How to encode each participant's perspective towards the multimodal meaning-making processes?

So far, only a few empirical studies have systematically approached how different interactants (teachers, students, artifacts) relate and coordinate in authentic teaching-learning situations (Thomas, 2018; Fillietaz et al., 2022, e.g.). Furthermore, although there are relevant methodological proposals in the field of multimodality research, there is still no widely standardized approach capable of dealing with the overall complexity of these interactions in real educational contexts. Given this scenario, this paper presents a proposal for a multimodal annotation scheme for real classroom interactions (Figure 1), based on the multimodal approach proposed by Bateman et al. (2017a) in combination with frame semantics (Fillmore, 1982), as systematized within the scope of FrameNet Brasil (Torrent et al., 2022). With a particular focus on the materiality and canvases of the communicative situation and by implementing the conceptual structure of Frame Semantics used in the FN-Br WebTool (Torrent et al., 2024), the proposal aims to capture multimodal interactions in an authentic educational environment in an integrated manner, promoting reproducibility, analytical consistency, and interoperability between annotation schemes. Moreover, by systematically annotating the same communicative events from different points of view, the scheme incorporates data perspectivism in its strong variety (Cabitza et al., 2023), to the extent that multiple annotation labels are assigned to one same piece of data depending on whose perspective — the teacher's, the student's, the group's — is being encoded.

2. Frame Semantics and FrameNet

The proposed annotation methodology adopts Frame Semantics as its main theoretical foundation, which includes context as a central aspect of the theory. According to this approach, frames are defined as the background scenes against which the meaning of a given event, situation, or linguistic utterance must be constructed. Such scenes include the participants and props that make up the situation, each of which is conceived as a Frame Element (FE). In this sense, FEs are related in such a way that the presence of one triggers the participation of the others (Fillmore, 1982, p.111).

Frames can be instantiated in two ways: by evoking Lexical Units (LUs) that trigger certain frames or by invocation performed by comprehenders based on the combination of contextual clues present in the text. Thus, the comprehension of a statement depends not only on linguistic structure but also on knowledge of the world, the immediate context, and the communicative conditions in which the message occurs (Torrent et al., 2022).

FrameNet was established as a computational resource for semantic annotation based on English language corpora (Baker et al., 1998), providing a formalized inventory of frames, FEs, and annotated LUs. Thus, the corpus analyzes performed in FrameNet follow the fundamental principle that *"meaning is relativized to scenes"* (Fillmore, 1977). In FrameNet, each frame is composed of a name, an in-prose definition, and a set of FEs and their definitions. Because it is a FrameNET, not a frame list, frames are linked to one another via frame-to-frame relations. Consider, as an example, the `Manipulate_into_doing` frame in (1).

(1) **`Manipulate_into_doing`**
Definition: A Manipulator gets a Patient to perform a Resulting_action.

Core Frame Elements:
Manipulator: The person who gets the Patient to act.
Patient: The person who the Manipulator gets to do something they would not have done had without some request or pressure.
Resulting_action: What the Manipulator intends to make the Patient do and the Patient actually does
Frame-to-Frame Relations:
inherits from: `Intentionally_affect`
is inherited by: `Talking_into`
uses: `Influencing_potential`
Lexical Units: *lure.v; manipulate.v...*

The Inheritance relation is a Parent-Child hierarchy in which the daughter frame represents an equally or more specific version of the scene portrayed by the mother frame, inheriting all its roles and constraints. Every FE in the mother frame must correspond to an FE in the daughter frame, although these elements may be assigned different names or more detailed definitions in the daughter. In this example, the `Manipulate_into_doing` frame inherits from the more general `Intentionally_affect` frame, and it is itself inherited by the even more specialized `Talking_into` frame. The MANIPULATOR FE in the `Manipulate_into_doing` frame corresponds to the AGENT in the `Intentionally_affect` frame and to the SPEAKER in the `Talking_into` frame. Similar relations hold for the other two FEs.

The Using relation, in turn, is a frame-to-frame connection indicating that a daughter frame presupposes the background knowledge of a mother frame as a necessary context for its own meaning. Unlike the Inheritance relation, where the child is a more specific version of the parent, the Using relation implies that only part of the scene in the daughter frame refers to the conceptual structure of the mother frame. In (1), the `Manipulate_into_doing` frame uses the `Influencing_potential` frame because the specific act of manipulation presupposes the broader concept of influence. Other types of relations in FrameNet include Perspective_on, Subframe, Preceedes, Inchoative_of and Causative_of (Ruppenhofer et al., 2016).

The FrameNet model has been implemented elsewhere in the world for several other languages. Among them, FrameNet Brasil was established for Brazilian Portuguese, considering the specific lexical, grammatical, and semantic properties of the language, ensuring interlinguistic comparability and theoretical consistency (Torrent and Ellsworth, 2013). More recently, the scope of semantic annotation within FrameNet Brasil has expanded beyond exclusively verbal data, giving rise to multimodal datasets that integrate linguistic and visual modes (Belcavello et al., 2024; Viridiano et al., 2024; Gamonal et al., 2025). To operationalize this integration, multimodal annotation tools have been developed that allow the alignment of frame-based semantic annotations with temporal and spatial representations of dynamic multimodal data, enabling the analysis of how frames are evoked through speech, gestures, gaze, visual objects, and other visual resources. This establishes conditions for the development of empirical research covering the study of event frames (Pinto, 2025), pragmatic frames (Abreu and Matos, 2025), and deictic center frames (Sigiliano, 2025) in audiovisual corpora. Such empirical studies have shown that, similar to the way words in a sentence evoke frames and organize their elements in syntactic locality, other elements in various communicative modes can also do so or work complementarily with frame evocation patterns across modalities (Belcavello et al., 2020).

3. Empirical Multimodality Research

In multimodality studies, there is currently a focus on empirical data analysis and the comprehensive examination of larger data sets (Pflaeging et al., 2021; Bateman et al., 2026). This trend builds on a productive context of theoretical, methodological, and analytical developments over three decades with an initial interest in the combination of written language with images and a rapid extension to all expressive forms and the question of how different types of meaning-making, previously studied separately in diverse disciplines, combine into an integrated, multimodal whole.

In order to capture the complex nature of multimodal communicative situations, many current works aim for a comprehensive analysis of larger corpora and therefore develop multi-level standoff annotation schemes that combine descriptions of the actual material with interpretations of the communicative structure of the analytical object(s). This method has been used for several kinds of artifacts, ranging from multi-dimensional page-based documents (Bateman, 2008), packaging (Thomas, 2009) and tourist brochures (Hiippala, 2015) to comics (Bateman et al., 2017b), TV series (Drummond and Wildfeuer, 2020), commercials (Wildfeuer and Coffie, 2022), news videos (Bateman and Tseng, 2023), and short form video content (Grzenkowicz and Wildfeuer, 2025).

As a starting point for the development of such annotation schemes, the concept of the material canvas (Bateman et al., 2017a) provides a way to capture the initial perceptual unit of analysis. A canvas functions as the material basis carrying and

realizing semiotic modes. Each canvas is structured and analyzed according to a classification scheme involving the following dimensions (Bateman et al., 2017a, chap. 3): (i) **temporality**, which defines whether the canvas is static or dynamic; (ii) **spatial dimensionality**, which determines whether the canvas is two-dimensional (a screen or paper) or three-dimensional (e.g. architectural spaces or face-to-face interactions); (iii) **transience**, concerning the permanence of information and distinguishing transient canvases, such as the dynamic body, hands, and face of a person speaking in a conversation, from permanent ones, such as a printed book page; (iv) **roles of participant/observer**, which considers how the message recipients interact with the canvas or act simply as observers; and (v) **ergodicity**[1], which addresses the level of effort or contribution of the reader/user in constructing the meanings of the text.

In relation to the analytical objects in question in this paper, Bateman et al. (2017a, chap. 7) discuss the communicative situation of classroom interaction as one of the most complex multimodal situations that can be "sliced" into several canvases and subcanvases. For example, a teacher may, at a given moment, explain content from a textbook while simultaneously pointing to the book and looking at a particular student. These are different activities that occur simultaneously on different canvases with different materialities that make space for different semiotic modes. The teacher's spoken language, for example, operates on a dynamic, three-dimensional, and transient canvas of interaction with the students where both the teacher and the students are participants that can interact and interrupt each other. In contrast, the book the teacher uses as a medium is a static, two-dimensional, permanent canvas for which teacher and students are observers and which features several other semiotic modes such as images, written language, page layout, etc.

In the annotation methodology proposed in this paper, we integrate the canvas analysis with FrameNet annotation in order to provide a detailed understanding of how meaning is constructed in classroom interactions. The main canvas, the overall classroom situation, includes several subcanvases that provide multiple dynamic interactional spaces — making visible the characteristics of the multimodal environment that shape social interaction. On the teacher-student interaction canvas, for example, speech and gesture operate as semiotic modes in a face-to-face interaction, which is mutable and transient. For a teacher, interacting with a whiteboard, in contrast, it is relevant to identify modes supported by the dynamic canvas of the whiteboard (e.g. moving images in a video or written language from a text displayed on the whiteboard screen). Frames establish semantic units that allow meanings to be mapped within this particular context of interaction. In order to analyze these frames as nodes in a network of relations, it is necessary to determine the different interactants on the canvas as well as the modes and processes that enable meanings to be dynamically constructed from the different connections between frames and FEs. The annotation scheme devised to support such an analysis is presented next.

[1] Ergodic canvases are those that require a non-trivial effort to be traversed. The reader of a complex comic page, for example, does not simply follow a predetermined linear sequence but actively participates in choosing the path for generating the content. This also happens in video games, for example, where the game also reacts to how players interact with the interface. Bateman et al. (2017a) therefore distinguish between different layers of ergodicity. Non-ergodic canvases are those in which the reading path is fixed, and the reader has a more passive role in determining the structure of the text, as in a traditional novel, for example.

4. A Multilevel Annotation Scheme for Classroom Interactions

The theoretical basis for the construction of the annotation scheme proposed in this paper is based on Systemic-Functional Linguistics (SFL), which conceives language as a resource for meaning-making, in which "*meaning resides in systemic patterns of choice*" (Halliday and Matthiessen, 2004, p.23). In this context, system networks present the structured options available to the language user, allowing for the selection of combinations to convey specific meanings. The application of this approach to multimodal artifacts was already proposed in early multimodality research and has served as the foundation for the development of several multilevel annotation schemes for the analysis of larger corpora (see Section 3).

Building on the concept of canvases introduced in Section 3 and on the advances of multimodal FrameNet analysis summarized in Section 2, the annotation methodology proposed here focuses on classroom interactions as the primary object of analysis. These interactions unfold across multiple layers of semiotic activity, including spoken language, gestures, visual resources, technological artifacts, and so on, each of which can take part in a subcanvas within the broader classroom interaction canvas. In addition, each interaction may be regarded from at least two different perspectives — see Figure 1.

Annotating and analyzing these perspectivized interactions presents several challenges. First, classroom communication is typically asymmetrical, with the teacher leading the progression of activities. However, multiple agents (the teacher

and the students) are simultaneously present in and, therefore, imposing their perspectives on the interaction, and varied teaching strategies create dynamic layouts that both reorganize and facilitate interaction in different ways. This allows for the creation of multiple layers of annotation. In addition, the main classroom canvas is mutable and fully ergodic, evolving over time as participants engage in different communicative activities (Bateman et al., 2017a). This significantly shapes the ways in which participants "*establish and maintain embodied copresence and mutual orientation to each other, relative to the unfolding activity and the multimodal and material environment*" (Bateman et al., 2026, p.404). By developing an annotation scheme that directly aligns with this hierarchy of canvases, our goal is to systematically and structurally capture the multimodal complexity of in-classroom interactions.

Addressing such complexities required the development of a multilevel annotation scheme capable of representing semiotic modes at different levels of granularity. To test the feasibility of implementing the proposed annotation scheme, we built a dataset of in-classroom interactions, developed a tagset, and implemented an annotation task by adapting the FN-Br WebTool (Torrent et al., 2024). We present each of these stages next.

4.1. Dataset

To test the feasibility of the implementation of the proposed methodology, we built a dataset comprising video recordings and eye-tracking data[2] from 10 Portuguese language lessons in elementary school, each lasting approximately 50 minutes.

In line with a strong perspectivist approach to dataset construction (Cabitza et al., 2023), five cameras were strategically positioned to capture the perspectives of the several participants involved in the classroom canvas: (i) one in the back of the room facing the front, capturing the point of view of ALL_STUDENTS; (ii) one in the front facing ALL_STUDENTS and capturing the point of view of the TEACHER; (iii) one mounted on a pair of eye-tracking glasses, which is worn by a randomly selected student in each class (the FOCAL_STUDENT); (iv) one focused on the FOCAL_STUDENT or on a GROUP_OF_STUDENTS, depending on the type of pedagogical activity being recorded and (v) one security camera installed on the ceiling, which was used only as a reference. The lessons were designed to provide a variety of teaching strategies, generating several forms of interaction — subcanvases — and allowing analyzes in terms of the interactants, modes and supports involved from different points of view.

The video files from all cameras were aligned for time intervals and the audio was automatically transcribed using an AI tool.[3] Audio transcriptions were then manually checked by two annotators with access to the original video files. Once validated, the transcriptions were uploaded to the FN-Br WebTool text annotation module (Torrent et al., 2024) to be annotated for frames and FEs, following FrameNet's fulltext annotation methodology (Ruppenhofer et al., 2016). The instructional materials used during the classes were scanned and uploaded to the same tool for multimodal annotation according to the methodology proposed in (Torrent et al., 2022). The annotation of the other communicative modes involved in classroom interactions, in turn, required both the development of a new tagset and adaptations to the WebTool, which are the core of the methodology presented here.

4.2. Units of analysis and the resulting annotation layers

The perspectivized nature of the dataset allows for the annotations to focus on the interactions occurring in the classroom from the viewpoint of different participants. The main canvas therefore corresponds to the complete physical situation in which the interaction takes place, constituted not only by the setting, objects, and environment, but also by verbal utterances, facial expressions, bodily postures, and distances maintained during interactions (Bateman et al., 2017a, p. 96).

The definition of the analytical focus articulates higher-level actions and levels of materiality through the concept of canvases. This directs the analysis toward the affordances available for meaning-making in the observed situation and allows for a clear delimitation of what is included or excluded from the analysis. In line with Bateman et al. (2017a) and Grzenkowicz and Wildfeuer (2025), the formulation of the annotation scheme begins with the identification of specific analytical units within these canvases. The scheme is then "*applied to such units, producing a set of classificatory features and, where necessary, additional segmentations, in order to ensure a satisfactory description of the investigated phenomenon*" (Bateman et al., 2017b, p. 13).

In this sense, the most general canvases distinguished in our annotation scheme are those of the ongoing **Interactions** in the classroom, which appear to be the most universal units of analysis within

[2] In the context of FrameNet multimodal annotation, eye-tracking data are used for the psycholinguistic validation of the image annotations. Research by Belcavello (2023) demonstrates that bounding boxes annotated for frames and FEs in a dataset composed of episodes of a TV Travel Series present higher fixation durations than other regions of the image less semantically-relevant.

[3] https://www.notta.ai

the communicative situation under study and are equally applicable to other face-to-face interaction contexts. The notion of **Interaction** also facilitates articulation with Frame Semantics, where it can be evoked through event frames, allowing a direct mapping to eventualities (Grzenkowicz and Wildfeuer, 2025) in the multimodal context.

At the level of **Interactions**, the subcanvases (layers) that proved most relevant for the analysis are the following:

Interactants The participants involved in the interactional situation, including vocally active participants and co-presence in the communicative situation, focusing specifically on interaction through speech, gaze, gestures, and other relevant semiotic modes supported by the canvas and its subcanvases. For this layer, we define the following labels:

- TEACHER: the participant fulfilling the role of mediator in the pedagogical situation, often guiding the interaction and regulating participation.
- FOCAL_STUDENT: the student specifically selected as the central observation point for analysis. This category is useful for examining, in detail, co-presence, mutual orientation, multimodal behaviors, and appropriation of semiotic modes by the focal student. Given the eye-tracking data, it also allows tracking of their actions, verbal and non-verbal responses, and interaction with other participants and objects in the classroom.
- SECONDARY_STUDENT: another student who is not the main focus of analysis but whose presence and participation influence or are influenced by the interactional dynamic.
- GROUP_OF_STUDENTS: a set of students (including the FOCAL_STUDENT) who interact around a specific activity, allowing analysis of group-level interaction, coordination of actions, role distribution, and shared use of multimodal and material resources (e.g., notebooks, laptops, educational games) in joint meaning-making.
- ALL_STUDENTS: the entire class.
- NONE: indicates the absence of an active participant in the analytical situation or moments when no meaningful action can be attributed to specific participants.

Mode The ways in which different semiotic modes are activated by interactants as socially and culturally organized systems of resources for meaning-making, endowed with specific materiality and their own expressive potentials. Within the scope of multimodal analysis, each mode constitutes a relatively stable set of conventions, forms, and possibilities of articulation that allow participants in interaction to perform communicative actions and construct meaning. In this sense, within our multilevel scheme, this category makes it possible to identify, segment, and describe which semiotic modes are in use at a given moment of the interaction and how they are articulated — whether mediated or not by an object — in the realization of pedagogical communicative events. Its attributes correspond to the main embodied and verbal resources identifiable in the interaction. This layer can be annotated for the following labels and sublabels:

- GAZE_DIRECTION: refers to the orientation of the interactant's gaze during the interaction. The direction of gaze functions as a semiotic marker of focus, engagement, and distribution of authority, allowing identification of how the participant positions themselves and others within the interactional space. The sublabels under this attribute are as follows: (i) LOOKING AT FOCAL_STUDENT; (ii) LOOKING AT SECONDARY_STUDENT; (iii) LOOKING AT GROUP_OF_STUDENTS; (iv) LOOKING AT ALL_STUDENTS; (v) LOOKING AT SUPPORT; (vi) LOOKING AWAY; (vii) NONE.
- HEAD_MOVEMENT: refers to the movements of the interactant's head during interaction, acting as a semiotic bodily mode that expresses attitudes and regulates interactional participation. The sublabels under this attribute are: (i) AGREEMENT; (ii) DISAGREEMENT; (iii) ATTENTION/ENGAGEMENT; (iv) DISBELIEF (v) SELF-REGULATION; (vi) NONE.
- BODY_MOVEMENT: refers to observable bodily movements of the interactant that assumes semiotic relevance in the interaction. This attribute encompasses broader displacements and postural changes involving the body as a whole or larger body segments such as the trunk (in positional changes), spatial displacement (e.g., walking around the classroom), and global movements. The values attributed are: (i) LEANING FORWARD; (ii) TURNING TO THE RIGHT; (iii) TURNING TO THE LEFT; (iv) MOVING IN SPACE; (v) FACING PARTICIPANT; (vi) NONE.
- GESTURES: specific observable movements performed with hands and arms that carry semiotic relevance in interaction, contributing to meaning construction in communication. The values of this attribute are: (i) DEICTIC; (ii) ICONIC; (iii) METAPHORIC; (iv) CONVENTIONAL; (v) SELF-REGULATORY; (vi) NONE.

- SPOKEN_LANGUAGE: everything that is said during the interactions. This layer is linked to the sentences transcribed from the audio recordings and is annotated for frames and FEs in the existing fulltext annotation module of the WebTool.

- WRITTEN_LANGUAGE: everything that is written in any of the support materials used during the class. Sentences in this layer are also annotated for frames and FEs.

- STATIC_IMAGE: pictures, drawings and other images in the support materials. This layer is annotated in the existing multimodal annotation module for frames and FEs.

- DYNAMIC_IMAGE: videos used during the classes. This layer is also annotated for frames and FEs.

Support Support refers to materials or technological resources visibly present and potentially mobilized in the interaction, functioning as support for participants. The values attributed are: PRINTED PEDAGOGICAL MATERIAL (teacher's book, student book, handout, poster, notebook); DIDACTIC GAME MATERIAL (board game, game cards, game manual, handheld choice boards); TECHNOLOGICAL SUPPORT (laptop, projector, projection screen); BOARD; or NONE.

Viewpoint This layer refers to the perspective from which the interaction is filmed, indicating the point of view represented in the framing (e.g. the focal student's perspective, teacher's perspective, whole-class perspective, or overhead view).

4.3. Annotation interface

In this section, we present how the proposed annotation scheme is used in the FN-Br WebTool (Torrent et al., 2024). As indicated in Section 4.2, different modules of the WebTool are used to annotate different communicative modes. The transcribed SPOKEN_LANGUAGE and WRITTEN_LANGUAGE modes are annotated for frames and FEs using the fulltext annotation module. For instance, a sentence such as (2), which is uttered by the teacher in the moment of one of the classes shown in Figure 3, can be annotated for the `Manipulate_into_doing` frame in (1).

(2) [Eu$_{\text{MANIPULATOR}}$] gostaria$^{\texttt{Manipulate_into_doing}}$ [que [vocês$_{\text{PATIENT}}$] abrissem o livro na página 16$_{\text{RESULTING_ACTION}}$].
[I$_{\text{MANIPULATOR}}$] would like$^{\texttt{Manipulate_into_doing}}$ [that [you$_{\text{PATIENT}}$] open the book on page 16$_{\text{RESULTING_ACTION}}$].

Visual aids, in turn, are annotated by means of bounding boxes which are labeled for frames, FEs and an additional label indicating the nature of the object inside the bounding box. Figure 2 shows one example of this type of annotation.

For the other layers described in Section 4.2 to be annotated, a new module had to be created in the WebTool. This module — Figure 3 — is organized in three panels:

- The top left panel has a video player from which annotators can watch the recordings from each of the cameras.

- The top right panel reproduces the video timeline and has four types annotation layers — interactant, modes, support and viewpoint. This panel records that, given the **viewpoint** of one of the participants — in the case of the example, the FOCAL_STUDENT —, one or more **interactants** — e.g. the TEACHER, the FOCAL_STUDENT and others — mobilize communicative **modes** — Body_movement, FACIAL_EXPRESSION, GAZE_DIRECTION, GESTURE, HEAD_MOVEMENT and SPOKEN_LANGUAGE — and a **support** — the TEXT_BOOK — in favor of meaning construction.

- The bottom left panel opens every time a new label is added to the panel on the right. In this panel, annotators can specify sublabels, associate frames and FEs to them and also indicate which visual element in the video triggers the annotation. In the example, the teacher's Deictic gesture of pointing to the book while speaking the sentence in (2) is annotated for the `Manipulate_into_doing` frame, with the TEACHER label for the **Interactant** layer being labeled as the MANIPULATOR FE. In addition, the correlation between the pointing gesture and the **support** is also noted, and the fact that the element of the image triggering the gesture annotation is the teacher's finger is noted in the CV Name field.[4]

The interface organized in multilevel layers allows for the simultaneous capture of different semiotic modes mobilized by the participants. Also, layers can be connected to one another, so that the mode layer can be associated with both the interactant and the support layers. Each interactive event is also associated with a frame and its corresponding FEs, which are evoked by these events. Additionally, a CV name is assigned to the image delimited by a bounding box, incorporated into a

[4]The CV Name field identifies the entity in the image as it would be labeled by a standard computer vision (CV) algorithm.

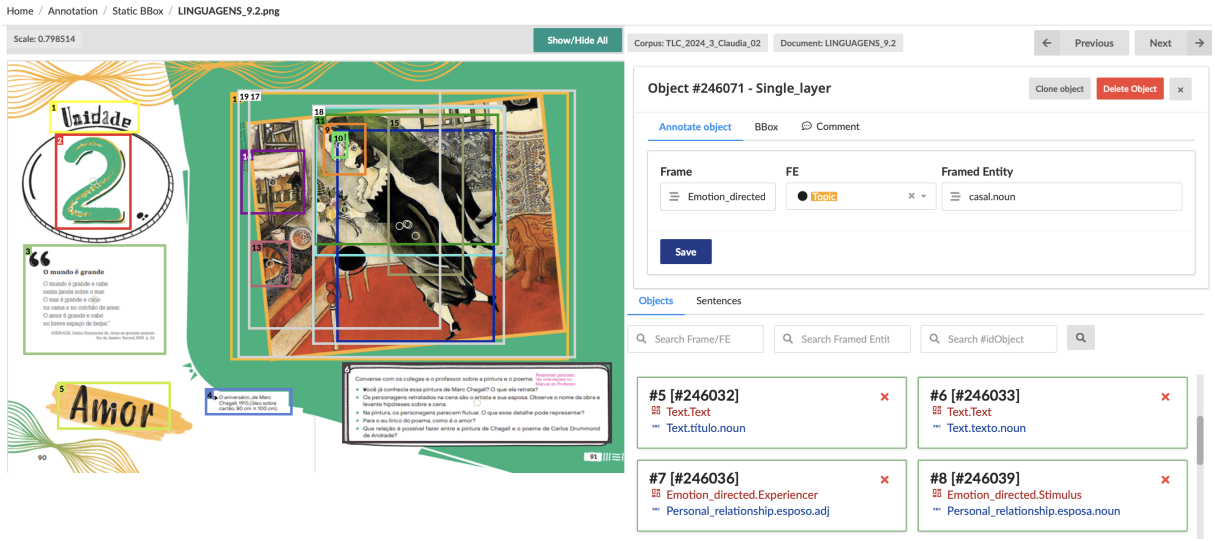

Figure 2: The visual annotation module

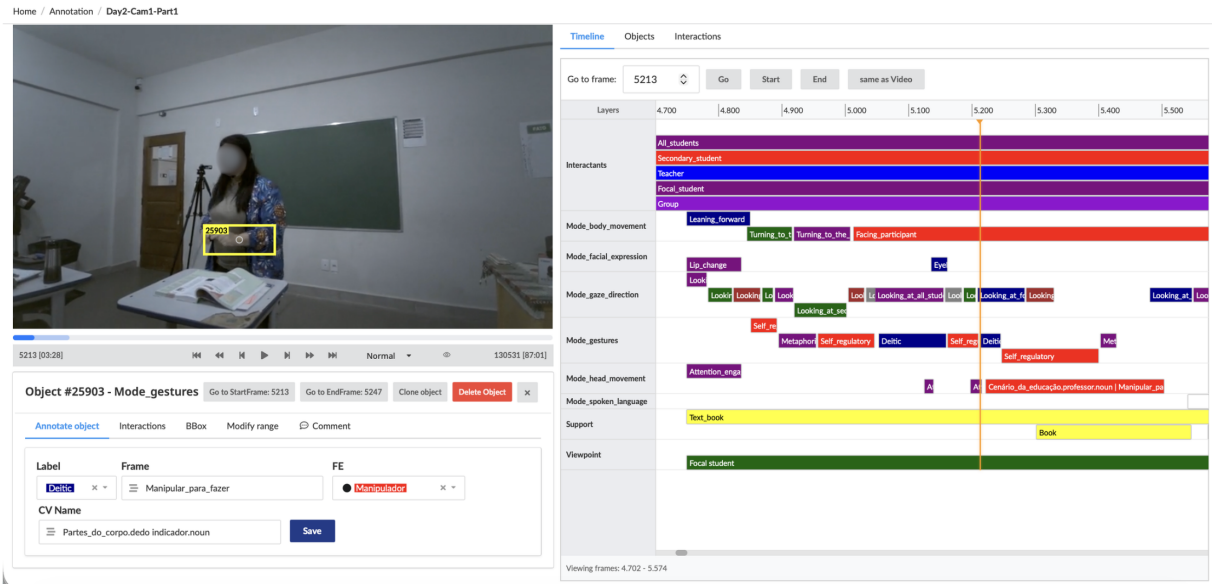

Figure 3: The canvas annotation module

frame-evoking LU. This bounding box spans the timeline until the event — for example, a gesture — ends.

Note also that the fact that frame, FE and CV Name labels are nested within the labels in the multiple layers allows the annotator to capture the complexities of multimodal communication in an organized fashion. In the example, the annotation scheme implemented in the WebTool records:

- In the BODY_MOVEMENT layer, that the TEACHER, annotated as the Manipulator in the `Manipulate_into_doing` frame is in the FACING PARTICIPANT position;
- In the GESTURE layer, that she is using her finger in a DEICTIC gesture to point to the TEXT_BOOK while using the SPOKEN_LANGUAGE mode in the form of the sentence in (2);
- In the GAZE_DIRECTION layer, that, some seconds later, the TEACHER will reinforce the instruction given in (2) by LOOKING_AT_FOCAL_STUDENT first, then by LOOKING_AT_SECONDARY_STUDENT and finally by LOOKING_AT_SUPPORT, that is, at the TEXT_BOOK.

Hence, the annotation methodology presented integrates two multimodal analysis methodologies, allowing for both the articulation of multilayers and the annotation of frames. As Fillmore (1985) argues, frames structure the understanding of events,

defining the roles and relationships between participants and objects. In our methodology, each interactive event evokes frames that organize the perception of meaning and articulate the mobilized semiotic modes, ensuring that the analysis captures not only the semiotic modes in isolation but also their potential for meaning-making within the communicative situation.

5. Conclusion

In this paper, we present an annotation scheme and a multimodal dataset of classroom interactions. Our work advanced in the construction of an annotation tool with multilevel layers, integration between the mode, interactant, and support layers, and association of events with frames and their elements, enabling the detailed capture of movements, gestures, speech, and support resources from a temporal and interactional perspective. These records will allow not only the qualitative analysis of interactions, but also an analysis of the frames evoked during the lessons. The annotation scheme also incorporates perspective, since one same interactional event can be annotated from different points of view.

Although the methodology has been devised for classroom settings, it can be extended to other types of interaction scenarios. As next steps, we plan to conduct inter-annotator reliability tests, evaluate the robustness of the scheme, and refine the tool to allow for more precise and replicable annotation. In the long term, our goal is to consolidate the methodology and datasets deriving from it as a gold-standard resource, which can be used in research on multimodality, classroom learning, and Machine Learning applications, such as the automatic analysis of semantic roles.

6. Ethical considerations and limitations

Research presented in this paper was approved by the Ethics in Research Committee of the Federal University of Juiz de Fora, process number 87876425.9.0000.5147. The research protocol includes strategies for participant anonymization. All students and the teacher taking part in the recorded classes, as well as the students' parents, signed a term of consent for volunteer participation in the experiment.

All annotation used in the experiments, including the revising of the audio transcription, was carried out by trained annotators who were paid a monthly stipend, which is, at least, equivalent to the minimum wage according to local regulations. All annotators involved in the annotation of the corpus used in the evaluation experiment reported here are co-authors of this paper.

Among the limitations of the methodology described, it is worth noting that the annotation categories were defined based on the configuration used for video-recording the classes. Different camera configurations may require different annotation categories.

7. Acknowledgements

Research reported in this paper was developed under the ReINVenTA—Research and Innovation Network for Vision and Text Analysis of Multimodal Objects—initiative, 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). The resulting dataset will be part of the data collection of the National Science and Technology Institute for Responsible Artificial Intelligence, Computational Linguistics and Information Treatment and Dissemination (INCT-TILDIAR, CNPq grant 408490/2024-1). Ferraz was supported by CAPES/PDSE (grant 88881.127655/2025-01). Torrent is a CNPq research productivity grantee (grant 311241/2025-5).

8. Bibliographical 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 *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

John Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2017a. *Multimodality: Foundations, research and analysis–A problem-oriented introduction*. Walter de Gruyter GmbH & Co KG, Berlin.

John A. Bateman. 2008. *The GeM Model: Treating the Multimodal Page as a Multilayered Semiotic Artefact*, pages 107–142. Palgrave Macmillan UK, London.

John A. Bateman and Chiao-I Tseng. 2023. Multimodal discourse analysis as a method for revealing narrative strategies in news videos. *Multimodal Communication*, 12(3):261–285.

John A Bateman, Francisco OD Veloso, Janina Wildfeuer, Felix HiuLaam Cheung, and Nancy Songdan Guo. 2017b. An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities*, 32(3):476–510.

John A. Bateman, Janina Wildfeuer, and Tuomo Hiippala. 2026. *Multimodality. A Hands-On Guide*. de Gruyter.

Frederico Belcavello. 2023. *FrameNet Annotation for Multimodal Corpora: devising a methodology for the semantic representation of text-image interactions in audiovisual productions*. Ph.D. Thesis in Linguistics, Universidade Federal de Juiz de Fora, Juiz de Fora.

Frederico Belcavello, Tiago Timponi Torrent, Ely E. Matos, Adriana S. Pagano, Maucha Gamonal, Natalia Sigiliano, Lívia 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.

Frederico Belcavello, Marcelo Viridiano, Alexandre Diniz da Costa, Ely Edison da Silva Matos, and Tiago Timponi Torrent. 2020. Frame-based annotation of multimodal corpora: Tracking (a)synchronies in meaning construction. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 23–30, Marseille, France. European Language Resources Association.

Frederico Belcavello, Marcelo Viridiano, Ely Matos, and Tiago Timponi Torrent. 2022. Charon: A FrameNet annotation tool for multimodal corpora. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 91–96, Marseille, France. European Language Resources Association.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Tamara Drummond and Janina Wildfeuer. 2020. *The Multimodal Annotation of Gender Differences in Contemporary TV Series*, pages 35–58. De Gruyter, Berlin, Boston.

Laurent Filliettaz, Stéphanie Garcia, and Marianne Zogmal. 2022. Video-based interaction analysis: A research and training method to understand workplace learning and professional development. In Michael Goller, Eva Kyndt, Susanna Paloniemi, and Crina Damşa, editors, *Methods for researching professional learning and development: Challenges, applications and empirical illustrations*, pages 419–440. Springer.

Charles Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di semantica*, 6:222–254.

Charles J. Fillmore. 1977. *The Case for Case Reopened*, pages 59 – 81. Brill, Leiden, The Netherlands.

Charles J. Fillmore. 1982. Frame Semantics. In Linguistics Society of Korea, editor, *Linguistics in the morning calm*. Hanshin Publishing Co., Seoul, South Korea.

Maucha Andrade Gamonal, Tiago Timponi Torrent, Ely Edison Matos, Adriana S. Pagano, Frederico Belcavello, Fl avia Affonso Mayer, Arthur Lorenzi, Natália S. Sigiliano, Helen de Andrade Abreu, Lívia Vicente Dutra, Marcelo Viridiano, André Coneglian, Victor A. S. Herbst, Franciany O. Campos, Kenneth Brown, Lívia Pádua Ruiz, Lisandra Carvalho Bonoto, Luiz Fernando Pereira, and Yulla Liquer Navarro. 2025. Audition: A Frame-Annotated Multimodal Dataset for Accessible Audiovisual Content. In *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)*, page 95—104.

Maciej Grzenkowicz and Janina Wildfeuer. 2025. Addressing tiktok's multimodal complexity: a multi-level annotation scheme for the audiovisual design of short video content. *Digital Scholarship in the Humanities*, 40(4):1143–1166.

Michael A. K. Halliday and Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar*, 3rd edition. Edward Arnold, London.

Tuomo Hiippala. 2015. *The Structure of Multimodal Documents: An Empirical Approach*, volume 13 of *Routledge Studies in Multimodality*. Routledge, London.

Jana Pflaeging, Janina Wildfeuer, and John A Bateman. 2021. *Empirical multimodality research:*

Methods, evaluations, implications. Walter de Gruyter GmbH & Co KG, Berlin.

Mariane de Carvalho Pinto. 2025. Anotação Multimodal para Copilotos de Produção de Tecnologias Assistivas: uma proposta para a audiodescrição. M.A. Thesis in Linguistics, Universidade Federal de Juiz de Fora, Juiz de Fora.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute (ICSI).

Natália Sathler Sigiliano. 2025. Multimodal Frame Semantics: Expanding the Analytical Categories of FrameNet Brasil Multimodal Datasets. *Caligrama: Revista de Estudos Românicos*, 30(1):110–138.

Chinchu Thomas. 2018. Multimodal teaching and learning analytics for classroom and online educational settings. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 542–545.

Martin Thomas. 2009. Developing multimodal texture. In Eija Ventola and Arsenio Jesús Moya Guijarro, editors, *The world told and the world shown: multisemiotic issues*. Palgrave Macmillan, Basingstoke.

Tiago Timponi Torrent and Michael Ellsworth. 2013. Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. *Revista Veredas*, 17(1).

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*, 13.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2024. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, pages 1–29.

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.

Janina Wildfeuer and Joseph Adika Coffie. 2022. # socialiseresponsibly. analyzing the rhetorical structure of heineken tv commercials during the pandemic. *Frontiers in Communication*, 7:887706.

Leveraging LLMs for Semantic Type Annotation of Verbs' Arguments

Elisabetta Jezek, Gabriele Errico
Department of Humanities, University of Pavia
Piazza del Lino 2, 27100 Pavia, Italy
elisabetta.jezek@unipv.it, gabriele.errico01@universitadipavia.it

Abstract

In this paper, we present the results of two small-scale experiments aimed at verifying whether part of the Corpus Pattern Analysis (CPA) procedure developed by Hanks (2004) to manually extract recurrent language patterns from texts, can be automated using LLMs. Specifically, we examine ChatGPT and Gemini performance in the task of semantic type tagging of arguments in 150 Italian sentences realising 30 verb patterns (5 sentences per pattern). We run two experiments. In the first, we prompt ChatGPT to use the CPA ontology (about 200 hierarchically organized semantic types) in the annotation task; we provide the model with 5 sentences per pattern and ask it to assign the most specific type to the argument(s) of each sentence. In the second, we prompt both ChatGPT and Gemini to perform the task without the ontology, and ask the models to assign a single label to the argument(s) of the 5 sentences. Both experiments are performed in a zero-shot setting. We evaluate the results using the existing Italian T-PAS pattern resource as benchmark. Our results show that LLMs perform comparably well on both concrete and abstract type tagging and can therefore be used in a pilot study to support analysts in acquiring verb patterns from text.

Keywords: semantic type annotation, corpus pattern analysis, large language models

1. Background and Motivation

According to Hanks (2004, 2013), meaning is activated when words are combined with other words in patterns. To gain insight into the patterns of a language, one needs to study actual language behaviour as recorded in corpora. Hanks provides guidelines in the form of the Corpus Pattern Analysis (CPA) methodology for sorting and classifying language data. In CPA, typically, experts annotate a sample of about 250 concordance lines for a target verb, identifying recurrent patterns of verb usage and determining the semantic types of its arguments. Semantic types are selected from a hierarchically structured ontology (CPA ontology) of about 200 semantic types that has been developed for this purpose (Pustejovsky et al., 2006; Jezek and Hanks, 2010).

CPA is applied in practice in various languages, with the goal of building pattern inventories, e.g., English (Pattern Dictionary of English Verbs (PDEV), (Hanks and Pustejovsky, 2005), Spanish (Verbario, (Renau et al., 2019), Italian (Typed Predicate-Argument Structures (T-PAS), (Jezek et al., 2014); PhraseBase, (Giacomini and Rebosio, 2024), Croatian (CroaTPAS, (Marini and Jezek, 2019)) and Dutch (WoordCombinatie, (Colman and Tiberius, 2018)). [1] These resources are essential because by analyzing patterns in language, researchers may gain indirect access to how cognition is structured and operates. At the same time, with an increasing proportion of texts generated by large language models (LLMs), studying authentic linguistic patterns is urgent. The dominance of English in LLM training data affects linguistic outputs in other languages, not only by introducing loan words but also by influencing grammar and discourse. This potential levelling and homogeneization effect is further enhanced when AI-generated texts are again used as input for training the next generation of LLM.

Creating such inventories, however, is still mainly a computer-assisted, manual process and therefore very time-consuming. In this paper, we present the results of two small-scale experiments we conducted to explore whether LLMs could support certain aspects of the pattern-editing process. Building on previous experiments (Moretta et al., 2018), we specifically considered the models' performance in semantic type annotation of verbs' arguments. While semantic role labelling via LLMs is a widely performed task (Li et al., 2025), leveraging LLMs for semantic type tagging remains underrepresented.

It is important to note that a fundamental theoretical distinction exists between semantic roles and semantic types: a semantic type is an inherent attribute of an entity, while a semantic role has the attribute thrust upon it by the context (Jezek and Hanks, 2010). For example, a *knife* is inherently an [[Artifact]] which may play different roles depending on the context of use (e.g *sharpen a knife* (Patient) vs *cut with a knife* (Instrument). Despite the availability of different sets of tags for roles and types in lexical resources such as Framenet and VerbNet, roles and types are still often confused by annotators.

[1] All these languages can be be considered under-resourced in comparison to English.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 33–44
12 May 2026.

In the following, we describe the experiments and report the results we obtained. The paper is organised as follows: in Section 2, we introduce the methodology we followed in our research and the resources we used as benchmarks for our experiments. In Sections 3 and 4, we present our experiments: we describe the experimental setup and report the results. In Section 5 we discuss our results. Section 6 concludes the paper.

2. Methodology

As referenced above, since the current CPA procedure is manual and labour-intensive, we examined whether an LLM could support some aspects of the pattern editing process. We specifically considered ChatGPT and Gemini, two of the most popular LLMs at present, and tested the models' performance in semantic type annotation of individual arguments. We devised two experiments: one in which ChatGPT is prompted to use the CPA ontology in the annotation task, and one in which both ChatGPT and Gemini are prompted without the ontology. For our experiment, we used a small selection of manually annotated data from the Italian T-PAS project as benchmarks (Gold Standard, GS) to evaluate the models' output. The resource includes 5,529 patterns for 1,164 Italian verbs and 252,943 corpus examples annotated with the verb pattern (https://tpas.unipv.it).

Based on related literature (Lakoff & Johnson, 1980; Han et al., 2023; FreedomIntelligence, 2023), we expected that semantic types related to concrete entities would be easier to assign than those related to abstract entities. To test this hypothesis, we constructed a dataset from the T-PAS data comprising 15 concrete semantic types and 15 abstract semantic types (see Table 1), each with varying levels of specificity/generality within the CPA hierarchy. For each type, we selected a verb together with 5 example sentences from one of its patterns, resulting in a dataset of 150 sentences in total.

All runs were performed through the LLMs' web interactive interface without API calls or fine tuning. In particular, Experiment 1 used ChatGPT-4o, while Experiment 2 used ChatGPT-5 and Gemini 1.5.

3. Experiment 1 (ChatGPT, with ontology)

The goal of the first experiment was to determine whether ChatGPT-4o could be used to annotate corpus data with semantic types from the hierarchically structured CPA ontology. We provided the model with the following information in the prompt: the role it should take on when performing the task (role-based prompting), some background on the project, and the comprehensive list of CPA types inserted in their taxonomic structure. No examples were given; the prompting method was zero-shot. For each verb in Table 1, the model was asked to assign the most specific semantic type to the verb's arguments in each sentence, choosing from the types present in the ontology. The prompt can be found in the Appendix. For the statistical evaluation of the results of the first experiment, we decided to assign the scores as follows:

- 1 point for a perfect match, i.e., when the LLM returns as output the same label of the Gold Standard (corresponding to the label in the benchmark, column 2 in Table 1), for example, when the label of the Gold Standard is [[Water Vehicle]] and the LLM gives as output [[Water Vehicle]].
- 1 point for a partial match, i.e., when the LLM returns as output a label that is a direct subtype of the output of the Gold Standard, for example, when the label of the Gold Standard is [[Beverage]] and the LLM gives as output [[Alcoholic Drink]].
- 0 points for all the other cases, including non direct subtypes, for example when the label of the Gold Strandard is [[Information]] and the LLM gives as output [[Abstract Entity]], because in the reference ontology ([[Information]] is a subtype of [[Concept]] which is a subtype of [[Abstract Entity]]).

The reason for assigning 1 point to both a perfect and a partial (more specific) match was that we explicitly instructed the model to choose the most specific semantic type for the argument in the individual sentences, and therefore felt it would be unfair to penalise the model for doing so. Using the above scoring method, we calculated accuracy, precision, recall, and F1 score using the Python scikit-learn library. We also created confusion matrices comparing the model's predicted labels with the Gold Standard labels to gain insight into the semantic types assigned by the model to arguments across sentences, and to compare these with those in the patterns. In the following subsections, we report the results we obtained.

3.1. Results for Concrete Types

For the annotation of concrete semantic types, we achieved a high accuracy (0.75, see Table 2). If we consider the weighted average, i.e., a mean that takes into account the relative importance (or weight) of each data point in both experiments, the precision is higher than the recall. This means that the number of false positives is lower than the number of false negatives: ChatGPT-4o is more likely to miss the correct label (false negatives) than to assign the wrong one (false positives).

TopType	SemType	Verb-IT	Translation	SynRole
Concrete	Alcoholic_Drink	trincare	swallow down	obj
Concrete	Water_Vehicle	affondare	sink	sub
Concrete	Flying_Vehicle	atterrare	land	sub
Concrete	Cloth	ricamare	embroider	obj
Concrete	Bomb	esplodere	explode	sub
Concrete	Vehicle	parcheggiare	park	obj
Concrete	Musical_Instrument	accordare	tune	obj
Concrete	Beverage	bere	drink	obj
Concrete	Garment	indossare	wear	obj
Concrete	Device	scattare	shoot	sub
Concrete	Food	mangiare	eat	obj
Concrete	Flag	sventolare	wave	sub
Concrete	Artifact	fabbricare	make	obj
Concrete	Human	addomesticare	tame	sub
Concrete	Inanimate	brillare	shine	sub
Abstract	Software	costruire	build	obj
Abstract	Document	scrivere	write	obj
Abstract	Musical_Composition	comporre	compose	obj
Abstract	Business_Enterprise	avviare	start	obj
Abstract	Emotion	annegare	drown	obj
Abstract	Information	diffondere	spread	obj
Abstract	Goal	raggiungere	reach	obj
Abstract	Deity	adorare	adore	obj
Abstract	Money	pagare	pay	obj
Abstract	Time Period	passare	pass	sub
Abstract	Time Point	suonare	sound	sub
Abstract	Concept	sposare	marry	sub
Abstract	Responsibility	addossare	put on	obj
Abstract	Opportunity	sprecare	waste	obj
Abstract	Abstract_ Entity	cancellare	delete	obj

Table 1: Dataset of Italian verbs with semantic types and syntactic roles

	precision	recall	f1-score	support
accuracy			**0.75**	75
macro avg	0.52	**0.45**	**0.46**	75
weighted avg	**0.87**	**0.75**	**0.76**	75

Table 2: Evaluation of Concrete Semantic Types

In Figure 1, we report the confusion matrix for the annotations of concrete types. As mentioned above, we compare the labels predicted by GPT-4o ("predicted labels" on the horizontal axis) with the "true labels" assigned by the human annotators in T-PAS (vertical axis) to gain insight into the semantic types the model assigns to arguments in individual sentences. The matrices use a gradient colour scale to represent frequency counts. In this case, the lighter the color, the lower the count in the cell; conversely, the darker the color, the higher the count. The cell with the darkest colour corresponds to 5 correct instances (which is the number of sentences we asked the model to process for each semantic type). For example, as shown in Fig. 1 [[Water Vehicle]] was always correctly annotated, and the corresponding cells in the matrices are dark. By contrast, [[Beverage]] was labelled as such only once, which is reflected by the pale colour of the corresponding cell in the confusion matrix. The objects in three of the five example sentences were labelled twice as [[Water]], once as [[Liquid]] and once as [[Alcoholic Drink]]. An intermediate example is [[Food]]: this semantic type, which is a complex type in the ontology (i.e., a type made up of two components, i.e. [[Meal [Food, Activity]], was labeled twice as [[Meal]], once as [[Animal]], once as [[Fish]], and only once as [[Food]] (see below for further comments on this annotation).

For the 15 concrete types that we tested, GPT-4o assigned 29 different labels. In addition to the 10 concrete types in Table 1, the following labels appear: [[Animal]], [[Fire]], [[Firearm]], [[Fish]], [[Furniture]], [[Human Group]], [[Light]], [[Light Source]], [[Liquid]], [[Meal]], [[Road Vehicle]][[System]], [[Water]], [[Weapon]].

As referenced above, in the prompt we asked the model to identify the most specific type for a lexical item in particular syntactic positions within a sentence (subject or object). This proved to be highly effective with verbs that select very specific semantic types in the pattern slots, as in the case of [[Water Vehicle]], [[Alcoholic Drink]], [[Bomb]], [[Flag]], [[Flying Vehicle]], [[Garment]] and [[Musi-

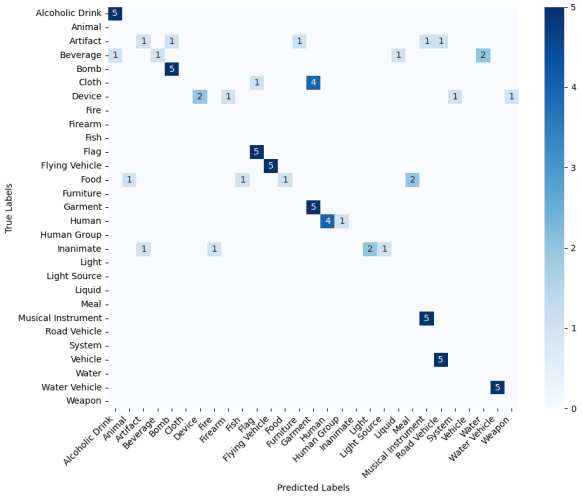

Figure 1: Confusion Matrix for Concrete Types (Experiment 1)

cal Instrument]]. On the other hand, some verbs do not select very specific semantic types in the pattern slots. For example, the verb *bere* 'to drink' does not select only one specific beverage, but rather all kinds of beverages: [[Beverage]] is the gold standard for the pattern. Considering that in the prompt we explicitly asked for the most specific type, it is only to be expected that the model would identify the type [[Water]] in cases where a sentence contains the word *acqua* 'water'. In the sentence *E' fondamentale ricordarsi di non bere acqua corrente* 'It is essential to remember not to drink tap water' the LLM labelled the object *acqua corrente* ('tap water') as [[Water]].

Similarly, the verb *fabbricare* 'to build, to make' can apply to several types of objects. The gold standard to capture this generality is the general type [[Artifact]], and the model identified more specific types in the different sentence, such as [[Furniture]] mobili per ingresso, [[Road Vehicle]] vagone di acciaio, [[Musical Instrument]] strumento musicale, which are all subtypes of [[Artifact]].

This raises an interesting issue about the relationship between the words in the texts and the types in the patterns (Jezek & Hanks, 2010): the type in the pattern is a generalisation over the word in the text. In a future experiment, it may therefore be of interest to include a further instruction in the prompt and ask the model to identify not only 'the most specific semantic type' but also 'the most specific semantic type that matches all the sentences provided'. This would also prove the ability of the model to generalize. For now, we adjusted the scoring method so that instances where the model chose a more specific type would not negatively affect the overall model's performance (see Table 2).

Polysemous lexical items are difficult for the LLM to identify. We identified cases of regular polysemy involving some of the objects of the Italian verb *mangiare* ('to eat'). In the sentence *Io proprio non ce la faccio a mangiare un animale, è più forte di me* 'I just can't eat an animal — I can't help it, it's just the way I am', the object *un animale* 'an animal' is assigned the semantic type [[Food]] in T-PAS, whereas the LLM assigns the label [[Animal]]. A similar case is found with the object *pesce freschissimo* 'incredibly fresh fish' in the sentence *Ti consiglio qualsiasi ristorantino nel Barrio Alto dove potrai mangiare pesce freschissimo a poco prezzo* 'I recommend any little restaurant in Bairro Alto, where you can eat incredibly fresh fish at a low price': while in T-PAS it is labelled as [[Food]], the LLM assigns the label [[Fish]]. *Animale* 'animal', like its hyponym *pesce* 'fish', is a polysemous word, as it can refer both to the living animal and to the dead one intended for consumption. Since this semantic relationship is currently not represented in the ontology, both cases received the score of 0. However, future implementations of the resource and the ontology could take this polysemous relationship into account.

Another issue is that some semantic types are interpreted differently by the LLM than they are intended in the ontology. For example, in the ontology, [[Human group]] is intended only for collective nouns, such as *family*, *class* or *police*. The LLM, by contrast, uses the label [[Human Group]] for everything that is 'more than one [Human]', assigning this label to plural nouns and pronouns, such as *they*. As a consequence, the matrix scores for [[Human]] and [[Human Group]] are lower (see Figure 1). For example, *gli uomini* ('the men'), the subject of the sentence *Gli uomini sapevano coltivare il grano e addomesticavano gli animali* ('The men knew how to cultivate wheat and domesticated animals'), was labelled as [[Human Group]].

3.2. Results for Abstract Types

As with the concrete types, the accuracy score for the annotation of the 15 abstract types is high: 0.79. In addition, as shown in Table 4, the precision score (weighted avg. Italian 0.81) is higher than the recall score (weighted avg. 0.79). The initial observation is that GPT-4o performed slightly better with abstract entities compared to concrete ones. On the one hand, this contradicts other studies: Han et al. (2023) and FreedomIntelligence (2023) state that ChatGPT shows promising skills in recognising concrete entities, while recognising abstract entities remains more challenging. This is probably because concrete entities are usually associated with more clearly defined sensory representations and richer contextual information (Peeters et al. 2023). On the other hand, we suggest that the level of polysemy for abstract types is lower. Moreover, the

level of correspondence between lexical items and types is higher. For example, in the sentence *Non sprechiamo questa opportunità di modificare una legge dannosa sia dal punto di vista fisico che psicologico* ('Let's not waste this opportunity to amend a law that is harmful both physically and psychologically'), there is a strong match between opportunità ('opportunity') and the label [[Opportunity]]. The lower number of abstract types in the CPA ontology compared to concrete types (i.e., 42 semantic types under [[Abstract Entity]] vs. 101 under [[Physical Entity]]) may also contribute to the higher scores.

	precision	recall	f1-score	support
accuracy			0.79	75
macro avg	0.58	0.56	0.56	75
weighted avg	0.81	0.79	0.79	75

Table 3: Evaluation of Abstract Semantic Types

Considering the confusion matrix in Figures 2, we observe that there are far fewer semantic types in the output for the abstract types than for the concrete types, i.e., only 16 compared to 26. The matrix also shows a higher number of dark cells compared to the confusion matrix for the concrete types. For the abstract types [[Business Enterprise]], [[Information]], [[Opportunity]], [[Responsibility]], and [[Time Period]], ChatGPT carried out the annotation completely correctly. At the same time, the number of pale cells is lower.

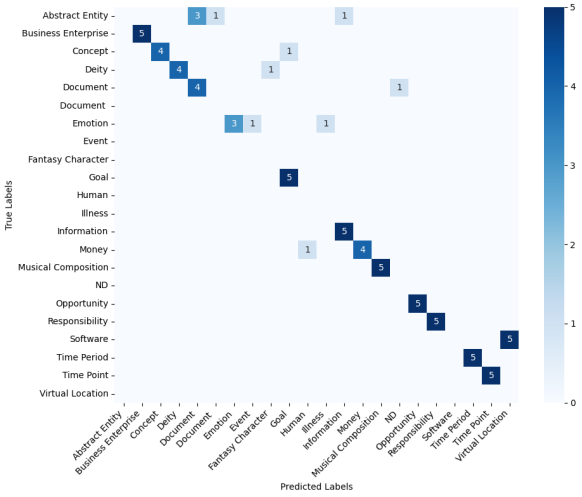

Figure 2: Confusion Matrix for Abstract Types, (Experiment 1)

Some semantic types, such as [[Software]], are challenging to annotate for both humans and LLMs. ChatGPT assigned the label [[Virtual Location]] to all five example sentences where the Gold Standard indicated [[Software]].

4. Experiment 2 (ChatGPT, no ontology)

Experiment 2 tests whether semantic type annotation can be performed without explicitly including the CPA ontology in the prompt. The dataset, interface, and workflow match those of Experiment 1 (interactive chat environment), so that the results remain directly comparable. However, it is important to note that the evaluation protocols differ across experiments. In Experiment 1, both perfect matches and direct subtype matches with the Gold Standard are scored as correct. By contrast, in Experiment 2, only perfect matches and accepted near-synonyms receive a correct score. Other outputs, including labels that are too general or too specific, are scored as incorrect. In Experiment 2, the model is then asked to identify the relevant argument(s) and assign the most specific semantic type that can extend across the full five-sentence set for the same verb slot.

4.1. Results for Concrete Types

Removing the ontology from the prompt led to a clear improvement in concrete-type annotation accuracy, increasing it to 0.80. These findings contrast with those from ontology-based approaches (Ouyang et al., 2023) but align with the notion of the predictable unpredictability of LLMs output, suggesting that prompt design is part of the method for annotating in the social sciences (Atreja et al., 2024). At the aggregate level, precision along with recall are balanced in this configuration. Figure 3 provides class-level evidence on remaining mismatches.

	precision	recall	f1-score	support
accuracy			**0.80**	75
macro avg	0.67	**0.67**	**0.67**	75
weighted avg	**0.80**	**0.80**	**0.80**	75

Table 4: Evaluation of Concrete Semantic Types

In Figure 3 we report the confusion matrix for concrete types in Experiment 2. The vertical axis contains the true labels from the GS, while the horizontal axis contains the adjudicated labels from the Score column (not predictions seen in Experiment 1). In this way rather than a gradient of blues the matrix shows either a complete alignment or a mismatch.

As we have seen in Experiment 1 for [[Beverage]] and for [[Food]], predictions included granularity shifts and polysemy effects (e.g., food-as-edible-entity vs animal-as-living-entity). Model errors on fine-grained hierarchy levels and no ontology prompt may help reach higher aggregates

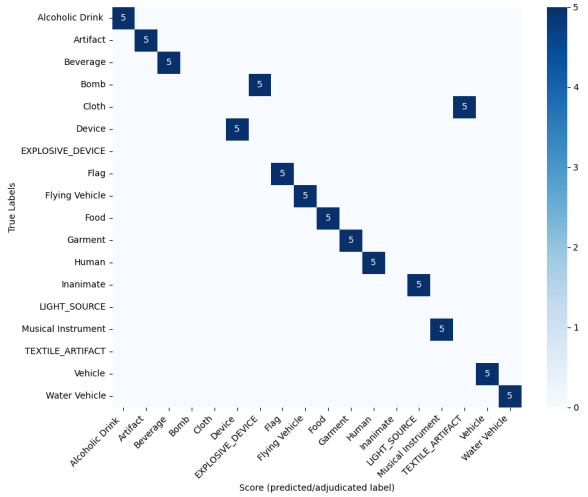

Figure 3: Confusion Matrix for Concrete Types (Experiment 2, ChatGPT

scores on semantic concrete types (see Section 3.1, accuracy 0.80 vs 0.75; macro F1 0.67 vs 0.46). Near-synonyms were accepted (e.g., Air Vehicle/Space Vehicle for [[Flying Vehicle]]), whereas super/sub typer were rejected (e.g., Textile Artifact for [[Cloth]]).

4.2. Results for Abstract Types

For abstract semantic types in Experiment 2 (ChatGPT, no ontology), accuracy is 0.73, with macro-average precision/recall/F1 at 0.58 and weighted precision/recall/F1 at 0.73. Similar results have also been demonstrated in legal studies (Savelka, 2023). At the class level, the error profile confirms that semantically related but over-general labels are penalized in the adjudicated Score: for example, Artifact / Digital Artifact for [[Software]] is scored as incorrect, and Music for [[Musical Composition]] is also scored as incorrect. As mentioned, only near-synonymous were retained, such as Text for [[Document]] and Organization / Business Activity for [[Business Enterprise]]. Figure 4 reports the confusion matrix for abstract types.

	precision	recall	f1-score	support
accuracy			**0.73**	75
macro avg	0.58	**0.58**	**0.58**	75
weighted avg	**0.73**	**0.73**	**0.73**	75

Table 5: Evaluation of Abstract Semantic Types

In Experiment 1, abstract types slightly outperformed concretes. In Experiment 2, the direction partly reverses (0.73 vs 0.79), while macro F1 slightly rises (0.58 vs 0.56), opening an interesting conversation about abstractness and specificity (Villani et al., 2024). A plausible interpretation, consistent with the same logic, is that removing the ontology and requiring a single label across five sentences increases pressure toward broader labels, especially in abstract domains (e.g., Artifact/Digital Artifact for [[Software]], Music for [[Musical Composition]]).

Concrete types, by contrast, remain more strongly anchored to sentence-level lexical evidence and therefore retain higher aggregate performance.

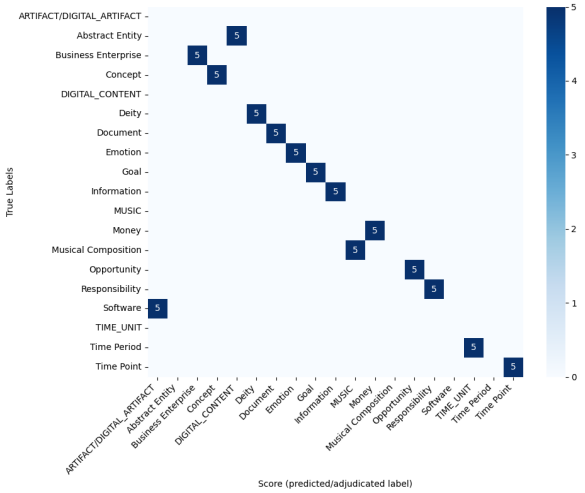

Figure 4: Confusion Matrix for Abstract Types (Experiment 2, ChatGPT)

5. Experiment 2 (Gemini, no ontology)

This parallel run mirrors Experiment 2 (ChatGPT) using the same dataset and no-ontology prompt objective. Predictions are logged in the same spreadsheet format and scored according to the Experiment 2 criteria described above.

5.1. Results for Concrete Types

	precision	recall	f1-score	support
accuracy			**0.73**	75
macro avg	0.61	**0.61**	**0.61**	75
weighted avg	**0.73**	**0.73**	**0.73**	75

Table 6: Evaluation of Concrete Semantic Types

For concrete semantic types the model reaches an accuracy of 0.73, with the other metrics at 0.61 and weighted averages at 0.73. As with the Experiment 2 (ChatGPT, no ontology), residual errors are mainly linked to label granularity. For example, Beverage was assigned where the GS was [[Alcoholic Drink]], too vague, so incorrect; by contrast, Aircraft for [[Flying Vehicle]] was retained acceptable

as considered a near synonym. A further mismatch concerns cases where [[Inanimate]] was rendered Light Source, i.e., very specific, so, rejected.

Compared with Experiment 2 (ChatGPT, no ontology), Gemini performs worse (accuracy 0.73 vs 0.80; macro F1 0.61 vs 0.67). This pattern is confirmed by the comparison against ChatGPT ontology: Gemini shows slightly lower accuracy (0.73 vs 0.75), even if it improves on macro F1 (0.61 vs 0.46). This pattern suggests a more balanced class-level behavior. Figure 5 reports the confusion matrix for Gemini concrete types, with Gold Standard labels on the Y-axis and adjudicated labels Score on the X-axis.

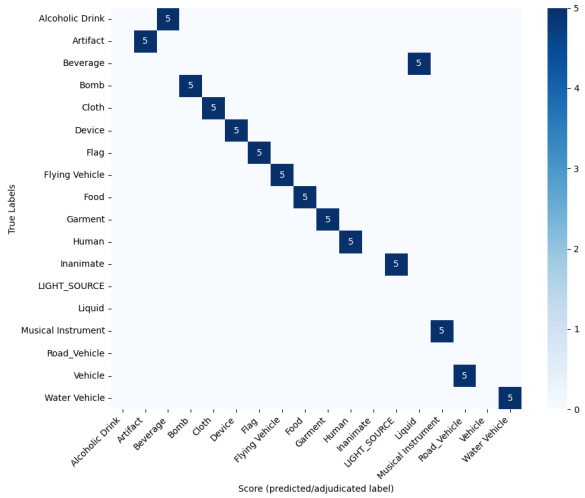

Figure 5: Confusion Matrix for Concrete Types (Experiment 2, Gemini)

5.2. Results for Abstract Types

For abstract semantic types in Experiment 2 (Gemini, no ontology), accuracy is 0.73, with macro-average precision/recall/F1 at 0.58 and weighted precision/recall/F1 at 0.73. [[Software]] was correctly captured, as well as Point in Time for [[Time Point]], accepted as near-synonym, troubling for ChatGPT; by contrast, Music for [[Musical Composition]], Psychological Feature for [[Emotion]], and Abstract Entity for [[Concept]] are plausibly incorrect with the current task.

Compared with Experiment 2 (ChatGPT, no ontology), aggregate indices are aligned (accuracy 0.73 and macro F1 0.58), with homogeneous performance overall and heterogeneous errors in different specific classes. Compared with Experiment 1 (ChatGPT, with ontology), abstract-type accuracy is lower in Gemini Experiment 2 (0.73 vs 0.79), while macro F1 is slightly higher (0.58 vs 0.56). These outcomes are consistent with the most recent research in the field. As above, this cross-experiment reading ought to remain careful because the scoring policies are not identical. Figure 6 presents the confusion matrix for Gemini abstract types, with Gold Standard labels on the Y-axis and adjudicated labels Score on the X-axis.

	precision	recall	f1-score	support
accuracy			**0.73**	75
macro avg	0.58	**0.58**	**0.58**	75
weighted avg	**0.73**	**0.73**	**0.73**	75

Table 7: Evaluation of Abstract Semantic Types

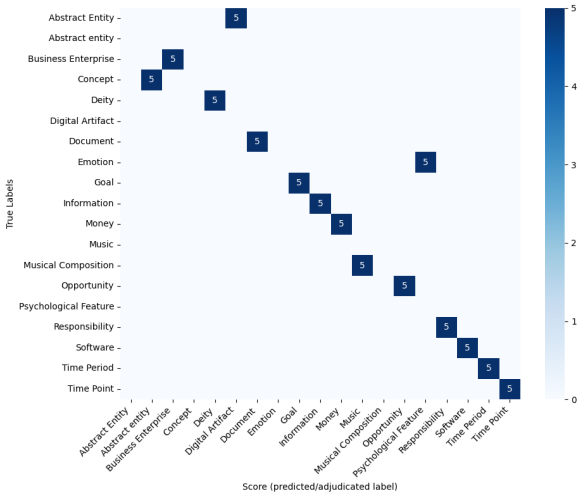

Figure 6: Confusion Matrix for Abstract Types (Experiment 2, Gemini)

Considering the confusion matrices in Figures 5 and 6, we observe that there are far fewer semantic types in the output for the abstract types than for the concrete types, i.e., only 16 compared to 26. The matrix also shows a higher number of dark cells compared to the confusion matrix for the concrete types. For the abstract types [[Business Enterprise]], [[Information]], [[Opportunity]], [[Responsibility]], and [[Time Period]], the model carried out the annotation completely correctly.

Some concepts are difficult to annotate, both for humans and for LLMs. The semantic type [[Software]] proved difficult. The model assigned the label [[Virtual Location]] to all 5 example sentences whereas the Gold Standard assigned [[Software]].

6. Discussion

From the Delta Snapshots in Table 9, we can notice that the contrast in concretes is greater, while abstracts are more aligned. In Table 8, under the tested settings, we can see a moderate split in semantic domains: for concretes (C-Acc) the accuracy metric puts ChatGPT without ontology at the top of the ranking, followed by its counterpart with

ontology, Gemini stays behind (0.73). For Abstracts (A-Acc) ChatGPT with ontology is the highest, while no ontologies setups converge (both 0.73). Delta snapshots are provided in Table 9 for visual aid of the differences across setups.

From these results, a reasonable and cautious explanation might be that prompting the LLM with packed information, such as ontologies, constrains the model on concrete labeling while stabilizing boundaries in abstract categorization. This is probably because the text of the ontology itself contains vectorial information that drives the model to assign an incorrect label due to the high pressure and high speed during the task, thereby 'distracting' the model from assigning the correct one. In this way, no ontology prompting scores higher on concretes assignments because lexical evidence, especially in stronger lexical binds across sentence sets occur (see, for instance, [[Alcholic Drink]]) does drive the model toward a less 'noisy' annotation task.

	C-Acc	C-F1	A-Acc	A-F1
GPT1	0.75	0.46	0.79	0.56
GPT2	0.80	0.67	0.73	0.58
Gem	0.73	0.61	0.73	0.58

Table 8: Summary Evaluation

	C-Acc Δ	C-F1 Δ	A-Acc Δ	A-F1 Δ
GPT2-GPT1	**+0.05**	+0.21	**-0.06**	+0.02
Gem-GPT1	-0.02	+0.15	-0.06	+0.02
GPT2-Gem	+0.07	+0.06	0.00	0.00

Table 9: Delta Snapshots across Experiments

Beyond the differences in aggregate scores, Gemini remains methodologically useful due to the fact that it provides a very user friendly explanation, guiding the user through why a label was selected or not (see Appendix c).

Taken into account that in Experiment 2 the models were not given any piece of information about the ontology, the results are quite astounding, they basically came up with their own types, which are, as shown by our research, directly comparable to the ones in the CPA ontology. The model didn't see the information, was not trained, and the possibilities that these models were trained through a specific fine-tuning for semantic types tagging, in respect to Hanks' ontology, are scarce.

7. Concluding remarks and further research

In this paper, we present the results of two small-scale experiments we conducted to verify whether part of the CPA procedure utilized to identify Italian verb patterns in corpora can be automated. In particular, we explore the performance of ChatGPT and Gemini on the task of semantic type annotation of verb arguments, using the existing T-PAS resource as benchmark. We set up the experiments to test the model's performance on both concrete and abstract semantic types. Overall, the results show that LLMs perform comparably well on both concrete and abstract type tagging.

From a practical standpoint, these results suggest that resource builders might markedly streamline their annotation workflow by incorporating LLM suggestions as a preliminary step. Similar workflows have also been implemented in historical studies (Celli and Mingazov, 2024). For example, a lexicographer might substantially reduce the time required for manual annotation by using an LLM to pre-annotate arguments' types. Then thereby, this allows to shift human expertise toward verification and correction of the model-proposed predictions. This approach could free up time for further linguistic analysis. In practical use cases, such as building computational lexicons, updating language resources, or assisting large-scale corpus studies, this workflow adjustment could improve scalability and budget efficiency, particularly when expert human resources are scarce or swift data processing is required. Remember that the interface and the models used were the free versions accessible to everyone, not giving any particular prompt-engineered advantage (Gilardi et al., 2023). Even small models can be helpful, from a mediocre resource to a better one (Simonetti et al., 2024). Considering this, prompt adjustment can sometimes take an unpredictable amount of time to measure, depending on the task.

Finally, human are better in performing label abstraction, but LLMs can be surprisingly accurate. Humans can be subject to tiredness, while LLMs have other constraints. For instance humans can directly infer that the verb to drink (it. *bere*) requires a general pattern label [[Beverage]]. This might denote a recurring mismatch between specific task a consistent generalization pattern. Polysemy, as seen with the difficulty to disambiguate between [[Animal]] (a living creature) instead of [[Food]] and group identification, as seen with plural label tagging resulting in generalization (collective nouns have not been contemplated by the model), clearly show that LLMs tagging should be carefully analyzed and realigned.

Ongoing research includes investigating whether LLMs are accurate in assigning these labels in other languages with CPA inventories: we already run preliminary experiments in Dutch with satisfying results. Future research includes: trying to use an open-source LLM such as Llama for the experi-

ments; expanding the dataset using more T-PAS data with the ultimate goal of developing an LLM-based semantic type labeler; studying the extent to which automatically pre-annotating sentences for patterns affects final annotation, by comparing LLMs assisted annotation with manual annotation from T-PAS data.

8. Acknowledgements

We are thankful to four reviewers for their insightful comments on an earlier version of this paper. Elisabetta Jezek has been partially funded by the University of Pavia with the ENHANCE SH project.

9. Limitations

The dataset we used is arguably a small dataset, and would benefit from expansion. While we considered existing resources and extracted linguistic data with the desired features from those, future work could expand it by collecting new data via human annotation, generation, or other data-driven approaches.

The authors initially performed multiple runs of Experiment 1 to observe the 'predictable unpredictability' of the models. Although, during the design of Experiment 2, a single-output was preferred to simulate a high-speed, efficient work scenario, instead of evaluating the unpredictability of outputs.

These results are considered very valuable by the authors and could possibly help not only in being less biasing during a human-in-the-loop pipeline work; or model-in-the-loop where humans construct the GS and the model are requested, without seeing the data first, to annotate semantic types, and only in the end human variation evaluation comparison checkups (GS vs, LLMs' output); but also address a variability that otherwise stays unchecked.

A further limitation interest hyperparameter setting: because the experiments were handled via web interfaces temperature was neither retrievable nor configurable. This binds pragmatically to our ready for use approach. However the authors are aware that unknowability may affect reproducibility.

While the present work performs an in-depth evaluation of how LLMs behave when faced with semantic typing, our research does not explore the inner mechanisms that underlie this capability. We acknowledge that doing so would provide complementary evidence that may be needed to shed full light on the phenomenon. Moreover, research could focus on how LLMs handle semantic typing in more naturalistic scenarios, e.g., in the context of real-world NLP applications, which is something the current work does not explore.

10. Ethical considerations

While this work presents no serious ethical concerns, a general consideration needs to be made about the use of pre-trained LLMs. As is commonly acknowledged, these models should be used with caution as they could perpetuate harmful biases present in their training data. Furthermore, there is a risk that they will generate false or misleading output. In our work, we minimize these risks as we do not use the LLMs to generate output, but only to score the plausibility of sentences fed as input.

11. Bibliographical References

S. Atreja, J. Ashkinaze, L. Li, J. Mendelsohn, and L. Hemphill. 2024. Prompt design matters for computational social science tasks but in unpredictable ways. *arXiv preprint arXiv:2406.11980*. https://doi.org/10.48550/arXiv.2406.11980.

F. Celli and D. Mingazov. 2024. Knowledge extraction from LLMs for scalable historical data annotation. *Electronics*, 13(24). https://doi.org/10.3390/electronics13244990.

L. Colman and C. Tiberius. 2018. A good match: A dutch collocation, idiom and pattern dictionary combined. In *Proceedings of the XVIII EURALEX International Congress: Lexicography in global contexts*, pages 233–246, Ljubljana, Slovenia.

L. Giacomini and L. Rebosio. 2024. Introducing PhraseBase: A linguistic information system for language learners and for NLP. In *PhrasaLex III Workshop on Phraseological Approaches to Lexicography*, Innsbruck, Austria. Conference presentation.

F. Gilardi, M. Alizadeh, and M. Kubli. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *arXiv*. https://doi.org/10.48550/arXiv.2303.15056.

P. Hanks. 2004. Corpus pattern analysis. In *Proceedings of the 11th EURALEX International Congress*, volume 1, pages 87–97. Université de Bretagne-Sud.

P. Hanks. 2013. *Lexical analysis: Norms and exploitations*. MIT Press, Cambridge, MA.

P. Hanks and J. Pustejovsky. 2005. A pattern dictionary for natural language processing. *Revue Française de Linguistica Appliquée*, 10(2):63–82.

E. Jezek and P. Hanks. 2010. What lexical sets tell us about conceptual categories. *LEXIS*, 4:7–22.

E. Jezek, B. Magnini, A. Feltracco, A. Bianchini, and O. 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, Paris. European Language Resources Association (ELRA).

Xinxin Li, Huiyao Chen, Chengjun Liu, Jing Li, Meishan Zhang, Jun Yu, and Min Zhang. 2025. LLMs can also do well! breaking barriers in semantic role labeling via large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23162–23180, Vienna, Austria. Association for Computational Linguistics.

C. Marini and E. Jezek. 2019. CroatPAS: A resource of corpus-derived typed predicate–argument structures for croatian. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, Torino. Accademia University Press.

F. Della Moretta, A. Feltracco, E. Jezek, and B. Magnini. 2018. Designing a methodology for semantic type tagging of argument positions. *Italian Journal of Computational Linguistics*, 4(2):1–16.

S. Ouyang, J. Huang, P. Pillai, Y. Zhang, Y. Zhang, and J. Han. 2023. Ontology enrichment for effective fine-grained entity typing. *arXiv preprint*. https://doi.org/10.48550/arXiv.2310.07795.

J. Pustejovsky, C. Havasi, J. Littman, A. Rumshisky, and M. Verhagen. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *Proceedings of LREC*, pages 1702–1705. European Language Resources Association (ELRA).

I. Renau, R. Nazar, A. Castro, B. López, and J. Obreque. 2019. Verbo y contexto de uso: Un análisis basado en corpus con métodos cualitativos y cuantitativos. *Revista Signos*, 52(101):878–901.

J. Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of GPT for zero-shot semantic annotation of legal texts. *arXiv preprint arXiv:2305.04417*. https://doi.org/10.48550/arXiv.2305.04417.

L. Simonetti, E. Jezek, and G. Vetere. 2024. Subcategorization of italian verbs with LLMs and T-PAS. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*. https://aclanthology.org/2024.clicit-1.99.pdf.

C. Villani, A. Loia, and M. M. Bolognesi. 2024. The semantic content of concrete, abstract, specific, and generic concepts. *Language and Cognition*, 16(4):867–894. Published online by Cambridge University Press on 18 January 2024. https://doi.org/10.1017/langcog.2023.64.

12. Appendix

12.1. Appendix A: Prompt Experiment 1 ChatGPT

Sei un linguista che sta lavorando per la risorsa T-PAS, un archivio di Typed Predicate Argument Structures, (anche dette pattern) per l'italiano. Tali pattern sono stati acquisiti manualmente da corpora attraverso l'analisi delle informazioni distribuzionali dei verbi italiani, seguendo la metodologia della Corpus Pattern Analysis descritta in Hanks (2004, 2013). Per ogni verbo presente nella risorsa, T-PAS mostra quali sono i suoi pattern più ricorrenti. I pattern sono costituiti dal verbo e dai suoi argomenti (soggetto, oggetto, complemento preposizionale, ecc.), a ognuno dei quali è assegnata una etichetta che ne identifica le proprietà semantiche (dette tipo semantico, ad esempio HUMAN, ARTIFACT, EVENT, LOCATION, ecc.). Per l'annotazione dei tipi semantici, gli annotatori si sono serviti di una ontologia, cioè di un elenco di etichette, organizzate in ordine gerarchico dal generale al particolare. L'ontologia che gli annotatori hanno utilizzato è la seguente: https://tpas.sketchengine.eu

Basandoti su quanto ti ho detto, per ciascuna delle seguenti frasi:

- 1. Avrebbe voluto vedere ancora quella luce verde brillare sulla lama e sentirsi tutt'uno con essa.
- 2. Un timido raggio di sole brillò su quei vetri.
- 3. La sua fronte si corruga ed i suoi anelli brillano, poi il bagliore si estende alla sua persona.
- 4. Dopo qualche istante una viva fiamma brillava dinanzi all' apertura.
- 5. Sembra quasi, che qui le stelle brillino di piu'.

Individua il soggetto e/o il complemento oggetto del verbo brillare.

Assegna a ciascuno di essi il tipo semantico più specifico a cui sono riconducibili, utilizzando come riferimento l'ontologia che ti ho fornito e la sua struttura gerarchica.

12.2. Appendix B: Prompt Experiment 2 ChatGPT /Gemini

Sei un linguista che sta lavorando per la risorsa T-PAS, un archivio di Typed Predicate Argument Structures, (anche dette pattern) per l'italiano. Tali pattern sono stati acquisiti manualmente da corpora attraverso l'analisi delle informazioni distribuzionali dei verbi italiani, seguendo la metodologia della Corpus Pattern Analysis descritta in Hanks (2004, 2013). Per ogni verbo presente nella risorsa, T-PAS mostra quali sono i suoi pattern più ricorrenti. I pattern sono costituiti dal verbo e dai suoi argomenti (soggetto, oggetto, complemento preposizionale, ecc.), a ognuno dei quali è assegnata un'etichetta che ne identifica le proprietà semantiche (dette tipo semantico, ad esempio HUMAN, ARTIFACT, EVENT, LOCATION, ecc.). Basandoti su quanto ti ho detto, per ciascuna delle seguenti frasi:

- 1. Partendo dal template scelto e' possibile costruire un sito web con script particolari, grafica aggiuntiva o richieste specifiche.
- 2. Ma come si fa a costruire un sito web che non faccia scappare chi lo visita?
- 3. Ho cominciato da poco a costruire siti web a livello amatoriale sopratutto per passatempo.
- 4. Attraverso lezioni passo per passo e progetti concreti imparerai a costruire siti Web in modo facile e veloce.
- 5. Nello zaino potremo trovare dunque ogni partecipante ha quindi illustrato il contenuto del suo zaino progettando e costruendo un proprio sito web.

Ti chiedo di: 1. Individuare il soggetto e/o il complemento oggetto del verbo costruire. 2. assegnare a ciascuno di essi il tipo semantico più specifico valido per tutte le frasi.

12.3. Appendix C: Example of Gemini output

3	**[[HUMAN]]** ("Homo Esageratissimus")	**[[BEVERAGE]]** (lista di cocktail e vodka)
4	**[[HUMAN]]** ("noi" - sottinteso)	**[[BEVERAGE]]** ("chardonnay" e "namabiru")
5	**[[HUMAN]]** ("i monatti")	**[[BEVERAGE]]** (sottinteso, nel "fiasco")

---

2. Assegnazione dei Tipi Semantici (Semantic Types)

Per identificare il pattern ricorrente in T-PAS, dobbiamo trovare le etichette semantiche più specifiche che accomunino tutti gli esempi (generalizzazione dai dati):

Soggetto (Arg0): [[HUMAN]]

Sebbene in alcuni contesti letterari o metaforici si possa trovare un animale (es. un gatto che "trinca" il latte), nel corpus analizzato l'agente è sistematicamente un essere umano. In T-PAS, l'etichetta **[[HUMAN]]** è quella corretta poiché implica la capacità intenzionale di compiere l'azione del bere con l'attitudine tipica di questo verbo.

Oggetto (Arg1): [[BEVERAGE]]

Il tipo semantico più specifico valido per tutte le frasi è **[[BEVERAGE]]** (bevanda).

- *Perché non [[LIQUID]]?* Sebbene ogni bevanda sia un liquido, il verbo *trincare* seleziona specificamente sostanze destinate al consumo umano (alcolici, birra, vino). Non si "trinca" solitamente l'olio motore o l'acqua piovana, a meno di usi altamente ironici.
- *Nota sulla specificità:* Molti esempi mostrano una preferenza distribuzionale per **[[ALCOHOLIC BEVERAGE]]**, ma per mantenere la validità universale del pattern in un archivio come T-PAS, si preferisce la categoria leggermente più ampia di **[[BEVERAGE]]**, eventualmente annotando la preferenza per gli alcolici nelle note d'uso.

Korean Quantification in Abstract Meaning Representation

Kiyong Lee, Harry Bunt,
Byong-Rae Ryu, Chongwon Park,
Younggyun Hahm
Korea University, Tilburg University,
Chungnam National University, University of Minnesota Duluth,
Teddysum AI
ikiyong@gmail.com, Harry.Bunt@tilburguniversity.edu,
ryu@cnu.ac.kr, cpark2@d.umn.edu,
hahmyg@teddysum.ai

Abstract

This paper explores the meaning of quantification in Korean and how it is encoded in Abstract Meaning Representation (Banarescu et al., 2019) and an enriched version AMR$^+$ accommodating Uniform Meaning Representation (UMR, 2022) and some of the contextual constraints proposed by Bos (2020). The extension makes five special references: Bunt et al. (2018), Bunt and Lee (2025), Pustejovsky et al. (2019), Bos (2020), and ISO (2025), the main reference. The aim of this paper is threefold. First, it focuses on implementing Korean AMR with the rich specification of QuantML (ISO, 2025) and its partially DRT-based semantics (Kamp and Reyle, 1993). Second, it supports the AMR multilingual development project by exploring methods for constructing a large-scale Korean AMR-annotated corpus. This line of research is necessary because Korean AMR resources remain severely underdeveloped. To date, only one study has examined a Korean AMR corpus (Choe et al., 2020). In addition, Korean's agglutinative morphology and head-final syntax challenge AMR frameworks that are largely based on the analytic inflectional language English. Third, it advances the current state of the UMR 2026 multilingual shared task by contributing more fine-grained annotations of quantification specified by ISO QantML ISO (2025) for resource domain, individuation, distributivity, and determinacy, as well as by treating coreference and lexical or scope ambiguities in Korean. Nonetheless, this paper postpones a review of the general semantics of Korean quantification until particular issues are addressed in future work.
Keywords: quantification, AMR, UMR, QuantML, contextual contraint

1. Aim and Motivation

This paper explores the meaning of quantified or polarity-sensitive quantifier-related sentences in Korean, not at a full scale, but at an experimental level, to examine how they are encoded in Abstract Meaning Representation (Banarescu et al., 2019) and (Knight et al., 2020), enriched by the ISO international standard QuantML (ISO, 2025). The enriched AMR$^+$ is also extended by accommodating Bos (2020) and Uniform Meaning Representation (UMR, 2022) with some modifications. Our extension draws on Bunt et al. (2018), Pustejovsky et al. (2019), Bunt (2021, 2024), Bunt and Lee (2025), and Bos (2020) to accommodate scope constraints within UMR and also to formulate the syntax of UMR in the manner of AMR.

Furthermore, unlike categorial grammar-based Montague semantics or phrase-structure dependent Discourse Representation Theory (DRT) (Kamp and Reyle, 1993), AMR-based semantics abstracts away from syntactic variations, having no syntactic pre-processing as required. On the other hand, Korean AMR is expected to be enriched with its rich morpho-syntactic information. For example, different case markings, such as the nominative marker -가 -ga or -이 i, in contrast to the topic marker -는 neun or -은 -eun, differentiate semantic interpretations.

The aim of this paper is threefold. First, it focuses on implementing Korean AMR with the rich specification of quantification in natural language by various efforts, such as Bunt (2021, 2024), Bunt et al. (2018, 2022), and Bunt and Lee (2025) to formulate ISO 24617-12:2025 on quantification as an ISO international standard. Second, this paper aims to support the AMR multilingual development project by exploring methods for constructing a large-scale Korean AMR-annotated corpus. This line of research is necessary because Korean AMR resources remain severely underdeveloped. To date, only one study has examined a Korean AMR corpus (Choe et al., 2020). In addition, Korean's agglutinative morphology and head-final syntax challenge AMR frameworks that are largely based on English. Third, it advances the current state of the UMR 2026 multilingual shared task[1] by contributing Korean data with more fine-grained annotations of quantification, including event distributivity and lexical or scope ambiguities in Korean.

The adoption of AMR for annotating quantification in Korean is motivated by both practical and theoretical considerations. From a practical perspective, AMR has gained wide acceptance in the

[1] See the homepage of the shared task (`https://ufal.mff.cuni.cz/umr-parsing/submission`) for instructions on how and where to submit a system output (as well as where to download the blind input).

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 45–53
12 May 2026.

computational linguistics community and scales well to large annotated corpora (Bos, 2016).

From a theoretical perspective, recent work, such as Bos (2020) and Pustejovsky et al. (2019), has shown the potential of AMR and UMR to encode quantificational phenomena in a principled yet compact way (Lee et al., 2025; Bunt and Lee, 2025). Focusing on events and their argument or adjunct structures, AMR performs most of the linguistic process of annotating data, while Bos (2020) and UMR (UMR, 2022) introduce discourse constraints on coreference, temporal or modal relations, and relative scopes of quantifiers and events to interpret AMR-based linguistic annotation structures.

These approaches demonstrate that quantificational meanings can be captured without relying on heavy meta-theoretic machinery or extensive inventories of categorial features, making them particularly suitable for large-scale annotation. These observations are relevant at the level of human annotators and programmers, and may only be insignificant at the level of computational operations. Nevertheless, the coverage of AMR at the current stage is not rich enough to accommodate the complex mechanism of quantification in natural language, as treated in the current version of QuantML (ISO, 2025).

It should be noted at the outset that the primary aim of this paper is to lay a preliminary basis for inviting shared work for developing the AMR-based construction of a Korean semantic database that allows the direct use of Hangul script in annotation, as is done extensively for Chinese with its complex but simplified Han-script. It thus postpones the task of formally introducing the general semantics of Korean quantification, even though more than a dozen excellent works on the subject exist; to cite one, see Lee (2018). We refer to them in depth as we treat particular relevant topics on Korean quantification in future projects.

The paper develops as follows: Section 2 outlines some issues in Korean quantification, chosen somewhat arbitrarily for this preliminary work; Section 3 illustrates how Korean data are annotated in AMR; Section 4 shows how AMR-annotated data is converted to logical form for interpretation. Here, we discuss how to treat various discourse constraints on scope ambiguity and other semantic phenomena. Section 5 sketches how to construct an AMR-based Korean semBank, and Section 6 concludes the paper.

2. Issues in Korean Quantification

2.1. Preamble

AMR 1.2.6 Specification (Banarescu et al., 2019) states: “AMR does not have a deep representation for quantifiers, but only *canonicalizes* their position." This paper, however, treats scope and other contextual constraints on variables related to quantification by extending the predicate-core level to another level, called *document level*, introduced by UMR (UMR, 2022). Following Lee et al. (2025), we call this level *discourse level*, treating scope, coreference, and other discourse constraints that have been discussed by Bos (2020) on variables, introduced at the predicate level of AMR.

As in English, Korean exhibits several challenges related to quantifiers, which may belong to nominal (e.g., “everyone"), adjectival (e.g., "all men"), or adverbial (e.g., "all done") categories. This paper addresses three such challenges. First, we propose a semantic classification of Korean quantifiers, as illustrated in 2.2. Second, we examine the lexical ambiguity of the universal quantifier 다 *all*, as shown in 2.2.3. Finally, we analyze issues of scope and distributivity, focusing on quantifiers such as 씩 ssik ‘each’. The semantics of Korean quantification, discussed here, follows QuantML (ISO, 2025), which makes use of Discourse Representations (DRSs) as defined in DRT (Kamp and Reyle, 1993) to represent logical forms in Neo-Davidsonian semantics (Davidson, 1967; Parsons, 1990), apart from the frequent use of second-order representations as in QuantML. The use of DRSs is preferred over the use of predicate-logic formulas, which are semantically equivalent, since they are more easily derived from the annotation structures of the QuantML abstract syntax.

2.2. Taxonomy of Korean Quantifiers

2.2.1. Overview

The semantic classification of Korean quantifiers is not much different from that of other languages. Notable differences, however, are shown in the morpho-syntactic variations of Korean quantifiers. These differences thus do not affect their lassification, except for how they may be implemented computationally or how they have developed historically. Nevertheless, we briefly introduce Korean quantifiers for references in the remaining part of this paper.

Based on the official eGovernment-supported open media Korean dictionary (kDictionary)[2], 우리말샘 (Urimalsaem, 2020), we classify quantifiers in Korean roughly into three types: universal, polarity-sensitive, and existential quantifiers.[3]

[2] https://opendict.korean.go.kr

[3] In this paper, we adopt the Korean standard romanization system, adopted for computational processing, instead of the Yale romanization of the Korean characters Hangul 한글, more often used in linguistic journals abroad.

2.2.2. Classification of Quantifiers (tentative)

1. Universal Quantifiers:
   다$_{003adv,007noun}$ da, “all"
   다$_{004adv}$ da,
   “reaching the end of an action or a state"
   Other items: 모두 modu, 모든 modeun.
   (a) Distributive universal (Choe, 1987):
   -씩$_{003sx}$ssik, “all each"
   “suffixed to a quantitative word in a distributive or divisive sense",
   Other items: 각각 gakgak, 마다 mada, 매 mae.
   (b) Partitive Universal:
   온$_{001adn}$ "whole", "entire"
   E.g.: 온 나라 "the whole (all parts of the) country""
   (c) Free choice universal:
   누구(나)$_{003,pro}$, nugu(na), “any"
   아무(나)$_{003,det}$, amu(na), “any".
2. Polarity sensitive:
   누구-도$_{001}$ nugu-do, “neg ... any"
   아무(도)$_{004,det}$ amu(do) “neg ... any"
   Note: 누구 nugu and 아무 have other senses, “who" and “someone", respectively, which are free of polarity.
3. Existential:
   - Logical existential:Lee (2002)
     적어도 하나 이상 jeogeodo hana isang “at least more than one"
     어떤 oddeon, 모 mo “a certain", “some"
   - Scalar existential:
     많다 manh-da, 많은 manh-eun “many"
     적다 jeog-da, 적은 jeog-un “a few"" or "a little"
     대부분 daebubun “most"
   - Numeric existential:
     하나 hana, 한 han “one"; 둘 “dul", 두 “du" “two"; 셋 ses 세 se “three"; etc.

The universal type has three subtypes: distributive -씩 (ssik), partitive 온 (on), and free choice 누구-나 (nugu-na). The existential type has logical, scalar, and numerical types. These quantifiers also have subtypes, depending on whether they are associated with countables or uncountable mass.

2.2.3. Senses of the Quantifier 다 da

The quantifier 다 da has at least three senses:

(1)
- 다$_{003}$ Adnominal (floating): “all"
- 다$_{004}$ Adverbial: meaning an action or state reaching its limit.[4]
- 다$_{007}$ Noun: “All came"

Quantifiers in general may quantify over either individual entities (e.g., 손님들 guests) or events (e.g., 왔다 came, 떠났습니다 leave). Here are examples:

(2) a. 모든$_{003}$ 손님들이 경주에 왔다.
modeun sonnim-deul-i gyeongju-e wass-da
all guest-Pl-Nom city-Loc come-Pst-Decl
“All the guests came to Gyeongju."
$\forall x[guest(x) \rightarrow \exists e[arrive(e), past(e), agent(e, x)]]$

b. 다(들)$_{007}$/모두가 떠났습니다.
da(deul)/modu-ga tteona-ss-seumni-da
all-Pl-Nom leave-Pl-Honorific-End
“All left."
$\forall x \exists e[person(x), arrive(e), past(e), agent(e, x)]$

c. 기차가 경주역에 (거의) 다$_{004}$/*모두 왔어.
train-Nom Gyeongju-yeog-e da/*modu wass-eo
“The train (almost) arrived at the Gyeongju Station."
$\exists\{x, y, e\}$
$[train(x), gyeongju(y), reach(e), past(e), goal(e, y)]$

d. 회의가 다$_{004}$ 끝났어.
hwe.eui-ga da ggunna-ss-eo
meeting-Nom all end-Pst-Ending
The meeting all ended.
$\exists e[meeting(e), end(e), past(e)]$

Examples (2a,b) illustrate how the individual entities are quantified by the universal quantifier. In Example (2c,d), the two occurrences of 다$_{004}$ are treated as quantifying over the events, as quantification is extended to events.

2.2.4. Scope

(3) 동네 총각들이 여자한테
다들 미쳤다.
dongne chongak-deul-i yeoja-hanthe
da-deul micheyeo-ss-da
village bachelor-Pl-Nom woman-Dat
all-Pl get-mad-Pst-Decl
“All the bachelors in the village got mad at a woman/women."

This example is interpreted either as indicating that there’s a specific woman that all the bachelors got mad at, or that each of them got mad at some woman, not neessarily at the same woman.

2.2.5. Case-dependent Meaning Differences

Case markers such as the topic marker “-는 -neun" or /“-은 -eun" and the nominative marker “-가 -ka" or “이 -i" differentiate meanings. Consider:

[4] This sense is controversial, as to whether it should be treated as a quantifier.

(4) a. 개가 짖는다. gae-ga jin-neun-da
dog-Nom bark-Pres-Decl
“A dog is barking (progressive)".
b. 개는 짖는다. gae-neun jin-neun-da
dog-Topic “Dogs bark."

Unlike (a), Sentence (b) with the subject marked with a topic marker has a *generic* reading. The verb in (a) is interpreted as progressive, whereas the verb in (b) is interpreted as present.

3. Annotating Korean Data in AMR

We annotate some of the examples given in Section 2 in the graph-based PENMAN format in AMR according to the guidelines in Banarescu et al. (2019). We, however, make some necessary modifications such as treating plurals or quantification. Note that AMR annotates raw data without any morphological or syntactic analysis. Note also that, as has been applied to Chinese large language data, AMR processors as well as the Python programming language are designed to accommodate Unicode-defined characters, including Chinese Han characters or Korean Hangul characters.

3.1. Existentials

(5) ::Snt-1 개가 짖는다.
:Interpretation: A dog is barking.
::id=dog1-2026-03-29

```
(b / 짖다 bark
   :agent (d / 개 dog)
   :aspect Progressive)
```

(6) ::Snt-2 개 두 마리가 짖고 있다.
::Interpretation: Two dogs are barking
::id=dog2-2026-03-29-KL

```
(b / 짖다 bark
   :agent (d / 개 dog
             :quant 2))
```

(7) ::Snt-3 여자 셋이서 맥주 한 캔씩 마셨다.
::Interpretation: There were three women. They each drank a can of beer.
::id=woman3-2026-03-29

```
(d / 마시다 drink
   :agent (w / 여자 woman
             :quant 3)
   :theme (b / 맥주 beer
             :quant 1
             :unit can))
```

3.2. Universal Quantifiers

There are three types of universal quantifiers: nominal, adnominal (adjectival), and adverbial. However, AMR does not differentiate them.

(8) :: snt: 세 사람 모두가 왔다.
:: interpretation: All of (the) three men came.
:: id:8, 2025-12-28

```
(c / 오다-01 come-01
 :: interpretation: \\
All of three men came.nt 3
        :quant A))
```

The universal quantifier is annotated as a logical constant `A`, like the negative polarity -: e.g., `:quant A`, and numeric quantities like 3 are treated as constants and annotated as `:quant 3`.

3.3. Negation and Polarity-sensitive Quantifiers

Korean has two types of negation: a short form and a long form. The short form 아니 “ani" or 안 “an" is adverbial like the English *not*, modifying a verb. The long form 않-다 “antha" or 아니하-다 “anihada" is an auxiliary verb requiring the stem of a main verb ending in a suffix -지 “-ji". This distinction is ignored in AMR-based annotation.

(9) a. 친구가 안 왔다.
chingu-ga an wa-ss-da
friend-NOM NEG come-Past-Declarative
“A friend did not come."
b. 친구가 오지 않았다.
chingu-ga o-ji anhae-ss-da
friend-Nom come-COMP NEG-Past-Decl
“A friend did not come."

(10) :: Snt: 친구가 안 왔다/ 오지 않았다.
:: Interpretation: A friend didn’t come.
:: id=neg1-2026-03-29

```
(c / 오다-01 come
   :agent (f / 친구 friend)
   :polarity -)
```

The polarity - above negates the event 오다 “come".

The universal quantifier 아무도 amudo is polarity sensitive, as in Example (11).

(11) :: snt: 아무도 나를 돕지 않았다.
:: interpretation: No one helped me.
:: id: k0011-2025-12-28

```
(h / 돕다-01 help
   :agent (p / 사람 person
             :quant
               (ex / E
                  :polarity -))
   :theme (m / 나 me))
```

`ex / E` stands for the existential quantifier, just as `A` stands for the universal quantifier.

(12) :: snt: 개 한 마리도 나한테 짖지 않았다.
:: interpretation: No dog barked at me.
:: id: k0012, 2025-12-28

```
(b / 짖다-01 bark
  :agent (d / 개 dog
           :quant
              (q / quantity
                :polarity -))
  :theme (m / 나 me))
```

(13) :: 손님들이 아무도/한 분도 오지 않았다.
:: interpretation: None of the guests came.
:: id: k0013, 2025-12-28

```
(c / 오다-01 come
    :agent (g / 손님 guest
        :quant (p / Pl
                :polarity -)))
```

Pl stands for plurality. As shown, many syntactic details are ignored in the annotation to keep AMR annotations simple and straightforward.

3.4. Senses of Adverbial 다 da

The adverbial form 다, which is generally treated as a universal in Korean, may also carry the sense of expressing the state of reaching the completion of an action or motion.

(14) :: Sentence: 기차가 서울에 다 왔습니다.
:: Interpretation: The train reached/all arrived in Seoul.
::: id: k0014, 21026-02-25

```
(c / 오다-01 come
   :agent (t / 기차 train)
   :goal (s / Seoul)
            :reached yes)
```

4. Interpreting AMR-annotated Korean Data

4.1. Conjunctive Logical Forms

As it is programmed, AMR is designed to automatically produce the graph-based conjunctive logical forms corresponding to the PENMAN format in annotating data. Consider Annotation (5) in the PENMAN format, copied here:[5]

(15) ::Snt-1 개가 짖는다.
:Interpretation: A dog is barking.
::id=dog1-2026-03-29

```
(b / 짖다 bark
   :agent (d / 개 dog))
```

[5] `aspect` is ignored here because the current version (Banarescu et al., 2019) does not annotate the progressive aspect.

This just says there is a dog that barks, but nothing else. AMR is rich enough to specify various meaningful attributes of a dog mentioned, as in QunatML.

This PENMAN annotation can be converted to a conjunctive logical form, as below:

(16) $\{b, d\}$
$[i(b,$짖다$\,bark), i(d,$개$\,dog), agent(b, d)]$

Here, there are two variables $\{e, d\}$ which may be bound by an existential quantifier $\exists$. The comma (,) stands for the logical connective $\wedge$, thus forming a sequence of conjunctive logical forms. i stands for a logical relation that instantiates a semantic concept, such as an event or a property, while relating a variable to it. $i(b,$짖다$\,bark)$, for instance, stands for $instance(b,$짖다$\,bark)$, where the relation $instance$ instantiates the event 짖다 of barking and b is its variable.

Consider another example, copied from (17), dealing with a negative polarity.

(17) :: Snt: 손님이 안 왔다/ 오지 않았다.
:: Interpretation: The guest didn't come.
:: id=neg1-2026-03-29

```
(c / 오다-01 come
   :agent (g / 손님 guest)
   :polarity -)
```

From this annotation, we can derive a sequence of conjunctive logical forms with negation.

(18) $\{e, x\}$
$[i(c,$오다$\,come), i(g,$손님$\,guest),$
$agent(c, g), polarity(c, -)]$

Here, the polarity negates the event of a friend's coming.

AMR cannot, however, treat plurality or quantification with its logical conjunction. We thus propose different ways to treat them.

4.2. Quantity and Plurality

(19) ::Snt-2 개 두 마리가 짖고 있다.
::Interpretation: Two dogs are barking
::id=dog2-2026-03-29-KL

```
(b / 짖다 bark
   :agent (d / 개 dog
             :quant 2))
```

Here is a sequence of conjunctive logical forms:

(20) $\{b, d\}$
$[i(b,$짖다$\,bark), i(d,$개$\,dog),$
$agent(b, d), quant(d, 2)]$

Here, $quant(x, 2)$ may be defined in set-theoretic terms.

(21) Definition of Quantity
Given a countable object x and a positive numeral n,
$quant(x, n)$ is defined as: $[x \in X, |X|{=}n]$.

With this definition, $quant(x, 2)$ can be replaced by (22), as in QuantML(ISO, 2025) below:

(22) $[X\,|\,|X|{=}2, x \in X \rightarrow dog(x)]$

Likewise, plurality (Pl) can be defined:

(23) Definition of Plurals
Given a countable object x and a positive numeral n,
$Pl(x, n)$ is defined as: $[x \in X, |X| \geq 2]$.

4.3. Universal Quantification

Consider:

(24) :: snt: 개들은 다 짖는다.
gae-deul-eun da jin-nun-da
dog-PL-topic ALL bark-Pres-Decl
"Dogs all bark."
:: interpretation:
All dogs bark.
:: id: k0024, 2026-03-28-KL

```
(b / 짖다-01 bark
   :agent (d / 개 dog
             :num Pl
             :quant A))
```

(25) Conjunctive Logical Form:
$\{b, d\}$
$[i(e,$짖다$\,bark), i(b,$개$\,dog),$
$agent(e, d), quant(d, Pl), quant(d, A)]$

Here, b and d are variables in AMR, and treated as discourse referents in DRT. The relation *instance*, which we abbreviate as i, instantiates each event or object. The comma "," is a conjunctive logical connective, usually represented by $\wedge$, making the entire list the sequence of conjunctive formulas.

The universal quantification $quant(d, A)$ may also be defined in set-theoretic terms, as in DRS.

(26) Definition of Universal Quantification
Given a set X, such that $X \geq 2$, of discourse referents, and a property P or its instantiation $i(x, P)$ in AMR:
$quant(x, A) =_{df} [X|x \in X \rightarrow i(x, P)]$.

With these definitions, Logical Form (25) is modified as below:

(27) Conjunctive Logical Form with Universal Quantification (for Individual distributivity):
$\{b, d\}$
$[i(b,$짖다$\,bark), i(d,$개$\,dog), agent(e, x)]$

$\{X\}$
$[[x{=}d, x \in X, |X| \geq 2] \rightarrow$
$\{E\}$
$[[e{=}b, e \in E] \rightarrow agent(e, x)]]$

Note here that $x{=}d$ and $e{=}b$ links the blower box of implicational logical forms for quantification to the first block with a sequence of conjunctive forms, created from AMR's PENMAN format.

In addition to the first block of AMR's conjunctive logical forms, quantification creates another block of logical forms, as shown in (27). This section mainly concerns how to relate this block to the basic block of AMR's conjunctive logical forms. At the same time, it treats scope and other semantic phenomena that depend on contextual constraints in discourse.

Like English and other languages, Korean data show how various discourse constraints, for instance, involving coreference and scope, can be represented at the discourse level. Each scopal constraint involving quantification can be represented by incorporating contextual constraints, such as those proposed by Bos (2020), into the document level, renamed *discourse level* (Lee et al., 2025), in UMR. Here are Bos's contextual constraints, some of which are directly incorporated into UMR's discourse level:

(28) 1. identical (=)
2. negated ($\neg$)
3. implication ($\Rightarrow$)
4. presuppositional ($<$)
5. inclusion (:)

The identity relation (=) represents the identity of a pronoun with its antecedent. The negation ($\neg$) negates the truth-value of a proposition, either by negating events of some properties, both of which are treated as *concepts* in AMR. The conditional ($\Rightarrow$) is used to represent universal quantification. The presupposition ($<$) involves the existential presupposition of names or definite descriptions with determinacy. The inclusion (:) relates contextual constraints. This paper illustrates some of these notions.

4.4. Logical Forms

Graph-based AMR annotation structures at the predicate level represent their logical forms as a list of conjuncts, where concepts are instantiated while relations relate the variable of a mother node to its daughters. Based on these conjunctive logical forms, the discourse-level constraints control the derivation of logical forms in first or second-order logic. Here is an example:

AMR produces a logical form corresponding to the PENMAN representation. Below, we produce a

minimal logical form that eliminates the representation of PL and A, which stand for plurality and the universal quantification.

(29) Minimal Logical Form
$i(b,$짖다$)$
$i(d,$개$)$
$agent(b,d)$

Here, i instantiates semantic concepts like the event 짖다 bark or the property 개 dog. By a contextual constrain, Bos (2020)'s implicational relation ($\Rightarrow$) holds between the property of being a dog and the event of barking in Figure 1.

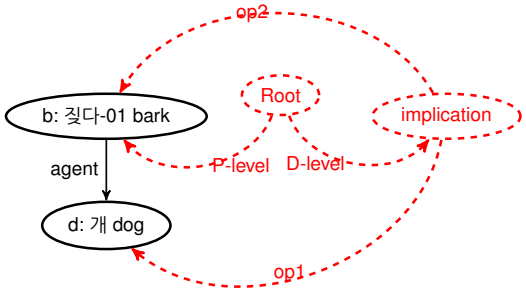

Figure 1:
Adding the Discourse Level to the AMR Graph

In Figure 1, the `Root` node is introduced as the top node dominating the AMR-based predicate level and the discourse level, marked in red. The *implication* ($\rightarrow$) relation dominates the AMR predicate structure under the root node. The op1 of the implicational constraint has a wider scope over the op2, as represented in the second block of Logical Form (30).

(30) $\{b,d\}$
$[i(b,bark),i(d,dog),agent(b,d)]$
$---------$
$\{X,E\}$
$[\,[|X| \geq 2, x=d, x \in X=] \rightarrow$
$[\,|E| \geq 2, e=b, e \in E] \rightarrow agent(e,x))]]$

This logical form is interpreted as asserting that there are dogs, and each of them bark.

4.5. Negative Quantification

Consider:

(31) :: snt: 개들이 다 안 짖고 있다.
:: two interpretations:
a. Not all of those dogs are barking.
b. All of those dogs are not barking.
:: id: k0031, 2026-01-13-KL

Interpretation (a) is annotated as in (32):

```
(32) (e / 짖다-01 bark
       :agent (x / 개 dog
            :quant (a / A
               :polarity -)))
             )))
     :discourse-level
     (c / constraints
        :negation (a / A
              :polarity -))
```

Here, the universally quantified 개 "dog" is negated, deriving the following logical form:

(33) Not all the dogs bark:
$[x|\neg i(x,dog) \rightarrow [e|i(e,bark),agent(e,x)]$

Interpretation (b) is represented as below:

```
(34) (e / 짖다-01 bark
        :agent (x / 개 dog
                 :quant (a / 다 all))
        :polarity -)))
```

From this, we derive the following logical form in first-order logic::

(35) All the dogs do not bark:
$[x|i(x,dog) \rightarrow \neg[e|i(e,bark),agent(e,x)]]$

5. Towards Constructing an AMR-based Korean SemBank

Designing a Korean SemBank (kAMR-SemBank) requires a sophisticated pipeline that can bridge the gap between Korean's rich morphosyntactic features and the abstract nature ofAMR. Unlike English, Korean's agglutinative morphology, where particles (Josa) and endings (Eomi) carry significant semantic weight, must be pre-processed to ensure accurate predicate-argument structures. Some of them also have morphologically complex derivational structures. Applying AMR-based annotation to data such as this requires time-consuming morpho-syntactic preprocessing to facilitate each annotation step.

In order to achieve this, our proposed design utilizes a multi-source data collection strategy, as depicted in Figure 2.

For correctly and efficientlyobtaining a balanced output of the AMR-based Korean SemBank (kAMR-based SemBank) einriched with QuantML for very large data, the proposed pipeline goes through the following three stages:

1. **Structured Resource Integration**: We prioritize the collection of data from the *Sejong Corpus* and the National Institute of Korean Language's *'Modu' Corpus* that already contain dependency annotations. Since Korean follows head-final syntax, the use of pre-structured syntactic data significantly reduces

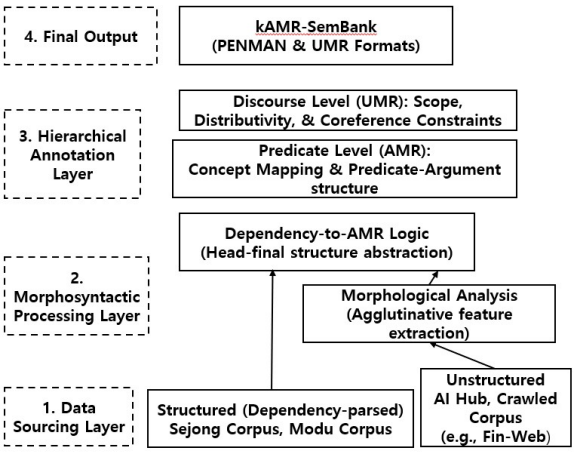

Figure 2: Proposed Pipeline for kAMR-SemBank Construction: Integrating Large-scale Raw Corpora with Morphosyntactic Enrichment

the time required to extract core argument structures for AMR.

2. **Large-scale Raw Corpus Acquisition**: To ensure the diversity and scalability of the SemBank, we incorporate extensive raw corpora from *AI Hub* (covering news, books, and web data) and crawled web-scale datasets such as *Fin-Web*. These resources provide a wealth of contemporary Korean quantification examples in various discourse contexts.

3. **Dual-layered Annotation Process**: The pipeline incorporates a hierarchical approach: a predicate-level for core semantic relations and a discourse-level (as defined in UMR) to handle Korean-specific quantification, distributivity features of events, and scope constraints.

6. Concluding Remarks

This paper argues that Korean quantification can be modeled in AMR while benefiting from UMR and Bos (2020)'s discourse-level constraints. We proposed a practical taxonomy of Korean quantifiers and used it to guide consistent annotation. We also treated the multiple senses of 다, including completion readings that quantify over events. At the predicate level, we normalized Korean surface variation, including nominal, adnominal, and adverbial forms, as well as short and long negation. We marked universals and existentials with logical constants and explicitly represented polarity, which keeps the graphs compact and readable. At the discourse level, we added constraints for presupposition, scope, and coreference, enabling first-order logical forms to be derived in a principled way.

This separation clarifies contrasts, such as *none* versus *not all*, and supports distinctions between *individual* and *collective distributivity* features, although we did not provide an adequate treatment of these fetures here. Finally, we sketched a corpus-building pipeline that combines structured resources with large-scale raw data and a dual-level annotation workflow. Future work will refine the guidelines and tools, expand coverage across genres, and validate the scheme through annotation studies and shared-task evaluation.

To initiate the project, TeddySum AI will first secure large-scale government funding to organize a shared task for constructing an AMR-based Korean semBank for quantification. The content of the shared work will be enriched by the ISO international standard 246217-12:2025 (ISO, 2025).

7. Acknowledgments

Kiyong Lee, the first author, is deeply grateful to the four reviewers for their detailed comments, which substantially improved this paper. He also sincerely thanks his coauthors, Chongwon Park, Harry Bunt, Byonrae Ryu, and Younggyun Hahm, for their patience, generosity, and sustained engagement with difficult issues, including many that extend beyond the Korean data itself. Their criticism, suggestions, and encouragement have been invaluable.

8. Limitations

The first author also regrets that he was not able to accommodate all of the reviewers' suggestions in the present version, especially those concerning the taxonomy of Korean quantifiers. He nevertheless remains deeply appreciative of their thoughtful and constructive comments.

In particular, the first author wishes to express special gratitude to Harry Bunt for his valuable comments on several of the paper's central theoretical issues, especially the inadequacy of AMR in the treatment of plurals, determinacy, individuation, and distributivity; in the use of DRSs to represent logical form; and in possible ways of relating AMR's predicate-structure level to UMR's discourse-structure level. Regrettably, not all of these issues could be covered in the present version of the paper due to limitations of time and space.

9. Bibliographical References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and

Nathan Schneider. 2019. *Abstract Meaning Representation (AMR) 1.2.6 Specification*. The DMR Working Group.

Johan Bos. 2016. Squib: Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Johan Bos. 2020. Separating argument structure from logical structure in AMR. In *Proceedings of the 2nd Interantional Conference on Designing Meaning Representations*, pages 13–20, Barcelona, Spain (online), December 13, 2020.

Harry Bunt. 2021. The ISA-17 quantification challenge: Background and introduction. In *Proceedings of the 17th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 33–40, Groningen, The Netherlands (online). Association for Computational Linguistics.

Harry Bunt. 2024. ISO 24617-12: A new standard for semantic annotation. In *Proceedings of LREC-COLING 2024*, pages 9361–9371, Turin, Italy. ELRA Language Resource Association: CC BY-NNC 4.0.

Harry Bunt, Maxime Amblard, Johan Bos, Karën Fort, Bruno Guillaume, Philippe de Groote, Chuyuan Li, Pierre Ludmann, Michel Musiol, Siyana Pavlova, Guy Perrier, and Sylvain Pogodalla. 2022. Quantification annotation in ISO 24617-12, second draft. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3407–3416, Marseille, France. European Language Resources Association.

Harry Bunt and Kiyong Lee. 2025. The representation of QuantML annotations in UMR – an exploration. In *Proceedings of the 21th Joint ACL–ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, Heinrich Heine University Düsseldorf, Germany. IWCS 2025.

Harry Bunt, James Pustejovsky, and Kiyong Lee. 2018. Towards an ISO standard for the annotation of quantification. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1787–17949, Miyazaki, Japan. ELRA.

Hyonsu Choe, Jiyoon Han, Hyejin Park, Tae Hwan Oh, and Hansaem Kim. 2020. Building Korean Abstract Meaning Representation corpus. *Proceedings of the Second International Workshop on Designing Meaning Representation*, pages 21–29.

Jae-Woong Choe. 1987. *Anti-quantifiers and a Theory of Distributivity*. Ph.D. dissertation, University of Massachussetts, Amherst.

Donald Davidson. 1967. The logical form of action sentences. In *The Logic of Decision and Action*, pages 181–120, Pittsburgh. Universityt of Pittsburgh.

ISO. 2025. *ISO 24617-12:2025 Language resource management – Semantic annotation framework (SemAF) – Part 12: Quantification*. The International Organization for Standardization, Geneva. Project leader: Harry Bunt.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT*. Kluwer, Dordrecht.

Kevin Knight, Bianca Badarau, Laura Baranescu, Claire Bonial, Madalina Bardocz, Kira Griffitt, Ulf Hermjakob, Daniel Marcu, Martha Palmer, Tim O'Gorman, and Nathan Schneider. 2020. *Abstract Meaning Representation Release 3.0*. Linguistic Data Consortium, Philadelpia, PA. Techinical Report LDC2020T02.

Chungmin Lee. 2002. (in)definites, case markers, classifiers and quantifiers in Korean. *Harvard Studies in Korean Linguistics*, 3:469–487.

Eunhee Lee. 2018. *Korean Syntax and Semantics*. Cambridge University of Press, Cambridge, UK.

Kiyong Lee, Harry Bunt, James Pustejovsky, Alex C. Fang, and Chongwon Park. 2025. Representing ISO-annotated dynamic information in UMR. In *Proceedings of the Sixth International Workshop on Designing Meaning Representations*, pages 49–58, Prague, Czechia. Association for Computational Lingustics.

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

James Pustejovsky, Nianwen Xue, and Kenneth Lai. 2019. Modeling quantification and scope in abstract meaning representations. In *Proceedings of the First International Workshop on Designing Meaning Representation*, pages 28–33. Association for Computational Linguistics. Florence, Italy, August 1, 2019.

Working Group UMR. 2022. *Uniform Meaning Representation (UMR) 0.9 Specification*. UMR Working Group for Guidelines.

Towards Corpus-Based Population and Visualization of ISO 24617-8 Ontology

Dariusz Czerski, Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
dariusz.czerski@ipipan.waw.pl, maciej.ogrodniczuk@ipipan.waw.pl

Abstract

This paper presents an extension of the ISO 24617-8 ontology for discourse relations through the integration of corpus-based examples and the development of a dedicated Ontology Viewer. The goal is to bridge the gap between formal ontological representations and practical corpus-based linguistic analysis, making discourse annotation frameworks more accessible to researchers. The proposed approach introduces a method for populating the ISO ontology with instances derived from three corpora (in Polish and English) compliant with the ISO 24617-8 standard. These instances formally connect discourse relations, argument roles, and explicit connectives within a unified semantic model. The Ontology Viewer enables intuitive browsing, filtering, and full-text searching of examples by language, relation type, and connective, offering both a relation-oriented and connective-oriented perspective. The experiment demonstrates the feasibility and effectiveness of this corpus-driven instantiation method and its visualization. The system provides a foundation for future integration of multilingual discourse corpora and contributes to the development of interoperable language resources for the Semantic Web and Natural Language Processing applications.

Keywords: discourse relations, ISO 24617-8, discourse ontology, ontology viewer

1. Introduction

While existing discourse formalisms provide conceptual schemata for discourse representation, their formalization into Web Ontology Language (OWL) ontologies is crucial for the Semantic Web, knowledge graph construction, and interoperable Natural Language Processing. A notable proposal of linking these two worlds is the set of Ontologies of Linguistic Annotation (OLiA; Chiarcos, 2014)[1], containing a Discourse Extensions[2] framework to cover discourse-related phenomena. The primary goal of the discourse part is to enhance the semantic interoperability of discourse annotations across major corpora and formalisms.

Although ontologies such as OLiA offer a powerful formalism for representing linguistic concepts and their interrelations, they can be difficult for linguists to use directly. Most linguists are not trained in ontology engineering or description logics, and the abstract syntax and formal constraints of OWL often obscure the linguistic content they aim to model. Specialized ontology editors such as Protégé[3] (Musen, 2015) provide useful visualization methods, but they remain too general-purpose, optimized for ontology builders rather than users who don't need to be aware of complex class hierarchies or logical restrictions. To make ontology-based linguistic resources more accessible and practically useful, there is a strong need for a dedicated interface abstracting away from the formal representation and instead presenting ontology content in user-friendly terms, supporting corpus-based exploration, and allowing intuitive linking between linguistic annotations, corpus examples and ontology concepts.

Following this path of thought, this article intends to propose two contributions: (1) the corpus-based method of ontology population limited to one discourse representation formalism – ISO 24617-8 (ISO, 2016), also with respect to discourse marker inventory induction, tested on two small ISO-compatible mini-corpora and one larger discourse corpus, and (2) presentation of the newly developed ontology viewer tailored to corpus-centered presentation of discourse relations as compared to previously proposed discourse marker-centered applications.

2. Related Work

Here we present three components that directly influenced our work: the ontology building framework concentrated on discourse, the ontology browser and a connective lexicon which can be also used as a connective-oriented discourse corpus browser.

The OLiA Discourse Extensions currently formalize nine annotation schemes covering corefer-

[1] https://acoli-repo.github.io/olia/
[2] https://acoli-repo.github.io/olia/discourse.html
[3] https://protege.stanford.edu/

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 54–61
12 May 2026.

ence, information status, information structure, and discourse structure for multiple languages. Each scheme is represented as an OWL/DL Annotation Model linked to a unified Reference Model via declarative linking models. This modular approach enables interoperability between different discourse annotation frameworks, serving as a terminological bridge and foundation for future integration with additional discourse and pragmatics resources. The work on the ontology is ongoing; since November 2024, the Reference Model for the Discourse Extensions has been integrated with the OLiA Reference Model.

One of the ontologies integrated in OLiA models is the ISO 24617-8 ontology, limited to listing all 20 discourse relations and their argument roles. Instead of directly annotating discourse relations, the ontology emphasizes argument roles (e.g. REASON and RESULT), from which the associated discourse relations (here: CAUSE) can be inferred via entailment. These entailment relations are formalized through `rdfs:subClassOf` links between argument roles and discourse relation classes, these in turn being subclasses of SYMMETRICDREL and ASYMMETRICDREL and grandchildren of DREL classes. For asymmetric discourse relations, argument roles are traditionally named ARG1 and ARG2 (subclasses of ASYMMETRICARGUMENTROLE).

The most widely used platform for viewing and editing OWL ontologies is PROTÉGÉ (Musen, 2015), an open-source framework developed at Stanford University and also available as a Web application[4] (Tudorache et al., 2013). WEBPROTÉGÉ offers intuitive access to classes, properties, and instances through an entity-centered interface. The browsing capabilities of the tool are offered by the Entity Description Browser which displays complete logical definitions and annotations of selected entities, Users can easily search and filter entities within projects, explore imported ontologies, and inspect both class hierarchies and instance data.

CONNECTIVE-LEX.INFO[5] (Stede et al., 2019) is a multilingual platform primarily designed to document and describe discourse connectives – words and expressions that signal semantic or pragmatic relations between discourse units (e.g. *because*, *although* or *however*). It provides fine-grained descriptions of connectives across multiple languages, linking them to standardized inventories of discourse relations. While concentrating on discourse connectives, the platform integrates information from various discourse-annotated sources, offering a view of corpus-based examples.

While general-purpose corpus query tools such as PML-Tree Query[6] (Štěpánek and Pajas, 2010) or ANNIS[7] (Krause and Zeldes, 2014) provide powerful multi-layer searching, they often require complex query languages (e.g., AQL). Our Ontology Viewer differs by being schema-native: it is built specifically to reflect the ISO 24617-8 hierarchy. Unlike Connective-Lex.info, which is connective-centric, our tool provides a balanced dual perspective: relation-oriented and connective-oriented, specifically tailored for validating the ISO semantic model.

3. Ontology Extension

OLiA defines ISO relations and their arguments[8] in a `discourse.ISO.owl` file as a series of classes following the naming from ISO 24617-8 standard. Figure 10 presents such definition for the PURPOSE relation. This asymmetric discourse relation takes two arguments: GOAL and ENABLEMENT which ISO standard describes as:

> Arg2 (Goal) is the goal or purpose of the situation described by Arg1 (Enablement).

Our goal was to extend this ontology with representation of corpus examples (from various corpora, in various languages) and all necessary constructs linking typed relations, arguments and connectives into a common formal description. To achieve that, the following design principles were assumed:

1. base definition of ISO relations are preserved in the original `discourse.ISO.owl` file
2. project extensions to the ontology are stored in `discourse.ISO.Ext.owl` file
3. corpus-based instance definitions are stored in a series `discourse.ISO.<corpus>.owl` files, e.g. `discourse.ISO.PDC.owl`.

3.1. Connective Property

The only project extension to the ISO ontology, independent of individual corpora, is the association of connectives (defined in the top-level ontology file `olia.owl`) to discourse relations. Contrary to

[4] http://webprotege.stanford.edu/
[5] http://connective-lex.info
[6] https://ufal.mff.cuni.cz/pmltq
[7] https://corpus-tools.org/annis/
[8] This unobvious decision is motivated by its better compatibility with other annotation models which reflects ''that most argument roles are recognized as independent discourse relations in other schemes" – see comment in the top `owl:Ontology` element in `discourse.ISO.owl` and (Bunt and Prasad, 2016). The authors of OLiA are of course well aware that ''In terms of ISO DR-Core, however, these are not in a hierarchical structure, but only in an entailment relation" (see another comment at DREL class definition).

relation arguments, which need to be assigned to a specific relation, connectives can be assigned to any relation so we use DRel class from `discourse.ISO.ow` file as the property domain:

```
<owl:ObjectProperty rdf:about=
  "http://purl.org/olia/discourse/
   discourse.ISO.Ext.owl
   #hasConnective">
   <rdfs:domain rdf:resource=
     "http://purl.org/olia/discourse/
      discourse.ISO.owl#DRel"/>
   <rdfs:range rdf:resource=
     "http://purl.org/olia/
      olia.owl#ExplicitConnective"/>
</owl:ObjectProperty>
```

3.2. Object Properties

To formally assign typed arguments to the corpus instance, we define properties of relation classes which would point to respective arguments (here: hasGoal and hasEnablement for the Purpose relation/class):

```
<owl:ObjectProperty rdf:about=
  "http://purl.org/olia/discourse/
   discourse.ISO.Ext.owl#hasGoal">
   <rdfs:domain rdf:resource=
     "http://purl.org/olia/
      discourse/discourse.ISO.owl
      #Purpose"/>
   <rdfs:range rdf:resource=
     "http://purl.org/olia/
      discourse/discourse.ISO.owl
      #Goal"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about=
  "http://purl.org/olia/discourse/
   discourse.ISO.Ext.owl
   #hasEnablement">
    <rdfs:domain rdf:resource=
      "http://purl.org/olia/
       discourse/discourse.ISO.owl
       #Purpose"/>
    <rdfs:range rdf:resource=
      "http://purl.org/olia/
       discourse/discourse.ISO.owl
       #Enablement"/>
</owl:ObjectProperty>
```

3.3. Instance Definitions

Finally, we construct corpus examples (instances of respective OLiA classes) using the original and newly defined building blocks. One example is an `owl:NamedIndividual` of a given type, carrying the text of the example split into relation arguments and (optionally) a discourse marker (connective).

An instance of a given relation carries its complete textual representation as the `label` and `skos:notation` value is used to represent the identifier of the sample. Then, the description uses previously defined object properties to in-line assign arguments to the relation. `skos:order` is used to preserve the linear sequence of text segments in the original discourse.

3.4. Design Rationale and Alternatives

The choice of `owl:NamedIndividual` with inline object properties was driven by the need for direct compatibility with existing Semantic Web reasoners. While RDF-star was also considered, we opted for a standard OWL 2 approach to ensure broader interoperability with legacy tools like Protégé. In the era of Large Language Models, where custom annotation tools can be generated rapidly, our approach prioritizes a standardized ontological schema. This model ensures that discourse data remains machine-actionable, queryable via SPARQL, and integrated into the global Linked Data cloud.

4. Ontology Population

Eventually the ontology is intended to be used to represent corpus examples from all future corpora converted to ISO representation in an ongoing project. Currently we decided to test its applicability to this task by representing three test corpora for two languages, based on the annotation guidelines (Ogrodniczuk, 2021) for the Polish Discourse Corpus (Ogrodniczuk, 2024), the ISO 24617-8 standard for English and the complete PDC (Ogrodniczuk et al., 2024). The samples from the two first sources were extracted from respective documents (the technical report and the ISO standard description) and stored as two mini-corpora (see their statistics in Table 4), with identifiers of examples automatically constructed. Polish examples were numbered according to their order within each relation description (and they are taking the form of `relation name-number`, e.g. `Cause-1`) while the ISO standard used letters for consecutive examples and they were also preserved (e.g. `Purpose-a`). Additionally, the complete PDC was encoded.

Dialogue act-related relations (Functional dependence and Feedback dependence) were not represented in the source document because these type of relations were omitted both from PDC annotation based on the assumption that their complexity exceeds the current capabilities of the annotation project. This decision goes in line with both the composition of ISO guidelines where dialogue acts are described by a separate part of the standard

Relation	PDC guidelines	ISO 24617-8	PDC (full)
Conjunction	3	2	8 286
Cause	1	7	1 753
Contrast	3	2	1 511
Asynchrony	1	4	1 057
Disjunction	3	3	815
Condition	1	2	807
Concession	1	2	712
Synchrony	3	3	528
Purpose	1	2	517
Exemplification	1	3	282
Substitution	1	3	227
Similarity	3	2	139
Manner	1	2	102
Restatement	3	2	78
Exception	1	2	32
Elaboration	1	3	12
Expansion	1	5	8
Negative Condition	1	3	5
Functional dependence	0	2	0
Feedback dependence	0	2	0
Total	**30**	**56**	**17 088**

Table 1: Number of examples in the ontology representing discourse relations of individual ISO types from the converted corpora

(ISO, 2020) and the decisions made by OLiA creators, deliberately excluding dialogue and speech act models from the ontologies.

5. Ontology Viewer

To facilitate exploration of discourse relation examples and annotations compatible with the ISO 24617-8 standard we have developed an interactive, Web-based Ontology Viewer powered by the ontologies described in the previous section. It integrates instances from available ISO 24617-8-compatible multilingual corpora, allowing users to browse and search through discourse relations, connectives and text samples across individual corpora and languages.

As for the final design of the tool, we consulted both Web-based versions of Protégé and Connective-Lex.info. While WebProtégé claims to be highly customizable with respect to user interface, its editing interface prevails over browsing and search capabilities which makes it less adaptive to our needs as the interface of Connective-Lex.info from which we decided to borrow the basic concepts of filter-oriented views and examples-centered search result lists

Figure 1 presents the main interface of the viewer. Each corpus entry in the left-hand panel shows the number of available annotated examples and the primary language of annotation. The top panel offers powerful filtering and search capabilities. Users can filter examples by language, connective, relation type, or relation symmetry. They can also perform full-text searches through all example sentences. This facilitates both relation-oriented analyses (e.g., inspecting how a relation such as Purpose or Contrast is realized in a specific language) and comparison across individual use of connectives within relations.

In the main panel individual discourse-relation instances are displayed with their internal structure: relation arguments (here: Enablement and Goal), the connective linking them (*so that*), and the relation type (Purpose). Each example links to its unique ontology identifier, ensuring traceability to the underlying OWL definition and corpus source.

6. Limitations and Replicability

Currently, the viewer focuses exclusively on the discourse layer. It does not yet visualize lower linguistic layers such as morphology or syntax (based on available CoNLL-U annotation). The current population is limited to three corpora, representing a pilot phase of the project.

To support open science, the Ontology Viewer source code is available at: https://github.com/ipipan/ontology-viewer. A live demo of the Ontology Viewer is accessible at: https://ontology-viewer.ipipan.waw.pl/

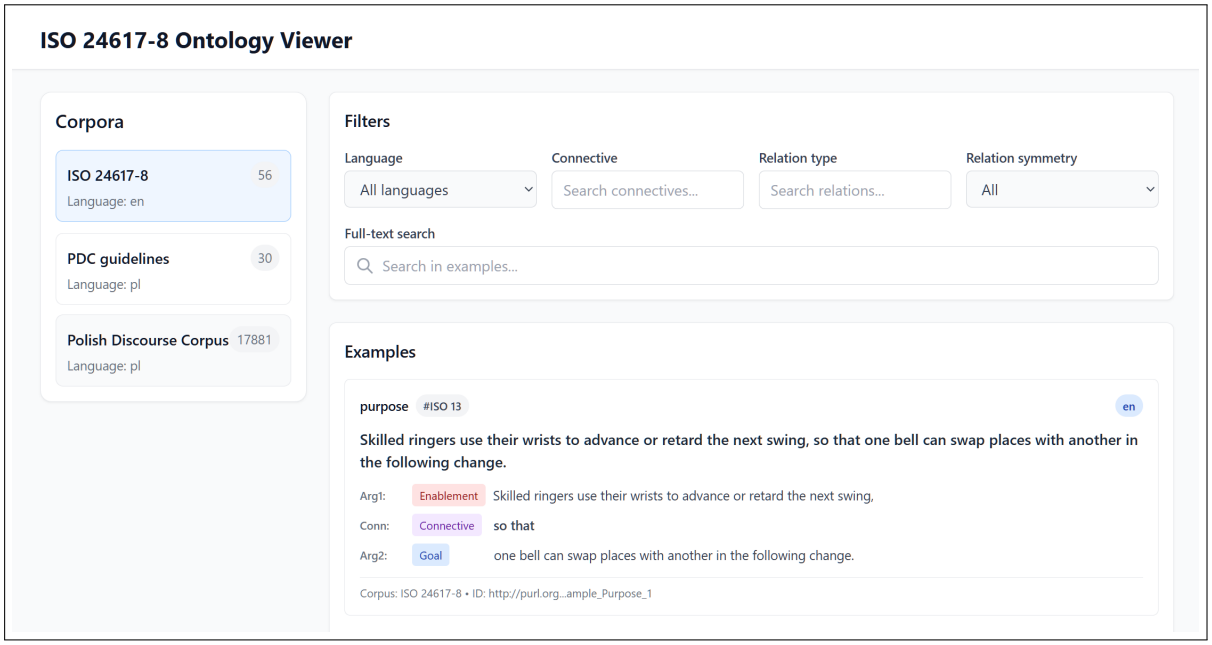

Figure 1: Interface of the ontology viewer

7. Conclusions and Future Work

We believe that the proposed population of ISO ontology with corpus examples and implementation of the ontology viewer can act as a bridge between ontology data and corpus-based linguistic evidence, providing both a formal description and an intuitive environment for linguists and annotators to explore, validate, and compare ISO 24617-8 discourse-relation annotations.

Both the representation of the instances of discourse relations from individual corpora in the ISO ontology and the flexible interface of the viewer were developed as a starting point for integration of a set of 10–15 discourse corpora in several formalisms and languages to be converted to the ISO 24617-8 standard in the ongoing Universal Discourse project[9] (see Acknowledgements).

The experiment showed that both the proposed method of providing corpus-based instantiation of ISO ontology classes and its presentation in a dedicated viewed are feasible and effective. We believe that the interface of the ontology viewer constitutes another convenient method of browsing through the corpus samples, filter them and offer and easy-to-use search across languages, relation types, connectives and textual content.

The viewer also offers another method of presenting the discourse corpora through the lens of discourse connectives, similar to ConnectiveLex.info and other corpus-based methods (see e.g. Das et al., 2018; Silvano et al., 2022) augmenting the inventory of discourse markers with reference material.

Even though it was not intended to be used for other discourse representation formalisms, the ontology viewer can be easily adapted to be used for displaying 'flat' relation models such as PDTB.

[9] http://udisc.org/

8. Acknowledgements

This research was funded in whole by the National Science Centre, Poland, grant 2023/50/A/HS2/00559 ("Universal Discourse: a multilingual model of discourse relations").

9. Bibliographical References

Harry Bunt and Rashmi Prasad. 2016. ISO DR-Core (ISO 24617-8): Core Concepts for the Annotation of Discourse Relations. In *Proceedings of the 10th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)*, pages 45–54.

Christian Chiarcos. 2014. Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 4569–4577, Reykjavik, Iceland. European Language Resources Association (ELRA).

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. 2018. Constructing a Lexicon of English Discourse Connectives. In *Pro-*

ceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue, pages 360–365, Melbourne, Australia. Association for Computational Linguistics.

ISO. 2016. ISO 24617-8:2016: Language resource management – Semantic annotation framework (SemAF) – Part 8: Semantic relations in discourse, core annotation schema (DR-core).

ISO. 2020. ISO 24617-2:2020: Language resource management – Semantic annotation framework (SemAF) – Part 2: Dialogue acts.

Thomas Krause and Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Mark A. Musen. 2015. The Protégé Project: A Look Back and a Look Forward. *AI Matters*, 1(4):4–12.

Maciej Ogrodniczuk. 2021. Założenia dla korpusu metatekstowego (EN: Discourse corpus guidelines). Technical report, CLARIN-PL. Institute of Computer Science, Polish Academy of Sciences.

Maciej Ogrodniczuk, Aleksandra Tomaszewska, Daniel Ziembicki, Sebastian Żurowski, Ryszard Tuora, and Aleksandra Zwierzchowska. 2024. Polish Discourse Corpus (PDC): Corpus design, ISO-compliant annotation, data highlights, and parser development. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12829–12835, Torino, Italy. ELRA and ICCL.

Purificação Silvano, Mariana Damova, Giedrė Valūnaitė Oleškevičienė, Chaya Liebeskind, Christian Chiarcos, Dimitar Trajanov, Ciprian-Octavian Truică, Elena-Simona Apostol, and Anna Baczkowska. 2022. ISO-based Annotated Multilingual Parallel Corpus for Discourse Markers. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 2739–2749, Marseille, France. European Language Resources Association.

Manfred Stede, Tatjana Scheffler, and Amália Mendes. 2019. Connective-Lex: A Web-Based Multilingual Lexical Resource for Connectives. *Discours*, pages 1–38.

Jan Štěpánek and Petr Pajas. 2010. Querying Diverse Treebanks in a Uniform Way. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Tania Tudorache, Csongor Nyulas, Natalya F. Noy, and Mark A. Musen. 2013. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic Web*, (1):89–99.

10. Language Resource References

Ogrodniczuk, Maciej. 2024. *Polish Discourse Corpus*. Institute of Compute Science, Polish Academy of Sciences. PID https://zil.ipipan.waw.pl/PolishDiscourseCorpus.

Appendix

A. Definition of PURPOSE discourse relation class in OLiA

```
<owl:Class rdf:about="http://purl.org/olia/discourse/
                      discourse.ISO.owl#Purpose">
    <rdfs:subClassOf rdf:resource="http://purl.org/olia/discourse/
                                   discourse.ISO.owl#AsymmetricDRel"/>
    <rdfs:comment>Arg2 is the goal or purpose of Arg1
                  Purpose: Goal, Enablement</rdfs:comment>
    <rdfs:label>Purpose</rdfs:label>
</owl:Class>

<owl:Class rdf:about="http://purl.org/olia/discourse/
                      discourse.ISO.owl#Goal">
    <rdfs:subClassOf rdf:resource="http://purl.org/olia/discourse/
                                   discourse.ISO.owl#Arg1"/>
    <rdfs:subClassOf rdf:resource="http://purl.org/olia/discourse/
                                   discourse.ISO.owl#Purpose"/>
    <rdfs:label>Goal</rdfs:label>
</owl:Class>

<owl:Class rdf:about="http://purl.org/olia/discourse/
                      discourse.ISO.owl#Enablement">
    <rdfs:subClassOf rdf:resource="http://purl.org/olia/discourse/
                                   discourse.ISO.owl#Arg2"/>
    <rdfs:subClassOf rdf:resource="http://purl.org/olia/discourse/
                                   discourse.ISO.owl#Purpose"/>
    <rdfs:label>Enablement</rdfs:label>
</owl:Class>
```

B. Example instances of PURPOSE relation from the test English discourse corpus

```
<owl:NamedIndividual rdf:about="http://purl.org/olia/discourse/
  discourse.ISO.EN.owl#Example_Purpose_1">
  <rdf:type rdf:resource="http://purl.org/olia/discourse/
                          discourse.ISO.owl#Purpose"/>
  <rdfs:label xml:lang="en">Skilled ringers use their wrists to advance
    or retard the next swing, so that one bell can swap places with
    another in the following change.</rdfs:label>

  <skos:notation>ISO 13</skos:notation>

  <hasEnablement>
    <owl:NamedIndividual rdf:about="http://purl.org/olia/discourse/
      discourse.ISO.EN.owl#Example_Enablement_1">
      <rdf:type rdf:resource="http://purl.org/olia/discourse/
        discourse.ISO.owl#Enablement"/>
      <rdfs:label>Skilled ringers use their wrists to advance
        or retard the next swing,</rdfs:label>
      <skos:order rdf:datatype="http://www.w3.org/2001/
        XMLSchema#integer">1</skos:order>
    </owl:NamedIndividual>
  </hasEnablement>
```

```
  <hasConnective>
    <owl:NamedIndividual rdf:about="http://purl.org/olia/discourse/
      discourse.ISO.EN.owl#Connective_Purpose_1">
      <rdf:type rdf:resource="http://purl.org/olia/olia.owl
        #ExplicitConnective"/>
      <rdfs:label>so that</rdfs:label>
      <skos:order rdf:datatype="http://www.w3.org/2001/
        XMLSchema#integer">2</skos:order>
    </owl:NamedIndividual>
  </hasConnective>

  <hasGoal>
    <owl:NamedIndividual rdf:about="http://purl.org/olia/discourse/
      ISO.EN.owl#Example_Goal_1">
    <rdf:type rdf:resource="http://purl.org/olia/discourse/
      discourse.ISO.owl#Goal"/>
    <rdfs:label>one bell can swap places with another in the following
      change.</rdfs:label>
    <skos:order rdf:datatype="http://www.w3.org/2001/
      XMLSchema#integer">3</skos:order>
    </owl:NamedIndividual>
  </hasGoal>
</owl:NamedIndividual>
```

Tracing consensus formation in meetings: Annotation and incremental decision modelling in the MEET Corpus

Ghazaleh Esfandiari-Baiat, Jens Edlund
KTH Royal Institute of Technology
Stockholm, Sweden
geb@kth.se, edlund@speech.kth.se

Abstract

We present an incremental annotation scheme and discourse model designed specifically for the study of consensus formation in collaborative meetings. By grounding the representation in observable contributions and enforcing a strict no-lookahead principle, the model provides a tractable way to analyse how decisions emerge over the course of interaction. The resulting structures are intentionally minimal yet expressive enough to capture the evolving task state and support dynamic visualisation and replay of the decision process. A web-based reference implementation of the model demonstrates how the evolving decision state can be inspected and replayed during analysis. Together with a suitable corpus, this framework provides a practical foundation for investigating the multimodal dynamics of collaborative decision-making in professional meetings.

Keywords: speech corpus, collaborative meetings, ranking tasks, annotation, dialogue modelling

1. Introduction

Meetings are a central mechanism for collaborative decision-making in professional settings. Meeting participants often work toward a shared outcome through extended discussion, and it is important to examine not only the final outcome but also the interactional process through which consensus emerges over the course of the meeting.

To investigate this process in a controlled setting, we analyse a corpus of small-group meetings in which participants solve a collaborative ranking task. In each meeting, the group must agree on an ordered list of items with respect to a given criterion. This task defines a clear decision space in which the emerging consensus can be observed as participants progressively construct and adjust the ranking.

Given this structured decision space, our aim is to track how the group's decision state evolves throughout the meeting. At any point in the interaction, we maintain a current estimate of the emerging consensus based solely on the discussion observed so far. As participants introduce, question, and refine propositions, this estimate is updated to reflect the state of the discussion. Ideally, the estimate converges toward the consensus ultimately reached by the participants.

To support this analysis, we introduce an annotation scheme and a discourse model designed to represent the evolution of the decision state as the meeting unfolds. The framework interprets the interaction incrementally, allowing the discussion to be represented as a sequence of updates to the emerging consensus. In this way, the interaction can be analysed not only as a collection of utterances, but as a structured process in which propositions about the ranking are proposed, examined, and progressively stabilised.

The framework is guided by a small set of practical design principles that constrain both the annotation procedure and the interpretation of the interaction. In particular, annotations are assigned without lookahead, reflecting only the information available at the current point in the meeting. The scheme also deliberately captures only those aspects of the interaction that are essential for modelling the evolving decision state. Annotations are currently produced by human annotators, forming a human-in-the-loop process that supports inspection and correction of the evolving interpretation. The approach is intended to remain compatible with existing dialogue annotation traditions, including the ISO framework for dialogue act annotation, while focusing specifically on the task-oriented modelling of consensus formation.

The annotation scheme and discourse model are intended not only to represent the interaction, but also to support the investigation of the modelling process itself. During replay of a meeting, the evolving decision state can be inspected as the model updates its estimate of the emerging consensus. A web-based reference implementation demonstrates how the incremental state of the model can be visualised and replayed during analysis. This allows the effects of modelling assumptions to be examined while following the interaction over time, making the framework a practical tool for exploring how consensus formation can be captured through incremental discourse modelling.

The present paper makes three main contributions. First, we introduce an incremental annota-

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 62–67

tion scheme designed specifically for collaborative ranking tasks in meetings. Second, we propose a minimal discourse model that derives the evolving decision state directly from these annotations while enforcing a strict no-lookahead interpretation. Third, we present a reference implementation that supports visualisation and replay of the decision process, enabling detailed investigation of how consensus emerges over the course of the interaction.

2. Background and related work

In this section we briefly point to prior work that is directly relevant to the present study, foregoing a comprehensive positioning within the many related research areas for reasons of space.

Research on meetings is closely related to the social psychology of small groups, which examined how groups coordinate activity and reach decisions through interaction (Davis et al., 1976). However, meetings constitute a more specific form of organised interaction than small groups in general. As noted by (Goffman, 1961), meetings are characterised by a sustained shared focus of activity among participants, which distinguishes them from more loosely structured forms of group interaction. Within the taxonomy of group tasks proposed by (McGrath, 1984), the meetings studied here fall into the category of decision-making tasks, in which participants must evaluate alternatives and converge on a joint choice.

Experimental studies of group decision making have often relied on structured ranking tasks, in which participants must agree on the relative importance of a fixed set of alternatives. A well-known family of such tasks are survival ranking exercises, where groups are asked to prioritise items that might help them survive in a hypothetical scenario. Participants must discuss the available items and collectively produce an ordered ranking based on their perceived utility. The meetings in the present corpus follow this experimental tradition and employ three such survival ranking tasks (Hall and Watson, 1970; Hare, 1952; Lafferty and Pond, 1974).

Several approaches to dialogue and discourse modelling have explored how the evolving state of an interaction can be represented incrementally as the dialogue unfolds. In spoken dialogue systems, incremental interpretation has long been used to maintain partial semantic representations that are updated as new input becomes available (Skantze and Edlund, 2004). Such approaches emphasise robustness and the ability to operate with no knowledge of future context and even without full knowledge of the past. At the same time, dialogue annotation frameworks have been developed to provide standardised ways of describing communicative acts in interaction, most notably the ISO framework for dialogue act annotation (Bunt et al., 2020). The present work aligns with these traditions while focusing specifically on modelling the emergence of consensus in a constrained decision-making task.

3. Annotation scheme

The annotation scheme design was guided by two methodological commitments. First, annotation is performed incrementally and without lookahead. Annotators record only the information that is available at the current point in the interaction, without using later context to reinterpret earlier contributions. The motivation is that participants themselves do not have access to future context, and our aim is to model the state of the decision process as it could plausibly appear to the participants at any given moment. The annotation procedure must operate under the same constraint.

Second, the scheme is deliberately selective. Only those aspects of the interaction that are directly relevant to the ranking task are annotated. The aim is not to reinvent a comprehensive representation of dialogue structure, but to produce a representation that remains closely aligned with the decision process that unfolds during the meeting.

The annotations are currently produced in the ELAN multimodal annotation environment. Separate tiers are used to represent different aspects of the interaction. TurnUnits provide the temporal segmentation of the speech signal, while additional tiers capture the task-oriented structure of the discussion. Of these, the proposition track, which references task entities (items and rank positions), is of particular relevance to this paper.

In the present scheme, any mention of a task entity (an ITEM, a RANK, or both) is annotated as a proposition. Each proposition represents the current association between ITEM and RANK introduced by the contribution, and a newly annotated proposition may modify or override the previously current proposition. ITEM and RANK values may each contain one or more values, reflecting cases where the speaker explicitly proposes multiple alternatives (e.g. “map should be first or second”). Annotators are trained to apply this principle consistently when identifying propositions in the dialogue. They record the surface form of the contribution according to the scheme, while resolution of references and inheritance of previously mentioned values are handled automatically during later processing. Interactional categories such as questions or acknowledgements are therefore not encoded directly; instead they can often be inferred from the form of the proposition.

Two special symbols distinguish underspecification from explicit unknowns. The underscore (_) indicates that the corresponding entity is not men-

tioned in the contribution. The question mark (?) indicates that the value is explicitly unknown or requested. Annotators are trained to apply this distinction when interpreting contributions: "_" reflects the absence of a mention, whereas "?" reflects an explicit request for, or removal of, a value. For example, prop(itemA, _) marks a contribution mentioning itemA without specifying a rank, whereas prop(itemA, ?) marks a contribution asking for the rank placement of itemA.

Where multiple alternatives are expressed, the corresponding argument may contain a set of values. For example, prop(map,[1,2]) represents a contribution proposing that the map may occupy either rank 1 or rank 2. Similarly, a contribution mentioning several items in relation to a rank may be represented as prop([itemA,itemB],3). Such sets are interpreted as explicit alternatives rather than as separate propositions.

Acknowledgements of propositions are encoded as the fully underspecified form prop(_, _), which represents an additional mention of the currently active proposition without introducing new task information. Negative acknowledgements function similarly to questions in that they remove the currently active association; they are therefore annotated using explicit unknowns (?). This treatment works well for the majority of cases observed in the data, but the precise handling of negation is still under investigation and the present representation may not capture all possible interactional patterns. Misunderstandings are not annotated explicitly. Each contribution is recorded according to its surface interpretation at the moment it occurs; discrepancies are resolved through later propositions in the dialogue.

4. Discourse model

The discourse model is designed as a minimal operational interpretation of the proposition annotation track. It treats the annotated propositions as observable constraints on the evolving task state of the meeting and derives the current decision state directly from these contributions as they occur. The model operates strictly incrementally and without lookahead: each new proposition updates the state of the model using only the information available at that point in the interaction. This mirrors the constraints imposed on the annotation process and ensures that the representation remains closely aligned with the unfolding discussion.

The model consists of three components. The proposition stack records the chronological sequence of annotated propositions and thus provides a complete incremental history of task-relevant contributions in the meeting. From this sequence, the model derives an Issue Under Discussion (IUD), which represents the currently active proposition after resolving underspecified references. Finally, a consensus ranking structure aggregates the successive IUD states into a running estimate of the emerging group ranking. Each component performs a minimal transformation of the previous one, preserving the incremental structure of the interaction while progressively deriving the task state.

Although the three components operate on the same propositions, they serve different roles in the model. The proposition stack functions as an event log of task-relevant contributions, the IUD represents the current conversational focus, and the consensus ranking tracks the evolving decision outcome.

4.1. The proposition stack

The proposition stack is the simplest component of the model. It is a chronological record of all propositions produced by the annotation procedure, stored in the order in which they occur in the interaction. Each entry corresponds directly to an annotated proposition and may optionally carry additional metadata such as speaker identity or temporal position in the dialogue.

The stack itself performs no interpretation: it simply records the observable task-related contributions as they appear. As a result, it functions as an event log of the decision process and provides the complete incremental history from which the remaining components of the model derive their state.

Because it preserves the complete incremental history of the discussion, the stack can also be replayed to reconstruct the evolving state of the model under different interpretations or analysis settings.

4.2. Issue Under Discussion tracker

The Issue Under Discussion (IUD) tracker represents the currently active proposition in the dialogue. While the proposition stack records the history of contributions, the IUD tracker maintains the present conversational focus derived from that history.

The term Issue Under Discussion is adopted from Larsson's dialogue management framework (Larsson, 2002), where it refers to the currently active conversational issue. In the present model, the IUD tracker operationalises this idea by maintaining a single proposition of the same form as the annotated propositions.

Each new entry in the proposition stack updates the IUD incrementally, resolving underspecified references by inheriting values from the previous state

and allowing explicit unknown values to clear previously established associations.

The IUD tracker is updated whenever a new proposition is pushed onto the proposition stack. Resolution follows two simple rules.

Underspecified values, marked with "_", do not overwrite the corresponding field in the current IUD and therefore inherit the previously active value. Explicit unknown values, marked with "?", clear the corresponding field and remove any previously established association.

All other values overwrite the current value in the corresponding field. Through these operations the IUD tracker maintains a continuously updated representation of the item–rank association currently under discussion.

The apparent simplicity of this mechanism is deliberate: in task-oriented dialogue within constrained domains, such lightweight resolution has proven both robust and efficient compared to more complex approaches.

4.3. Consensus ranking

The consensus ranking represents the model's current best estimate of the group's decision state. While the proposition stack records the history of contributions and the IUD tracker represents the current conversational focus, the consensus structure aggregates these successive IUD states into a running representation of the emerging ranking.

The structure maintains two parallel representations of the ranking: one for unambiguous placements and one for ambiguous placements. The unambiguous representation contains rank–item associations in which a single item is assigned to a rank. The ambiguous representation captures cases in which the discussion leaves multiple alternatives open, either because an item may occupy several possible ranks or because several items are considered for the same rank. Together, these structures allow the model to represent both resolved and unresolved parts of the ranking at any given point in the meeting.

The consensus ranking is updated incrementally whenever the IUD tracker produces a new proposition. Updates follow a small set of consistency constraints. First, the model follows a last-mention-wins policy: a new placement overrides earlier placements that involve the same item or rank. Second, ranks are exclusive: assigning an item to a rank removes any previous assignments for that rank. Third, items are unique: when an item receives a new placement, any previous placements involving that item are removed.

Ambiguous propositions introduce alternative placements into the ambiguous representation. For example, a proposition such as prop(water,[1,2]) indicates that the item may occupy either of the two ranks, while prop([map,gun],2) indicates that either item may occupy the same rank. Such alternatives remain represented until subsequent propositions resolve the ambiguity or introduce new constraints.

Through these incremental updates, the consensus structure provides a compact view of how agreement gradually develops during the meeting, while still preserving the uncertainty that characterises intermediate stages of the decision process.

5. Examples and reference implementation

The discourse model described above has been implemented in a web-based reference system using web components. The current implementation fully supports the incremental processing of sequences of propositions and the equally incremental computation of the IUD and consensus ranking structures from these sequences.

The system also provides visualisation of these structures during analysis. Replay functionality is currently manual, but is under development and will support both step-wise replay based solely on the annotation sequence and time-aligned replay using the original audio recordings and time aligned annotations.

The following example illustrates how the model represents alternative proposals and their resolution during the discussion. In the dialogue, participants first establish tentative placements for two items. A subsequent contribution introduces an explicit alternative for the second rank by proposing two possible items for that position. The final contribution resolves this ambiguity by selecting one of the alternatives. Because the rank is not mentioned in the final utterance, the annotation uses the underspecified value '_', allowing the IUD to inherit the previously active rank. Table 1 shows the dialogue and the corresponding proposition annotations.

	Utterance	Annotation
A	The map should be first.	`prop( 1,map)`
B	The knife should be second.	`prop( 2,knife)`
C	Knife or compass for second place.	`prop( 2, [knife, compass])`
A	Compass.	`prop(_, compass)`

Table 1: Example dialogue fragment and corresponding proposition annotations.

The sequence of propositions produced by this exchange is processed incrementally by the discourse model. The resulting proposition stack, cur-

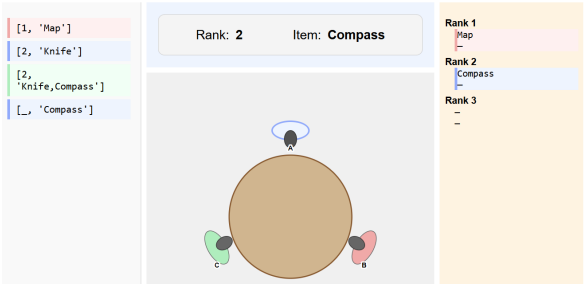

Figure 1: State of the discourse model after processing the dialogue shown in Table 1. The interface displays the proposition stack (left pane), the current IUD (centre top pane), and the evolving consensus ranking (right pane).

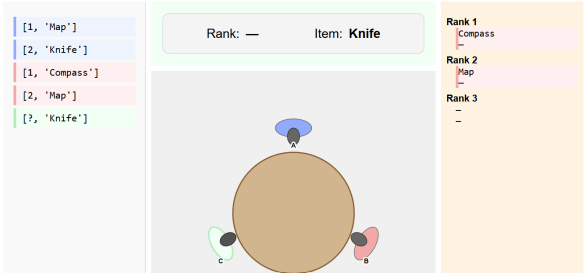

Figure 2: State of the discourse model after processing the dialogue shown in Table 2. The interface displays the proposition stack (left pane), the current IUD (centre top pane), and the evolving consensus ranking (right pane).

rent Issue Under Discussion (IUD), and consensus ranking are shown in Figure 1.

The following example illustrates how the model handles revisions of previously proposed rankings and the role of explicit unknown values. In this exchange, the participants first establish tentative placements for two items, after which a new proposal revises the ordering. The final contribution reopens the question of the placement of the knife. In this case the annotation uses the explicit unknown marker '¿, which clears the previously associated rank in the IUD. If the utterance were instead annotated using the underspecified value '_', the rank would be inherited from the current IUD state and the contribution would be interpreted as a renewed proposal for the previously active rank. The '¿ marker therefore allows the model to represent the contribution as reopening the placement decision rather than reinforcing or editing an earlier proposal. able 2 shows the dialogue and the corresponding proposition annotations.

	Utterance	Annotation
A	The map should go first and the knife second.	`prop( 1, map)` `prop( 2, knife)`
B	No compass first, map second.	`prop( 1, compass)` `prop( 2, map)`
C	But what about the knife?	`prop( ?, knife)`

Table 2: Example dialogue illustrating revision of previous proposals and the use of the explicit unknown marker '¿.

Processing the propositions derived from the dialogue in Table 2 yields the discourse model state in Figure 2, including the proposition stack, the current IUD, and the evolving consensus ranking.

6. Discussion

The modelling approach presented here is fully incremental and strictly constrained by the information available at each point in the interaction. This design reflects the methodological goal of studying consensus formation as an emergent process, rather than reconstructing it retrospectively using knowledge of the final outcome. By avoiding lookahead and global reinterpretation, the model approximates the informational conditions under which the participants themselves operate during the meeting.

The model is also narrow in scope. Rather than attempting to represent the full structure of dialogue, the annotation scheme and discourse model focus specifically on the ranking task that structures the meetings in the corpus. This restriction makes it possible to obtain a representation that is both robust and computationally tractable, while still capturing the aspects of the interaction that are most relevant to the collaborative decision process.

In this sense, the work draws on experience from other task-oriented dialogue domains, where lightweight incremental representations have proven effective for modelling constrained conversational tasks. Applying similar principles to collaborative human–human meetings provides a useful bridge between dialogue system research and the empirical study of group decision-making.

The resulting representation provides a structured account of the evolving decision process that can serve as a foundation for further analysis. Because the model tracks the state of the ranking incrementally over time, it can be aligned with other modalities present in the corpus, including acoustic-prosodic cues, gesture, and interactional dynamics across different meeting settings such as remote, co-located, or hybrid interaction.

Finally, the replayable structure of the proposition stack and discourse model enables dynamic visualisation of the decision process. When analysing

recordings of real collaborative interaction, such visualisations provide valuable support for exploration and interpretation, allowing researchers to inspect how consensus develops over time. The reference implementation described above demonstrates how the incremental state of the model can be inspected and replayed, making it possible to evaluate modelling decisions and incorporate human expertise directly into the analytical process.

7. Conclusion and future work

This paper presented an annotation scheme and discourse model for studying consensus formation in collaborative meetings. The approach represents the evolving decision state of the meeting incrementally through a sequence of annotated propositions, an Issue Under Discussion tracker, and a consensus ranking structure. Together these components provide a minimal operational account of how group decisions emerge over the course of interaction.

Future work will focus on extending the reference implementation with full replay functionality and on applying the model to larger portions of the MEET corpus. The incremental representation also lays the ground for integrating additional modalities such as prosodic, gestural, and interactional signals in order to study how different communicative resources contribute to the formation of consensus in meetings.

Acknowledgements

The results of this work will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2023-00161_VR).

8. References

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 *Proc. of LREC'20*, pages 549–558, Marseille, France. European Language Resources Association.

J H Davis, P R Laughlin, and S S Komorita. 1976. The social psychology of small groups: Cooperative and mixed-motive interaction. *Annual Review of Psychology*, 27(1):501–541.

Erving Goffman. 1961. *Encounters: Two Studies in the Sociology of Interaction*. Encounters: Two Studies in the Sociology of Interaction. Bobbs-Merrill, Oxford, England.

Ernest James (Jay) Hall and W. H. Watson. 1970. The effects of a normative intervention on group decision-making performance. *Human Relations*, 23(4):299–317.

A. Paul Hare. 1952. A study of interaction and consensus in different sized groups. *American Sociological Review*, 17(3):261–267.

J. Clayton Lafferty and Alonzo William Pond. 1974. *The Desert Survival Situation: A Group Decision Making Experience for Examining and Increasing Individual and Team Effectiveness*. Human Synergistics, Plymoth, Michigan, US.

Staffan Larsson. 2002. *Issue-Based Dialogue Management*. Ph.D. thesis, Department of Linguistics, Göteborg University.

Joseph Edward McGrath. 1984. *Groups: Interaction and Performance*. Prentice-Hall.

Gabriel Skantze and Jens Edlund. 2004. Robust interpretation in the Higgins spoken dialogue system. In *Proc. of ROBUST 2004*, page 4, Norwich, UK. ISCA.

GeoAffect: A Multi-Layer Annotation Schema and Few-Shot LLM Evaluation for Geoaffective Analysis of Literary Texts

Fotini Koidaki, Stergios Chatzikyriakidis
Department of Philology, University of Crete
coidacis@gmail.com, stergios.chatzikyriakidis@uoc.gr
{Fotini Koidaki,Stergios Chatzikyriakidis}@uoc.gr

Abstract

GeoAffect is an annotation framework that has been especially developed to capture how places are emotionally framed in literary narrative. The project focuses on nineteenth-century Greek prose fiction and brings together named entity recognition with an affect schema that distinguishes experiential, appraisal, and identity-oriented relations to place. The annotation design linked entities, emotion spans, and rhetorical devices, allowing us to model not only sentiment but also forms of belonging, alienation, and longing. To test the schema, we created a manually annotated gold dataset of approximately 360 sentences and evaluated thirteen Large Language Models in a few-shot setting for both entity recognition and affect classification. The results indicate that, with carefully designed prompts and selection strategies, LLMs can support structured geoaffective annotation even in low-resource historical language contexts.

Keywords: geoaffective annotation, semantic annotation schema, named entity recognition, emotion classification, large language models, few-shot learning, computational literary analysis

1. Introduction

Computational literary studies have shown growing interest in both emotion and place over the last decade — but the relationship between the two remains underdeveloped, particularly for non-English, low-resource historical corpora. The corpus of 19th-century Greek fiction from the Ionian Islands is a case in point. This corpus has only recently come to light since modern Greek literary studies have for a long time focused on Ionian poetry.[1] The Ionian authors came from a region shaped by centuries of Venetian rule, where Greek, Italian, and European cultural presences have long coexisted. A foundational question for this literary tradition remains open: do these texts share the regional identity of the Ionian Islands, the so-called *Heptanese*, or merely a shared provenance? Place and identity are closely bound (Tuan, 1977; Hall, 1996), which means that how place is emotionally framed in these texts is not a marginal question. GeoAffect was built to investigate exactly this — though the full corpus analysis lies ahead. This paper presents the GeoAffect framework and its evaluation against a manually annotated gold standard.

Existing computational approaches to emotion in literary texts rely on psychological taxonomies (Ekman, Plutchik) or dimensional models such as valence-arousal, developed to capture individual affective states in response to events or persons (Kim and Klinger, 2019). Even the most closely related work - studies that relate toponyms with sentiment in historical fiction (Heuser et al., 2016)- operates with binary or coarse-grained polarity. These approaches fail to capture the identity dimension of place-affect relations: the sense of belonging, absence of belonging, and longing to belong that characterizes how people relate to geographical space. Crucially, in literature the affective charge of a place reference is rarely expressed directly. It is constructed through narrative voice, rhetorical figuration, and implicit cultural associations. Existing tools — whether categorical emotion models or lexical resources — are not designed to capture this. To our knowledge, no annotation schema or evaluation framework exists for geoaffective classification in literary texts — let alone for a low-resource language such as 19th-century Greek. GeoAffect is our attempt to address this. We propose a multi-layer annotation schema grounded in a three-level affect taxonomy, which we evaluate against a manually annotated gold standard through systematic few-shot prompting of thirteen Large Language Models.

[1] On the relative neglect of Ionian prose fiction in favour of poetry, see Tziovas (2017), p. 83.

2. Related Work

Heuser, Moretti, and Steiner (Heuser et al., 2016) were the first to systematically map emotional associations with place names in literary texts at scale using 5,000 British novels (1700–1900), and combining named entity recognition with crowdsourced sentiment annotation. This work remains a landmark. Yet it operates within a binary polarity framework, not by choice, but because attempts to capture a broader emotional spectrum failed to produce sufficient annotator agreement.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 68–76

This line of inquiry has been extended in recent work to incorporate other national literatures. Grisot and Herrmann (2023), for example, extend this line to German-Swiss fiction (1840–1940) through a dictionary-based approach, identifying rural and urban spatial terms via curated word lists and computing sentiment through lexicon-based scoring in fixed textual windows. Karlińska et al. (2022) examine Polish fiction (1864–1939) by tracking the emotional valence of broad spatial concepts such as "city" and "country." Together, these studies demonstrate how spatial references can be linked to affect at scale, they rely on lexicon-driven polarity measures. Our work builds on this direction by introducing a span-based annotation framework that captures the relational and identity-oriented dimensions of place in narrative.

Computational emotion research has followed mainly three approaches: categorical taxonomies such as Ekman's (1992) or Plutchik's (2001), which label discrete emotions; dimensional models that describe affect in terms of valence and arousal; and appraisal-based approaches, which explain emotion as a result of how –individuals evaluate events. These approaches are well suited to modeling clearly expressed emotional states and evaluations, and we draw on this insight in structuring our affect layer. However, they are less equipped to account for how emotion becomes attached to place as a marker of belonging or identity. In literary prose, such relations often emerge indirectly, through narrative voice or figurative language rather than explicit emotion terms.

Troiano et al. (2023) introduced an annotation schema that was grounded on appraisal theories for emotion analysis and proved that appraisals can be reliably inferred from text and improve emotion classification. However, the focus remains on event-based evaluation and does not address how places function as objects of attachment or identity within narrative. Bozia et al. (2024), studying ancient Greek and Latin corpora with a focus on identity and belonging, also draw attention to the role of place. Their analysis, however, works largely with polarity distinctions, as no specific schema is introduced for modelling place-related affect.

Large language models have recently been tested as annotation tools, with few-shot prompting approaching human performance on some classification tasks (Ziems et al., 2024). Performance becomes less reliable, however, when the task requires fine-grained theoretical distinctions: Imamovic et al. (2024) find that ChatGPT achieves high precision on Appraisal Theory labels but poor recall, with systematic errors at the level of fine-grained distinctions — a result that points to the sensitivity of theory-driven annotation to prompt design. Historical NER adds a further layer of difficulty: diachronic language variation, orthographic instability, and scarce annotated data compound extraction errors in ways that standard benchmarks do not capture (Ehrmann et al., 2023) — and 19th-century Greek prose sits squarely within these conditions. No existing work addresses all three challenges together.

3. The GeoAffect Schema

GeoAffect is a multi-layer annotation schema designed to capture the affective encoding of place references in 19th-century literary narrative. Rather than applying generic sentiment analysis to spatial entities, the schema operationalizes place-related affect as a structured, hierarchical phenomenon grounded in distinct theoretical traditions. It comprises four interdependent layers visible in Figure 1: (i) named entity recognition, (ii) emotion classification, (iii) rhetorical device annotation, and (iv) relational linking.

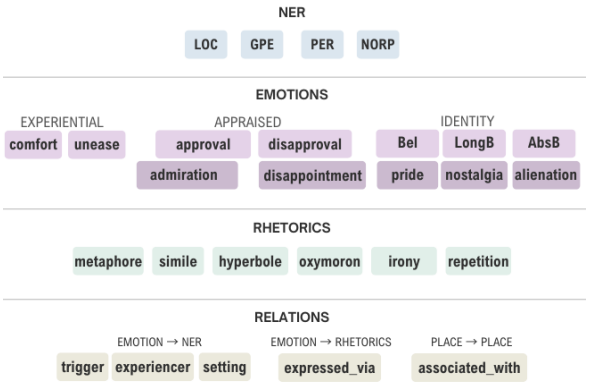

Figure 1: The four layers of the GeoAffect schema.

3.1. Named Entity Layer

The NER layer targets four entity types: locations (LOC), geopolitical entities (GPE), persons (PER), and nationalities, or religious groups (NORP). The LOC/GPE distinction ensures annotation precision for place references, separating physical or geographical spaces from entities defined by political and administrative boundaries. Entity types are annotated on the basis of their referent rather than their surface form, using adjectival or toponymic modifiers as anchors for entity identification when no proper name is present. An expression such as "the Greek state" (*elliniko kratos*, lit. "the Greek-ADJ state") is tagged GPE because it denotes a geopolitical entity, while "the lake of Kastoria" (*limni Kastorias*) is tagged LOC regardless of whether a proper name is present. The same principle applies to NORP: a collective reference such as "Italian nation" (*italikos laos*, lit. "Italian-ADJ nation") is tagged

based on what it denotes, not on whether a proper ethnonym is used.

NORP entities are treated as full-fledged targets of emotion annotation alongside place references: ethnic, national, and religious collectivities are not merely the backdrop of identity expression but one of its primary sites - the groups through which subjects define who they are, who they are not, and where they belong. This is particularly salient in the case of the Ionian (Heptanese) corpus, where collective identity is articulated at the intersection of Greek, Venetian, and Italian presences, and where belonging to a people and belonging to a place are frequently co-constructed in the same utterance.[2]

3.2. Emotion Layer

The emotion layer organizes place-related affect into three hierarchical levels. The guiding principle is that the relationship between a subject and a place is constituted experientially, evaluated culturally, and transformed into an identity narrative - each level presupposing the one below it.

The first level, ExperientialAffect, captures the immediate, pre-reflective affective tone of a place encounter: `comfort` and `unease`. Grounded in Tuan's phenomenology of place (Tuan, 1977), this level registers the bodily and sensory attunement through which a narrative subject inhabits a space before evaluation or reflection intervenes.

The second level, AppraisedAffect, captures the evaluative stance toward place as a sociopolitical and cultural space: `approval` and `disapproval`, with more intense children `admiration` and `disappointment`. Grounded in appraisal theory (Lazarus, 1991), this level treats emotion as cognitive evaluation - the judgment of a place as worthy, degraded, or disappointing within a social and ideological field. The parent/child structure reflects the intensity gradient central to appraisal models: `approval` intensifies into `admiration` when the stance takes on an aesthetic or moral dimension; `disapproval` deepens into `disappointment` when a gap between expectation and reality is affectively marked.

The third level, IdentityAffect, captures the relational orientation of a narrative subject toward place as a constitutive dimension of selfhood: belonging (`Bel`) with child `pride`, absence of belonging (`AbsB`) with child `alienation`, and longing for belonging (`LongB`) with child `nostalgia`. This level draws on Hall's account of identity as constructed through difference and exclusion (Hall, 1996) - belonging is only legible against the possibility of its absence - and on Ricoeur's theory of narrative identity, in which the self is constituted through the act of narrating experience in space and time (Ricoeur, 1992). A subject does not merely feel at ease in or approve of a place; they may recognise themselves in it, find themselves excluded from it, or orient themselves toward it as a lost or longed-for anchor of selfhood. In this sense, the three levels form a progression: a place is first experienced, then evaluated, and eventually becomes part of how the self is narrated and understood.

- **Experiential:**
  `comfort` (+)
  `unease` (–)
- **Appraised:**
  `approval` → `admiration` (+)
  `disapproval` →`disappointment` (–)
- **Identity:**
  `Bel` → `pride` (+)
  `AbsB` → `alienation` (–)
  `LongB` → `nostalgia` (+/–)

3.3. Rhetorical Layer

The rhetorical layer is a dependent sublayer: it is activated only when an emotion span has already been identified, and annotates the expressive mechanism through which affect is conveyed. The six devices recognised - `irony`, `simile`, `metaphor`, `hyperbole`, `repetition`, and `oxymoron` - were selected on the basis of their attested frequency in the corpus and their capacity to mediate place-related affect in nineteenth-century literary prose, where emotional meaning is frequently constructed through figuration rather than direct lexical expression. The layer is designed to support analysis of the relationship between rhetorical choice and affective register, without presupposing a fixed inventory of devices.

3.4. Relations Layer

The relations layer maps the structural dependencies between annotated spans. Three emotion-to-entity relations are defined: `experiencer` identifies the entity that undergoes the emotion; `trigger` identifies the entity that causally provokes it; and `setting` marks entities that function as background context - they spatially or socially locate the emotional expression without participating in it as agent or recipient. Any entity type (`LOC`, `GPE`, `PER`, `NORP`) may occupy any of these three roles. An `expressed_via` relation links emotion spans to rhetorical device spans. An `associated_with`

[2] Pronominal and anaphoric references to place are not currently tracked. Coreference resolution tools for 19th-century Greek are not available, which makes systematic annotation of such references impractical at this stage.

relation links entity spatial spans to each other, capturing co-reference and geographical association between named entities.

3.5. Annotation example

The following sentence illustrates how the schema captures identity-oriented place affect in practice. In this example, drawn from a narrative about Venetian nobles who had settled in Crete, two place references appear in the same sentence carrying distinct identity labels: BEL marks an active, present-tense relation to place, while NOSTALGIA marks one that has been displaced into memory. The relations layer, in this case, captures the entities that act as triggers for each emotion.

> *"Crete having become the common homeland for most of the nobles, Italy was consigned to the land of genealogical memories."*[3]

NER:
`Crete`: GPE | `Italy`: GPE

EMOTIONS:
`common homeland`→BEL | `land of genealogical memories`→NOSTALGIA

RELATIONS:
BEL $\overset{\text{trigger}}{\rightarrow}$ `Crete` | NOSTALGIA $\overset{\text{trigger}}{\rightarrow}$ `Italy`

To assign relations, we asked three questions for each annotated emotion span: what triggers it, who experiences it, and where it is located. When the answer was a named entity already marked in the NER layer, the corresponding relation was encoded — which is why in our representation relations are directed from the emotion span outward. The `associated_with` relation was reserved for cases where two spans refer to different names for the same place.

4. Pipeline & Evaluation Methodology

4.1. Gold Standard Dataset

The gold standard dataset comprises 363 manually annotated sentences drawn from a corpus of 19th-century Ionian prose fiction spanning approximately 80 texts and 11,695 pages. Very few of these were available in digital or scanned form. Source texts were digitised using a Transkribus model trained specifically on the corpus (19th-century Greek 8.0) — rather than an off-the-shelf OCR solution — in order to preserve the linguistic particularity and typographic conventions of each text, which will be relevant for future analyses.

The digitized texts were subsequently normalised from polytonic to monotonic Greek to reduce OCR error rates. After sentence segmentation, sentences were filtered for the presence of spatial entities either `LOC` or `GPE` references - using a combination of a domain-adapted spaCy model (`el_core_news_lg`) fine-tuned for this purpose,[4] and manual inspection across randomly selected texts from the Ionian corpus. The resulting sentences were stored in JSON format without any annotation and subsequently annotated in Label Studio using the full GeoAffect schema.

Annotation was carried out in two sequential phases. [5] In the first phase, all four named entity types were annotated. In the second phase, emotion spans, rhetorical devices, and relational links were annotated. The six rhetorical devices included in the schema were not selected a priori from a theoretical inventory but emerged inductively from the corpus during this phase, reflecting the devices actually attested in the material. Across the 363 sentences, annotation yielded 1,410 NER spans (GPE: 779, NORP: 277, PER: 257, LOC: 97), 836 emotion spans, and 728 relational links. Emotion labels are unevenly distributed: disapproval and unease account for over 40% of all emotion annotations, while identity-tier labels such as alienation (7) and nostalgia (11) are rare — a distribution that reflects the corpus rather than annotation bias, and that directly shaped our few-shot pool design.

Regarding span overlap, a non-overlap policy was applied within each layer: NER spans do not overlap with other NER spans, and emotion spans do not overlap with other emotion spans. NER and emotion spans likewise do not overlap with each other, as their function is relational rather than co-referential: the two layers are designed to enter into structured relations (`experiencer`, `trigger`, `setting`) rather than to describe the same textual unit simultaneously. Rare exceptions occur in cases of literary creative language where the narrative mode renders a place reference itself an affective expression. In the sentence "*It was finally printed in the Bulletin of the Medical Academy of Paris… of Paris!*", the repeated toponym is anno-

[3] Translated from a gold standard sentence. The original is written in formal 19th-century Greek prose and is considerably denser than the translation suggests.

[4] The domain-adapted `el_core_news_lg` model was fine-tuned on 150 annotated samples. It performed adequately on simple toponyms but did not generalise to compound spans, shared-head constructions, or the `GPE`/`LOC` distinction. For this reason it was used only for sentence filtering, not for entity annotation. `NORP` entities — which in this corpus include adjectival ethnic references — were excluded from its scope entirely.

[5] Annotation was carried out by a single annotator with expertise in 19th-century Greek literature and computational linguistics. Sentences where label assignment was uncertain were excluded from the test sets, which were drawn exclusively from cases of high annotation confidence.

tated as a GPE span; the repetition of this toponym functions simultaneously as a repetition span and an admiration span, since the rhetorical device is itself the vehicle of the affective expression.

4.2. Test Set Construction

Few-shot evaluation was conducted using manually curated test sets as part of a prompt optimization process aimed at producing frozen prompt configurations for subsequent large-scale corpus analysis. Random sampling from a dataset with rare labels and compound annotation structures would not provide a principled basis for evaluating the model's understanding of schema-specific phenomena.

For the NER task, test samples were selected to include the most challenging cases attested in the corpus: compound entity spans, entities referenced through adjectival or toponymic modifiers without a proper name (e.g. *Italian nation*), and elliptical coordination where a shared head noun is not repeated across conjoined entities (e.g. *the people of Greece and [of] Turkey*). This design tests whether models have internalized the referent-based annotation principle described in Section 3.1 rather than defaulting to surface-form heuristics.

For the geoaffective emotion task, priority was given to sentences with unambiguous emotional triggers and high annotation confidence - that is, sentences in which the annotator had no doubt about the label assignment. Since geoaffective analysis of literary texts is an understudied task with no prior benchmark, this selection criterion was adopted deliberately: establishing model performance under favourable conditions provides a meaningful baseline before evaluating harder cases.

Two test configurations were constructed. T1 comprises sentences with multiple co-occurring emotion annotations (two to six per sentence), selected to cover the full range of emotion labels in the schema, with preference given to sentences in which emotions are associated with place references rather than persons. T2 comprises sentences with a single emotion annotation, designed to isolate model behaviour in the absence of competing labels and contextual noise. In both configurations, all selected sentences contain at least one LOC, GPE, or even NORP entity to which the emotional expression is grounded.

4.3. Few-Shot Evaluation Pipeline

Models were evaluated via the OpenRouter API across thirteen[6] large language models spanning different model families and parameter scales, from lightweight open-weight models to frontier systems (see Table 1).[7]

Model	Provider	Scale
Claude 3.5 Sonnet	Anthropic	frontier
Claude Sonnet 4	Anthropic	frontier
Claude 3.5 Haiku	Anthropic	lightweight
GPT-4o	OpenAI	frontier
GPT-4o-mini	OpenAI	lightweight
Gemini 2.0 Flash	Google	lightweight
Gemma-2/3 27B[8]	Google	mid-scale
DeepSeek-V3	DeepSeek	frontier
DeepSeek-R1	DeepSeek	frontier
Llama 3.1 8B	Meta	lightweight
Llama 3.3 70B	Meta	mid-scale
Mixtral 8x7B	Mistral	mid-scale
Qwen-2.5 72B	Alibaba	mid-scale

Table 1: Models included in the evaluation.

All prompts were written in Modern Greek to avoid code-switching overhead introduced by translating 19th-century Greek text into English for processing. Temperature was set to 0.01 to minimise output variability across runs. Each model received the same system prompt loaded from an external file, followed by a fixed set of few-shot examples and the target sentence; no fine-tuning or parameter updates were performed. Few-shot examples for NER comprised 17 samples selected to cover the full typology of challenging cases described in Section 4.2. Few-shot examples for the affect task comprised 12 samples per configuration (T1 and T2), selected to provide clear geoaffective triggers across the emotion label hierarchy.[9]

4.4. Evaluation Metrics

The two annotation tasks were evaluated using different metrics, reflecting the different nature of their output spans.

For the NER task, the primary metrics are precision, recall, and F1 computed over exact-match entity spans, where both the text and the label must match. Exact match was chosen deliberately as a strict criterion: given that the NER test set was designed to stress-test boundary cases, a relaxed matching criterion would obscure precisely the er-

[6] Gemma 2 was used for NER few shot prompting, while Gemma 3 was used for SA.

[7] KriKri, a Greek-specific large language model, was initially considered for inclusion in the evaluation. However, response latency under long prompts with multiple few-shot examples rendered it impractical for systematic evaluation at this stage; it remains a candidate for future work.

[8] Gemma 2 27B was used for NER prompting and evaluation, and Gemma 3 27B was used for SA prompting

[9] Significance testing was not conducted given the small size of the test sets, which were designed for targeted evaluation rather than statistical inference.

rors that are most informative about model behavior.

For the geoaffective emotion task, three complementary metrics are reported. The primary metric is subset/superset F1 (SubF1): a predicted span is counted as a match if its text is a substring or superstring of the gold span with the same label. This relaxed criterion acknowledges that emotion span boundaries in literary prose are inherently less determinate than named entity boundaries, and that partial overlap with correct label assignment constitutes meaningful performance. Exact F1 is reported as a stricter reference point, and token-overlap F1 is reported to capture the degree of lexical overlap between predicted and gold spans regardless of boundary alignment. A hallucination rate - the proportion of predicted spans with no gold counterpart - is reported as a secondary metric across all models to track false positive generation, which is a known tendency of instruction-tuned models on open-ended span extraction tasks.

5. Results

Prompt optimisation for the NER task proceeded across eleven iterative versions, grouped here into three phases visible in Figure 2. From these versions, three major optimisation phases can be identified: schema rule elaboration (v0–v3.1), few-shot refinement (v4–v4.2), and structural reordering (v5–v5.1).

At baseline (v0), F1 scores ranged from 39.3% (Mixtral 8x7B) to 79.6% (DeepSeek-R1), with most frontier models clustering between 66% and 80%. The introduction of entity category definitions and general rules (v1–v2) produced modest but consistent gains across models, with Claude Sonnet 4 rising from 73.2% to 84.7%. The addition of span boundary rules, compound entity handling, and the shared head rule (v3) brought further improvement for most models, but the effect of the step-by-step recognition instructions added at this stage was non-monotonic: the verbose formulation (v3-lg) underperformed the minimal formulation (v3-mini) for ten out of fourteen models, with differences ranging from +1.0 to +16.5 percentage points in favour of the minimal style. DeepSeek-V3 and Qwen-2.5-72B showed the largest sensitivity to this choice (+16.5 and +15.2 respectively), while Claude 3.5 Sonnet was largely unaffected (–0.7). This finding suggests that procedural verbosity introduces instruction-following overhead that smaller or less instruction-tuned models cannot absorb. The most significant gains came from targeted few-shot replacement in v4–v4.2. Replacing four few-shot examples with samples specifically covering shared-head and compound entity cases raised Claude Sonnet 4 from 88.3% to 89.9%; adding one further compound example brought it to 94.5%; reordering the prompt components - placing span rules before general rules and the procedure last - produced the largest single-step gain, reaching 97.2%. This final configuration remained stable through v5 and v5.1.

Mixtral 8x7B showed a distinct pattern. Its score fell from 51.4% (v3-mini) to 29.6% at v4, , then to 26.7% and 19.6% in v4.1 and v4.2, before rising again to 47.1% at v5. No similar drop was observed in the other models. The change coincides with the few-shot replacement introduced at v4 suggesting that Mixtral is particularly sensitive to the specific few-shot examples used and that the replacement set was incompatible with its instruction-following behaviour.

5.1. NER Final Results

Table 2 reports precision, recall, and F1 for all models at v5.1. Claude Sonnet 4 achieves the highest F1 (97.2%, P=96.4, R=98.1), followed by DeepSeek-R1 (93.6%, P=92.7, R=94.4). Claude 3.5 Sonnet and GPT-4o are tied at 89.9%. Gemini 2.0 Flash reaches 87.0% at a cost of €0.005 per run - the most cost-efficient result in the upper tier. DeepSeek-V3 achieves 84.4% at €0.015, making it the strongest value proposition among mid-scale models. Mixtral 8x7B finishes last at 32.7%, confirming that its v4 collapse was not fully recovered.

Notably, Claude Sonnet 4 ranked fourth at v0 (73.2%), behind DeepSeek-R1 (79.6%), Claude 3.5 Haiku (77.1%), and Claude 3.5 Sonnet (75.9%). Its emergence as the top performer reflects a higher capacity to exploit structured prompt information - a property that was not predictable from baseline performance alone.

Model	F1	€	s
Claude Sonnet 4	**.972**	0.227	26.3
DeepSeek-R1	.936	0.043	270.8
Claude 3.5 Sonnet	.899	0.227	42.3
GPT-4o	.899	0.124	17.2
Gemini 2.0 Flash	.870	0.005	12.2
DeepSeek-V3	.844	0.015	53.6
Claude 3.5 Haiku	.833	0.063	41.3
Qwen-2.5-72B	.756	0.029	50.8
Llama 3.3-70B	.733	0.019	42.8
Llama 3.1-70B	.685	0.017	38.7
Llama 3.1-8B	.614	0.002	44.9
GPT-4o-mini	.600	0.007	23.5
Gemma-2-27B	.523	0.004	31.5
Mixtral 8x7B*	.327	0.022	30.5

Table 2: NER results at v5.1 (final prompt), sorted by F1. € = cost per run; s = avg. response time. *Mixtral collapsed at v4–v4.2; see Section 5.1.

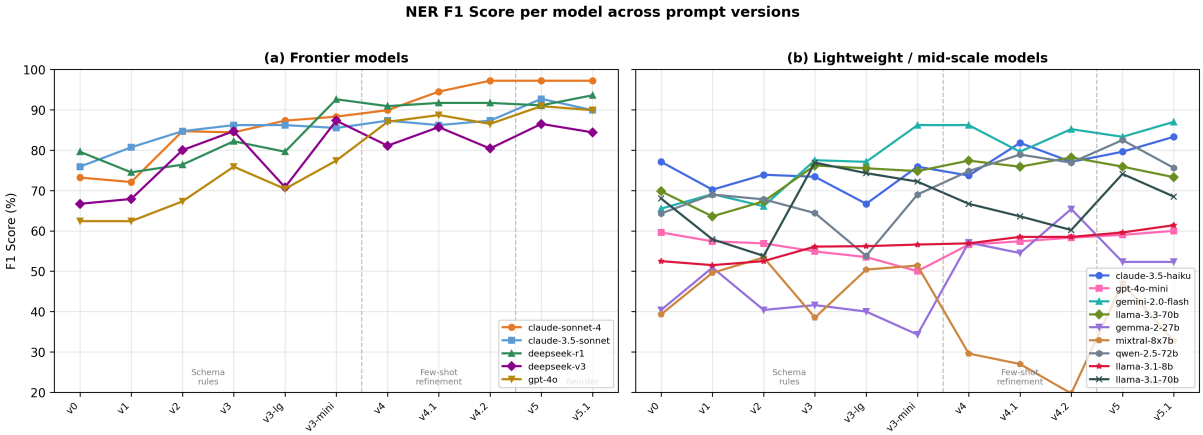

Figure 2: NER F1 score per model across prompt versions. Dashed vertical lines delimit the three optimisation phases: schema rule elaboration (v0-v3.1), few-shot refinement (v4-v4.2), and structural reordering (v5 v5.1).

5.2. Geoaffective Analysis: Test T1

Building an effective few-shot pool required several rounds of adjustment. We began with 12 examples, ensuring that each emotion label was represented at least twice, and conducted an initial evaluation to identify systematically weak categories. `LongB`, `AbsB`, and `pride` showed lower recognition rates, prompting the addition of targeted examples, first expanding the pool to 13 and then to 15 samples. All three configurations were tested across four prompt versions, ranging from a minimal schema description (v1) to a more elaborated formulation with explicit guidance for potentially confusable labels (v4).

Exploratory trials with pools exceeding 15 examples were also conducted. In these cases, performance decreased across models. Since T1 already consists of compound, multi-label sentences, increasing the number of examples appeared to add contextual load without improving discrimination. For this reason, 13 and 15 examples were treated as the upper bounds in this setting.

What we found was that adding more examples did not reliably help. For Claude Sonnet 4 and Claude 3.5 Sonnet, pool_13 gave the optimal (91.4% and 89.2% SubF1), while pool_15 resulted in lower performance for both. DeepSeek-V3 peaked at pool_12. Beyond a certain threshold, adding further examples appear to introduce noise rather than useful signal — and where that threshold falls differs by model.

What proved more important, however, was how prompt version and pool size interacted with each other. For Claude Sonnet 4, GPT-4o, and Gemini 2.0 Flash, v4 improved consistently over v1 regardless of pool size. For Claude 3.5 Sonnet, DeepSeek-V3, and DeepSeek-R1 the outcome depended heavily on which pool was used: at pool_13, v4 caused notable drops — 9.8 points for Claude 3.5 Sonnet, 13.9 for DeepSeek-R1, 18.4 for DeepSeek-V3 — while at pool_15 the same prompt helped or had negligible effect. A richer prompt does not simply add guidance; it interacts with the few-shot composition in ways that are not easy to predict.

Model	$v1_{12}$	$v1_{13}$	$v1_{15}$	$v4_{12}$	$v4_{13}$	$v4_{15}$
Consistent improvement with v4						
Sonnet 4	84.1	91.4	90.1	91.4	**93.2**	**93.2**
GPT-4o	69.0	72.4	66.7	73.7	67.9	**74.2**
Gemini Flash	69.8	71.9	67.7	70.8	**76.9**	**74.6**
Pool-dependent response to v4						
Sonnet 3.5	83.6	**89.2**	81.8	85.3	79.4	80.0
DeepSeek-V3	**83.9**	80.6	69.8	65.5	74.2	74.6
DeepSeek-R1	70.8	**77.4**	63.3	68.8	63.5	66.7

Table 3: SubF1 (%) for prompt versions v1 and v4 across pool sizes (12, 13, 15), Test T1. Bold: best per model.

5.3. Geoaffective Analysis: Test T2

Test T2 used the same prompts as T1, but this time each sentence carried a single emotion label — though not necessarily a single span. Running this simpler setting alongside T1 let us see where the models genuinely struggle with a label, and those per-label weaknesses were precisely what guided our few-shot pool expansions in T1.The difference in overall performance was clear: most models dropped to zero hallucination across all prompt versions, and results improved steadily from v1 to v4.

We also saw a different set of models at the top. Gemini 2.0 Flash and Claude 3.5 Sonnet reached the highest scores, with Gemini being the only

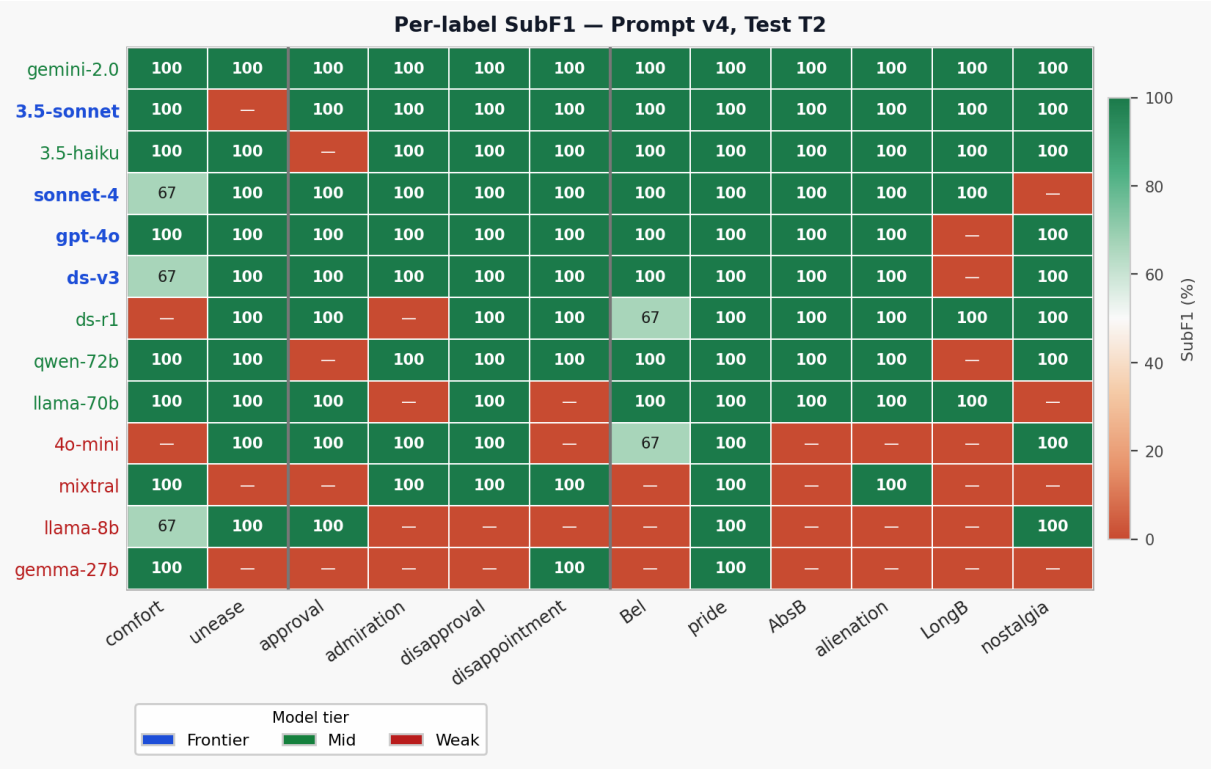

Figure 3: SubF1 per label and model — prompt v4, Test T2.

model to achieve 100% SubF1 at v4 — a result we ran multiple times to confirm. What makes this particularly interesting is the trajectory: Gemini started at 85.7% at v1 and improved consistently with each prompt version, ending at a perfect score. Its Token F1 of 84.7%, however, shows that perfect label recognition does not always mean perfect span boundaries. The model identified the right labels consistently but did not always capture the exact extent of the span. Claude 3.5 Haiku also performed well above expectations for its size. Claude Sonnet 4, our best performer in T1, came in at 92.3%. When the task is cleaner, the advantage of the largest models seems to become less pronounced.

Model	Sub F1	Tok F1	Halluc
Gemini 2.0 Flash	**1.00**	.847	.000
Claude 3.5 Sonnet	.957	.720	.000
Claude 3.5 Haiku	.957	.700	.000
Claude Sonnet 4	.917	.785	.083

Table 4: T2 results at v4, top models.

At the label level, most categories posed no particular difficulty by v4. `LongB` stood out as the hardest, given that 7 out of 13 models missed it entirely, across all versions. A second group of labels (`approval`, `admiration`, `AbsB`, `nostalgia`, and `Bel`) also showed complete misses in 3–4 models, suggesting that identity-tier and finer appraised distinctions remain genuinely hard even in a single-label setting. `comfort` was a different case: most models recognised it but some failed to capture the exact span, producing partial matches rather than complete misses.

6. Conclusions

GeoAffect was designed to study how place is affectively framed in nineteenth-century Greek prose and to test whether such distinctions can be supported through few-shot LLM annotation.

The experiments show that performance depends less on longer or detailed prompt descriptions and more on the targeted selection of few-shot examples, which appear to be the most effective learning source for LLMs. At the same time, we observe a ceiling effect, since after about 12–15 examples, improvements tended to level off, and in some cases additional examples introduced noise. This also has practical implications, given the cost associated with longer prompts.

Model behaviour also varied in consistency. Even at low temperature, some systems produced unstable outputs. At the same time, lighter models proved competitive in simpler configurations. Gemini 2.0 Flash stood out in T2, reaching the highest SubF1, while remaining faster and cheaper than frontier models. With further prompt and example refinement, it appears suitable for scaling annota-

tion to the full Ionian corpus.

Overall, the experiments indicate that LLMs can function as annotation and information extraction tools for digital humanities tasks beyond their primary training domains, provided that the schema is clearly specified and evaluation distinguishes label recognition from span accuracy.

The GeoAffect framework is still under active development. Having a clearer picture of how models respond to different prompt configurations, our next step is to select the best candidate — in terms of accuracy, speed, and cost — and apply it to the full Ionian corpus.

7. Acknowledgements

Funded by the European Union under Horizon Europe (project TALOS-AI4SSH, G.A. 101087269)

8. Bibliographical References

Eleni Bozia, Austin Stein, Wavid Bowman, Annie Gjineci, Gillian Vilela, Zachary Hracho, Rohan Prasad, Neema Owji, Niloufar Saririan, Aidan Burrowes, Aarushi Jain, and Nitaicandra Stevens. 2024. Performing sentiment analysis to trace the history of identity and belonging in ancient greek literature. *Digital Scholarship in the Humanities*, 39(4):1019–1025.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Giulia Grisot and Berenike Herrmann. 2023. Examining the representation of landscape and its emotional value in german-swiss fiction between 1840 and 1940. *Journal of Cultural Analytics*, 8.

Stuart Hall. 1996. *Who Needs Indentity*, 3rd edition, pages 19–33. SAGE Publications LTD, London, Thaousand Oaks, New Delhi.

Ryan Heuser, Franco Moretti, and Erik Steiner. 2016. The emotions of london. *Stanford Literary Lab Pamphlets*, 13(101).

Mirela Imamovic, Silvana Deilen, Dylan Glynn, and Ekaterina Lapshinova-Koltunski. 2024. Using ChatGPT for annotation of attitude within the appraisal theory: Lessons learned. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 112–123, St. Julians, Malta. Association for Computational Linguistics.

Matthew Jockers. 2015. The ancient world in nineteenth-century fiction; or, correlating theme, geography, and sentiment in the nineteenth century literary imagination. *Digital Humanities Quarterly*, 10(2).

Agnieszka Karlińska, Cezary Rosiński, Jan Wieczorek, Patryk Hubar, Jan Kocoń, Marek Kubis, Stanisław Woźniak, Arkadiusz Margraf, and Wiktor Walentynowicz. 2022. Towards a contextualised spatial-diachronic history of literature: mapping emotional representations of the city and the country in Polish fiction from 1864 to 1939. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 115–125, Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Richard S Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press, New York.

Niu Minxue, Jaiswal Mimansa, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. *ArXiv*, abs/2408.17026.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Paul Ricoeur. 1992. *Oneself as Another*. University Of Chicago Press, Chicago.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

Yi-Fu Tuan. 1977. *Space and Place: The Perspective of Experience*. University of Minnesota Press, Minneapolis, MN.

Dimitris Tziovas. 2017. *I Politismiki Poiitiki tis Ellinikis Pezografias: Apo tin Ermineia stin Ithiki*. Crete University Press, Heraklio.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Evaluating the Impact of LLM-Assisted Annotation in a Perspectivized Setting: the Case of FrameNet Annotation

Frederico Belcavello[1], Ely Matos[1], Arthur Lorenzi[1], Lisandra Bonoto[1], Lívia Ruiz[1], Luiz Fernando Pereira[1], Victor Herbst[1], Yulla Navarro[1], Helen de Andrade Abreu[1], Lívia Dutra[1,2], Tiago Timponi Torrent[1,3]

[1] Federal University of Juiz de Fora | FrameNet Brasil, [2] Gothenburg University,
[3] Brazilian National Council for Scientific and Technological Development - CNPq
{fred.belcavello, ely.matos, tiago.torrent}@ufjf.br, {arthur.lorenzi, lisandra.bonoto, livia.padua, luizfernando.pereira, victor.herbst, yulla.liquer}@estudante.ufjf.br, livia.vicente.dutra@svenska.gu.se, helen.abreu@visitante.ufjf.br

Abstract

The use of LLM-based applications as a means to accelerate and/or substitute human labor in the creation of language resources and datasets is a reality. Nonetheless, despite the potential of such tools for linguistic research, an evaluation of their performance and impact on the creation of annotated datasets, especially under a perspectivized approach to NLP, is still missing. This paper contributes to the reduction of this gap by reporting on an extensive evaluation of the (semi-)automatization of FrameNet-like semantic annotation by the use of an LLM-based semantic role labeler. The methodology employed compares annotation time, coverage, and diversity in three experimental settings: manual, automatic, and semi-automatic annotation. Results show that the hybrid, semi-automatic annotation setting leads to increased frame diversity and similar annotation coverage, when compared to the human-only setting, while the automatic setting performs considerably worse in all metrics, except for annotation time, which remains similar.

Keywords: FrameNet, LLM-assisted annotation, evaluation

1. Introduction

FrameNet is the implementation of the theory of Frame Semantics (Fillmore, 1982) in the form of a language resource that clusters together lexical units whose meaning require the same background scene, called a frame, to be evoked for their comprehension (Fillmore et al., 2003). Annotation serves as the empirical backbone of FrameNet, providing the evidence necessary to support the conceptual and linguistic analysis within the FrameNet model: it both validates how Lexical Units (LUs), defined as the pairing of a word with a specific frame, instantiate the frames they evoke, and yields semantically labeled datasets, which can be further used in several applications.

FrameNet-style semantic annotation remains an essential yet labor intensive process. Since its creation (Baker et al., 1998), FrameNet has relied on meticulous manual curation, requiring trained linguists to identify frame-evoking elements and label their corresponding frame elements in context. This fine-grained approach has produced a high-quality resource, but at the cost of scalability: expanding coverage across domains and languages demands considerable amounts of human effort and time. Early studies on FrameNet-based semantic role labeling (Gildea and Jurafsky, 2000) have already pointed out that the lack of large annotated corpora constrained model performance.

Recent advances in large language models

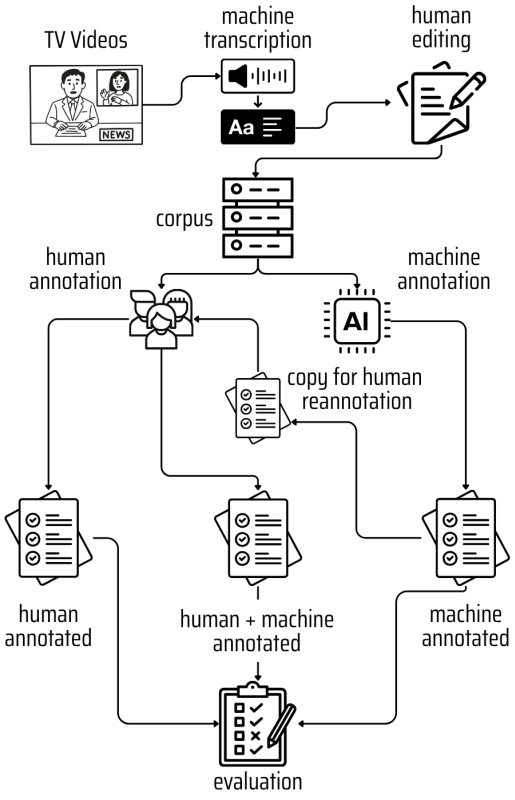

Figure 1: Experiment design

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 77–87

(LLMs) have opened new possibilities to reduce human workload in annotation tasks. Various research studies claim that such models demonstrate strong zero-shot and few-shot performance in annotation tasks (Brown et al., 2020). Others compare model performance to that of crowd workers in different annotation tasks (Gilardi et al., 2023).

Despite their potential, LLM-based annotation systems also pose significant risks. Baumann et al. (2025) show that subtle prompt or configuration changes can distort labels and introduce biases, a phenomenon they call LLM hacking. In large-scale tests, even state-of-the-art models produced incorrect or misleading annotations in roughly one third of cases. Such errors can propagate into downstream analyses, leading to false or exaggerated findings. The results by Baumann et al. (2025) highlight the need for rigorous human oversight and validation in LLM-assisted settings.

The question under investigation in this paper is whether LLMs can be used as assistants to facilitate, accelerate, and improve the quality of FrameNet annotation. Integrating these capabilities into FrameNet workflows could, in theory, significantly lower the annotation barrier, aligning FrameNet with contemporary data-driven NLP while preserving its linguistic depth and interpretability. This semi-automatic approach, leveraging LLM-based automated annotation with human validation, may represent a viable path toward sustainable, large-scale FrameNet growth in the era of generative AI.

In this paper, we present an experiment designed to explore the semi-automation of FrameNet annotation through the use of LOME (Large Ontology Multilingual Extraction) (Xia et al., 2021). LOME is an open-source LLM-based system which includes a FrameNet-semantic parser based on XLM-RoBERTa (Conneau et al., 2020). Our approach integrates LOME-generated suggestions into the human annotation workflow, allowing annotators to validate, correct, refine, or delete the automatically proposed frame and frame element labels, as shown in Figure 1. We hypothesize that such semi-automatic annotation can accelerate the FrameNet annotation process while maintaining quality. Beyond time efficiency, we also investigate whether machine-assisted annotation influences the diversity of frame interpretations and the perspectivization inherent to FrameNet annotation. Because FrameNet frames often encode viewpoint and conceptual stance, evaluating how automatic suggestions interact with these perspectivist aspects is essential to understand the epistemological impact of AI in the construction of human-curated language resources and datasets.

2. FrameNet Annotation: a Perspectivized Approach to Semantic Role Labeling

Frame Semantics (Fillmore, 1982) proposes that linguistic meaning emerges from our ability to interpret expressions against structured conceptual backgrounds known as frames. A frame represents a schematic scene, such as a commercial transaction, a motion event, or a perception event, along with its participants, props, and internal relations. Words evoke these frames, activating the knowledge structures necessary for comprehension.

Within this theoretical model, perspective plays a central role. Language does not simply encode objective situations; it construes them from specific points of view (Trott et al., 2020). In FrameNet, such perspectival distinctions are formally represented through frame-to-frame relations, for example, `Commerce_buy` and `Commerce_sell` are distinct perspectives on the more general `Commerce_goods-transfer` frame.

The Berkeley FrameNet project (Fillmore et al., 2003) implemented these ideas computationally by building a lexicon organized around frames and their frame elements (FEs), which correspond to the participants and attributes of each situation. For each frame, the database stores: (i) the set of core and non-core FEs in each frame; (ii) the lexical units (LUs)–pairings of words and senses that can evoke the frame; and (iii) a network of frame-to-frame relations that encode conceptual dependencies such as inheritance or perspective. This whole data structure is based on corpus evidence obtained from annotation.

There are two types of text annotation in FrameNet: lexicographic and full-text. In lexicographic annotation, sentences are queried in corpora with the aim of attesting the syntactic and semantic affordances of a given LU, which is known to evoke a given frame. In the process of selecting examples to be annotated, it is common that annotators choose those in which the frame being annotated is clearly represented. On the other hand, in full-text annotation, the goal is to annotate every LU in a given text. Annotators should read the text and select the appropriate frame for every lexeme which is a potential annotation target. Each LU thus yields its own Annotation Set (AS), so that a single sentence often generates multiple ASs targeting different LUs. This method places the focus on the broader discourse context, ensuring that all frame-evoking expressions in the text are represented. Because Frame Semantics assumes that different lexical items referring to the same conceptual scene impose distinct perspectives, full-text annotation naturally reveals these perspectival contrasts within coherent discourse. Lexicographic

annotation, on the contrary, is restricted to sentences selected to exemplify the valence patterns of a single predetermined LU, which limits the emergence of cross-frame perspective relations within the same text.

Crucially, FrameNet annotation differs from other approaches to semantic role labeling by treating meaning as interpretive rather than categorical. Because frames are prototype-based and overlap, a single expression may evoke more than one plausible frame depending on context. Take the sentence in (1) as an example:

(1) Children who started going to school this year may have jobs that are yet to be invented.

The noun *school* may instantiate either the Goal FE in the `Motion` frame or the Activity FE in the `Activity_start` frame, depending on how one interprets the semantics of *going* in the sentence. Instead of enforcing a single correct label, the annotation process acknowledges that such variation reflects legitimate differences in interpretation, making FrameNet annotation inherently perspectivized (Cabitza et al., 2023).[1]

From this standpoint, FrameNet annotation is not merely a task of tagging arguments, but an act of making explicit the interpretive stance that language adopts toward experience. This perspectivized understanding of semantic roles underlies the epistemological strength of the FrameNet model: it reveals how meaning is structured, rather than merely what is labeled. In the context of semi-automatic annotation, preserving this interpretive dimension is essential. By emphasizing perspective as an integral part of meaning representation, FrameNet offers both a theoretical and methodological foundation for the semi-automatic annotation approach explored in this study, one that combines the scalability of computational methods with the interpretive rigor of expert linguistic analysis.

3. Related Work

Recent research has explored the use of LLM-based tools for a series of FrameNet-related tasks.

Torrent et al. (2024a) propose and evaluate a series of prompts for using conversational LLMs for the creation of LUs and frames. The authors recognize the potential of LLMs to augment FrameNet coverage and expand it to other languages. However, they do not explore frame annotation.

Cui and Swayamdipta (2024), in turn, propose a methodology for generating synthetic annotated sentences from original human annotated ones. They claim that their method can be of use in low-resource settings but recognize its limitations in substituting human-annotated data where available.

Another group of methods approaches FrameNet annotation more directly. Chundru et al. (2025), Devasier et al. (2025) and Garat et al. (2025) propose methods to leverage conversational LLMs for that purpose. However, all three proposals require the frame and the frame elements to be included in the prompt along with the exemplar annotations. This type of methodology does not solve the problem of full-text annotation focused by the research presented in this paper, since it requires sentences to be annotated by the LLM to be previously sorted out and organized by frame-evoking targets. Moreover, they require LUs to be already modeled in the FrameNet database, which precludes LLM-based automatic annotation to contribute to the expansion of the FrameNet for which the annotation is being performed.

Finally, several frame-semantic role labelers have been proposed over the years, using different computational techniques, as well as requiring different levels of complexity for their training and deployment, over the years (Das et al., 2014; Hartmann et al., 2017; Swayamdipta et al., 2017; Kalyanpur et al., 2020; Jiang and Riloff, 2021; Xia et al., 2021; Tamburini, 2022; An et al., 2023). Despite assessing the performance of each system, these papers did not evaluate the incorporation of their frame parsers into an annotation pipeline. Therefore, they are not able to provide any evaluation of the impact of LLM assistance on the annotator's work.

To our knowledge, the only paper that studied the extent to which semi-automatization affects annotation's time and quality is the one by Rehbein et al. (2009).[2] In their paper, authors evaluate the impact of automatically pre-annotating sentences for frames and frame elements in terms of the amount of time needed to perform the annotations and the precision, recall and f-score for the annotations. They evaluated three conditions of semi-automatization: one in which no pre-annotation was performed and annotators performed the whole process, one in which a state-of-the-art frame and FE labeler was used to pre-annotate the sentences, and a third in which errors were manually inserted in a gold-standard annotation. All three conditions were tested on lexicographic annotation. Authors concluded that pre-annotation did not have a sta-

[1] For this same reason, FrameNet annotation is usually not evaluated for inter-annotator agreement, since annotations of one same sentence are expected to differ without the need of one of them being wrong.

[2] Rehbein et al. (2012) present a revised version of their 2009 experiment. The only parameter whose evaluation changes in comparison with the original paper is annotation time.

tistically significant effect on annotation time, despite the fact that annotators were faster under the third annotation condition than when presented with the unannotated sentences. As for annotation quality, the authors conclude that pre-annotation does improve the overall quality of the annotations. Moreover, they investigate the influence of pre-annotation on the types of errors human annotators make, and conclude that the pre-annotation does not seem to corrupt human judgments, since annotators make the same kinds of deviate judgment in all three conditions, when compared to the gold standard annotation.

The evaluation presented in this paper differs from that by Rehbein et al. (2009) in several ways: (a) because we recognize the perspectivized nature of FrameNet annotation, we do not compare the outcomes of either human-only, machine-only and hybrid annotation to one gold standard reference dataset; (b) instead, we focus on measuring the impacts of LLM-assisted pre-annotation on the coverage, diversity, observance of minimal requirements and time spent on human annotation; and (c) we conduct the evaluation on a full-text annotation task, which yields the kind of annotated dataset that is more useful for downstream tasks. [3] In the next section, we detail the experiment design used.

4. Experiment Design

The experiment whose results are reported in this paper aimed to evaluate the potential of LLM-based tools for semi-automating FrameNet annotation.

The experiment was structured in two phases. In Phase 1, each annotator manually annotated their assigned sentences from scratch, following the aforementioned guidelines. In Phase 2, sentences were automatically annotated by LOME before being given to annotators. In this phase, each annotator was asked to revise the automatic annotations made to a set of sentences from the same corpus. Annotators did not revise the same set of sentences they annotated in Phase 1, to avoid biases derived from previous judgments regarding some specific sentence they had previously encountered. Figure 1 summarizes the experiment design, which is detailed in the following paragraphs.

Corpus The experiment used sentences from a corpus of TV news media in Brazilian Portuguese (Br-Pt). The corpus comprises a total of 178 documents and 3,442 sentences. A subset of the corpus containing 12 documents and a total of 311 sentences was randomly selected. Three versions of this subset were created: (i) one for Human annotation, (ii) one for Machine annotation, and (iii) a copy of the product of the machine annotated data for the Machine plus Human annotation.

Human Annotation A group of five annotators, each with intermediate experience in FrameNet annotation, participated in the study. Each annotator was assigned approximately 60 sentences, distributed over two or three different news stories from the corpus. All annotations included frame and FE assignment, following the FrameNet guidelines for full-text annotation (Ruppenhofer et al., 2016) and the FrameNet Brasil guidelines for multimodal annotation of audio-oriented videos (Belcavello, 2023). The former indicates that annotators should watch each news story while annotating the sentences as a way to make sure that they pick the frame considering the multimodal context. The latter means that the annotator creates ASs for each lexeme for which there is a LU in FrameNet, sentence by sentence.

Machine Annotation Machine annotation was done using LOME (Xia et al., 2021). The choice for LOME is justified because it combines four important features: first, it is built upon an LLM, meaning that it leverages the potential of such models for frame semantic role labeling; second, it does not require any preprocessing of the sentences to be annotated to be made before they are submitted to the system, as it is the case with the conversational LLMs reviewed in section 3; third, it can be trained for any language for which there are corpora annotated for FrameNet categories; fourth, it provides annotations for both frames and FEs.[4]

For the experiment, we used a version of LOME trained on existing full-text annotation in both Br-Pt and English (En). In total, 18,170 annotated sentences were used—12,240 in Br-Pt and 5,930 in En—, with a rough average of 5.5 ASs per sentence (Dutra, 2024).

The sentence annotation pipeline is as follows:

1. The sentence is parsed using the Trankit UD parser (Nguyen et al., 2021), allowing the extraction of the tokens to be used as input in the next step. Each token has an associated lemma and part of speech (POS).

2. The tokens are inputted into LOME for processing. The result is a variable number of

[3] Full-text annotation is more useful for the task of automatic semantic labeling because lexicographic annotation biases the model's performance toward valency patterns that were annotated based on the annotator's choice, which did not necessarily cover all possible valency patterns. FrameNet makes no assumptions about statistical representativeness of lexical annotation.

[4] DAISY (Costa et al., 2022), another FrameNet-based parser does not do FE labeling, and, therefore, could not be used for comparison.

sets of frames and FEs, when assigned. For each frame or FE we have the text span from the sentence where either the target LU or the material instantiating the FE is located.

3. Using the span position, the system recovers the lemma/POS associated to each word in it.
4. Using lemma/POS and the attributed frame, the system checks if there is a LU already created. If not, a new LU is added, evoking the frame.
5. Finally, with the LU, the system creates a new AS, indicating the FEs automatically assigned using the LOME-defined span.

The ASs created are associated with two copies of each sentence: one that joins the Machine annotation subset of the corpus, and another that serves as starting point for the Machine plus Human annotation. The latter will be edited by the human annotators, and the former is preserved for comparison.

Machine plus Human Annotation Each annotator received a set of sentences different from the ones they annotated in the first phase. This time, the sentences were pre-annotated by LOME. Annotators were instructed to review and correct these LOME-generated labels, focusing on the adequacy of frame and FE identification.

The annotator had the following options (with their corresponding statuses): (a) fully accept the automatic annotation without making any corrections (ACCEPTED); (b) completely reject the machine annotation by removing the AS (DELETED); (c) replace the frame suggested in the Machine annotation while keeping the same lemma, or accept the LU suggested by the Machine annotation but modify (add or remove) some or all of the FEs (UPDATED); and finally, (d) create new ASs (CREATED).

This design allowed us to compare human and machine plus human annotation not only in terms of speed but also in terms of adherence to FrameNet methodology and diversity. In the following section, we present the evaluation metrics used.

5. Evaluation Metrics

The methodology created to compare the three annotation configurations—human only, machine only, and machine plus human—in the experiment sought to address two relevant aspects: the product and the process of annotation.

The evaluation of the annotation **product** consisted of counting how many elements were annotated (manually and automatically) and comparing these annotations. These elements include documents, sentences, annotation sets (ASs), and frame elements (FE). Still addressing the product aspect of the annotation and aiming to identify possible impacts of each annotation configuration on the diversity of labels used, we measured how many unique frames are associated with each document for each type of annotation, as well as the average number of frames occurring in each annotated sentence under each configuration. Finally, to qualitatively compare the annotation product, the cosine similarity between the semantic representations of each sentence (Viridiano et al., 2024) was used. Once again, the comparison was performed pairwise for all possible combinations of the three annotation configurations.

Regarding the annotation **process**, three measures were evaluated: (i) minimal number of core FEs in the annotation, (ii) the time spent by the annotators in the annotation, and (iii) the types of edits made to correct the automatic annotation in the machine plus human condition.

Core FEs are those that directly express the semantics of the frame, and they should all be present in the annotation, except in two cases: (a) when one core FE is in an **excludes** relation with one or more FEs, and (b) when two or more FEs are in a core set. The `Self_motion` frame in (2) exemplifies both cases.

(2) `Self_motion`

Definition: The Self_mover, a living being, moves under its own direction along a Path. Alternatively or in addition to Path, an Area, Direction, Source, or Goal for the movement may be mentioned.

Core Frame Elements:

Area: It is used for expressions which describe a general area in which motion takes place when the motion is understood to be irregular and not to consist of a single linear path.

Excludes: Direction, Goal, Path, Source.

Direction: The direction that the Self_mover heads in during the motion.

Goal: It is used for any expression which tells where the Self_mover ends up as a result of the motion.

Path: It is used for any description of a trajectory of motion which is neither a Source nor a Goal.

Self_mover: It is the living being which moves under its own power.

Source: It is used for any expression which implies a definite starting-point of motion.

FE Core Set(s): {Source, Goal, Path,

Direction}

Note that, although the `Self_motion` frame has eight core FEs, there can be one complete annotation for this frame with only two FEs: Self_mover and Area or Self_mover and either Source, Path, Goal or Direction. This is so, first, because the presence of the Area FE excludes the possibility of any of the other locative FEs, as the odd example in (3) demonstrates. Second, because Source, Path, Goal and Direction are in a core set, only one of them must be present for the frame to instantiate, as (4) shows: the sentence is well formed with only one of the FEs following the verb, with any combination of two of them and also with all three of them.

(3) Mark was **running** around to school.

(4) Mark was **running** from home to school along the road.

Hence, the minimum number of FEs that must necessarily be annotated was calculated by taking into consideration the total number of FEs in a frame minus the ones in exclude relations and core sets, where only one of the possible FEs was counted. The percentage of core FEs annotated indicates how complete an annotation is.

The time measure was taken for each AS. In the case of human annotation, this measure considered the time spent by the annotator both to define a new AS to be annotated—by selecting the appropriate LU—and to record the FEs in the AS. For the editing of the LOME output under the Machine plus Human configuration, the ASs had already been created, and thus the recorded time refers to the edits (or removals) made to each AS.

Another measure related to the annotation process concerns the edits made to the Machine Annotation during the Machine plus Human annotation, according to the possibilities described in section 4.

6. Results and discussion

As indicated in section 4, a total of 12 documents comprising 311 sentences were used in the experiment. Each sentence has a variable number of associated ASs. An AS indicates the Lexical Unit (LU) associated with an expression in the sentence and, at the same time, the frame evoked by that LU. In turn, each AS is associated with a variable number of FEs. On average, the number of AS per document was 129 in the Human annotation condition, 126 in the Machine annotation, and 160 in the Machine plus Human annotation. This first measure reveals that, while the number of ASs varies very little in the comparison between the human-only and the machine-only conditions, it increases sensibly in the human plus machine condition, presenting a 24% increase compared to the human-only scenario and a 26.9% increase compared to the machine-only setting.

Beyond the average number of ASs per sentence, we also measured the frame diversity in each condition. The motivation behind this metric is to assess whether the use of LLMs could interfere with human judgment when annotating, similarly to what is investigated by Rehbein et al. (2009). Table 1 shows the number of unique frames used in each annotation setting.[5] A higher number of unique frames may be indicative that not only more AS were created—which was the case for the Human plus Machine condition—but also that different perspectives were adopted by annotators for one same lexeme, and, therefore, that more possibilities provided by the vast number of annotation labels available in FrameNet were used.[6] Naturally, those two aspects are interconnected. Data presented in Table 1 indicate that Machine annotation is still very limited in terms of frame diversity. Considering that, on average, the number of frames per document in the Human annotation and in the Machine annotation is very similar, the difference in the average number of **unique** frames per document—67.91 for the Human annotation versus 52.66 for the Machine annotation—and per sentence—3.80 for the Human annotation versus 2.50 for the Machine annotation—is considerable. Moreover, the Machine plus Human condition led to a higher average of unique frames per document and a similar one per sentence when compared with the Human annotation scenario.

Table 2 can be broadly understood as representing the degree of agreement regarding the semantic representation of each sentence in each document between the three types of annotation. Although the automatic annotation identified fewer frames, the data suggest that these frames were mostly kept in the Machine plus Human annotation condition. On the other hand, the also high similarity between the Human and the Machine plus Human conditions indicates that human judgments were preserved in the pre-annotated configuration. This data will be reinforced by the findings discussed in

[5] The number of sentences in each document varies across annotation conditions because, in the comparison, only sentences with at least one annotation set are considered. Variations in the number of sentences can indicate that either LOME was not able to find any frames for a given sentence, or that a human annotator did not associate any frame to a sentence, for any reason.

[6] The FrameNet Brasil database (Torrent et al., 2022), which was used in this experiment, offers 1,429 different frames and 13,071 different FEs for annotation.

	Human			Machine			Machine + Human		
Doc	Sent	Frames	Avg F/S	Sent	Frames	Avg F/S	Sent	Frames	Avg F/S
02_13	22	71	3.23	22	53	2.41	22	80	3.64
02_14	13	71	5.46	23	56	2.43	23	105	4.57
03_11	27	77	2.85	26	56	2.15	28	88	3.14
03_12	19	47	2.47	26	57	2.19	27	85	3.15
04_01	50	114	2.28	46	87	1.89	49	99	2.02
04_06	9	34	3.78	9	20	2.22	10	26	2.60
05_01	14	54	3.86	15	26	1.73	15	51	3.40
05_02	26	80	3.08	23	54	2.35	26	93	3.58
05_03	3	12	4.00	20	59	2.95	20	83	4.15
07_02	21	97	4.62	22	58	2.64	22	104	4.73
07_03	13	80	6.15	13	46	3.54	13	75	5.77
07_07	20	78	3.90	17	60	3.53	20	82	4.10
Avg	19.75	67.91	3.80	21.83	52.66	2.50	22.91	80.91	3.74

Table 1: Frame diversity across annotated documents

Doc	Human vs Machine	Human vs Machine + Human	Machine vs Machine + Human
02_13	0.7199	0.7763	0.8461
02_14	0.6918	0.8267	0.8509
03_11	0.5625	0.7768	0.6927
03_12	0.6153	0.7288	0.7547
04_01	0.5686	0.6757	0.8322
04_06	0.5982	0.7080	0.7668
05_01	0.6167	0.7193	0.7520
05_02	0.6672	0.7264	0.8175
05_03	0.6835	0.7250	0.9116
07_02	0.6182	0.7752	0.7106
07_03	0.6053	0.7843	0.7186
07_07	0.6370	0.7355	0.7279
Avg	0.6320	0.7465	0.7818

Table 2: Cosine similarity between annotation methods

the end of this section.

Table 3 shows the percentage of the minimal number of core FEs annotated per document. This metric assesses whether the annotation respected FrameNet methodology regarding the requirement that core FEs are present or inferred in the sentence for a frame to be instantiated. Data shows that human annotators excel in following the guidelines, with 95.79% of the minimal number of FEs being annotated. On the contrary, LOME falls short in this metric, with only 34.20% of the minimal number of FEs present in the annotation. This is mainly due to the fact that LOME does not indicate inferrable FEs—what FrameNet calls null instantiations (Ruppenhofer et al., 2016)—in the annotation. This is so because, as noted in section 4, LOME needs to assign FEs to spans of text in the sentence. In the Machine plus Human configuration, the percentage of minimal core FEs annotated decreases to 90.65%, which may be due some minor influence of the pre-annotation on the human annotator's judgment about the adequacy of the annotation to the FrameNet policy. This effect does not seem to be very relevant, though.

Table 4 presents the average time each annotator spent for the annotation of each sentence in each document. This metric compares the human part the Human-only condition with the human part in the Machine plus Human annotation. For the total of the experiment it shows that using pre-annotation leads to a small decrease of average time per sentence: 1.99 minutes. This indicates that reducing annotation time seems to be not the most compelling argument in favor of using LLM-based pre-annotation. This conclusion replicates the results in Rehbein et al. (2009).

Finally, Table 5 presents, in absolute and percentage terms, the edits annotators made to LOME pre-annotation during phase 2. Note that the annotators completely discarded 19.68% of the automatic annotations and fully accepted only 6.61% of them. This indicates that pre-annotation was far from being judged as perfect in this setting. However, while 17.5% of the ASs in the Machine plus Human condition were created from scratch, the majority of the ASs in the final dataset, 65.45%, were partially used and improved by the annotators. This finding correlates with the one presented in Ta-

	Human			Machine			Machine + Human		
Doc	Core	Min	%	Core	Min	%	Core	Min	%
02_13	229	246	93.09	92	244	37.70	338	299	100.00
02_14	309	301	100.00	120	352	34.09	523	491	100.00
03_11	268	296	90.54	106	267	39.70	350	384	91.15
03_12	193	248	77.82	111	360	30.83	358	412	86.89
04_01	578	624	92.63	186	543	34.25	407	567	71.78
04_06	159	128	100.00	27	99	27.27	130	152	85.53
05_01	210	194	100.00	49	132	37.12	173	213	81.22
05_02	447	332	100.00	90	296	30.41	346	428	80.84
05_03	38	29	100.00	107	291	36.77	284	314	90.45
07_02	438	400	100.00	108	321	33.64	465	386	100.00
07_03	294	308	95.45	66	197	33.50	318	284	100.00
07_07	323	312	100.00	87	248	35.08	364	314	100.00
Avg	290.5	284.83	95.79	95.75	279.17	34.20	338	353.67	90.65

Table 3: Percentage of core FEs annotated

Doc	Sent	Avg Length	Human Anno	Machine +Human Anno	Diff
02_13	20	82.1	9.36	9.37	-0.1
02_14	14	152.64	19.01	11.03	7.98
03_11	26	101.58	4.57	9.89	-5.32
03_12	21	80.14	2.76	4.61	-1.85
04_01	43	88.07	19.17	3.23	15.94
04_06	7	101.29	16.15	6.34	9.81
05_01	13	91.08	26.91	38.12	-11.21
05_02	26	104.5	23.6	18.13	5.47
05_03	3	129.33	20.45	5.72	14.73
07_02	20	116.0	13.61	19.61	-6
07_03	13	135.85	13.77	17.68	-3.91
07_07	19	107.11	10.19	11.89	-1.7
Avg	18.75	107.47	14.96	12.97	1.99

Table 4: Average annotation time per sentence in minutes

Doc	Total	ACCEPTED	%	CREATED	%	DELETED	%	UPDATED	%
03_11	230	7	3.04	18	10.84	64	27.83	141	61.30
03_12	228	2	0.88	18	10.59	58	25.44	150	65.79
02_13	161	0	0.00	25	17.12	15	9.32	121	75.16
02_14	276	0	0.00	18	7.59	39	14.13	219	79.35
04_01	320	20	6.25	68	26.05	59	18.44	173	54.06
04_06	80	2	2.50	4	5.80	11	13.75	63	78.75
05_01	113	24	21.24	7	6.80	10	8.85	72	63.72
05_02	222	47	21.17	5	2.76	41	18.47	129	58.11
05_03	139	27	19.42	19	14.73	10	7.19	83	59.71
07_02	253	4	1.58	8	4.60	79	31.23	162	64.03
07_03	182	2	1.10	9	7.03	54	29.67	117	64.29
07_07	229	5	2.18	11	7.05	73	31.88	140	61.14
Avg	202.75	11.67	6.61	17.5	10.08	42.75	19.68	130.83	65.45

Table 5: Human edits on LOME annotations

ble 2, since the similar cosine similarity between the Human and Machine plus Human conditions, and the one between the Machine and Machine plus Human conditions indicate partial preservation of both the frames obtained in the pre-annotation and the original human judgment of the sentences.

Data from Table 5, together with those in the previous tables, reinforce the idea that pre-annotation is a valid strategy to increase the number and diversity of ASs, while having little impact on both annotation time and on the observance of FrameNet annotation guidelines—with a small decrease in the percentage of the minimal number of core FEs annotated.

7. Conclusions and outlook

The experiment reported in this paper showed that LLM-based pre-annotation can be useful for improving the coverage of perspectivized FrameNet annotation, while preserving human judgment. Although no sensible improvement in annotation speed was observed, the use of pre-annotation validated by humans seems to be a viable path for fine-grained semantic annotation.

Future work should look at least into two extensions that should be adopted in the short term for a new evaluation of the impact of LLM assistants on FrameNet annotation.

The first is the inclusion of other types of semantic role labelers—such as DAISY (Torrent et al., 2024b), for example—in the automatic annotation pipeline. Additional parsers could serve as a post-processing step for LOME, aimed at adding annotations LOME was unable to perform.

A second aspect concerns the adoption of stricter policies in the annotation process to ensure that at least the minimum number of core FEs have been annotated (both in human and automatic annotation). This implies enabling the automatic system to record the occurrence of null instantiations when necessary.

8. Ethical considerations and limitations

All annotation used in the experiments, including for the annotation sets used for training the Br-Pt instance of LOME, was carried out by trained annotators who were paid a monthly stipend, which is, at least, equivalent to the minimum wage according to local regulations. All annotators involved in the annotation of the corpus used in the evaluation experiment reported here are co-authors of this paper.

Among the limitations of the experiment described, it is worth noting that all annotated sentences are written in Br-Pt. However, LOME is language-agnostic and its components are designed to prioritize multilinguality. LOME employs XLM-RoBERTa (Conneau et al., 2020) as the underlying encoder, which allows the experiment to be easily extended to other languages. Furthermore, the experiment relied exclusively on LOME as the frame-semantic parser, but other LLM-based semantic role labelers may be evaluated in future work.

9. Acknowledgements

Research reported in this paper was developed under the ReINVenTA—Research and Innovation Network for Vision and Text Analysis of Multimodal Objects—initiative, 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). The resulting dataset will be part of the data collection of the National Science and Technology Institute for Responsible Artificial Intelligence, Computational Linguistics and Information Treatment and Dissemination (INCT-TILDIAR, CNPq grant 408490/2024-1). Belcavello was supported by CNPq (grant 200270/2023-0). Lorenzi, Abreu and Pereira were supported by FAPEMIG. Bonoto, Ruiz, Herbst and Navarro were supported by CNPq. Torrent is a CNPq research productivity grantee (grant 311241/2025-5). The presentation of this paper at LREC 2026 is supported by the Institute of Artificial Intelligence at the National Laboratory for Scientific Computing (IIA-LNCC), an initiative led by the Ministry of Science, Technology, and Innovation (MCTI) as part of the Brazilian Artificial Intelligence Plan (PBIA).

10. Bibliographical References

Kaikai An, Ce Zheng, Bofei Gao, Haozhe Zhao, and Baobao Chang. 2023. Coarse-to-fine dual encoders are better frame identification learners. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13455–13466, Singapore. Association for Computational Linguistics.

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.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation.

Frederico Belcavello. 2023. *FrameNet annotation for multimodal corpora: Devising a methodology for the semantic representation of text-image interactions in audiovisual productions*. Doctoral dissertation, Universidade Federal de Juiz de Fora, Juiz de Fora, Brazil. Faculdade de Letras, Programa de Pós-Graduação em Linguística.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. 2020. Language models are few-shot

learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Jayanth Krishna Chundru, Rudrashis Poddar, Jie Cao, and Tianyu Jiang. 2025. Do llms encode frame semantics? evidence from frame identification. *arXiv preprint arxiv: 2509.19540*.

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.

Alexandre Diniz da Costa, Mateus Coutinho Marim, Ely Matos, and Tiago Timponi Torrent. 2022. Domain adaptation in neural machine translation using a qualia-enriched FrameNet. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1–12, Marseille, France. European Language Resources Association.

Xinyue Cui and Swabha Swayamdipta. 2024. Annotating FrameNet via structure-conditioned language generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 681–692, Bangkok, Thailand. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Jacob Devasier, Rishabh Mediratta, and Chengkai Li. 2025. Can llms extract frame-semantic arguments? *arXiv preprint arxiv: 2502.12516*.

Lívia Dutra. 2024. Evaluating the contribution of framenet to gender-based violence identification: How semantic annotation can be used as a resource for identifying patterns of violence. Masters Thesis in Language Technology, Göteborgs Universitet, Gothenburg.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, pages 111–138. Hanshin Publishing Co., Seoul, South Korea.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3):235–250.

Diego Garat, Guillermo Moncecchi, and Dina Wonsever. 2025. Exploring in-context learning for frame-semantic parsing. *arXiv preprint arxiv: 2507.23082*.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

Silvana Hartmann, Ilia Kuznetsov, Teresa Martin, and Iryna Gurevych. 2017. Out-of-domain FrameNet semantic role labeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 471–482, Valencia, Spain. Association for Computational Linguistics.

Tianyu Jiang and Ellen Riloff. 2021. Exploiting definitions for frame identification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2429–2434, Online. Association for Computational Linguistics.

Aditya Kalyanpur, Or Biran, Tom Breloff, Jennifer Chu-Carroll, Ariel Diertani, Owen Rambow, and Mark Sammons. 2020. Open-domain frame semantic parsing using transformers. *arXiv preprint arxiv: 2010.10998*.

Minh Van Nguyen, Viet Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.

Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. 2009. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 19–26, Suntec, Singapore. Association for Computational Linguistics.

Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. 2012. Is it worth the effort? Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. *Language resources and evaluation*, 46(1):1–23.

Josef Ruppenhofer, Michael Ellsworth, Miriam Petruck, Christopher Johnson, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute (ICSI).

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arxiv: 1706.09528*.

Fabio Tamburini. 2022. Combining ELECTRA and adaptive graph encoding for frame identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1671–1679, Marseille, France. European Language Resources Association.

Tiago Timponi Torrent, Thomas Hoffmann, Arthur Lorenzi Almeida, and Mark Turner. 2024a. *Copilots for Linguists: AI, Constructions, and Frames*. Elements in Construction Grammar. Cambridge University Press, Cambridge.

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*, 13.

Tiago Timponi Torrent, Ely Edison da Silva Matos, Alexandre Diniz da Costa, Maucha Andrade Gamonal, Simone Peron-Corrêa, and Vanessa Maria Ramos Lopes Paiva. 2024b. A flexible tool for a qualia-enriched FrameNet: the FrameNet Brasil WebTool. *Language Resources and Evaluation*, pages 1–29.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

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.

Patrick Xia, Guanghui Qin, Siddharth Vashishtha, Yunmo Chen, Tongfei Chen, Chandler May, Craig Harman, Kyle Rawlins, Aaron Steven White, and Benjamin Van Durme. 2021. LOME: Large ontology multilingual extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 149–159, Online. Association for Computational Linguistics.

Annotating Word Meanings Over Time: The Trade-off Between Scalability, Reliability and Expressivity Power

Pierluigi Cassotti, Nina Tahmasebi
University of Gothenburg
{pierluigi.cassotti,nina.tahmasebi}@gu.se

Abstract

Annotating the meanings of a word over time in order to document their emergence or disappearance presents substantial implementation challenges. These difficulties arise for several reasons, notably the need for sufficient expressive power in the annotation paradigm to capture unconventional or rare meanings, as well as issues of scalability related to the number of annotations required. The first challenge is particularly acute in the context of historical texts, where modern annotators must interpret word meanings in sources that are temporally distant and often absent from contemporary dictionaries and language use. The second challenge is inherent to the distribution of word meanings, which tend to occur sparsely and intermittently over long time spans. In this paper, we examine several annotation paradigms, discussing their respective advantages and limitations. We also present a pilot study on English and Swedish. Our results indicate that a usage-sense inventory based annotation paradigm can be adopted in place of a usage-pairs-based approach while maintaining expressivity power and reducing the complexity from quadratic to linear.

Keywords: word sense disambiguation, lexical semantics, diachronic linguistics

1. Introduction

The creation of datasets capturing word meanings by means of human annotation has long been, and continues to be, fundamental to the development of language technologies capable of understanding human language and its dynamics. However, a persistent challenge in annotation is that word meanings are latent, fuzzy variables whose boundaries are difficult to define. There is rarely a clear point at which one meaning ends and another begins, neither within a single period nor across time periods. Often, the set of meanings present in text vary across sources; different genre will use words in different ways (e.g., think of a *worm* in magazines on gardens or computers).

As a result of the difficulty to capture meanings, researchers commonly rely on an approximation that equates meanings with *senses* by which we refer to word meanings that have been defined by expert linguists or lexicographers as discrete, tangible entities, such as the sense definitions found in dictionaries. In the annotation task, people then get to label words in naturally occurring text with one (or more) senses from the sense repository. A task often called word sense disambiguation. As with most standard annotation tasks, each instance is annotated by several annotators and the majority label is chosen.

But even using a sense repository as an approximation, significant challenges remain. One such challenge is that the sense inventories, although created by experts, reflect the experts' conceptualizations of meaning. This means that the way in which the semantic space is divided is not necessarily deterministic, nor are the resulting senses fully reflected in natural language. Indeed, even among experts, such as lexicographers, there is often disagreement about how meanings should be delineated. This issue is commonly referred to as the lumpers versus splitters problem where on the one hand, lumpers tend to group similar meanings together, and on the other, splitters aim to document even the smallest variations. When a dictionary is used as a reference for sense inventories, this dichotomy is reflected in the representation of word meanings. And because dictionaries are typically the product of the work of multiple teams, compiled over long periods of time and subject to updates and revisions, the final resource is often inherently heterogeneous affecting the resulting annotations.

Also on the computational models, this has an effect. Models trained and evaluated on such resources tend to be effective at recognizing clear semantic distinctions, such as homonymy or cases of polysemy involving evident shifts in domain, for example in metaphorical extensions of meaning. However, when the same models are required to recognize subtle differences in meaning, their performance deteriorates significantly.[1]

An alternative way of annotating for word meaning (both synchronically and diachronically), has gained popularity in the past decade. Instead of basing itself in the use of sense repositories from external sources, it makes no assumptions on which meanings exist for a word in the corpus. The modeling paradigm instead compares pairs of usages

[1] It should be noted that even humans tend to show lower agreement in such cases, suggesting that the underlying reason is likely that individual perception plays a large role. Consequently, the internal models of meaning that each individual holds become more influential in these instances.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 88–98

of a word sampled from the corpus and allows the user to make decisions on the similarity of a target word across two usages. Based on judgments on a larger set of usage pairs, senses can be induced. Such datasets exist for many languages under the name Word-In-Context (WiC) which refers to synchronic data with binary annotation (same meaning/ different meaning), or (Diachronic) Word Usage Graphs (DWUG). In the WiC case, the aim is often to use the data for evaluation of large language models and hence to have many words with few usage pairs for each. In the DWUG case, however, the aim is to study words and their meanings over time, and thus many more usage pairs are included and consequently, many more annotations are needed.

In fact, when we want to annotate word meanings *over time*, as opposed to traditional word sense disambiguation that takes place synchronically, we must take into account the volume of data needed to answer the question: when did a meaning appear, change, or disappear? This necessitates sampling from large scale historical archives, frequently involving thousands of occurrences of a word for each time period to reduce the number of annotations needed, while maintaining the distribution of the senses present in the text. Because of the scalability issue, often a gross simplification has been used: two far apart time periods are compared to each other to determine change in the senses found for a word. However, this view is very limited and does not allow us to build realistic models that can trace the dynamics of individual senses over time. But, when we take further time periods into account, the scale of annotation explodes and becomes infeasible.

In this paper, we address the challenge of annotation scalability through a pilot study in which we compare different annotation paradigms. Based on our results, we argue that a more efficient approach can be adopted while maintaining a level of quality comparable to that of existing methods. These results will then feed into our large-scale annotations of meaning change across multiple time periods.

2. Annotation Paradigms

In this section, we describe the different paradigms for annotating word meanings and introduce the terminology used throughout the paper. A usage of a word refers to a specific instance of that word in-context, that is, a word instantiated within a naturally occurring sentence. A sense is a discrete entity that captures a particular interpretation of a word, and a sense inventory is the set of all possible senses associated with a word. In this article, references to definitions generally denote those provided in dictionaries. When we discuss scalability, these have to be multiplied by the number of desired annotations per instance. Standard is to have three annotators and take as the majority vote as the label in the case of binary labels, and an average of the (ordinal) values otherwise.

Example of Annotation We use an example to illustrate, in numerical terms, how many annotations are required. In our example, we consider a specific word, with 3 senses, 20 usages per time period (the minimum adopted so far in (Zamora-Reina et al., 2022)), and 10 time points.

2.1. Usage-Usage (U-U)

This paradigm follows a two step procedure for annotating word meanings. Annotators are presented with a single pair of word uses at the time, and are asked to assess how similar these uses are, that is, how similar the meaning of the word is across the two contexts. As an example, we can use the target word *rock* in the sentences *I listened to rock.* and *I threw a rock.* where the similarity is low, while *I listened to rock.* and *I went to a rock concert.* has a high similarity. The resulting similarity judgments across multiple pairs of usages for the same target word are then aggregated and can be thought of as a graph. In each graph representing a single target word, nodes are the usages of the target word while edges between the nodes correspond to the similarity judgments. The edges of the graph group together usages that are most similar to one another, where each group can be said to correspond to a distinct meaning of the target word.

Similarity between pairs of uses can be assessed either in a binary manner or on a continuous scale. In the binary setting, corresponding to the Word in Context task, annotators choose between two options: either the two uses express the same meaning or they express different meanings. When similarity is expressed on a graded scale, the task is often referred to as a Graded Word in Context task.

An alternative to binary judgments is the semantic relatedness scale introduced in the DUREL framework, which consists of four values ranging from 1 to 4. Here, 1 denotes Unrelated, 2 Distantly Related, 3 Closely Related, and 4 Identical. This scale is inspired by Blank's notion of semantic proximity and establishes a correspondence between the four values and the categories of homonymy, polysemy, context variance, and identity.

Scalability: This annotation paradigm scales **quadratically** since each usage has to be compared to each other usage. In diachronic annotation, this becomes problematic. Consider our example, one word with 20 usages for 10 time points, we have a total of 200 usages. The total number of possible pairwise comparisons (annotations)

↑	4:	Identical	Identity
	3:	Closely related	Context variance
	2:	Distantly related	Polysemy
	1:	Unrelated	Homonymy

Table 1: The DURel relatedness scale (Schlechtweg et al., 2018) and the respective Continuum of semantic proximity proposed by Blank (1997).

would then be $(U(U-1))/2 =$ **19,900**. If each comparison is annotated by at least three annotators, this amounts to 59,700 individual annotations. In practice however, it is infeasible to annotate each instance with every other (and has only been done for ChiWUG where 40 Chinese words with 40 usages each were annotated by two annotators, resulting in about 60,000 annotations (Chen et al., 2023a)). Instead, common practice is to sample among the usage pairs while ensuring that sufficient connectivity remains among the usages, see e.g., (Schlechtweg et al., 2021a).

Previous annotations and experiences Most notably, this style of annotation was used for the SemEval-2020 Task 1 on Unsupervised Lexical Semantic Change Detection (Schlechtweg et al., 2020). The work was followed up by benchmarks for multiple languages, including Russian (Kutuzov and Pivovarova, 2021a,b), Spanish (Zamora-Reina et al., 2022), Norwegian (Kutuzov et al., 2022), and Chinese (Chen et al., 2023b, 2022), as well as reannotations for Swedish, German and English with smaller set of usages and higher density of usage pair annotations (Schlechtweg et al., 2024). For binary Usage-Usage annotations with a temporal dimension, also TempoWiC (Loureiro et al., 2022) is available, focusing on shorter time spans within social media. In general, all WiC data where the labels are assigned by human annotators (as opposed to derived from external resources e.g. WordNet) fall in this category, see e.g., (Cassotti et al., 2023).

In terms of temporal bins, at least for the original SemEval corpora, each subcorpus was around 40-50 years in size[2] and thus significant amount of change occurred already in within a single time period. With limited annotation budgets, there is a natural trade-off between a high number of usages (regardless of time period) versus a high density of annotation between the usage pairs. If we can afford a limited set of annotations, the more usages we have, the less dense becomes the graph. For 200 usages, there are 19,900 pairs. A 5,000-annotation budget will result in a density of 0.25. While for a smaller set of 100 usages, 5,000 annotations covers the whole graph. This trade-off is discussed by Schlechtweg et al. (2024) who showed that it is preferable to have a higher degree of annotated pairs rather than a larger number of usages, as it leads to more robust sense clusters.

[2]This choice was made for several reasons. Firstly, corpora tend to grow in size over time. So to ensure that a sufficient number of usages exist for most words included in the benchmark, the initial time period needed to be broad. Secondly, the choice to have large time bins for each subcorpus was made to increase representation of each sense.

2.2. Usage-Sense (U-S)

In this paradigm, annotators are presented with a word usage and the definition of one of its senses, and they are asked to determine whether the sense definition corresponds to the meaning expressed in that particular usage. As before, this judgment can be made either in a binary fashion or on a continuous scale. In the binary setting, annotators decide whether the use and the definition reflect the same meaning or not. This task is commonly referred to as Target Sense Verification (TSV). When a continuous scale is employed, annotators assess how similar the meaning expressed by the word use is to the meaning described by the proposed sense definition, typically using the four value semantic relatedness scale.

We included this annotation paradigm in our pilot to allow for each sense to be considered independently from the other senses, as to avoid potential bias where users do not assign a sense because another one is a better fit. Another advantage is that users do not need to be overwhelmed with information, making the task complex. If a word has 20 senses as opposed to three, the choice of which one fits the best becomes much more difficult. If instead the annotator sees a single sense at the time, the choice is independent of the number of senses a word has.

Scalability: This annotation paradigm scales **linearly** with the number of usages. For each word, we will have kU instances to annotate, where k is the number of senses of the target word. In our example, this results in **600 annotations** (each of the 200 usages is paired with each of the 3 senses). That is two orders of magnitude less the than usage-usage annotation framework.

Previous annotations and experiences: To the best of our knowledge, this annotation schema has not been tested in previous research as the obvious downside is an increasing annotation need without obvious benefits. Erk et al. (2013) and Cassotti and Tahmasebi (2025), however follow this paradigm to some extent as they ask annotators to provide for a usage, a rating with respect to each sense in the sense inventory. While these are not independent, and can be randomized and mixed with annotations for other usages to maximize anno-

Lemma	Sense Definitions
graft	**1)** A shoot or scion inserted in a groove or slit made in another stock, [..] **2)** Surgery. 'A portion of living tissue transplanted [..] **3)** A ditch; a moat; [..] **4)** The depth of earth that may be thrown up at once with a spade; [..] **5)** A kind of spade, used in digging drains. **6)** Work, esp. hard work. / A trade, craft. **7)** The obtaining of profit or advantage by dishonest or shady means; [..]
twist	**1)** A divided object or part. **2)** The twisting of threads into a cord, and derived senses. **3)** A slight or weak support upon which something depends; [..] **4)** An act or the action of turning on or as on an axis; [..] **5)** A dance in which the body is twisted from side to side; [..] **6)** A young woman, a girl. [..]
konduktör	**1)** Person som säljer och kontrollerar biljetter och andra färdbevis på tåg. **2)** Elektroteknisk ledare, överförare av kraft, [..] Kan också användas bildligt. **3)** Titel för arbetsledare eller uppsyningsman vid slott, [..] **4)** En i militärutbildad byggnads- eller arbetsledare vid fästning och dylikt; [..] **5)** Arbetsledare vid byggnadsverk [..]
motiv	**1)** Underliggande orsak till viss handling **2)** Ämne för konstnärlig framställning särskilt inom bildkonst och litteratur; [..] **3)** Minsta melodisk-rytmiska enhet av musikstycke, vanligen återkommande och lätt igenkännlig **4)** Spets- eller broderiarbete i avpassade delar till påsättning på underkläder, [..]

Table 2: U-S and U-SI sense inventories for graft, twist, konduktör, and motiv.

tator objectivity, the results are as close as currently exist.

2.3. Usage-Sense Inventory (U-SI)

In this paradigm, annotators are presented with a word usage and the definitions of all the word's senses at once. It follows a standard word sense disambiguation setting, where the annotators choose either the best fit, or alternatively, the set of best fitting senses. Like for prior annotation paradigms, this judgment can be made either in a binary fashion or on a continuous scale. Prior research has shown that there are great benefits to allowing annotators to provide a graded rating to *every* sense, rather than choosing top senses. Erk et al. (2013) found that the latter can lead to a bias in assigning a single best sense (and by implication, not assigning the rest). However, grading all senses increases the number of annotations and becomes equivalent to the U-S scenario above.

Scalability: This annotation paradigm is the most scalable one (assuming one does not require a rating for each and every sense). The schema scales with the number of usages and is thus linear, as each usage is judged only once with respect to *all* senses (in our annotation example this corresponds to **200 annotations**). Despite the linear scaling, the annotation task becomes complex if each usage is to be judged with respect to a large set of senses. The cognitive effort for keeping all senses in memory to make a good choice increases with the number of senses. After annotating a larger set of usages, the annotator typically becomes familiar with the senses, which reduces the required effort. However, a cold-start problem remains if the annotator pauses for some time and later returns to the task. Thus, to have maximum gain from this annotation schema, it is important not to overload the user with too many, or too long, sense descriptions.

Previous annotations and experiences: This annotation paradigm is standard for word sense disambiguation tasks. While typically, in WSD, the temporal dimension is lacking, the task itself is identical. Prominent examples of WSD data are SemCor (Miller et al., 1994), Sense-Eval (Snyder and Palmer, 2004; Edmonds and Cotton, 2001a,b) and SemEval (Moro and Navigli, 2015; Navigli et al., 2013; Agirre and Soroa, 2007a). A notable difference between some of these corpora and the paradigm we are proposing above is that some of these corpora disambiguate not only a target word but all words in the sentence. We propose disambiguating only a target word.

2.4. Other Paradigms

Beyond the paradigms described above, prior work has explored bottom-up sense induction through context clustering (Agirre and Soroa, 2007b; Schütze, 1998), substitute-based approaches (McCarthy and Navigli, 2007), and direct semantic change judgments across time (Cook et al., 2014).

2.5. Reliability

In the U-U paradigm, previous work (Schlechtweg et al., 2021b; Cassotti and Tahmasebi, 2025) clearly shows that annotators struggle to reach agreement on adjacent categories (e.g., Identical vs. Closely Related). This error can be systematic, stemming from differing perceptions of semantic proximity among annotators, or it may arise from the intrinsic difficulty of the task. In particular, the scale values

are not anchored to easily interpretable or objective reference points, which can even lead to inconsistencies within the same annotator's judgments.

3. Expressivity Power

In general, U–U is a paradigm with greater expressive power than U–S and U–SI, because the annotation scheme does not impose explicit constraints on which meanings can be labeled, leaving the representation of meaning implicit. However, in practice this theoretical advantage does not fully materialize, due to the clustering choices.

Clustering is typically performed using correlation clustering (Bansal et al., 2004) with a fixed threshold to binarize the relations between nodes, commonly set at 2.5. Under this scheme, all node pairs with a score above the threshold are treated as valid links, while those below it are discarded. As a consequence, clusters may include node pairs with an average score of 3, grouping them together despite only moderate relatedness.

Regarding the senses used in U-S and U-SI paradigms, we can make several choices. The first one is to use all senses available at all time periods. This results in redundant senses presented to the user at some time periods, as, e.g., the sense of e.g., computer virus did not exist for the word *worm* in year 1900. However, to reduce the number of senses to only those presumed active at the time from which the sentence stems, has several consequences. Firstly, we would need to make assumptions about validity periods of senses on the basis of dictionaries. However, often dictionaries are normative and have poor correspondence with empirical data; words can be used with specific senses before the senses conventionalize and thus are attested in the dictionary. Secondly, the order of the sentences annotated should be random, also across time periods, to avoid that the annotators make a priori choices based on knowledge of the time period. The set of senses presented to the annotators can thus be a biasing factor and should therefore include all senses. Finally, from a cognitive perspective, we assume that the annotators get aquatinted with the senses and their order, thus changing the set, or order, of the senses between instances can lead to additional cognitive load.

4. Pilot

Our long-term goal is to annotate a set of words across 10 time periods. However, our limited budget constrains the total *annotation hours* we can fund, which in turn limits the number of annotations we can obtain. The higher the price of an annotation paradigm, the fewer words and/or usages we can afford to annotate. Before we scale up the annotations, we therefore need a good evaluation of the complexity of the different annotation paradigms together with an estimate of how many annotations per hour that can be done.

	t_1	t_2
English	CCOHA 1810–1860	CCOHA 1960–2010
Swedish	Kubhist 1790–1830	Kubhist 1895–1903

Table 3: Time periods of subcorpora for English and Swedish from which annotation data was sampled.

4.1. Data Annotation

To test the different annotation paradigms, we started from the existing resampled DWUG dataset for English and Swedish (Schlechtweg et al., 2024). In particular, we focused on two English target words (*graft* and *twist*) and two Swedish target words (*motiv* 'motif' and *konduktör* 'conductor'). The choice of languages, and thus dataset, was pragmatic as we plan on annotating both in our future work.

In the *resampled* DWUG dataset (DWUG henceforth), each of these four words includes up to 50 usages, one half drawn from period T1 and T2 respectively. The corresponding time spans and corpora are reported in Table 3. For each word there is a maximum of 1,225 usage pairs that can be annotated. However, only a subset of these pairs were actually annotated: on average, approximately 35% for English and 60% for Swedish. In total, the number of annotated usage pairs amounts to 855 for *motiv*, 981 for *konduktör*, 342 for *graft*, and 509 for *twist*.

We extended the DWUG dataset (U-U) by annotating the usages of the four target words, employing the U-S and U-SI paradigms[3]. For Swedish, we recruited six native Swedish annotators. For English, we selected three annotators: two native speakers of British English and one native speaker of Canadian English.[4] We recruited both sets of annotators such that they had studied either linguistics or languages at the university level. Swedish annotators were paid standard hourly wages for research assistants, while English annotators received a lump sum voucher.

Sense Inventories One of the challenges in both the U-S and U-SI paradigms is defining the set of senses to be used for annotation. In particular, for diachronic annotation it is essential to ensure that the sense inventory also includes senses that are now obsolete or have disappeared, as well as those that have been introduced more recently.

[3]The dataset and the code are available on Github

[4]Our English annotators were a convenience sample.

Dictionaries rarely capture all the senses of a word; in practice, it is often necessary to aggregate information from multiple lexicographic sources. Moreover, sense definitions should be as precise as possible, while minimizing overlap between definitions.

It is also necessary to adjust the granularity of the senses to the specific research objective. Some dictionaries adopt an extreme splitter approach, listing dozens of senses for a single word. This creates difficulties in both paradigms. In the U-S paradigm, the number of examples to annotate with additional senses. In the U-SI paradigm instead, the annotator's cognitive load increases, as they must select the correct sense from a large set of closely related definitions.

Given the sensitivity and complexity of constructing a sense inventory, we believe that the only way to ensure its reliability is through careful manual curation[5]. For this paper, a member of our team curated the sense inventories for the four target words, drawing on a combination of definitions from SO and SAOB for Swedish [6] and from the OED for English [7]. In some cases, the definitions were slightly modified to make them more general and to cover a broader semantic field. The definitions for the four words are reported in Table 2.

Annotation Setting For the annotation process, we used Qualtrics[8]. For each language, we created a survey including the two target words. The survey presents all examples sequentially, first for the U-S paradigm and then for the U-SI paradigm. For all annotators of a language, we use the same order of paradigms, words, and usages so that, when annotators begin with the U-S paradigm, they are not yet familiar with the complete sense inventory but instead discover them progressively. The annotation order was as follows:

- Swedish:
  Usage-Sense (konduktör) → Usage-Sense (motiv) → Usage-Sense Inventory (konduktör) → Usage-Sense Inventory (motiv)

- English:
  Usage-Sense (twist) → Usage-Sense (graft) → Usage-Sense Inventory (twist) → Usage-Sense Inventory (graft)

[5]Manual curation of the sense inventory also allows a hypothesis-driven approach: If the aim is to study certain aspects of change, then more fine-grained senses can be included to reflect those aspects while other senses can remain more coarse.

[6]https://svenska.se

[7]https://www.oed.com

[8]https://www.qualtrics.com

Word	**Lang.**	α**(U-S)**	α**(U-SI)**
konduktör	swedish	0.818	0.759
motiv	swedish	0.888	0.882
twist	english	0.552	0.593
graft	english	0.739	0.795

Table 4: **Inter-annotator agreement**. Krippendorff's α for Usage-Sense (U-S) and Usage-Sense Inventory (U-SI) annotation.

Given the high number of examples to be annotated, annotators were given two weeks to complete the entire survey. During this period, they were allowed to pause the annotation process and resume it later from where they had left off. We recorded several metadata variables, including the time elapsed between opening an example and the first click, the last click, and the final submission, as well as the total number of clicks.

In the U-S paradigm, annotators are provided with a word's usage and a single candidate sense of that word. They are asked to answer yes or no as to whether the definition corresponds to the meaning of the word in that specific usage. Annotators may also select *Cannot decide*. In that case, they must choose between two options: *I cannot decide because the text is noisy (e.g., OCR errors)* or *I cannot decide because the text is ambiguous*. This choice is exclusive: if annotators select Cannot decide, they cannot simultaneously answer yes or no.

In the U-SI paradigm, annotators are provided with a word in context and the full inventory of senses. They are also given the option to indicate that none of the provided senses adequately captures the meaning of the word (*Other*), in which case they must supply a textual explanation of the intended meaning.

Annotators are allowed to select *Other* simultaneously with one or more of the predefined senses. By contrast, the choice of Cannot decide (again divided into *noise* and *ambiguity*) remains exclusive in this paradigm as well. The annotation guidelines and the interface were provided in the annotators' native language.

4.2. U-S vs U-SI paradigm

Agreement To assess whether the U-S and U-SI paradigms yield comparable annotations, we examine the bootstrap estimates of Cohen's κ of the intra-annotator agreement reported in Tables 5 and 6. These estimates directly quantify cross-paradigm agreement at the annotator level, clustered by usage with 95% confidence intervals. The inter-annotator agreement instead is reported in Table 4

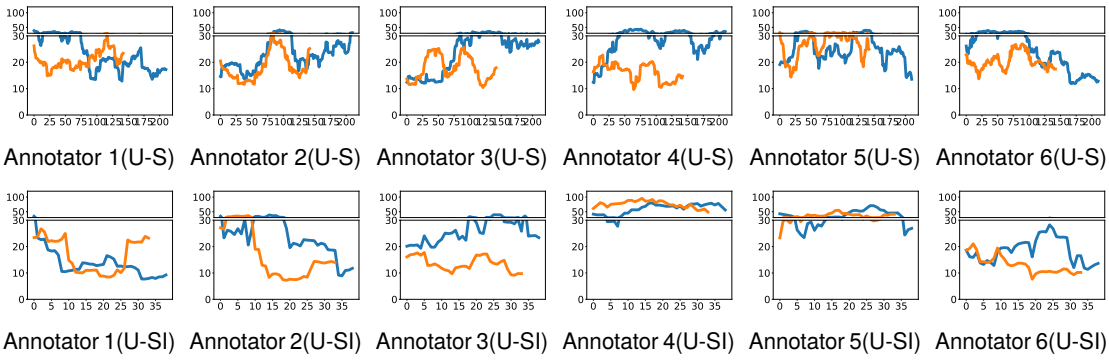

Figure 1: Time in seconds (y-axis) spent by each annotator (shown in the columns) and annotation paradigm (U-S first row, U-SI second row), relative to the number of annotated examples (x-axis) for the words konduktor and motiv.

	konduktör	motiv
ann1	1.000 [1.000, 1.000]	0.981 [0.935, 1.000]
ann2	0.853 [0.763, 0.929]	1.000 [1.000, 1.000]
ann3	0.797 [0.691, 0.892]	0.953 [0.892, 1.000]
ann4	0.891 [0.805, 0.956]	0.907 [0.812, 0.979]
ann5	0.762 [0.656, 0.860]	0.938 [0.880, 0.985]
ann6	0.946 [0.882, 1.000]	0.915 [0.809, 0.983]

Table 5: **Intra-annotator agreement.** Bootstrap estimates of Cohen's κ (U-S vs U-SI), clustered by sentence, with 95% confidence intervals ($B_{\text{eff}} = 2000$).

	twist	graft
ann1	0.633 [0.455, 0.787]	0.948 [0.896, 0.987]
ann2	0.780 [0.651, 0.891]	0.966 [0.926, 1.000]
ann3	0.905 [0.791, 0.980]	1.000 [1.000, 1.000]

Table 6: **Intra-annotator agreement.** Bootstrap estimates of Cohen's κ (U-S vs U-SI), clustered by sentence, with 95% confidence intervals ($B_{\text{eff}} = 2000$).

For the Swedish targets, intra-anntoator agreement between paradigms is consistently high. For *konduktör*, κ ranges from 0.762 to 1.000 across annotators. For *motiv*, κ values are even higher, ranging from 0.907 to 1.000. These values indicate substantial to almost perfect agreement, showing that U-S and U-SI produce highly similar annotation outcomes for these words. A similar pattern holds for the English word *graft*, where κ values range from 0.948 to 1.000.

The only notable deviation appears with *twist*, where κ ranges from 0.633 to 0.905. While two annotators show high agreement (≥ 0.78), one annotator exhibits only moderate agreement ($\kappa = 0.633$). Overall, the results do not prove strict equivalence between U-S and U-SI. However, the consistently high κ values for three of the four target words

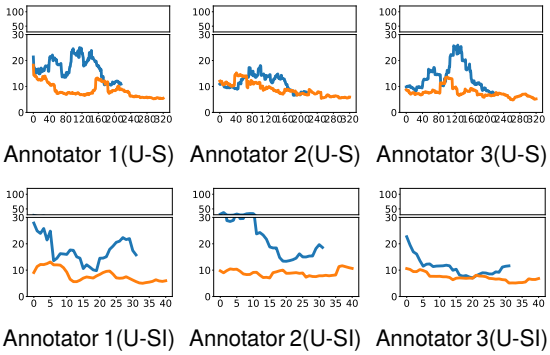

Figure 2: Time in seconds (y-axis) spent by each annotator (shown in the columns) and annotation paradigm (U-S first row, U-SI second row), relative to the number of annotated examples (x-axis) for the words twist and graft.

demonstrate strong practical comparability. Additionally, we note that almost all annotators are more consistent between U-S and U-SI for the second word they annotate, which could be an effect of learning the task. E.g., Swedish annotator 5 has a 0.762 agreement rate for *konduktör* but 0.938 for *motiv*.

Annotation Time To analyze annotator effort over time, we extracted page-level completion times from Qualtrics logs for both paradigms (U-S, U-SI) and both languages (English, Swedish). For each item, the system records the time (in seconds) between page load and submission, which we use as a proxy for annotation duration. We aggregate these times per annotator and per target word, forming sequences of annotation times across items.

Two limitations apply. First, task order was fixed (U-S before U-SI), which may create an order effect if familiarity gained during U-S facilitates U-SI. Second, recorded times likely overestimate true

effort, as pages could be left open before submission. The timings should therefore be interpreted as approximate. To mitigate extreme delays, annotation times were capped at 120 seconds. We then applied moving-average smoothing separately per paradigm: a trailing window of 10 observations for U-SI and 30 for U-S, reflecting longer sequences in the latter. Separate time-course plots were generated per annotator and language (Figures 1 and 2).

Across languages, a consistent pattern emerges after the warm-up phase: U-SI is generally faster and more stable than U-S. Once the sense inventory is internalized, U-SI becomes a direct selection task. In contrast, U-S requires sequential evaluation of each sense, increasing the number of micro-decisions per item.

In Swedish (Figure 1), convergence is relatively smooth, particularly for motiv, where half of the annotators quickly reach low, stable times under U-SI. In English (Figure 2), graft shows similarly efficient behavior, especially under U-SI. By contrast, twist exhibits greater variability and spikes, particularly under U-S, consistent with its lower cross-paradigm agreement.

4.3. Comparison to the U-U paradigm

For each usage, we derive sense labels by aggregating individual annotator judgments through majority voting. For both U-S and U-SI, we count the frequency of each individual sense across annotators and assign the sense with the highest frequency to the usage, applying a random tie-breaking where necessary.

The special label *cannot decide* is not treated as valid assignment in the comparative analysis. We restrict evaluation to usages that receive a valid cluster assignment in all three paradigms (U-U, U-S, U-SI).

Each sense has a corresponding group consisting of all usages that have the sense as a label. To quantify similarity between the diverse usage groupings, we compute the Adjusted Rand Index (ARI) pairwise between U-S, U-S, and U-SI groupings. To assess how the different paradigms capture meaning change, we compute the Graded Change Detection (GCD) score by first deriving, for each method, the distribution of groupings across two time periods and then calculating the Jensen–Shannon divergence between these distributions; higher divergence indicates stronger redistribution of senses across time and is interpreted as greater semantic change.

The results reported in Table 7 show that twist yields the lowest ARI scores across all pairwise comparisons, which mirrors the generally lower agreement observed for this word in all annotation paradigms. In contrast, graft presents a different pattern: while U-S and U-SI are in perfect agreement with each other, their clustering diverges from U-U, resulting in only moderate ARI scores. A close qualitative analysis reveals that in the U-U paradigm one cluster contains a mixture of examples belonging to sense 1 and sense 2, which, although conceptually related through the notion of transplantation, refer to clearly distinct domains, horticulture and medicine. This merging is likely due to annotators frequently assigning an average relatedness score of 3 to such cross domain pairs, which, given the clustering threshold, leads to their aggregation into a single cluster. As a consequence, the semantic change scores differ substantially, with a GCD value of 0.467 in U-U compared to 0.815 in both U-S and U-SI, indicating that U-U underestimates the degree of change for graft by smoothing over a redistribution between domain specific senses across time.

5. Conclusion

In this paper, we examined three annotation paradigms for capturing word meanings over time: Usage–Usage (U–U), Usage–Sense (U–S), and Usage–Sense Inventory (U–SI), with particular focus on the trade-off between scalability, reliability, and expressive power. Through a pilot study on two English and two Swedish target words, we empirically compared these paradigms with respect to inter- and intra- annotator agreement, annotation effort, and their ability to model diachronic semantic change.

Our findings demonstrate that the U–SI paradigm offers a compelling balance between efficiency and quality. While U–U provides high theoretical expressive power by not constraining annotators to predefined sense inventories, in practice this advantage is limited by relatedness labels and the selection of thresholds as well as parameters for the clustering based on heuristics.

Between the two inventory-based paradigms, U–SI emerges as the more efficient option. Cross-paradigm agreement between U–S and U–SI was consistently high for three of the four target words, indicating that the two methods produce largely comparable sense assignments. At the same time, annotation time analyses indicate that U–SI requires less annotator effort once the sense inventory has been internalized, reflecting its single-decision structure compared to the repeated micro-decisions required in U–S. Given that both paradigms scale linearly in the number of usages, the lower cognitive and temporal cost of U–SI makes it preferable for large-scale diachronic annotation projects.

Importantly, **our results do not establish strict methodological equivalence between**

Word	ARI			GCD		
	U-S vs U-U	U-SI vs U-U	U-S vs U-SI	U-U	U-S	U-SI
konduktör	0.703	0.731	0.954	0.603	0.748	0.748
motiv	0.789	0.877	0.895	0.202	0.255	0.307
twist	0.368	0.434	0.717	0.508	0.461	0.445
graft	0.508	0.508	1.000	0.467	0.815	0.815

Table 7: ARI and GCD scores by word and annotation paradigm.

paradigms. Variation across words, most notably for twist, highlights that lexical complexity and sense granularity influence human agreement.

Overall, our pilot study supports replacing usage-pair-based annotation with a usage–sense-inventory framework in large-scale diachronic semantic annotation. U–SI maintains high reliability, reduces annotation time, scales linearly, and preserves meaningful distinctions in semantic change modeling. These properties make it particularly well suited for our long-term objective of annotating word meanings across multiple time periods while operating under realistic budget constraints.

Acknowledgements

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021). We further thank Felix Morger for choosing the senses.

6. Bibliographical References

Eneko Agirre and Aitor Soroa. 2007a. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Eneko Agirre and Aitor Soroa. 2007b. SemEval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 7–12, Prague, Czech Republic. Association for Computational Linguistics.

Nikhil Bansal, Avrim Blum, and Shuchi Chawla. 2004. Correlation clustering. *Machine learning*, 56:89–113.

Andreas Blank. 1997. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, volume 285 of *Beihefte zur Zeitschrift für romanische Philologie*. Niemeyer, Tübingen.

Pierluigi Cassotti, Lucia Siciliani, Lucia Passaro, Maristella Gatto, and Pierpaolo Basile. 2023. WiC-ITA at EVALITA2023: Overview of the EVALITA2023 Word-in-Context for ITAlian Task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2023)*, Parma, Italy. CEUR.org.

Pierluigi Cassotti and Nina Tahmasebi. 2025. Sense-specific historical word usage generation. *Transactions of the Association for Computational Linguistics*, 13:690–708.

Jing Chen, Emmanuele Chersoni, and Chu-ren Huang. 2022. Lexicon of Changes: Towards the Evaluation of Diachronic Semantic Shift in Chinese. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 113–118, Dublin, Ireland. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023a. ChiWUG: A graph-based evaluation dataset for Chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023b. ChiWUG: A Graph-based Evaluation Dataset for Chinese Lexical Semantic Change Detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99, Singapore. Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Philip Edmonds and Scott Cotton. 2001a. Senseval-2: Overview. In *Proceedings of The Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, pages 1–6.

Philip Edmonds and Scott Cotton. 2001b. SENSEVAL-2: Overview. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 1–5, Toulouse, France. Association for Computational Linguistics.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Andrey Kutuzov and Lidia Pivovarova. 2021a. RuShiftEval: A Shared Task on Semantic Shift Detection for Russian. In *Proceedings of the Conference on Computational Linguistics and Intellectual Technologies (Dialogue)*, 20, (online). RSUH.

Andrey Kutuzov and Lidia Pivovarova. 2021b. Three-part Diachronic Semantic Change Dataset for Russian. In *Proceedings of the 2nd International Workshop on Computational Approaches to Historical Language Change 2021*, pages 7–13, Online. Association for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic Semantic Change Dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Daniel Loureiro, Aminette D'Souza, Areej Nasser Muhajab, Isabella A. White, Gabriel Wong, Luis Espinosa-Anke, Leonardo Neves, Francesco Barbieri, and Jose Camacho-Collados. 2022. TempoWiC: An evaluation benchmark for detecting meaning shift in social media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3353–3359, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 240–243. Morgan Kaufmann / MIT Press. SemCor: a WordNet-sense tagged corpus based on the Brown Corpus, manually annotated.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 Task 12: Multilingual Word Sense Disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Pierluigi Cassotti, Bill Noble, David Alfter, Sabine Schulte Im Walde, and Nina Tahmasebi. 2024. More DWUGs: Extending and evaluating word usage graph datasets in multiple languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14379–14393, Miami, Florida, USA. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic Usage Relatedness (DURel): A Framework for the Annotation of Lexical Semantic Change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021a. DWUG: A large resource of diachronic word usage graphs in four languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021b. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task in senseval-3. In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pages 41–43.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A Shared Task on Semantic Change Discovery and Detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Gaze Behaviour & Conversation Unfolding in the HCRC Map Task Corpus

Anaïs Claire Murat and Carl Vogel
Trinity Centre for Computing and Language Studies
Trinity College Dublin, the University of Dublin
{murata,vogel}@tcd.ie

Abstract

Dialogue interactions have varied internal structure, with flow varying, *inter alia*, in face of both difficulty and agreement. This study investigates eye-gaze in the linguistic progression of interactions. We observe the relation of gaze to illocutionary functions of turns through dialogue acts, and to how turns present "new" or "old" content, through lexical entropy and repetition. Results on the HCRC Map Task corpus, enabled by an event alignment annotation method described, show how gaze is related to linguistic progression. A gaze towards the conversation partner at the end of a turn tends to align with complexity and difficulties being expressed in the turn, while keeping gaze down at the map is more typical of obstacle- and disagreement-free interactions. Addressees who look up or off at the start of a turn show evidence of lexicon adaptation to gaze values.

Keywords: gaze, entropy, repetition, dialogue acts, cross-modal event alignment, interaction analysis

1. Introduction

Most Human-Human interactions can rely on at least two modalities: the auditory and the visual ones. These modalities complement each other, and within the visual modality, eye gaze has been shown to be a catalyst for smoother interactions. Gaze has been explored as a tool for turn-taking and for signalling attention or emotional states – multiple functions are evident. This paper examines eye-gaze behaviour in dyadic conversation in relation to the overall purpose of turns through dialogue acts, and abstractions of their content; through computed entropy and repetition.To do so, we explore the HCRC Map Task corpus and, more specifically, the gaze behaviour of both participants at the start and end of turns, to statistically assess the co-presence of gaze and turn qualities. We are interested in whether the choice of dialogue act and two measures of linguistic unfolding – turn by turn repetition and categories entropy behaviour (i.e. increase, decrease, or stable states) interact with gaze behaviour. Fundamental to this analysis is the temporal alignment of the distinct modalities within a representation framework that supports counting and other forms of measurement of events within each modality and their overlap; in support of this we use an annotation method called Beginning-Middle-End (BME), for the facets of events noted. We use observational methods to develop generalisations over statistically significant interactions in categorial representations of gaze types and in sec. 5 discuss their coherence, both in this research work and with respect to the current literature.

2. Background Literature

Gaze has been studied as a paralinguistic tool. It has been noted in turn taking management (Bavelas et al., 2002). Initially, it can signal communicative intent (Cary, 1978; Pfeiffer et al., 2013) and resolve cases of simultaneous starts (Zima et al., 2019). Although Kendon (1967) and Ho et al. (2015) reported that speakers avert their gaze briefly after turn-taking and look back at the listener at the turn end, Beattie (1978) reported finding more averted gaze at turn ends, and that turns containing gaze up at their end were not correlated with prompt answers. In fact, Degutyte and Astell (2021) reviewed 29 papers from the past 50 years in the gaze literature and found no single pattern, as had been believed since Kendon (1967). Additionally, there seem to be influential factors beyond the mere difference in dialogue nature (free flowing conversations or question/answers (Rossano, 2010; Beattie, 1978)) or study design (Kendon, 1978).

Gaze may also provide insights into the emotional state of participants. A direct gaze at the conversational partner may signal attention, engagement (Bavelas et al., 2002; Kendon, 1967; Goodwin, 1980; Argyle and Cook, 1976). Gaze favours success in task-based conversations by supposedly preventing and repairing conversation breakdowns (Boyle et al., 1994; Streeck, 2014; Nicholson et al., 2005). In map task corpora, Boyle et al. (1994) notably showed that direct gaze were more frequent in instances of difficulty (i.e., mismatch between two maps) than otherwise, Doherty-Sneddon (1995) that one of the gaze function was to access feedback, and Nicholson et al. (2005) that instruction givers were more likely to look up at the follower

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 99–110
12 May 2026.

after the latter expressed difficulties. On the other hand, an averted gaze could help identify problematic interruptions by the listener (Brône et al., 2017; Spaniol et al., 2023) and signal unwillingness to pursue an interaction in the current terms (Jokinen et al., 2010).

Additionally, gaze has been scrutinised in relation to the content of the conversation. The deictic function of referential gaze has been used to disambiguate and ground text (Mehlmann et al., 2014), and has been shown to be more efficient than speech for solving referring expressions or searching tasks (Brennan et al., 2008; Neider et al., 2010; Hanna et al., 2020). Directed gaze has been shown to mark questions in Italian (Rossano, 2010), and to indicate preferred answers in polar questions (Kendrick and Holler, 2017). Information may also be encoded in gaze. Mutual gaze windows were shown to accompany minimal responses (such as back-channels or continuers) (Bavelas et al., 2002), and Edlund et al. (2012) showed that, indeed, 3^{rd} party's gaze of a conversation is likely to reveal which one of the speakers is providing the most information.

In relation to entropy, it has been proposed by Genzel and Charniak (2002) that entropy measures ignoring context should increase for sentences in relation to their ordinal position in a text, as a consequence of an entropy constancy rate principle. Evidence for this effect has been provided using gaze during reading as an indicator of cognitive effort associated with higher entropy but not the position of a sentence within a text (reading of sentences does not ignore contexts) (Keller, 2004). Using gaze events as tokens and gaze categories as types, the entropy rate constancy principle has also been demonstrated for listener gaze behaviours in interactive dialogue (Wang et al., 2024). Our work relates specific gaze behaviours of speakers and listeners to categories of lexical entropy change, local repetition, and dialogue acts that interact with cognitive effort (e.g. Doubt vs Explanation).

3. Method

3.1. Description of the HCRC Corpus

This research uses the HCRC Map Task Corpus (Thompson et al., 1993). It is made of dyadic conversations whose task involves closing an information gap between the interlocutors: one designated party is to linguistically indicate a path to follow (the Giver "G") and the dialogue partner is to reproduce that path (the Follower "F"), both with reference to imperfectly matching maps. The entire corpus includes 128 dialogues; half of the corpus had participants having access to mutual eye contact, in the other half, eye contact was obscured with a screen. 46 dialogues had gaze annotations but we focus in this paper on the 31 annotated conversations which allowed for direct eye-contact.

We reused the timestamped annotations as provided in the XML version of the dataset. [1] The gaze tier is continuously annotated and annotations can take one of the three following values: "gaze down" when looking at the map, "gaze up" when looking at the other participant, "gaze off" when looking elsewhere. We examine tuples of co-occurring G and F gazes, also referred as "paired gaze" in what follows.

Similarly, we reused the turns annotations as available in the dataset. Due to the hierarchical organisation of the HCRC annotations, the timestamps for the turn transcripts were retrieved from the timed-units (i.e. words): matching the onset of the first word of a turn, and the offset of the last.

The corpus also includes dialogue act (DA) augmentation for every turn.

As a matter of indication, the HCRC corpus includes 64 talkers (half male, half female), mostly from Glasgow, Scotland; half of the conversations included acquaintances, and each participant participated in 4 conversations, alternating between the role of Giver and Follower.

3.2. Data selection

We used the 31 conversations which were annotated with gaze values and which allowed for direct eye-contact between the participants (in total, 10894 paired gazes, 5943 turns). Some gaze annotation anomalies were spotted. For instance, we noticed several gaze values for a single millisecond, and some starting times occurring after the endtimes. For privacy reasons, the video recordings of the corpus are not available outside the HCRC community, we therefore could not resolve such anomalies and had to remove these gazes from the dataset, along with any turn and other gaze that could have been impacted. The final dataset comprised 10761 paired gaze and 5889 turns.

3.3. BME Modality Alignment

The BME method is an alignment technique which focuses on events (Beginnings, Middles, and Endings of annotations) to reveal the structure of multimodal datasets (Murat et al., 2026). The method is similar to that used by other researchers investigating multi-modal dialogue (Magnusson, 2000; Hunyadi et al., 2018; Magnusson, 2020). Assuming each modality is captured on a different tier, such as what annotation software like Elan (Wittenburg

[1] Data available https://groups.inf.ed.ac.uk/MapTask/index.html, last accessed in March 2026.

Turns	Gaze
	B - Gaze 1
B - Turn 1	M - Gaze 1
M - Turn 1	E - Gaze 1
M - Turn 1	B - Gaze 2
E - Turn 1	M - Gaze 2
	E- Gaze 2

Table 1: Example of alignment of 1 Turn and 2 gaze annotations using the BME method

et al., 2006) produces, it arranges multi-modal data based on the relative onsets and offsets of the tiers' annotations. It retrieves the beginning time (B) and ending time (E) of each annotation and organises them relatively to each other. Middles (M) mark, for a given annotation, the B and E lines of another tier occurring in-between its onset and offset. Ms can then be enumerated as a measure of duration to account for how many events happened during a given annotation: that is, event durations and overlaps may still be reckoned in terms of absolute time elapsed, or by the count of intervening events. Table 1 provides an example of how three overlapping annotations could be organised. This fictive example reads "Gaze 1 started outside a turn, Turn 1 started inside Gaze 1, Gaze 1 ended inside Turn 1, Gaze 2 started inside Turn 1, Turn 1 ended inside Gaze 2, Gaze 2 ended outside a turn." It can also be read as "1 turn event (Middle) occurred in Gaze 1, 2 gaze events occurred in Turn 1, and 1 Turn event occurred in Gaze 2".

The annotation alignment technique supports analysis event type interactions via the counts of the sorts of things that happen in one tier at the same time as starts or stops in another tier, such as used in contingency tables.

Timestamps of interest are those of turns and gaze annotations, as described in section 3.1.

3.4. Illocutionary function through Dialogue Acts

To generalise dialogue act (DA) behaviour during conversations and to have sufficient observations to conduct meaningful statistical tests (i.e., using χ^2-tests, limiting the number of expected values lower than 5), we classify basic DAs into supra-categories INSTRUCTION, CONFIRMATION, EXPLANATION, DOUBT, and REJECTION (see Table 2). INSTRUCTION is only representative of the Giver's utterances as it represent instances where the Giver commands the Follower to carry out an action (e.g., take a certain route, move to a certain landmark, etc.), CONFIRMATION corresponds to instances where one validates the position of the other, EXPLANATION where one develops informative answers, DOUBT where one expresses the need to confront their understanding or position, and REJECTION when one answers the other by the negative. Further description of the individual dialogue acts can be found in the manual Carletta et al. (1996).

3.5. Characterizing the Lexical Development of a Conversation

In order to confront the gaze behaviour to the unfolding of the conversation, we need to find ways to measure lexical choice over time. Similarly to the method (Murat and Vogel, 2025) used to track engagement throughout conversations via linguistic cues, we study word use on two scales: firstly, Lexical entropy from the start to the end of a conversation, measured at turn boundaries; secondly, Linguistic repetitions from one turn to another.

3.5.1. Transcript Processing

In a pre-processing step the HCRC transcripts were normalised: we wrote a Python script to remove punctuation, lowercase letters, split "pronoun + verb" contractions (e.g "you're" $\Rightarrow$ "you", "are"), and split contracted negated auxiliaries (e.g "can't" $\Rightarrow$ "can", "not"). Additionally, using the Spacy small model for English (Montani et al., 2023), we noted for each lexical token its part of speech category and, from it, determined whether the token belonged to open-class (OC) or closed-class (CC) words (as exhaustive and mutually exclusive possibilities; i.e., if their part-of-speech tag was either "NOUN", "VERB", "ADJ" (adjective), or "PROPN" (proper noun) they were considered as open-class; and closed-class otherwise). We distinguish thusly because, by their frequency, closed-class items are more likely to be repeated than open-class items.

3.5.2. Entropy

Initially adapted from Gnjatović et al. (2018), we reuse the method developed in Murat and Vogel (2025) to measure entropy evolution in sequential lexical choice. For a dialogue D and lexical types (v), $V = \langle v_1, v_2, ..., v_k \rangle$ is the *sequence* of lexical items used. D is defined as a sequence of turns T; $D = T_1, T_2, ..., T_n$ (abbreviated T_1^n) which are *sets* T of lexical types produced in the turn. The (cumulative) entropy H of a dialogue from its opening (1) to its nth turn (T_1^n) is based on Shannon Entropy:

$$H(T_1^m) = -\sum_{1}^{k} P(v_i) \log_e P(v_i) \quad (1)$$

The probability $P(v_i)$ is estimated from the relative frequency of the type v_i in the dialogue sequence from the start to the point of measurement.

CONFIRMATION (10888)	DOUBT (6437)	EXPLANATION (4267)	INSTRUCTION (4267)	REJECTION (884)
Acknowledge (5599) Reply_y (3229) Ready (2060)	Check (2133) Align (1778) Query_yn (1758) Query_w (768)	Clarify (1192) Explain (2160) Reply_w (915)	Instruct (4267)	Reply_n (884)

Table 2: Summarised DA into the supra-categories CONFIRMATION, DOUBT, EXPLANATION, INSTRUCTION, and REJECTION. In brackets are the numbers of turns of each kind.

These calculations can be evaluated on a sequential basis and assessed either word after word, or turn after turn. It is relevant to examine the difference between the entropy at the end of any turn h_n and at the end of the immediately prior turn h_{n-1}:

$$\Delta H = \Delta(h_n - h_{n-1}) \quad (2)$$

Eq. 2 thus allows to tell that a certain turn led to an increase or a decrease in entropy. An increase is typical of scenarios adding up new or relatively rare words, while a decrease in entropy suggests that the turn reuses well-known and frequent words. In this work, we look into entropy calculated for the overall conversation vocabulary (H_C), but also at the speaker's level by only accounting for the vocabulary *they* have produced (H_S), and investigate the impact of a turn on both (respectively ΔH_C , and ΔH_S). For this study, we categorise turns (exhaustively and mutually exclusively) thusly :

- $\Delta H_C < 0$, $\Delta H_S < 0$: the turn leads to a decrease in both conversational (ΔH_C) and speaker entropy (ΔH_S).
- $\Delta H_S < 0 < \Delta H_C$: the turn only leads to a decrease in speaker entropy.
- $\Delta H_C < 0 < \Delta H_S$: the turn only leads to a decrease in conversational entropy.
- $\Delta H_C = \Delta H_S$: the turn leads to a similar increase in conversational and speaker entropy.
- $\Delta H_S > \Delta H_C > 0$: the turn leads to an increase in both conversational and speaker entropy, greater increase for the speaker.
- $\Delta H_C > \Delta H_S > 0$: the turn leads to an increase in both conversational and speaker entropy, greater increase for the conversation.

In a nutshell, this measure of entropy allows us to track the overall reuse of tokens in a dialogue. Closed-class words tend to be often used and have a negative impact on entropy; they add little information. In contrast, novel content words lead to entropy increase, and their impact on entropy diminishes as novelty fades. To address conversation dynamics that related to closer events, such as short-term planning and spontaneous reactions, we add another dimension, whose window is smaller: direct repetitions from one turn to another.

3.5.3. Repetitions

There is not one way to quantify repetition in the literature. Distinctions occur in the nature of what is being counted (lexical vs syntactic repetition (Reitter and Moore, 2007; Healey et al., 2014), bag of words vs set of words (Murat et al., 2022), lemma vs tokens (Reverdy and Vogel, 2017), different size of n-grams, etc.), but also in terms of calculations where there can be strict counts as well as length- or frequency-based normalisations (Mekhaldi, 2006). Furthermore, assessing the overall similarities of all interlocutors' productions is not enough: it does not allow one to draw any conclusion on the impact of what one previously said on the current production (Mehler et al., 2011). Temporal cues are therefore crucial and we do so by comparing strictly ordered turns (Reverdy et al., 2020; Murat et al., 2022).

Given the nature of this work, where repetition is an object of study in its relation to gaze, we qualify repetition by the strict presence or absence of a repeated token between the current turn and the preceding one, while also relying on sub-classes of repetition. We make a distinction between, on the one hand, self-repetitions (SR), the speaker is repeating content from their last turn, which might signal grounding or involvement; and on the other hand, other-repetitions (OR), when the speaker is repeating what the addressee said in their last turn, which can be an example of discourse planning or floor-holding (Koutsombogera and Vogel, 2019). We also track as a separate category turns accounting for both self- and other-repetitions (OR-SR) as they might signal even further efforts to co-build through the conversation. As closed-class words are more likely to be repeated than content words, we also made a distinction between repetitions which only counted closed-class words (CC) and turns which also accounted for open-class repetition (OC).

4. Results

The data mining was conducted through identification of statistically significant co-occuring behaviours through the use of χ^2-tests. As such, 12 χ^2-tests were performed (F or G's turns $\times$ DA, entropy or repetitions $\times$ at the start or at the end of a turn). We test cross-categorial interactions only

Table 3: Abbreviation key

Short form	Concept
CC	tokens in closed-class categories
OC	tokens in open-class categories
OR	other-repetition
SR	self-repetition
G	information giver
F	information follower
H_C	entropy for conversation
H_S	entropy for speaker
Δ	change

where less than 20% of expected values were less than 5 (the maximum we have is 13% of cells < 5 for the entropy test in relation to F's turns), and none was less than 1. To palliate this exigence, we had to merge all pairs of gaze including at least one gaze off into one single category (may they be from the Follower or the Giver).[2] Furthermore, considering the number of tests and the relatively high number of categories, we adjusted the p-value of 0.05 by the number of dimensions d ($\alpha = 0.05/d$) in each table for determining significance. Similarly, the critical value z for residual was calculated $P(Z > z) = \frac{\alpha}{2}$. For the sake of space, we do not report non-significant tests. The 8 tables across fig. 4, 5 & 6 report the 8 χ^2-tests showing significant interactions in these conservative conditions.

4.1. Dialogue Acts & Gaze

Exploring dialogue acts enables assessment of whether the evident purpose of an utterance is reflected in the gaze behaviour of interlocutors. The INSTRUCTION category was not included in tests about Follower's turns as, by definition, Followers had little chance to utter such turns.

All 4 χ^2-tests (DA $\times$ Gaze-pairings for $\{F, G\} \times$ {Start, End}; Table 4) showed significant interactions: Giver and Follower's paired gazes at the start and at the end of each turn interact with the DA being performed. When G is speaking, INSTRUCTIONS are very likely to start while both F and G are looking down ($r = 7.82$) and unlikely in any other context ($r < -3.48$). The end of an INSTRUCTION turn, however, is not typical of one gaze category, but has significant dearth where G is looking down and F, up ($r = -4.94$). An expression of DOUBT by G is rather expected when G is looking up but F is looking down ($r = 8.80$) and rather not expected if G is looking down ($r < -3.46$). F's DOUBT turns, however, can also appear when both are looking down ($r = 3.75$), on top of not being likely to occur when F is looking down but G is looking up ($r = -6.49$). For both F and G, the end of a DOUBT is likely occurring in the situation where the non-speaking participant is now looking down and the other looking up ($r > 10.47$). EXPLANATION turns are mostly significantly characterised by their ending behaviour: the speaker tends to look up ($r = 1.24$ for G when both look up, but $r > 3.11$ in all other cases). For CONFIRMATION turns – which intervene in response to the other participant's content, they tend to start in a context where the addressee is looking up but the speaker is already looking down (particularly significant for G's case $r = 6.03$), and they end in a context where the speaker is looking down and the other participant is looking either down or up ($r > 3.86$). On the other hand, REJECTION turns have similar behaviour, to the exception that in G's case, they might both be looking up (hence, making eye contact) at the start ($r = 4.42$).

These results suggest that both participants looking down at the object of interest (the map) is a sign of conversation progressing without obstacles: it is the context in which further INSTRUCTIONS are likely provided ($r = 7.82$) or CONFIRMATION turns can end ($r > 3.86$). F's DOUBT turns are also likely to start in that category ($r = 3.75$). This distinction might come from the need for F to interrupt and express doubts as they lose track of G's elaborations.[3]

Gazes up seem to sometimes signal conversation flow disruption. DOUBT turns end with their speaker gazing up at the partner, and the potential responses, a CONFIRMATION or a REJECTION still start during a look up, while the responder looks down at the map. At the end of such turns, CONFIRMATIONS can lead the non-speaking person to look back down at the map ($r > 3.8$) or not ($r > 5.94$) while REJECTIONS, which are more divisive answers, only favour the non-speaking participant to look up ($r > 6.09$). EXPLANATIONS are also characterised by the speaker looking up to the other participant at the end, possibly checking for feedback ($r > 3.11$).

This section has focused on dialogue acts, i.e, the apparent intentions with every turn start or stop, in relation to the gaze behaviours of the interlocutors. Shared gaze behaviours (both up or both down) distribute differently across DAs than asymmetric behaviours, varying by whether the turn owner is the information Follower or Giver. The next two sections examine gaze in relation to lexical novelty.

4.2. Entropy & Gaze

Entropy allows quantification of the novelty of vocabulary of one turn over the conversation: are the words that have just been uttered relatively new to

[2] In practice, 332 of the gaze-offs at the start of a turn were due to F looking off; against only 73 due to G. At turn ends, these numbers are respectively 327 and 71.

[3] A Pearson's χ^2 test revealed that, indeed, F's DOUBT turns are more likely than G's to occur in the middle of the other participant's turns ($p = 3.775e^{-12}, r = 6.98$).

Start	G's Turns					F's Turns				
G's Gaze	Down		Up		Off	Down		Up		Off
F's Gaze	Down	Up	Down	Up		Down	Up	Down	Up	
Instruction	**7.82**	**-5.52**	**-4.93**	**-3.48**	0.74					
Doubt	**-3.46**	**-4.07**	**8.80**	1.15	0.90	**3.75**	3.05	**-6.49**	1.31	0.68
Explanation	-0.50	1.04	-0.87	-0.50	0.84	-0.49	-1.51	2.08	-1.30	-0.56
Confirmation	-2.52	**6.03**	-2.10	1.42	-1.76	-1.70	-1.69	2.75	-0.63	0.29
Rejection	**-5.38**	**6.90**	-0.22	**4.42**	-1.68	**-3.67**	0.03	**4.45**	1.04	-1.35

(a) **(b)**

End	G's Turns					F's Turns				
G's Gaze	Down		Up		Off	Down		Up		Off
F's Gaze	Down	Up	Down	Up		Down	Up	Down	Up	
Instruction	2.96	**-4.94**	0.03	-0.58	0.31					
Doubt	**-7.08**	**-4.75**	**10.47**	0.99	1.04	-2.27	**10.64**	**-7.19**	2.71	0.46
Explanation	**-3.39**	-1.41	**3.70**	1.24	1.08	-0.98	**3.11**	-2.24	**3.21**	-0.89
Confirmation	**7.45**	**7.67**	**-12.23**	-1.61	-1.69	**3.86**	**-11.35**	**5.94**	**-4.30**	0.55
Rejection	-2.53	**8.77**	-2.31	0.45	-1.63	-2.95	-1.62	**6.09**	-1.38	-0.88

(c) **(d)**

Table 4: Standard residuals of 4 significant χ^2 testing the likelihood of a certain dialogue act given the gaze values of the Follower (F) and the Giver (G) at the start of G's (sub-table **a**, $p < 2.2e^{-16}$) or F's turns (sub-table **b**, $p = 1.396e^{-09}$) and the end of G's (sub-table **c**, $p =< 2.2e^{-16}$) and F's turns (sub-table **d**, $p =< 2.2e^{-16}$). In **bold** are significant residuals.

End	G's Turns					F's Turns				
G's Gaze	Down		Up		Off	Down		Up		Off
F's Gaze	Down	Up	Down	Up		Down	Up	Down	Up	
$\Delta H_S > \Delta H_C > 0$	**-5.13**	2.42	**4.03**	0.64	0.01	**-3.81**	**4.96**	-0.61	**3.01**	-0.16
$\Delta H_C < 0 < \Delta H_S$	0.16	0.67	-1.46	0.75	0.74	-0.68	1.83	-1.53	0.12	1.37
$\Delta H_C = \Delta H_S$	-0.51	-1.58	2.42	1.57	-2.42	1.07	-0.65	-0.68	-0.70	0.36
$\Delta H_S < 0 < \Delta H_C$	1.60	-0.57	-1.77	1.60	-0.74	1.20	-2.22	2.33	-2.66	-1.18
$\Delta H_C > \Delta H_S > 0$	2.88	0.32	-2.64	-1.39	-0.50	1.02	-1.81	0.21	-0.85	0.60
$\Delta H_C < 0, \Delta H_S < 0$	**4.27**	-1.50	**-4.11**	-2.45	2.12	**3.16**	**-4.33**	0.30	-1.26	-0.22

(a) **(b)**

Table 5: Standard residuals of 2 significant χ^2-tests testing the likelihood of a certain entropy ratios given the gaze values of the Giver (G) and the Follower (F) at the end of G's (sub-table **a**, $p = 5.62e^{-08}$) or F's turns (sub-table **b**, $p = 2.48e^{-06}$). In **bold** are significant residuals.

the context (increase in entropy), or have they already been extensively used (decrease in entropy)? Similarly to the DA results, we conducted four (turn edge $\times$ turn owner) χ^2-tests. Unlike the DA results, only the ones matching the turns with the gaze value at their end were shown significant (Table 5).

Turns which had the speaker's entropy increase greater than the conversation's entropy ($\Delta H_S > \Delta H_C > 0$) tend to end in a context where the speaker looks up and the addressee looks down ($r > 4.03$). In the case of F's turns, it can even result in a mutual gaze ($r = 3.01$). In any case, it is very unlikely to result in both gazes converging down to the map ($r \leq -3.81$).

On the opposite, both entropy decreasing ($\Delta H_C < 0$, $\Delta H_S < 0$) are rather positively correlated to turns ending in a context where both participants look down ($r \geq 3.16$), and negatively correlated to the speaker looking up while the other participant looks down ($r \leq -4.11$).

These results show only partial alignment with observations made about DA. Only gaze values at turn ends revealed significant interactions with entropy categories: entropy behaviour types are more equally distributed across gaze types at turn starts. However, turns with specific entropy leading to significantly different gaze behaviour at the end of a turn is a good indicator that lexical novelty can

be assessed through gaze. $\Delta H_S > \Delta H_C > 0$ corresponds to turns where the vocabulary is relatively more novel to the speaker than to the conversation (e.g. the speaker reusing words introduced by their interlocutor). It is, similarly to DOUBT turns, a potential marker of hesitation and need for feedback on the present utterance. The opposite $\Delta H_C < 0$, $\Delta H_S < 0$ turns, where the utterance does not increase entropy for the speaker nor the conversation, behave similar to CONFIRMATION turns: both look down at the map, likely not inviting a break in flow.

4.3. Repetitions & Gaze

The next results address the most tangible aspect of lexical spread: direct repetitions (other and/or self) as described in section 3.5.3. We again realised a set of four χ^2-tests, and Table 6 shows the standard residuals of the two tests which were significant. Amongst these tests are the gaze values at the start of the G's turns (sub-table **a**) and the gaze values at the end of F's turns (sub-table **b**).

At the Start of the Giver's Turns: G is more likely to repeat F when G is looking at the map while F is looking at G ($r = 3.04$) but also if either one of them is looking off ($r = 3.05$) – in practice, 75% of the gazes away in the "off" bin are produced by F, it might hence be a strategy for G to get F's attention back when they are not looking at the map. However, even if this tendency is also significant when running a χ^2-test over a contingency table differentiating between when open-class tokens are being repeated or only closed-class ones (Table 7), none of the residuals returned an amplitude greater than the significance threshold ($|r| \geq 3.19$). Thus, it cannot be established whether OC or CC repeated content played a greater role than the other.

At the End of the Follower's Turns: On the other hand, F's turns are significantly correlated with the gaze values present at the end of their turns (Table 6). When F does not repeat anything, they are more likely to end on a look down at the map, while G looks up ($r = 4.65$). G looking down while F looks up, on the contrary, is unlikely for such turns ($r = -4.21$). Yet, the perfect opposite is likely for F's turns which include both self and other repetitions: the Follower then looks up and the Giver down ($r = 3.62$), and the reversed behaviour is unlikely ($r = -3.45$). Differentiating between OC and CC content (Table 7b), it seems that such a behaviour holds mostly for OC repetitions ($r = 3.22$), and is also visible for self-repetition only ($r = 3.69$).

These results align with previous results. F's repeating both G and themselves results in F gazing up while G looks down at the map ($r > 3.22$), similarly to EXPLAIN or DOUBT turns, as if they were looking for feedback. When F does not repeat any, however, and likely moves forwards, they tend to look down only while the Giver is likely looking at them ($r = 4.65$), similarly to CONFIRMATION or REJECTION turns. The gaze behaviour correlated with the start of the Giver's turns is however singular enough. It is the only one which significantly relates with off-gazes ($r = 3.05$) along with F's gaze up when G is focusing on the map ($r = 3.04$). Other-repetition might then be used to shape the new turn in a way that addresses the needs of the Follower: either to solve their hesitation, as looks up suggest, or maybe return attention to the map.

5. Discussion

Our findings suggest that paired gaze behaviours meaningfully interact with distinct categorisations of dialogue turn contents. If we consider the types of paired gazes that were significantly correlated with the turn qualities, we observe a disparity between symmetrical and asymmetrical gaze. Similarly to a finding of Sbranna et al. (2025), our results suggest that a gaze does not have to be mutual to impact the conversation. In fact, mutual gaze was rarely significantly correlated with a certain type of turn, as opposed to one-sided gazes to the other participants (5 out of 40 vs 28 out of 80; $\chi^2 = 6.7$, $p = 0.009$). As they did, we note that this distinction may be an artefact of the conversation being *task*-based and with a visual competitor (the map).

We investigated the communicative intent of gaze by matching it with dialogue acts.[4] Foundational work (Kendon, 1967; Ho et al., 2015) suggested that, at the end of turns, speakers look up at the addressees to check their availability to speak. While our research does not oppose this claim, and arguably will, in this discussion, agree that looks up are inviting a response, not all turns end with a look up in our dataset. Rather, we have to call for a more qualified answer where gaze behaviour at the end of a turn is dependent on the intent of the participants. Bavelas et al. (2002) found most backchannelling to be happening during "gaze windows" (i.e. mutual gaze instances). While our data does not suggest that CONFIRMATION turns – nor $\Delta H_C < 0$, $\Delta H_S < 0$ which tend to host such turns – start in such windows, the Giver's CONFIRMATION turns were indeed more likely when the Follower was looking at them. The corresponding tendency for the Follower's turns was, however, not found significant. Kendrick and Holler (2017) found that dis-preferred answers were associated with averted gaze, while preferred answers were associated with looks at the conversational partner. Only based on the contrast between CONFIRMATION and REJECTION turns, our observation does not match either as both DAs seemed to display similar behaviours: a tendency for the answerer to look down while

[4] In work under review we study sequential measurements of entropy and repetition in interaction with DAs.

	Start					End				
	G's Turns					F's Turns				
G's Gaze	Down		Up		Off	Down		Up		Off
F's Gaze	Down	Up	Down	Up		Down	Up	Down	Up	
OR	**-3.65**	**3.04**	-0.42	0.05	**3.05**	0.69	0.85	-1.69	1.26	-0.69
SR	2.44	-2.65	1.49	-2.41	-1.17	-0.60	1.76	-1.52	-0.55	1.81
OR-SR	-1.26	1.69	-0.73	0.44	0.62	-0.82	**3.62**	**-3.45**	1.01	1.78
ø	1.54	-1.20	-0.43	1.70	-1.72	0.34	**-4.21**	**4.65**	-1.32	-1.79
	(a)					**(b)**				

Table 6: Standard residuals for 2 significant χ^2-tests testing the likelihood of a certain repetition type given the gaze values of the Follower (F) and the Giver (G) at the start of G's turns (sub-table **a**, $p = 1.37e^{-04}$) and end of F's turns (sub-table **b**, $p = 8.51e^{-07}$). In **bold** are significant residuals.

		Start					End				
		G's Turns					F's Turns				
G's Gaze		Down		Up		Off	Down		Up		Off
F's Gaze		Down	Up	Down	Up		Down	Up	Down	Up	
OR	CC	-2.80	1.60	1.22	-1.52	2.48	-0.25	0.07	-0.76	1.75	0.25
	OC	-2.08	2.66	-2.18	1.98	1.57	1.28	1.13	-1.53	-0.22	-1.27
SR	CC	2.04	-3.13	2.20	-1.71	-1.07	0.52	0.02	-1.21	-0.69	1.42
	OC	1.09	-0.04	-0.54	-1.48	-0.42	-2.28	**3.69**	-0.89	0.17	1.09
OR-SR	CC	-0.80	-0.85	2.02	0.49	-0.03	-0.54	1.64	-1.71	0.71	1.17
	OC	-0.91	2.76	-2.51	0.15	0.78	-0.58	**3.22**	-2.95	0.68	1.27
ø	ø	1.54	-1.20	-0.43	1.70	-1.72	0.34	**-4.21**	**4.65**	-1.32	-1.79
		(a)					**(b)**				

Table 7: Standard residuals for 2 significant χ^2-tests testing the likelihood of a certain repetition type, differentiating between open-class (OC) and closed-class (CC) tokens, given the gaze values of the Follower (F) and the Giver (G) at the start of G's turns (sub-table **a**, $p = 3.82e^{-06}$) and end of G's turns (sub-table **b**, $p = 4.71e^{-06}$). In **bold** are significant residuals.

the addressee looks up. However, our findings match studies which take into account interaction difficulty (Boyle et al., 1994; Nicholson et al., 2005). For instance, Beattie (1978) reported that "Questions terminating with gaze were judged to be more difficult than questions without gaze" (abstract). Although we do not directly work with questions, this notion of difficulty seems to extend to our analyses. DOUBT and EXPLANATION, corresponding to instances in which the participants had to discuss and negotiate meaning, as opposed to the conversation flowing with no obstacle, but also F's turns containing self-repetitions of open-class content – suggesting more planning, are indeed correlated with looks at the addressee at their end.

Beyond abstract intentions observable through dialogue acts, we also find a correlation between the way turns are shaped and gaze behaviour. Sentences which lead to a decrease of both the speaker and conversation entropy tend not to divert gazes from the object of attention – the map – while those with particularly greater speaker's than conversation entropy – that is, those where the speaker uses vocabulary that is newer to them than to the conversation – are correlated with the speaker gazing up at the end. Furthermore, the study of repetition and gaze, shows that the Giver tends to spontaneously – understand from one turn to another – re-use lexical items from the Follower when starting while the latter is looking up. This might be evidence that the shape of a turn is also adapted depending on the gaze of the addressee. Differences in gaze behaviours based on the participant's role were also noted. For instance, DOUBT turns uttered by the Follower can start while both are looking down at the map, while the Giver's DOUBT turns are unlikely to occur in such a context. This can also be explained by the higher likelihood for F's DOUBT to be expressed during G's turns ($p = 3.775e^{-12}$).

Interpreting these results provides further insights into mutual understanding. CONFIRMATIONS seem to lead the speaker to keep looking down at turn ends. REJECTION turns only seem to end with the speaker looking down and the other participant looking up – a finding in favour of analysing referential looks down as a sign that the understanding is mutual. The alignment theory (Pickering and Garrod, 2004) states that participants will align their

beliefs as communicative intent unfolds, and this is visible through different means, linguistic and paralinguistic (Garrod and Pickering, 2007). Such a result adds to past work on repetitions and mutual gaze (Murat et al., 2022) which hypothesised mutual gaze to convey evidence of mutual understanding. In fact, our results are further evidence that mutual gaze is relatively rare in a conversation which includes a gaze competitor (here, the map) and that it occurs in situations of struggle: at the start of a REJECTION turn, at the end of an EXPLANATION or at the end of a turn in which the Follower tries to make the conversation's words theirs ($\Delta H_S > \Delta H_C > 0$). This matches previous literature which sees gaze up as a way to check the availability of the other participant to speak – and hence contribute (Kendon, 1967; Ho et al., 2015). It is also similar to the early gaze studies reported on the HCRC Map Task Corpus (Boyle et al., 1994), which highlighted that gazes up tended to mark moments of difficulty. Interestingly, the method they had then used differed considerably from ours. They focused their analysis on the mismatch between the maps and found significantly more gazes up when non-matching references were discussed than when matching references were discussed. On our end, the methods were purely quantitative. We did not inspect the maps nor the details of the lexicon being used beyond their frequency and their part-of-speech. Yet, we were able to highlight moments of difficulty and draw similar conclusions (presumably because the language used was *about* the maps).

We have reported on a reduction of temporal information to categories and counts based on events and their overlap, without direct attention to durations associated with the relevant boundaries. These have been shown to be important (e.g., Oertel et al. (2012)). Our approach does not destroy information about duration beyond counts of events; rather, the method may complement such analyses.

In sum, exploring gaze behaviour at the start and end of each turn allowed us to study anteriority of events (is the gaze or the turn coming first) and, thus, contributes insights to the turn-taking literature (Degutyte and Astell, 2021) (see §2). The relationship between gaze and turn qualities shows different coherent patterns: as the property of a turn display a need for feedback, the speaker's gaze ends up, and as the gazes are up, the content gets more reused in response, as a way to clarify the situation.

6. Conclusion

This paper investigates the relationship between eye-gaze behaviour in the dyadic conversations and the linguistic progression of the interaction. To do so, we studied the correlation between paired gaze and dialogue acts to reveal the intention behind turns, along with their lexical properties, as measured by entropy and repetition, using a cross-modality event annotation scheme that enables count-based interaction analysis. Our results on the HCRC Map Task corpus show that gaze at the turn level (their start or end) is correlated with the progression of the interaction. A gaze towards the conversation partner at the end of a turn tends to align with complexity and difficulties expressed in the turn, while keeping a look down at the map is more typical of situations without obstacles, and neither participant shows signs of disagreement. There is also evidence that the Guide may adapt the content of their turns to the Follower when the latter is looking up at the start of the turn, as such a case is correlated with the presence of other-repetition.

7. Ethics Statement & Limitations

This observational study uses part of only one – albeit large – corpus: the HCRC corpus (Thompson et al., 1993). Its strict set-up – a task having to be performed by strictly two persons with a visual competitor that is the map, is more natural than some settings, but still far from representative of all face-to-face human interactions. Rather, this paper has to be read as an incremental contribution, and this aspect is accented in the discussion where the results are compared with prior findings.

While the HCRC Map Task corpus is widely studied and this observational study relied on its ethical provenance, the results here have some bias. Even if the corpus is gender-balanced, its tightly controlled design records only native English speakers, most of them Scots. We have not yet analysed Map Task data constructed in other languages. The raw HCRC recordings are not available for privacy reasons, so all the gaze annotations have to be trusted – in fact, due to contradictions in some gaze annotations and absence of material to reconcile them, we had to omit excerpts (see sec. 3.2). We have found prior descriptions of gaze annotation in this corpus (e.g. Boyle et al. (1994); Anderson et al. (1997); Doherty-Sneddon (1995), but they imply a different label inventory and temporal precision than is in the corpus. We have not spotted a clear discussion of the released annotations.

8. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant No. 18/CRT/6223. We are grateful to the anonymous reviewers for their constructive feedback. We think this paper is improved by having addressed their comments.

9. Bibliographical References

Anne H Anderson, Ellen Gurman Bard, Cathy Sotillo, Alison Newlands, and Gwyneth Doherty-Sneddon. 1997. Limited visual control of the intelligibility of speech in face-to-face dialogue. *Perception & Psychophysics*, 59(4):580–592.

Michael Argyle and Mark Cook. 1976. *Gaze and mutual gaze*. Cambridge Univ. Press, Cambridge.

Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication*, 52(3):566–580.

Geoffrey W. Beattie. 1978. Floor apportionment and gaze in conversational dyads. *British Journal of Social and Clinical Psychology*, 17(1):7–15.

Elizabeth A. Boyle, Anne H. Anderson, and Alison Newlands. 1994. The Effects of Visibility on Dialogue and Performance in a Cooperative Problem Solving Task. *Language and Speech*, 37(1):1–20.

Susan E. Brennan, Xin Chen, Christopher A. Dickinson, Mark B. Neider, and Gregory J. Zelinsky. 2008. Coordinating cognition: the costs and benefits of shared gaze during collaborative search. *Cognition*, 106(3):1465–1477.

Geert Brône, Bert Oben, Annelies Jehoul, Jelena Vranjes, and Kurt Feyaerts. 2017. Eye gaze and viewpoint in multimodal interaction management. *Cognitive Linguistics*, 28(3):449–483. Publisher: De Gruyter Mouton.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Alison Newlands, Gwyneth Doherty-Sneddon, and Anne Anderson. 1996. HCRC dialogue structure coding manual. Technical report, University of Edinburgh. HCRC Technical Report 82.

Mark S. Cary. 1978. The Role of Gaze in the Initiation of Conversation. *Social Psychology*, 41(3):269.

Ziedune Degutyte and Arlene Astell. 2021. The Role of Eye Gaze in Regulating Turn Taking in Conversations: A Systematized Review of Methods and Findings. *Frontiers in Psychology*, 12.

Gwyneth Doherty-Sneddon. 1995. *The development of conversational and communication skills*. Ph.D. thesis, University of Glasgow.

Jens Edlund, Simon Alexandersson, Jonas Beskow, Lisa Gustavsson, Mattias Heldner, Anna Hjalmarsson, Petter Kallionen, and Ellen Marklund. 2012. 3rd party observer gaze as a continuous measure of dialogue flow. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1354–1358, Istanbul, Turkey. European Language Resources Association (ELRA).

Simon Garrod and Martin J. Pickering. 2007. Alignment in dialogue. In M. Gareth Gaskell, editor, *The Oxford Handbook of Psycholinguistics*, pages 442–452. Oxford University Press.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 199–206.

Milan Gnjatović, Jovica Tasevski, Branislav Borovac, and Nemanja Maček. 2018. An Entropy-Based Approach to Automatic Detection of Critical Changes in Human-Machine Interaction. In *2018 9th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000175–000178. ISSN: 2380-7350.

Charles Goodwin. 1980. Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry*, 50(3-4):272–302.

Joy E. Hanna, Susan E. Brennan, and Kelly J. Savietta. 2020. Eye Gaze and Head Orientation Cues in Face-to-Face Referential Communication. *Discourse Processes*, 57(3):201–223. _eprint: https://doi.org/10.1080/0163853X.2019.1675467.

Patrick G. T. Healey, Matthew Purver, and Christine Howes. 2014. Divergence in dialogue. *PloS one*, 9(6):e98598. Https://doi.org/10.1371/journal.pone.0098598.

Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions. *PLOS ONE*, 10(8):e0136905. Publisher: Public Library of Science.

Laszlo Hunyadi, Tamás Váradi, Gy Kovács, István Szekrényes, Hermina Kiss, and Karolina Takács. 2018. Human-human, human-machine communication: on the HuComTech multimodal corpus. *Selected papers from the CLARIN Annual Conference 2018. Linköping Electronic Conference Proceedings*, 159:56–65.

Kristiina Jokinen, Kazuaki Harada, Masafumi Nishida, and Seiichi Yamamoto. 2010. Turn-Alignment Using Eye-Gaze and Speech in Conversational Interaction. *Interspeech*, pages 2018–2021.

Frank Keller. 2004. The entropy rate principle as a predictor of processing effort: An evaluation against eye-tracking data. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 317–324.

Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63.

Adam Kendon. 1978. Looking in conversation and the regulation of turns at talk: A comment on the papers of G Beattie and D. R. Rutter *et al* . *British Journal of Social and Clinical Psychology*, 17(1):23–24.

Kobin H. Kendrick and Judith Holler. 2017. Gaze Direction Signals Response Preference in Conversation. *Research on Language and Social Interaction*, 50(1):12–32.

Maria Koutsombogera and Carl Vogel. 2019. Observing Collaboration in Small-Group Interaction. *Multimodal Technologies and Interaction*, 3(3):45.

Magnus S. Magnusson. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32(1):93–110.

Magnus S. Magnusson. 2020. T-Pattern Detection and Analysis (TPA) With THEMETM: A Mixed Methods Approach. *Frontiers in Psychology*, 10.

Alexander Mehler, Andy Lücking, and Peter Menke. 2011. Assessing Lexical Alignment in Spontaneous Direction Dialogue Data by Means of a Lexicon Network Model. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6608, pages 368–379. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a Model of Gaze for Grounding in Multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 247–254, Istanbul Turkey. ACM.

Dalila Mekhaldi. 2006. *A study on multimodal document alignment: bridging the gap between textual documents and spoken language*. Ph.D. thesis, Faculty of Science, University of Fribourg (Switzerland).

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Anaïs Claire Murat, Maria Koutsombogera, and Carl Vogel. 2026. Event chronography in multimodal data: the BME method for quantitative analyses. In *Proceedings of the 2026 Language Resources and Evaluation Conference*. To appear.

Anaïs Claire Murat, Maria Koutsombogera, and Carl Vogel. 2022. Mutual Gaze and Linguistic Repetition in a Multimodal Corpus. In *Proceedings of the Language Resources and Evaluation Conference*, pages 2771–2780, Marseille, France. European Language Resources Association (ELRA).

Anaïs Claire Murat and Carl Vogel. 2025. Online meetings: When repetition signals engagement. In *2025 IEEE 16th International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 37–42, Vienna, Austria. IEEE.

Mark B. Neider, Xin Chen, Christopher A. Dickinson, Susan E. Brennan, and Gregory J. Zelinsky. 2010. Coordinating spatial referencing using shared gaze. *Psychonomic Bulletin & Review*, 17(5):718–724.

Hannele Nicholson, Ellen Gurman Bard, Robin Lickley, Anne H. Anderson, and Catriona Havard (3) Yiya Chen. 2005. Disfluency and behaviour in dialogue: evidence from eye-gaze. In *Disfluency in Spontaneous Speech (DiSS 2005)*, pages 133–138.

Catharine Oertel, Marcin Włodarczak, Jens Edlund, Petra Wagner, and Joakim Gustafson. 2012. Gaze patterns in turn-taking. In *13th Annual Conference of the International Speech Communication Association 2012 (INTERSPEECH 2012)*, pages 2246–2249.

Ulrich J. Pfeiffer, Kai Vogeley, and Leonhard Schilbach. 2013. From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, 37(10):2516–2528.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(02).

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Comp utational Linguistics*, pages 808–815. Association for Computational Linguistics.

Justine Reverdy, Maria Koutsombogera, and Carl Vogel. 2020. Linguistic Repetition in Three-Party Conversations. In Anna Esposito, Marcos Faundez-Zanuy, Francesco Carlo Morabito, and

Eros Pasero, editors, *Neural Approaches to Dynamics of Signal Exchanges*, volume 151, pages 359–370. Springer Singapore.

Justine Reverdy and Carl Vogel. 2017. Measuring Synchrony in Task-Based Dialogues. In *Interspeech 2017*, pages 1701–1705, Stockholm, Sweden. ISCA.

Federico Rossano. 2010. Questioning and responding in Italian. *Journal of Pragmatics*, 42(10):2756–2771.

Simona Sbranna, Michelina Savino, Florence Baills, and Martine Grice. 2025. Gaze behaviour and vocal feedback in task-based dyadic conversations with and without eye contact. *Frontiers in Communication*, 10. Publisher: Frontiers.

Malin Spaniol, Alicia Janz, Simon Wehrle, Kai Vogeley, and Martine Grice. 2023. Multimodal signalling: The interplay of oral and visual feedback in conversation. In *Proceedings of the 20th International Congress of Phonetic Sciences. ICPhS*.

Jürgen Streeck. 2014. Mutual gaze and recognition: Revisiting Kendon's "Gaze direction in two-person conversation". In Mandana Seyfeddinipur and Marianne Gullberg, editors, *From Gesture in Conversation to Visible Action as Utterance: Essays in honor of Adam Kendon*, pages 35–56. John Benjamins Publishing Company.

Yu Wang, Yang Xu, Gabriel Skantze, and Hendrik Buschmeier. 2024. How much does nonverbal communication conform to entropy rate constancy?: A case study on listener gaze in interaction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3533–3545.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Elisabeth Zima, Clarissa Weiß, and Geert Brône. 2019. Gaze and overlap resolution in triadic interactions. *Journal of Pragmatics*, 140:49–69.

10. Language Resource References

Thompson, Henry S. and Anderson, Anne and Bard, Ellen Gurman and Doherty-Sneddon, Gwyneth and Newlands, Alison and Sotillo, Cathy. 1993. *The HCRC Map Task corpus: natural dialogue for speech recognition*. Association for Computational Linguistics, ISLRN 777-455-577-608-1.

CATS: An annotation scheme of causality and temporal structure

Nana Yu[*], **Purificação Silvano**[*‡§], **Luís Filipe Cunha**[†§] , **Alípio Jorge**[†§]
[*]Faculty of Arts and Humanities, University of Porto
[†] Faculty of Sciences, University of Porto
[‡]Centre for Linguistics of the University of Porto (CLUP)
[§] INESC TEC
Porto, Portugal
robertananayu@hotmail.com, {purificacao.silvano, luis.f.cunha, amjorge}@inesctec.pt

Abstract

This paper presents CATS, a causal and temporal annotation scheme designed to jointly represent causal relations and temporal structures in news texts. The proposed framework integrates components of ISO 24617 Semantic Annotation Framework (SemAF), drawing in particular on Part 1 (Time and Events) (ISO 24617-1: 2012) and Part 8 (Semantic Relations in Discourse) (ISO 24617-8: 2016). Building on the Text2Story annotation framework (Silvano et al., 2021), the scheme adapts and extends its principles for representing temporal information while introducing new entities and links for modeling causal relations. The resulting annotation model enables the integrated representation of causal arguments, events, temporal relations, and causal signals within a unified structure. By jointly capturing causal and temporal dependencies, CATS provides a resource for studying the interaction between causality and temporality in discourse and supports downstream NLP tasks such as event extraction, temporal ordering, and causal reasoning.

Keywords: causal relations, temporal relations, discourse relations, semantic annotation schemes, ISO 24617

1. Introduction

Causality constitutes one of the fundamental principles guiding human reasoning, as knowledge about the world tends to be organized through causal networks that explain, justify, or constrain the occurrence of phenomena (Pearl, 2009; Woodward, 2003; Sloman, 2005). For this reason, causality has been the subject of study across several areas of social sciences, humanities, and natural sciences, and, more recently, has also become an important topic in the field of Natural Language Processing (NLP)(Yang et al., 2022; Feder et al., 2022; Koupaee et al., 2025).

Language widely reflects causality relations, but linguistic studies acknowledge that defining causality is a complex task (Talmy, 2000). In Linguistics, causality has been widely studied within the framework of discourse relations established between textual segments/utterances (Hobbs, 1985; Mann and Thompson, 1988; Asher and Lascarides, 2003; Kehler, 2002). These relations, also referred to as rhetorical relations, constitute an important object of study in semantics, since they explain how different discourse units are connected in the construction of the overall meaning of a text (Mann and Thompson, 1988; Asher and Lascarides, 2003). Several theoretical and annotation frameworks integrate such relations into their discourse models, including proposals by Hobbs (1985), Mann and Thompson (1988), Asher and Lascarides (2003), Prasad et al. (2008), and the ISO discourse-relation standard ISO 24617-8: 2016 (Prasad and Bunt, 2015; Bunt and Prasad, 2016). Within this perspective, linguistic elements and structures play a central role in inferring discourse relations. In practice, these relations can be signaled or inferred from multiple types of linguistic cues, such as lexical choices (nouns, verbs, adverbs, prepositions), discourse connectives, and syntactic constructions of coordination and subordination (Prasad et al., 2008). Beyond these elements, other linguistic dimensions, such as tense and aspect, also contribute substantially to the interpretation of relations between events (Moens and Steedman, 1988; Lascarides and Asher, 1993). Accordingly, characterizing temporal relations and the aspectual properties of the involved situations is an essential dimension for the analysis of causality. Their interaction has been explicitly discussed in both linguistic theory and NLP-oriented studies of event relations (Lascarides and Asher, 1993; Mirza and Tonelli, 2014; Ning et al., 2018a).

The representation and processing of causal and temporal information have received increasing attention in recent years. Nevertheless, many studies tend to focus primarily on one of these dimensions rather than the other, and the number of works addressing temporal and causal relations jointly remains relatively limited (Bethard et al., 2008; Ning et al., 2018a; Mirza and Tonelli, 2014). Within the domain of semantic annotation schemes, several proposals include labels and attributes designed to represent temporal, aspectual, and discourse-

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 111–122
12 May 2026.

related properties. In the domain of temporality, relevant work includes Pustejovsky et al. (2003a), Setzer (2001), Mani and Schiffman (2004), Ning et al. (2018b), and ISO 24617-1: 2012 (Pustejovsky et al., 2010), and, more recent ISO-based, multilayer efforts applied to narrative/news data such as Silvano et al. (2024). In the domain of causality, reference points include the SemEval annotation task that explicitly covers Cause–Effect relations (Girju et al., 2007), corpus-based causal annotation schemes grounded in linguistic analysis (Dunietz et al., 2015, 2017), and recent event-causality resources such as the Causal News Corpus (Tan et al., 2022). Finally, some approaches seek to integrate temporal and causal relations simultaneously, including Bethard et al. (2008) and Mirza and Tonelli (2014).

Despite these contributions, there remains a need for annotation schemes that systematically model how causality and temporality interact in discourse, motivating the development of the proposal presented in this work. The need for a unified treatment of causality and temporality arises from the fact that these two dimensions are deeply interdependent in discourse, yet are typically treated in isolation in existing annotation frameworks. A unified annotation scheme enables the systematic investigation of questions such as how causal relations are temporally anchored, how different types of causal relations correlate with temporal configurations, and how ambiguities between mere temporal succession and causality are resolved in discourse. It also allows the identification of the linguistic and structural cues that signal these interactions, including connectives, tense–aspect marking, and discourse organization. By explicitly representing the interaction between causality and temporality, the present work provides a framework for both fine-grained linguistic analysis of causality and more robust modeling in downstream applications, such as event ordering, causal inference, and narrative understanding.

The main contributions of this paper are the following:

- an integrated scheme for the annotation of causal and temporal relations grounded in ISO 24617;
- a description of the methodology used to develop the scheme;
- a critical analysis of several incompatibilities between the two ISO parts and suggestions for addressing these issues.

The remainder of the paper is organized as follows. Section 2 reviews related work on causal and temporal annotation. Section 3 describes the process of designing the annotation scheme, starting with the methodology for developing the scheme (Section 3.1), and the motivation for using ISO 24617 (Section 3.2.). Section 3.3 discusses the integration of both layers and presents the main challenges encountered during the process, together with the proposed solutions. Section 4 introduces the final causal and temporal annotation scheme. Finally, Section 5 concludes the paper and outlines directions for future research.

2. Related work

A major milestone in the annotation of causal relations in discourse is the work of Prasad et al. (2008) associated with the Penn Discourse Treebank (PDTB). Built from texts from the *Wall Street Journal*, the PDTB includes thousands of annotated instances of causal relations within the semantic class CONTINGENCY. The framework distinguishes between explicit and implicit discourse relations. In explicit relations, causality is marked by lexical connectives (e.g., *because*, *as a result*), which are classified according to their syntactic and functional properties. Annotation follows the syntactic attachment of the connective: the argument syntactically linked to the marker is labeled *Arg2* (typically interpreted as the cause), while the other segment is labeled *Arg1* (often corresponding to the effect). In implicit relations in the PDTB, where no lexical connective is present, annotators identify two arguments and insert an implicit connective between them. *Arg2* corresponds to the clause where the connective would appear, while *Arg1* is the other argument, typically preceding *Arg2* in the text. This approach highlights the discourse-based nature of causal interpretation and the methodological challenges posed by non-lexicalized relations.

An alternative perspective focusing on the linguistic realization of causality is proposed by Dunietz et al. (2015). The authors argue that no single representation scheme can capture the full semantic diversity of causal expressions, since causality is a cognitive construct interacting with dimensions such as temporality, counterfactuality, factuality, and negation. Their annotation framework is grounded in Construction Grammar and defines causal language as any construction that presents one event or state as promoting or hindering another, provided that at least one lexical trigger is present. Consequently, the annotation scheme is restricted to explicit causal relations, excluding implicit cases or constructions whose causal interpretation depends primarily on extralinguistic knowledge. The resulting corpus, *BECauSE 1.0*, identifies three core components, *connective*, *cause*, and *effect*, and distinguishes four causal types: *consequence*, *motivation*, *purpose*, and *inference*.

This line of work is further extended in *BE-*

CauSE 2.0 (Dunietz et al., 2017), which conceptualizes causality as causal networks in which phenomena may cause, enable, or prevent other events. The updated scheme expands the scope of annotation by capturing overlapping semantic relations frequently associated with causal expressions, including temporal, hypothetical, and contextual relations. Compared to the first version, the scheme reorganizes causal categories, retaining *consequence*, *motivation*, and *purpose* while removing *inference*, and introducing a distinction between facilitating and inhibitory causation. The resulting corpus contains sentences with annotated causal instances and overlapping relations, illustrating the complexity of causal meaning in natural language.

From an event-oriented perspective, Tan et al. (2022) emphasize the central role of causality in natural language understanding and inference tasks. They propose annotation guidelines that classify sentences as either causal or non-causal, allowing both explicit and implicit causal relations as long as both cause and effect are present. Event identification follows the TimeML framework, while the notion of causality is inspired by the CONTINGENCY relation in PDTB 3.0. The resulting *Causal News Corpus* contains news texts annotated with binary sentence-level labels indicating the presence of causal relations.

In contrast to the relatively limited number of causal annotation frameworks, temporal annotation has been extensively studied. A foundational proposal is TimeML (Pustejovsky et al., 2003b), designed to represent events, temporal expressions, and relations between them through a structured set of entities and links. TimeML has been widely adopted and later standardized as ISO-TimeML (ISO 24617-1: 2012). Adaptations for other languages include *TimeBankPT* (Costa and Branco, 2012a,b) for European Portuguese. Building on the ISO Semantic Annotation Framework, the Text2Story scheme (Silvano et al., 2021; Leal et al., 2022) introduces a multilayer annotation model integrating temporal, referential, spatial, and semantic-role information, enabling the annotation of complex narrative structures.

Despite the maturity of temporal annotation schemes, the joint modeling of causal and temporal relations remains relatively rare. These dimensions are strongly interrelated, since causal relations always imply a temporal ordering between events, specifically, that the cause precedes its effect (Tan et al., 2022). However, our investigation confirms that certain cases exhibit alternative temporal configurations, including simultaneity and inclusion, where one event can occur at the same time as the other, encompass, or be encompassed by the other. Some attempts to integrate both causal and temporal dimensions include the annotation frameworks proposed by Bethard and Martin (2008) and by Mirza and Tonelli (2014). While the former jointly annotates causal and temporal relations between event pairs, the latter combines TimeML-style temporal annotation with lexicalized causal relations based on force-dynamics categories such as *Cause*, *Enable*, and *Prevent*. Nevertheless, most existing approaches still treat causal and temporal information in separate annotation layers.

Some semantic representation formalisms, such as Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), Uniform Meaning Representation (UMR) (Gysel et al., 2021), and YARN (Pavlova et al., 2024), adopt a layered architecture in which different semantic phenomena are encoded in distinct but interrelated levels, enabling a structured and modular treatment of meaning. Although these frameworks are not specifically designed to model causal and temporal phenomena in isolation, nor explicitly targeted at causality, their architectures are nonetheless capable of supporting the representation of such relations. In some cases, however, their aim of providing a comprehensive semantic representation leads to considerable complexity, as they incorporate a wide range of linguistic phenomena.

Although temporal annotation frameworks are relatively well established, causal annotation schemes remain less consolidated, more heterogeneous, and in need of broader validation. Moreover, integrated representations that jointly capture causality and temporality are still limited, particularly when considering the need for interoperable frameworks. This gap motivates the development of approaches that model both dimensions simultaneously in order to provide a more comprehensive account of how events are organized in discourse. In this context, our proposal not only integrates these two dimensions, but also adopts an interoperable framework that enables implementation across different semantic formalisms, thus facilitating broader applicability and reuse.

3. The process of designing the annotation scheme

The scheme proposed in this work aims to represent causal relations in discourse while simultaneously capturing the temporal relations that hold between the events involved in those relations. The development process involved the integration of two semantic layers, causality and temporality, grounded in the ISO 24617 framework and implemented within the Text2Story semantic annotation architecture. Although the annotation scheme is exemplified with data from European Portuguese, it is, in principle, applicable across languages.

3.1. Methodology for developing the annotation scheme

The development of the proposed annotation scheme followed the methodology for the construction and validation of annotation frameworks proposed by Fernandes et al. (2025). This methodology combines conceptual modeling and empirical validation through successive phases, namely: (i) literature review, (ii) design and specification, (iii) empirical validation, and (iv) consolidation and refinement. The initial phase consisted of a comprehensive review of existing approaches to the annotation of causal and temporal relations, including semantic annotation standards and discourse and temporal annotation frameworks. The second phase focused on defining the conceptual structure of the proposed scheme, including its entities, attributes, and relations. The third phase involved pilot annotation experiments on real corpus data, which enabled the systematic identification of inconsistencies, gaps, and incompatibilities. In this phase, the development process followed the MATTER methodology proposed by Pustejovsky and Stubbs (2012). In particular, the iterative subcycle *Model–Annotate* was adopted and operationalized into four stages: *Model*, *Annotate*, *Evaluate*, and *Revise*. This iterative process supported continuous refinement of the scheme based on corpus evidence and annotation feedback. During the development process, this cycle was applied repeatedly because several issues emerged when annotating real data, requiring continuous refinement of the scheme in order to adequately represent causal and temporal relations. As noted by Silvano et al. (2021), building a bootstrapping annotation scheme is a complex and time-consuming task that involves several iterative phases. Finally, a consolidation phase was conducted to refine the model and stabilize the guidelines.

3.2. Motivation for using the ISO 24617 framework

The annotation scheme proposed in this work is grounded in the ISO 24617 series of standards for semantic annotation. This framework provides a modular architecture for representing multiple semantic dimensions of discourse, including temporal information (ISO 24617-1: 2012) and discourse relations (ISO 24617-8: 2016). The decision to adopt this framework was motivated by several factors. First, ISO 24617 provides a standardized and interoperable model for semantic annotation, allowing the integration of different semantic layers within a unified structure. Second, the framework has been successfully applied in previous annotation projects, including the Text2Story multilayer annotation scheme proposed by Silvano et al. (2021) and Leal et al. (2022). Third, the modular architecture of the ISO framework facilitates the integration of discourse-relation and temporal information.

3.3. Development of the annotation scheme

The proposed annotation scheme was developed from the Text2Story framework (Silvano et al., 2021), which has already demonstrated positive results in the annotation of journalistic narratives (Silvano et al., 2024). Since Text2Story is grounded in the ISO 24617 semantic annotation framework, it provides a suitable basis for extending the representation of narrative structures toward causal information. In this work, the temporal layer inherited from ISO 24617-1: 2012 was preserved with adaptations from Text2Story framework, while a causal layer inspired by ISO 24617-8: 2016 was integrated into the architecture in order to represent both causal and temporal relations within the same annotation scheme.

In our proposal, we adopt a broader notion of *entity* in order to harmonize the temporal and causal layers within a unified annotation model. Following the architectural principles of ISO 24617-1: 2012, the annotation scheme is organized into two complementary structures: *entity structures* and *link structures*. Entity structures contain semantic information associated with a segment of the source text, while link structures represent semantic relations between such segments by establishing links between entity structures. Formally, an annotation structure consists of two sets, M and L. The set M contains pairs $\langle m, a \rangle$, where m corresponds to a markable (a segment of the source text) and a corresponds to the associated entity structure describing its semantic properties. The set L contains triples $\langle r, a_i, a_j \rangle$, where r represents a relational annotation and a_i and a_j correspond to entity structures participating in that relation. Each element of L links two entity structures contained in M, ensuring that relations are defined over annotated textual segments. More complex annotation structures consist of sets of entity structures linked through temporal, causal, and auxiliary relations, thereby allowing the representation of interactions between events, discourse arguments, and explicit causal signals in the text.

Figure 1 summarizes the interaction between the entity structure (events, situations, signals) and the link structure (discourse, temporal, and auxiliary links).

The temporal component of the scheme derives from ISO 24617-1: 2012 (SemAF-Time), which defines a semantic framework for the annotation of time and events. Within this layer, the label *Event* is used to represent eventualities expressed in the

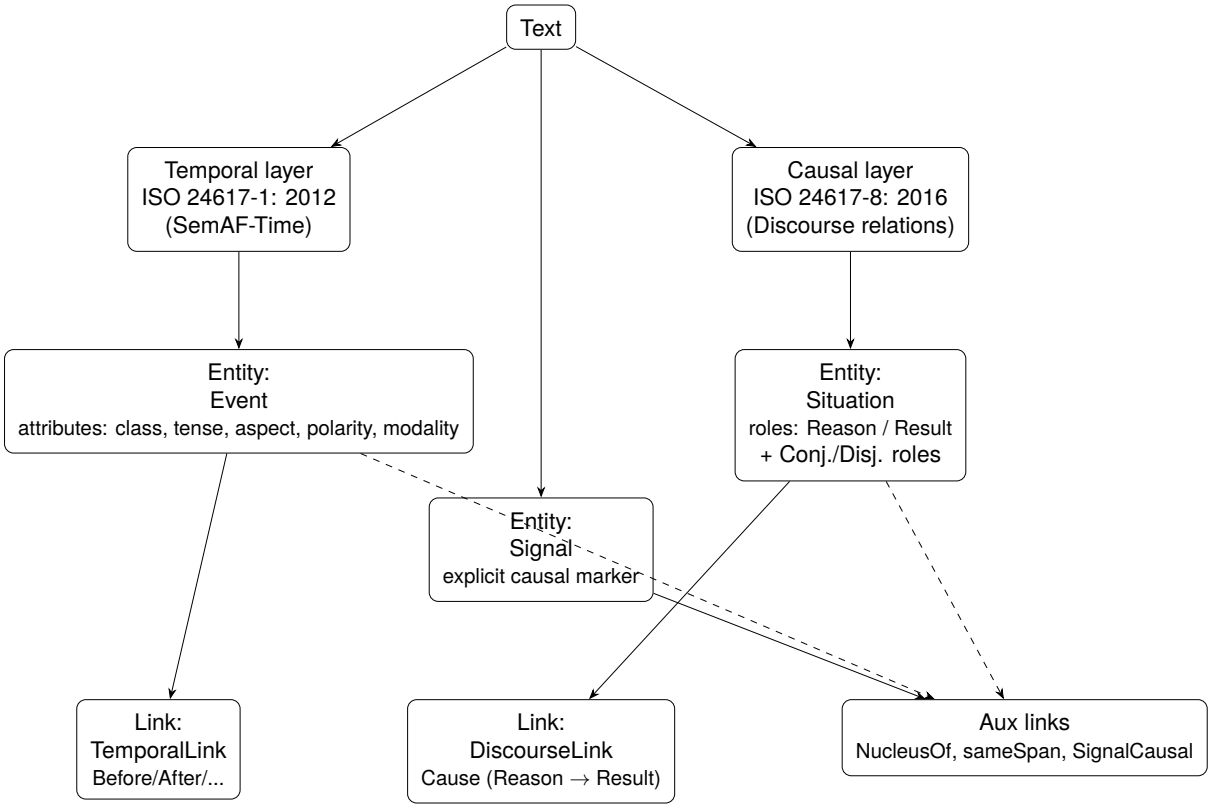

Figure 1: Overall architecture integrating ISO 24617-8: 2016 (causal discourse relations) and ISO 24617-1: 2012 (temporal layer) within Text2Story.

text. Events correspond to occurrences, states, or circumstances and are typically realized by verbs, event nominals, or nominalizations. Each event is characterized through a set of semantic and grammatical attributes, including *Class*, *Type*, *Part of Speech*, *Tense*, following the annotation guidelines of the Text2Story framework (Silvano et al., 2023). Temporal relations between events are represented through temporal links encoding ordering relations such as *Before*, *After*, *Simultaneous*, *Includes*, and *Is_Included*. These relations allow the chronological organization of the eventualities expressed in the text to be represented.

On the basis of this temporal layer, a causal component was introduced following the discourse-relation framework defined in ISO 24617-8: 2016. This standard provides an interoperable approach to the annotation of discourse relations and establishes correspondences between different theoretical frameworks, including Rhetorical Structure Theory (RST) (Mann and Thompson, 1988), Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003), the Penn Discourse Treebank (PDTB) (Prasad et al., 2008), Hobbs' Theory of Discourse Coherence (HTDC) (Hobbs, 1985), and the Cognitive Approach to Coherence Relations (CCR) (Sanders et al., 1992). In the ISO framework, discourse relations are defined as semantic links between two situations, which may correspond to clauses, sentences, nominalizations, or larger discourse segments. To incorporate causal relations into the proposed scheme, a new entity label called *Situation* was introduced in the entity structure. This label allows the annotation of situations expressed in texts as clauses, sentences, noun phrases, nouns, or verbs. Within *Situation*, the two core semantic roles proposed by ISO 24617-8: 2016 were adopted: *Reason* and *Result*. These roles correspond to the two arguments of the causal relation and enable the explicit identification of the textual segments that function as cause and effect. To establish the causal relation between these arguments, a new link type called *Discourse* was introduced in the link structure. Thus, according to ISO 24617-8:2016, the relevant discourse relation is labeled *Cause* and involves two arguments associated with distinct semantic roles, namely *Reason* and *Result*[1], where these two terms are interpreted as having the same underlying meaning but different functions: one as a discourse relation and the other as a semantic role within that relation.

The scheme also includes the annotation of ex-

[1]Although one may argue that a distinction may be drawn between 'cause' and 'reason' in terms of their semantic contribution, this work adopts the terminology and conceptualization provided by ISO 24617-8:2016.

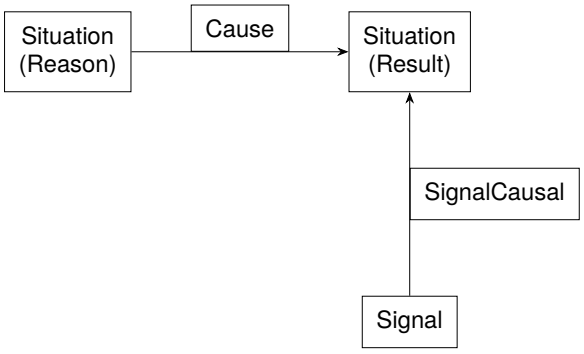

Figure 2: Causal layer: *Cause* links reason and result situations; explicit *Signal* is associated with the result argument.

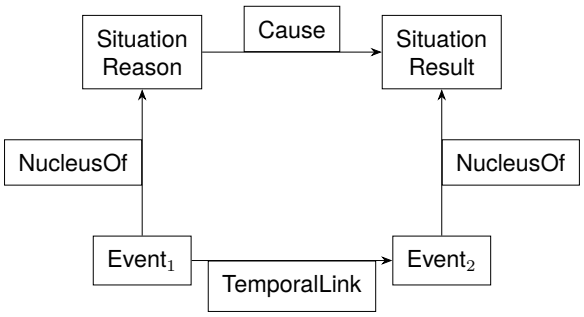

Figure 3: Integration: causal relations link *Situations*, while temporal relations are encoded between the main *Events* associated with each argument.

plicit causal markers. For this purpose, a label called *Signal* was introduced to identify lexical or multiword expressions that explicitly indicate a causal relation between two arguments. Based on previous work, particularly the Penn Discourse Treebank and linguistic studies on causal markers, explicit signals were categorized into grammatical classes. A specific link type, also called *Signal*, was added to the link structure to associate the signal with the result argument of the causal relation.

Figure 2 summarizes the core representation of the causal layer, including the association of an explicit signal with the result argument via a dedicated link type.

Figure 3 illustrates the adopted integration strategy.

The integration of the temporal and causal layers required several modeling decisions. In the ISO framework, causal relations are defined between arguments, whereas temporal relations are defined between events. However, the standard does not explicitly define how these two layers should interact when both causal and temporal information must be represented simultaneously. This limitation becomes particularly evident when applying the annotation scheme to real data, as illustrated in the following example[2].

(1) *O fogo provocou danos no recheio e na estrutura do edifício.* (Lusa2_36)
The fire caused damage to the interior and the structure of the building.

In this example, the sentence clearly expresses both a causal relation and a temporal ordering between the events involved. To address this issue, the proposed scheme establishes temporal relations between the main events associated with each causal argument rather than directly between the arguments themselves. Thus, the causal relation is maintained between the *Reason* and the *Result*, while the temporal relation is encoded between the events that instantiate these arguments. To support this interaction between layers, a relation called *NucleusOf* was introduced. This relation links an *Event* to the *Situation* to which it belongs, making it possible to explicitly identify the main event associated with each causal argument.

During the annotation experiments conducted on a set of news articles from the Lusa news agency, several issues emerged that required adjustments to the scheme. One major challenge concerned the representation of complex causal configurations. Although ISO 24617-8: 2016 defines causal relations as binary relations between a reason and a result, the corpus revealed more complex structures. In some cases, a single reason leads to multiple results, whereas in other cases several reasons converge toward a single result.

The following examples illustrate these situations:

(2) *Um fenómeno meteorológico geralmente associado ao aumento das temperaturas, a secas em algumas partes do mundo e a chuvas fortes noutras.* (Lusa2_100)
A meteorological phenomenon generally associated with rising temperatures, droughts in some parts of the world, and heavy rains in others.

(3) *Pelo menos quatro pessoas continuam hoje desaparecidas e cerca de 20 mil foram afetadas por inundações no sul da Tailândia, na sequência de uma forte tempestade.* (Lusa2_116)
At least four people remain missing today and around twenty thousand have been af-

[2]For illustrative purposes, we use European Portuguese data collected from the Lusa news agency, the largest Portuguese-language news agency. The examples presented in this paper are drawn from a dataset currently under preparation for publication. The original Portuguese sentences are provided alongside their English translations, and each example is identified by its corresponding reference (e.g., Lusa2_36).

fected by floods in southern Thailand following a strong storm.

(4) *As outras mortes foram causadas pela queda de ramos ou desmoronamento de estruturas.* (Lusa2_109)
The other deaths were caused by falling branches or collapsing structures.

To represent these configurations while preserving the ISO requirement that causal relations involve only two arguments, four additional roles were introduced within the *Situation* entity: *Conjunction1*, *Conjunction2*, *Disjunction1*, and *Disjunction2*. These roles instantiate the ISO definitions of the Conjunction relation (both arguments hold: Arg1 Arg2) and the Disjunction relation (non-exclusive alternatives: Arg1 Arg2). Specifically, *Conjunction1* and *Conjunction2* capture multiple elements converging on a shared situation (multiple reasons → one result; one reason → multiple results), while *Disjunction1*and *Disjunction2* capture alternative elements (multiple alternative reason → one result; one reason → multiple alternative results). This approach allows coordinated or alternative causal elements to be represented as components of a single argument, while still allowing more complex causal configurations to be represented.

A further difficulty concerns the annotation of discontinuous spans due to limitations of the INCEpTION annotation tool, used in this project, which does not allow non-contiguous segments to be annotated as a single markable. This situation occurs in examples such as:

(5) *a morte por eletrocussão de um homem de 50 anos* (Lusa2_109)
the death by electrocution of a 50-year-old man

In this example, the argument with the role of *Result* includes both *a morte* and *de um homem de 50 anos*. To preserve the semantic unity of the argument, a relation called *sameSpan* was introduced to link the discontinuous segments belonging to the same argument.

Finally, difficulties also arose in identifying the events associated with each causal relation in sentences containing multiple events. For instance:

(6) *Cerca de 1.000 pessoas ficaram feridas e mais de 14 mil casas foram destruídas na sequência do sismo mais mortífero dos últimos nove anos na China.*
Around 1,000 people were injured and more than 14,000 houses were destroyed following the deadliest earthquake in China in the last nine years.

In this case, the events *feridas* and *destruídas* must be explicitly associated with the corresponding causal argument. The introduction of the *NucleusOf* relation made it possible to link these events to their corresponding *Situation* argument and thus to determine their role in the causal relation.

4. CATS: Causal and Temporal Annotation Scheme

After several rounds of experimentation and refinement, the final version of the annotation scheme was established. Figure 4 presents the complete CATS scheme: an annotation scheme of causality and temporal structure.

As shown in Figure 4, the entity structure includes the labels *Event*, *Situation*, and *Signal*, each associated with a set of attributes and values. The link structure includes the relations *SameSpan*, *NucleusOf*, two technical links, and *Signal*, and *Discourse*, the latter encompassing both causal and temporal relations.

The newly proposed label *Situation* tag corresponds to the arguments of discourse relations and is realized by simple or complex linguistic expressions such as clauses, nominalizations, phrases, or sets of phrases. The roles associated with *Situation* are the following: *Reason* and *Result*, corresponding to the two arguments of a causal relation; *Conjunction1* and *Conjunction2*, used to represent coordinated causes or effects; *Disjunction1* and *Disjunction2*, used to represent alternative causal configurations.

The *Signal* label refers to lexical items or multiword expressions that explicitly mark a causal relation. In the annotation scheme, signals are not considered part of the causal arguments but function exclusively as markers linking the arguments. The attributes associated with this label correspond to different grammatical categories of causal signals: *Verb* - verbs that encode causality in their semantics (e.g., *provoke*, *cause*, *lead to*); *Noun* - nouns referring to the cause or origin of an event (e.g., *cause*, *origin*, *motive*); *Conjunction* - conjunctions and conjunctive expressions linking clauses to express cause or explanation (e.g., *because*, *since*, *as*); *Preposition* - prepositions or prepositional expressions introducing causal relations (e.g., *due to*, *because of*, *thanks to*); *Adverb* - adverbs or adverbial expressions expressing causal relations (e.g., *thus*, *consequently*).

The *Event* tag represents occurrences, states, or circumstances described in the text. Following ISO 24617-1: 2012 and the annotation manual by Silvano et al. (2023), events are characterized by several attributes: *Class* - occurrence, state, reporting, perception, aspectual, IState, IAction; *Type* - aspectual type (state, process, or transition); *Pos* - grammatical category (verb, noun, adjective, or preposition);*Tense* - temporal location of the event

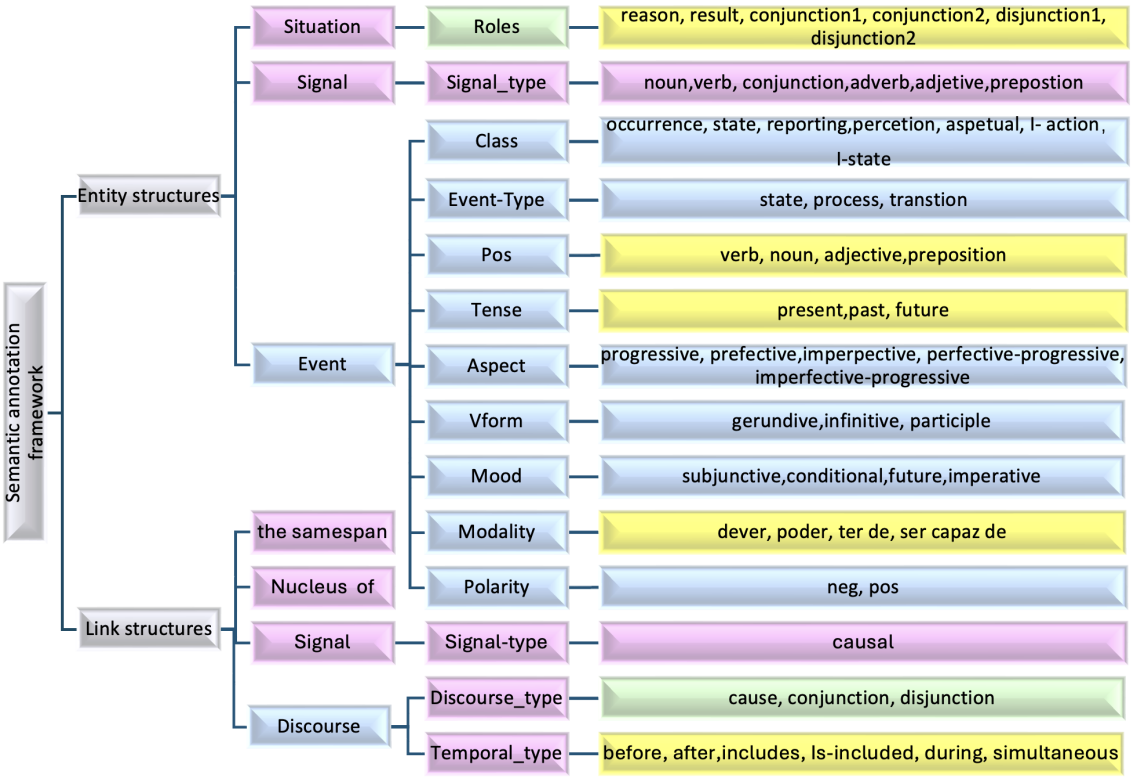

Figure 4: CATS: Annotation scheme for causality and temporal structure. Color coding: blue indicates labels/attributes from ISO 24617-1:2012; green indicates labels/attributes from ISO 24617-8:2016; yellow indicates labels/attributes adapted from either standard; pink indicates newly proposed labels/attributes.

(present, past, future); *Aspect* - aspectual value (progressive, perfective, imperfective, imperfective progressive); *VForm* - non-finite verb forms (gerund, infinitive, participle); *Mood* - verbal mood (subjunctive, conditional, future); *Modality* - modal value (e.g., *dever* 'must, *poder* 'can'); *Polarity* - positive or negative.

The link structure includes relations representing both semantic and technical connections between annotated elements. The *Discourse* relation includes both causal and temporal relations. A causal relation is established when one argument provides an explanation for another. In the annotation scheme, the causal relation connects a *Reason* argument to a *Result* argument, with directionality from *Reason* to *Result*. Temporal relations represent the ordering between events associated with the causal arguments. The values of temporal relations include: *Before* - one event precedes another; *After* - one event follows another; *Simultaneous* - two events occur at the same time; *Includes* - one event temporally includes another; *IsIncluded* - one event is temporally included in another.

In addition to these semantic relations, the scheme includes three technical relations: *Signal* - links a causal signal to the corresponding causal relation, connecting the signal to the *Result* argument; *NucleusOf* - links events to the situations (*Reason* or *Result* arguments) in which they occur; *the sameSpan* - connects non-contiguous segments belonging to the same causal argument.

In our proposal, the annotation procedure should be the following:

1. Identification of textual segments instantiating a causal relation;
2. Annotation of the *Reason* and *Result* arguments;
3. Identification and annotation of the main events of the *Reason* and *Result* arguments;
4. Identification and annotation of *causal signals*;
5. Annotation of *NucleusOf* relations linking events to arguments;
6. Annotation of *causal* and *temporal relations*;
7. Annotation of *signal relations*.

To identify causal relations, we draw on five tests proposed by (Grivaz, 2010; Dunietz et al., 2017): *Why* test, *Temporal Order* test, *Counterfactual* test, *Ontological Asymmetry* test, and *Linguistic* test. The *Why* test assesses whether the effect can be

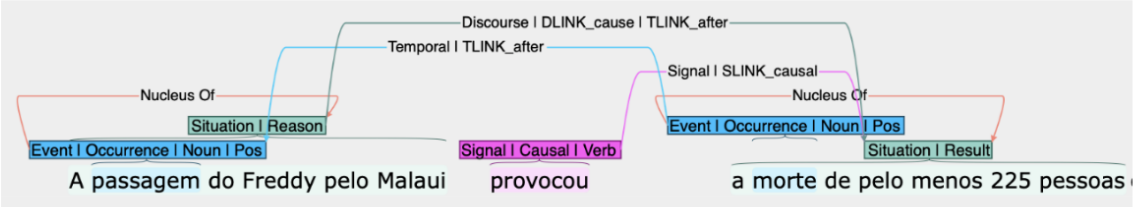

Figure 5: Annotated example according to the CATS scheme.

explained by asking why it occurred; *Temporal Order* checks whether the cause precedes the effect; *Counterfactual* test evaluates whether the effect would not occur in the absence of the cause; *Ontological Asymmetry* captures the non-symmetrical nature of causality; and *Linguistic* verifies whether the relation can be paraphrased as "It is because (of) X that Y" or "X causes Y." In this work, we adopt the *Why*, *Ontological Asymmetry*, and *Linguistic* tests as the most relevant. The *Temporal Order* test is excluded, as causal relations may involve simultaneity or posteriority. The *Counterfactual* test is also excluded, since the notion of causality adopted in ISO 24617-8:2016 is broader and does not always entail strict counterfactual dependence.

Example (7) illustrates the annotation of a causal relation in a news sentence.

(7) *A passagem do Freddy pelo Malaui provocou a morte de pelo menos 225 pessoas.*
Freddy's passage through Malawi caused the death of at least 225 people.

The annotation of this sentence includes the following elements:

- ***Reason* argument**: "A passagem do Freddy pelo Malaui" (Freddy's passage through Malawi);
- ***Result* argument**: "a morte de pelo menos 225 pessoas" (the death of at least 225 people);
- **Events**: "passagem" (passage) and "morte" (death);
- **Causal signal**: "provocou" (caused);
- **NucleusOf**: linking the event "passagem" to the *Reason* argument and the event "morte" to the *Result* argument;
- **Causal and temporal relations**: the *Reason* argument is linked to the *Result* argument through the relation *Cause*, and a temporal relation *After* is established between the corresponding events;
- **Signal relation**: linking the signal "provocou" to the *Result* argument.

Figure 5 presents an example of annotation carried out in the INCEpTION tool (Klie et al., 2018) using the CATS annotation scheme.

5. Conclusion

This article presents a new annotation framework designed to jointly capture causal and temporal information. To achieve this objective, we integrate elements from the temporal layer defined in ISO 24617-1: 2012 and the discourse relation framework of ISO 24617-8: 2016, resulting in CATS, an annotation scheme of causality and temporal structure.

Building on the Text2Story annotation framework, the proposed model extends the entity structure with additional labels and attributes for representing events, situations, and causal signals, while the link structure incorporates relations required to encode causal and temporal dependencies between annotated elements. This integration enables a coherent representation of the interaction between causal relations and the temporal ordering of events.

CATS therefore provides a unified framework for the integrated annotation of causality and temporality, while remaining compatible with the broader ISO 24617 architecture. In addition to causal relations, the framework also supports the annotation of other discourse relations defined in ISO 24617-8: 2016.

Future work will focus on presenting the results of applying the proposed annotation scheme to a dataset of Portuguese news articles to provide evidence of its effectiveness in real annotation scenarios. Furthermore, we plan to extend the scheme by incorporating additional discourse relations, thereby broadening its applicability for the semantic annotation of complex discourse structures.

6. Acknowledgements

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

7. Bibliographical References

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press.

Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 908–915. LREC 2008.

Steven Bethard and James H. Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT*, pages 177–185. Association for Computational Linguistics.

Harry Bunt and Rashmi Prasad. 2016. ISO DR-core (ISO 24617-8): Core concepts for the annotation of discourse relations. In *Proceedings of the Joint ACL-ISO Workshop on Interoperable Semantic Annotation*.

Francisco Costa and António Branco. 2012a. Extracting temporal information from Portuguese texts. In Helena Caseli, Aline Villavicencio, António Teixeira, and Fernando Perdigão, editors, *Computational Processing of the Portuguese Language*, volume 7243 of *Lecture Notes in Computer Science*, pages 105–116. Springer, Berlin, Heidelberg.

Francisco Costa and António Branco. 2012b. TimeBankPT: A TimeML annotated corpus of Portuguese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3727–3734, Istanbul, Turkey. European Language Resources Association (ELRA).

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 188–196.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Ana Luísa Fernandes, Purificação Silvano, António Leal, Nuno Guimarães, Rita Rb-Silva, Luís Filipe Cunha, and Alípio Jorge. 2025. The incremental process of building an annotation scheme for clinical narratives in Portuguese: the contribution of human variation analysis. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 332–343, Vienna, Austria. Association for Computational Linguistics.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18.

Cécile Grivaz. 2010. Human judgements on causation in French texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah R. Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajic, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *Künstliche Intell.*, 35(3):343–360.

Jerry R. Hobbs. 1985. On the coherence and structure of discourse. Technical Report CSLI-85-37, Center for the Study of Language and Information (CSLI), Stanford University.

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

ISO. 2016. ISO 24617-8: 2016, Language Resource Management - Semantic Annotation

Framework (SemAF) - Part 8: Semantic RRelations in Discourse (DR-Core). Technical report, International Organisation for Standardisation ISO, Geneva.

Andrew Kehler. 2002. *Coherence, Reference, and the Theory of Grammar*. CSLI Publications.

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, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mahnaz Koupaee, Xueying Bai, Mudan Chen, Greg Durrett, Nathanael Chambers, and Niranjan Balasubramanian. 2025. Causal graph based event reasoning using semantic relation experts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26169–26199.

Alex Lascarides and Nicholas Asher. 1993. Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy*, 16(5):437–493.

António Leal, Purificação Silvano, Evelin Amorim, Inês Cantante, Fátima Silva, Alípio Mario Jorge, and Ricardo Campos. 2022. The place of ISO-space in Text2Story multilayer annotation scheme. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 61–70, Marseille, France. European Language Resources Association.

Inderjeet Mani and Barry Schiffman. 2004. Temporally anchoring and ordering events in news.

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

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Marc Moens and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288.

Qiang Ning, Hao Wu, and Dan Roth. 2018b. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Siyana Pavlova, Maxime Amblard, and Bruno Guillaume. 2024. YARN is all you knit: Encoding multiple semantic phenomena with layers. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations @ LREC-COLING 2024*, pages 66–76, Torino, Italia. ELRA and ICCL.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press, Cambridge.

Rashmi Prasad and Harry Bunt. 2015. Semantic relations in discourse: The current state of ISO 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pages 2961–2968, Marrakech, Morocco. European Language Resources Association.

James Pustejovsky, Jose M. Castano, Robert Ingria, Roser Sauri, Robert Gaizauskas, Andrea Setzer, and Graham Katz. 2003a. TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The timebank corpus. In *Proceedings of Corpus Linguistics 2003*, pages 647–656.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

(LREC'10), Valletta, Malta. European Language Resources Association (ELRA).

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Ted Sanders, Wilbert Spooren, and Leo Noordman. 1992. Toward a taxonomy of coherence relations. *Discourse Processes*, 15:1–35.

Andrea Setzer. 2001. *Temporal Information in Newswire Articles: an Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield.

Purificação Silvano, Evelin Amorim, António Leal, Inês Cantante, Alípio Jorge, Ricardo Campos, and Nana Yu. 2024. Untangling a web of temporal relations in news articles. In *Proceedings of the Text2Story'24 Workshop*. Glasgow, Scotland, 24 March 2024 (workshop proceedings).

Purificação Silvano, Alípio Mário Jorge, António Leal, Evelin Amorim, Hugo Sousa, Inês Cantante, Ricardo Campos, and Sérgio Nunes. 2023. Text2story lusa annotated corpus. Dataset.

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.

Steven A. Sloman. 2005. *Causal Models: How People Think about the World and Its Alternatives*. Oxford University Press.

Leonard Talmy. 2000. *Toward a Cognitive Semantics, Volume II: Typology and Process in Concept Structuring*. MIT Press.

Fiona Anting Tan, Liyuan Lee, and Yejin Choi. 2022. Causal news corpus: Annotating causality in event sentences from news. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, pages 2298–2308, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

James Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, 64:1161–1186.

ISO-TimeML semantics for interlinking annotations

Harry Bunt[1]
[1]Tilburg University, Netherlands
harry.bunt@tilburguniversity.edu

Alex Fang[2]
[2]City University of Hong Kong
alex.fang@cityu.edu.hk

Kiyong Lee[3]
[3]Korea University, Seoul
ikiyong@gmail.com

Volha Petukhova[4]
[4]Saarland University, Saarbrücken
v.v.petukhova@gmail.com

James Pustejovsky[5]
[5]Brandeis U., Waldham, USA
jamesp@brandeis.edu

Purificação Silvano[6]
[6]University of Porto
puri.msilvano@gmail.com

Abstract

This paper describes a step in the development of a methodology for combining annotation made with different annotation schemes. The methodology, called 'interlinking', assumes that different annotations of the same data will contain certain elements that refer to the same entities. This can be represented by a set of 'identity links' . These links are used for constructing a single, integrated annotation structure at the level of abstract syntax with a semantic interpretation. In this paper we focus on the interlinking of annotations of time and events with ISO-TimeML (ISO 24617-1:2012) and quantification with QuantML (ISO 24617-12:2025). interlinking annotations is in practice only feasible if the respective annotation schemes use the same or convertible representation formalisms. Since QuantML and ISO-TimeML use different formalisms and QuantML has a more fully developed semantics than ISO-TimeML, we developed a new, DRT–based semantics for ISO-TimeML which is presented and discussed in this paper.

Keywords: Semantic annotation, combining annotations, interlinking, QuantML, ISO-TimeML

1. Introduction

This paper reports on activities within the ISO 'preliminary work item' PWI 254617-17, 'Interlinking of annotations using the 'interlinking' approach (Bunt, 2024). which investigates the possibility to combine annotations made with different annotation schemes, This approach has been suggested for the joint application of different parts of the ISO Semantic Annotation Framework ('SemAF'), a family of standards each of which was developed for representing a certain type of semantic information, such as semantic roles, discourse relations, coreference, and information about events, time and space. Combining annotations from different parts would lead to richer, more powerful semantic annotations.

The annotations of two annotation schemes can in principle be combined easily if they are disjoint, resulting in a single annotation if they use the same representation format. This is the case for various SemAF parts, but some parts have domains that are not entirely disjoint. In particular, certain semantic phenomena that form the focus of one SemAF-part often also play a role, albeit of secondary importance, in another. In such a case the application of two annotation schemes results in two annotations that capture partly the same semantic information in different and possibly incompatible ways.

Early in 2025 a long desired new member was born in the SemAF family. The new member, called QuantML after the markup language that it adopts, was officially registered as ISO 24617-12:2025, Quantification. In view of the ubiquitous nature of quantification phenomena in natural language, the new member was most welcome. QuantML and ISO-TimeML exemplify the phenomenon of application domains that are not entirely disjoint. Consider, for example, the sentence *"Most of the guests arrive on Friday"*. Using QuantML, this sentence is analysed as expressing the occurrence of *arrive* events with guests as agents and with Friday as the entity that plays the semantic role *Time*. In ISO-TimeML, by contrast, events are viewed as anchored in time with special temporal relations such as *before, during, simultaneous*, and *included*. The combination of annotations of these two schemes is therefore not straightforward, while it would be very interesting because of their joint coverage of aspects of meaning. This paper explores the use of interlinking for combining annotations of these two schemes, with a focus on the semantic interpretation of the combined annotations.

The paper is structured as follows. Section 2 describes the interlinking approach in relation to the architecture of SemAF parts. Section 3 summarizes those aspects of QuantML and ISO-TimeML that are of particular interest for interlinking. Section 4 presents the semantic interpretation of ISO-TimeML annotations using the same DRT-based framework as used in QuantML. Section 5 discusses the semantic interpretation of interlinked QuantML - ISO-TimeML annotations. Section 6 closes with concluding remarks. Finally, Appendix A contains a detailed example of interlinking the annotations of the two schemes.

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 123–133
12 May 2026.

2. Interlinking

Interlinking aims at combining the annotations of textual, spoken or multimodal data according to different parts of the SemAF family without altering the original annotations. It exploits the fact that the two (or more) separate scheme-specific annotations apply to the same source material by identifying the elements in these annotations that refer to the same semantic entity and adding identity links between these elements.

In order to see what is involved in interlinking annotations of two SemAF parts, we consider the architecture of a SemAF scheme as shown schematically in Fig. 1. The following three levels are distinguished:

- the concrete syntax, which specifies a representation format. Any representation format that is expressive enough can be chosen (see Fig. 1), but XML is most commonly used In SemAF.
- the abstract syntax, which expresses the semantically relevant information of the annotations in the form of pairs, triples, and other set-theoretical structures.
- the semantics, which defines the meaning of annotations in terms of representations with a well-defined semantics.

These levels are interrelated through decoding and interpretation functions.

At the level of concrete syntax, interlinking consists of adding identity links between components of representations from different schemes. This is a rather straightforward matter (although there may be some issues in the identification and use of markables). The real challenges lie at the other two levels. In particular, sitting in between the levels of concrete representation and semantics, the combination of annotations at the level of abstract syntax faces a dual challenge. On the one hand, the expressions at this level should have a systematic encoding-decoding relation to concrete representations; 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 and QuantML are two of the best developed and most complex SemAF parts, a sensible strategy for exploring the possibilities and limitations of interlinking is to explore the possible combination of their respective annotations, in particular at the levels of abstract syntax and semantics.

Table 1 indicates the development of applying the interlinking approach to ISO-TimeML and QuantML. Both schemes have a fully developed

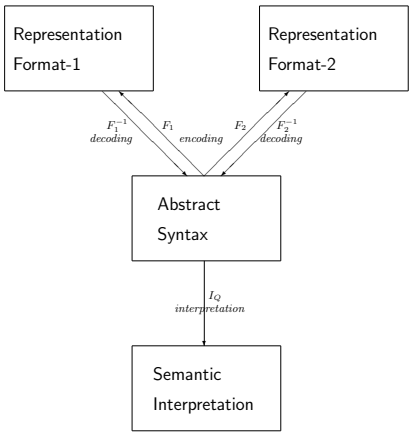

Figure 1: Levels and interrelations in SemAF annotation schemes.

	ISO-Time TimeML	QuantML	Inter-linking
concrete syntax	√	√	√
decoding	1	√	1
abstract syntax	1	√	1
inter-pretation	2	√	2
semantics	2	√	2

Table 1: √ = done before start of PWI, 1 = presented in XXXX (YYYY), 2 = topic of this paper

concrete syntax with XML-based reference representation format. QuantML is also fully developed at the other levels and their interrelations. ISO-TimeML, being the oldest SemAF part, has a partially developed abstract syntax and lacks the specification of a decoding function; proposals for these components have been presented in Bunt et al. (2025) together with the abstract syntax and decoding function for interlinking elements. At the level of semantics and interpretation function, ISO-TimeML is partly developed and uses different forms of semantics than QuantML. This paper aims to provide the remaining ingredients for effective interlinking by (1) defining a semantics for ISO-TimeML annotations in the same DRT-based style as the QuantML semantics (Section 4), and (2) outling the semantic interpretation of integrated annotation structures (Section 5).

3. Two linked annotation schemes

3.1. Theoretical background

The ISO-TimeML and QuantML annotation schemes are both theoretically rooted in Davidsonian event semantics (Davidson, 1967; Parsons, 1990) in which verbs are viewed as denoting events with participants (rather than as predicates with arguments). This view is extended in ISO-TimeML, acknowledging that certain nouns, like *"concert", "meeting"* and certain adjectives *"the dormant volcano"* also denote events. The focus is on the temporal anchoring of events and on temporal relations among events, as expressed by temporal, aspectual and subordination relations.

QuantML, on the other hand, focuses on the involvement of participants in events. Participants have semantic roles, like *Agent* and *Theme*, and may be involved individually as well as collectively. Time of occurrence is treated as a participant which is a temporal object, with *Time* as semantic role. Quantification arises when a participant is not a single individual but a set of individuals or when a verb or other event-denoting expression does not refer to a single event but to a set of events, which may be quantified over (as in *"Tom called twice every Sunday"*).

ISO-TimeML additionally takes inspiration from the general framework of Allen's algebra of temporal intervals (Allen, 1984), which uses 13 relations between intervals, including equality. For QuantML, the main source of inspiration has been the theory of generalised quantifiers (GQT, Barwise and Cooper, 1981, Cooper, 1983). In this theory, natural language quantifiers are (mostly) NPs like *"Each of the students"* and *"Three kilos of sugar"*.

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. relative participant scopes
10. event scope
11. repetitiveness
12. size of reference domain
13. restrictiveness of modifiers
14. linking of modifiers (inverse or linear)

These categories correspond to attributes of XML elements in the concrete syntax of QuantML. Some of these attributes are optional and have a default value: polarity is by default 'positive', exhaustivity is 'negative', event scope is 'narrow', and repetiveness is 'at least once'. For explanations and discussion of all the categories see.(Bunt, 2020) and ISO (2025).

3.2. Concrete syntax

QuantML and ISO-TimeML both have a concrete syntax that defines a reference representation format using XML. While in the SemAF architecture shown in Figure 1 (See ISO 2015) the choice of representation format is of secondary importance, having the same representation format facilitates the interlinking at the level of concrete syntax. This is illustrated in (2), which shows an interlinked annotation, consisting of three parts: (1) the ISO-TimeML annotation, (2) the QuantML annotation, and (3) the identity statements that link the two.

(2) Some of the students will protest on Friday.

Markables:
m0 = "Some of the students will protest on Friday", m1 = "Some of the students", m2 = "the students", m3 = "students", m4 = "will protest", m5 = "on Friday"

```
<linkedRep> xml:id="tq1" target="#m0">
  <tmeML xml:id="crT">
    <EVENT xml:id="eT1" target= “#m4" pred=
      "protest" class="OCCURRENCE" type=
      "PROCESS'"/>
    <TIMEX3 xml:id="tT1" target= "#m5"
      type=“DATE" value="XXXX-YY-05"/>
    <TLINK eventID="#eT1" relatedToTime=
      "#tT1" relType="IS-INCLUDED"/>
    </timeML>
 <quantML xml:id="crQ">
    <event xml:id="eQ1" target="#m4"
      pred="protest"/>
    <entity xml:id="xQ1" target="#m1" ref-
      Domain="#xQ2" individuation="count"
      involvement="some"/>
    <refdomain txml:id="xQ2" target="#m2"
      pred="student" determinacy="det"/>
    <participation event="#eQ1"
      participant="#xQ1" semRole="agent"
      distr="collective" evScope="narrow" />
    <entity xml:id="xQ3" target="#m5" ref-
      Domain="#xQ4" individuation="count"
      involvement="some"/>
    <refdomain xml:id="xQ4" target="#m5"
      pred="friday" determinacy="det"/>
    <participation event="#eQ1"
      participant="#xQ3" semRole="time"
      distr="individual" evScope=
      "narrow"/>
    </quantML>
 <idL xml:id="crIL">
  <idLink xml:id="i1" arg1="#eQ1" arg2="#eT1"/>
```

<idLink xml:id="i2" arg1="#xQ2" arg2="#tT1"/>
</idL>
</linkedRep>

A linked annotation structure can be represented schematically as follows, where X_T and X_Q are the linked ISO-TimeML and QuantML annotations, and X_{IL} is the set of linking identity specifications.

(3) <linkedRep> xml:id="tq1" target="#m0">
X_T
X_Q
X_{IL}
</linkedRep>

The example shows how QuantML uses the core elements <event>, <entity> and <participation> for representing semantic information about participants and the way they participate in events. In addition, QuantML has a number of elements for representing modification by adjectives, prepositional phrases, possessive phrase, and relative clauses occurring in natural language expressions that describe quantified participants.

ISO-TImeML uses the core elements <EVENT>, <TIMEX3> and <TLINK> for representing information about events, temporal objects and temporal relations, respectively. Three types of temporal objects are distinguished: instants, dates, and periods, which are all represented by <TIMEX3> elements with different values of the @TYPE attribute and with different attributes depending on the @TYPE value.

From a semantic point of view, the elements <TIMEX3> and <TLINK> are rather overloaded. For example, a <TLINK> element is used to represent the relation between an event and its time of occurrence, or a temporal relation between events, or a temporal relation between periods. This causes ambiguities for their decoding into abstract annotation structures.

3.3. Abstract syntax and decoding

The specification of an abstract syntax consists of (1) a list of primitive concepts, called the 'conceptual inventory', and (2) the recursive specification of annotation structures in the form of set-theoretic structures like pairs and triples.

The conceptual inventory of ISO-TimeML contains the following time-related basic concepts (1) named temporal objects like *January, Christmas*; (2) temporal relations, subordination relations, and aspectual relations, like *during, before, conditional, initiates*; and (3) units for measuring durations, like *hour, minute, day,*

The conceptual inventory of QuantML includes (1) quantitative and proportional predicates such as *some, several, most*; (2) dimension predicates such as *weight, length, volume* and corresponding units of measurement; (3) the possessive relation 'Poss' (due to Peters & Westerståhl, 2013).

In addition, both conceptual inventories contain (1) predicates that correspond to (senses of) natural language content words, such as *teach, concert, student*; (2) linguistic categories like *process, past, mass, indeterminate*; (3) the real numbers and numerical relations like *greater than*.

The structures defined by the abstract syntax come in two forms: (a) *entity structures*, which contain semantic information about a stretch of source data, and (b) *link structures*, which express semantic relations between two or more entity structures.

The variety of ways of representing temporal objects in the concrete syntax of ISO-TimeML requires a way of distinguishing different types of n-tuples with different meanings, as illustrated by example (4). In (4a) the phrase *"two hours before lunch"* designates an instant at a certain distance in time from an event, while in (4b) it indicates a period of a certain length. This means that the corresponding triples of the abstract syntax should be interpreted differently. In order to enable this, we introduce the use of labels such as *inst*, *per* and *val* to mark the semantic type of a structure.

(4) a. Tom arrived two hours before lunch.
⟨*inst* ⟨lunch, ⟨*amt*⟨tdistance, 2, hour⟩, before⟩⟩⟩
b. Tom taught (for) two hours before lunch.
⟨*per* ⟨lunch, ⟨*amt*⟨tdistance, 2, hour⟩, before⟩⟩⟩

Apart from the introduction of labels, the abstract syntax and the decoding of concrete representations into abstract annotations of ISO-TimeML are as defined in (NN4)). We add labels in the same way to the annotation structures of QuantML. For identity statements of the interlinking we use the label *id*. Example (5) illustrates this.

Using labels in the abstract syntax has the effect shown in (5) when the decoding functions dF_T of ISO-TimeML (see Bunt et al, 2025) and dF_Q of QuantML are applied to the annotation representation of the sentence *"John called at midnight"*, which is analysed in detail in Appendix A. This formulation of the decoding functions makes use of an auxiliary function, $IdEl$, which, applied to an annotation representation X of the concrete syntax, selects the element with identifier x_i. For example, $IdEl(X_{IL}, i2)$ = <idLink xml:id="i2" arg1="#xQ2" arg2="#tTt"/>.

(5) **a. QuantML**
$A_Q = dF_Q(X_Q) = \langle \epsilon_{Qe}, \langle \epsilon_{xQ1}, \epsilon_{xQ2} \rangle, \langle pL_{Q1}, pL_{Q2} \rangle, \langle scL_1 \rangle \rangle$
$\epsilon_{Qe} = dF_Q(\text{IdEl}(X_Q, e_Q)) = \langle$*evq*, ⟨call, past⟩$\rangle$
$\epsilon_{Qx1} = dF_Q(\text{IdEl}(X_Q, xQ1))$
⟨*ent*, ⟨student, determinate, count⟩⟩
$\epsilon_{xQ2} = dF_Q(\text{IdEl}(X_Q, xQ2))$

⟨*ent*, ⟨midnight, determinate, count, 1⟩⟩
pL_{Q1} = ⟨*pLink*, ⟨ϵ_{Qe}, ϵ_{Qx1}, agent, individual⟩⟩
pL_{Q2} = ⟨*pLink*, ⟨ϵ_{Qe}, ϵ_{Qxt}, time, individual⟩⟩,
scL_{Q1} = ⟨*scope*, ⟨pL_{Q1}, pL_{Q2},wider⟩⟩

b. ISO-TimeML
$A_T = dF_T(X_T)$ = ⟨ϵ_{Te}, ⟨ϵ_t⟩, ⟨tL_1⟩ ⟨⟩⟩
$\epsilon_{Te} = dF_T(IdEl(X_T, e_t))$
= ⟨*evt*, ⟨call, occurrence, transition, past⟩⟩
$\epsilon_{Tt} = dF_T(IdEl(X_T, tT1)$ = ⟨*inst,* ⟨T00:00⟩⟩
tL_{T1} = ⟨*tLink*, ⟨ϵ_{Te}, ϵ_{Tt}, included⟩⟩

c. Interlinking:
$A_{IL} = dF_{IL}(X_{Il})$
= ⟨dF_{IL}(IdEl($X_I, i1$), dF_{IL}(IdEl($X_{IL}, i2$))⟩
= ⟨⟨$\epsilon_{Qe}, \epsilon_{Te}$⟩, ⟨$\epsilon_{xQ2}, \epsilon_{Tt}$⟩⟩

The interlinking of QuantML and ISO-TimeML annotations includes the construction of an integrated abstract annotation structure from the two concrete annotation representations and their decodings. To this end a function dF_{QT} is defined which combines the results of the decoding functions dF_Q and dF_T, plus the function dF_{IL} defined for the <*idLink*> structures.

This latter function embodies a key aspect of the interlinking approach, namely that identity statements correspond to pairs of the entity and event structures involved. The function dF_{QT} substitutes these pairs for their elements in link structures (participation structures, scope relations structures, and temporal as well as subordination and aspectual relation structures). Auxiliary functions (PLink, ScopeLink, SubLink, and AspLink) implement this for the various types of link structures.

PLink, for example, combines two participation link structures into a single link structure in which the participants of the two arguments are paired, their semantic roles are paired, and their parameters, extracted by the auxiliary function Params, are checked for being compatible. These parameters are the distributivity, event scope, exhaustiveness and polarity which occur in a participation link structure. Example (6) illustrates this.

(6) **Interlinked annotation structure:**
A_{QT} = ⟨ϵ_e, ϵ_x, pL, scL⟩, where
$\epsilon_e = dF_{IL}$(IdEl(X_{Il}), $i1$) = ⟨$\epsilon_{Qe}, \epsilon_{Te}$⟩
ϵ_x = ⟨$dF_{QT}(\epsilon_{Qx1}), dF_{QT}(\epsilon_{xQ2}), dF_{QT}(\epsilon_{Tt})$⟩
= ⟨ϵ_{Qx1}, ⟨$\epsilon_{xQ2}, \epsilon_{T1}$⟩⟩
$=_D$ ⟨ϵ_1, ϵ_2⟩
$pL = \{pL_1, pL_2\}$
$pL_1 = pL_{Q1}$
pL_2 = PLink(⟨pL_{Q2}, tL_{T1}⟩) =
= ⟨ϵ_e, ϵ_2, ⟨SRole(pL_{Q2}), SRole(tL_{T1}⟩, Params(pL_{Q2})⟩⟩
= ⟨ϵ_e, ϵ_2, ⟨time, simultaneous⟩, ⟨individual, narrow, non-exh, positive⟩⟩
$scL = \{dF_{QT}(scL_{Q1}\})$⟩

4. ISO-TimeML semantics using DRT

In this section we develop a semantic interpretation-by-translation for ISO-TimeML annotations in terms of second-order discourse representation structures (DRSs, Kamp & Reyle 1993). The discourse referents in this semantics are non-empty sets of individuals (events and temporal objects included). We use a simplified linear notation for DRSs, for example, [X | x ∈ X → α], suppressing the condition |X| ≥ 1.

To keep the semantics as simple as possible, the elements of the conceptual inventories are assumed to be available as building blocks of semantic representations under the same name, therefore: $I_T(c) = c$ and $I_Q(c) = c$.

4.1. Event structures

An event structure is a labeled 6-tuple ⟨evt, ⟨P, Ty, C, Te, As, V⟩⟩, consisting of an event predicate, an event type, an event class, a tense, an aspect, and a veracity. Semantic interpretation:

I_T(⟨evt, ⟨P, T_y, C, Te, As, positive⟩⟩) =
= [E | e ∈ E → P(e), Ty(e), C(e), Te(e), As(e))]

I_T(⟨evt, ⟨P, T_y, C, T, A, negative⟩⟩) =
= ¬[E | e ∈ E → P(e), Ty(e), C(e), Te(e), As(e))]

Example: "wrote":
I_T(⟨evt, ⟨write, occurrence, process, past, imperfective, positive⟩⟩ =
= [E | e ∈ E → write(e), occurrence(e), process(e), imperfective(e), past(e)]

4.2. Temporal entities

Temporal entities fall into six categories: (1) instant, (2) date, (3) period, (4) set of any of these, (5) amount of time, and (6) frequency. Entities of the respective categories are distinguished in the abstract syntax by the labels *inst, date, per, qset, amt* and *freq*. Formally, a label can be viewed as just an additional element in an n-tuple, so in the abstract syntax all entity structures have the form ⟨type, P⟩, where 'P' is a sequence of properties.

Named temporal entities can be divided into those that function exclusively as proper names and hose that can additionally be quantified over. The first category contains calendar years and clock times, the second contains calendar months (*"January"*), calendar weekdays (*"Wednesday"*), and named special days (*"Christmas", "Thanksgiving"*). 'For the use as a proper name, as in (7a), ISO-TimeML annotation makes use in the concrete syntax of the @temporalFunction attribute, giving it the value "TRUE" to indicate that a specific temporal entity has to be determined in a postprocessing

stage. The abstract syntax and semantics of a named temporal entity used as a quantifier is illustrated in (7b,c).

(7) a. *"Friday"* used as a proper name:
"Jelle will graduate on Friday"
A_T = ⟨*date*,⟨*val*,⟨day, friday⟩⟩⟩
Semantics:
$I_T(A_T)$ = [X | |X|=1, x ∈ X → day(x)=friday$_0$],
where 'friday-0' is the specific friday calculated in postprocessing.

b. *"Friday"* used as an existential quantifier:
"Most students graduate on (a) Friday":
A = ⟨$qset$, ⟨*val*⟨day, friday⟩, some⟩⟩
Semantics:
$I_T(A_T)$ = [X| x ∈ X → day(x)=friday]

c. *"Friday"* used as a universal quantifier:
"Bill teaches every Friday":
A = ⟨$qset$, ⟨*val*⟨day, friday⟩, all⟩⟩
Semantics:
$I_T(A_T)$ = [X| x ∈ X ↔ day(x)=friday]

In the remainder of this section we specifiy the semantic interpretation of the annotations of the six categories listed above.

4.2.1. Instants

The four possible forms of an instant structure have the following semantic interpretation.

1. A single clock time: a constant, such as "16:45" or *"midnight"*. Semantic interpretation:
I_T(⟨*inst*, t_1⟩) =
= [T | t ∈ T → clocktime(t) = I_T(t$_1$)]
= [T | t ∈ T → clocktime(t) = t_1]

2. A labeled pair ⟨day name, clock time⟩⟩:
I_T(⟨*inst*, ⟨d_1t_1⟩⟩) =
= [T | t ∈ T → [day(t) = I_T(d$_1$),
clocktime(t) = I_T(t$_1$)]

3. A labeled triple ⟨instant, amount of time, relation⟩ or ⟨*inst*, ⟨event, amount of time, relation⟩⟩ (*"half an hour before midnight/departure"*).

The interpretation of such a triple makes use of a temporal instance of the 'glue merge' operation defined in QuantML, where it is used to interpret participation link structures. This operation is defined as follows:

$\cup^t$(t$_1$, a$_1$, R_1) = $[T|t \in T \rightarrow [X|x \in X \rightarrow I_T(t_1)(x), I_T(a_1)(t,x), I_T(R_1)(t,x)]]$

Using this operator:
I_T(⟨*inst*, ⟨t$_1$, a$_1$, R_1⟩⟩) = $\cup^t(I_T$(t$_1$), I_T(a$_1$), I_T(R$_1$))

Example: *"half an hour before midnight."*
I_T(⟨*inst*⟨midnight$_0$, ⟨*amt*, ⟨tdistance, 0.5, hour⟩⟩,
before⟩⟩ =
= [T | t ∈ T →] [Z | $z \in Z \rightarrow$ clocktime(z)=T00:00,
before(t, z)), tdistance((t, z), hour) = 0.5]]

4.2.2. Dates

Date structures, having four possible forms, have the following semantic interpretation.

1. A labeled triple ⟨*date*, ⟨year, month, day⟩⟩:
I_T(⟨*date*, ⟨y1, m1, d1⟩⟩ =
= [T| $t \in T \rightarrow$ year(t) = $I_T(y1)$,
month(t) = $I_T(m1)$, day(t) = $I_T(d1)$]

Example:
I_T(⟨*date*, ⟨2025, may, 25⟩⟩⟩)
= [T| $t \in T \rightarrow$ year(t) = 2025, month(t) = may,
day(t) = 25]

2. A labeled pair ⟨year, month⟩ (*"May 2026"*).
I_T(⟨*date*, t⟨y1, m1⟩⟩ =
= [T| $t \in T \rightarrow$ year(t) = $I_T(y1)$, month(t) = $I_T(m1)$
]
Example:
I_T(⟨*date*, ⟨⟨*val*, ⟨year, 2026⟩⟩, ⟨*val*⟨month, may⟩⟩⟩⟩⟩)
= [T| $t \in T \rightarrow$ year(t) = 2026, month(t) = may

3. A labeled pair ⟨month, day number⟩:
I_T(⟨*date*, ⟨m1, d1⟩⟩ =
= [T| $t \in T \rightarrow$ month(t) = $I_T(m1)$, day(t) = $I_T(d1)$]

Example: (*"December 25"*)
I_T(⟨*date*, ⟨december, 25⟩⟩)
= [T| $t \in T \rightarrow$ day(t) = 25, month(t) = december]

4. A labeled pair ⟨week, day name⟩:
I_T(⟨*date*, ⟨w1, d1⟩⟩ =
= [T| $t \in T \rightarrow$ week(t)=$I_T(w1)$,dayname(t)=$I_T(d1)$]

Example: (*"week 20 on Tuesday"*).
I_T(⟨*date*, ⟨w20, tuesday⟩⟩)
= [T| $t \in T \rightarrow$ week(t) = w20, dayname(t) = tuesday]

4.2.3. Periods

The semantics of period structures makes use of an auxiliary function 'AbsInt' which, given a DRS $]A|C_1, a \in A \rightarrow \alpha]$ and a binary relation R, moves the relation inside the DRS while introducing a lambda abstraction as follows:
AbsInt($]A|C_1, a \in A \rightarrow \alpha], R$) =
= $\lambda x.]A|C_1, a \in A \rightarrow \alpha, R(x, a)]$.

Using this function, the possible forms of period structures have the following semantic interpretation.

1. A labeled pair ⟨begin point, end point⟩:
I_T(⟨*per*, ⟨t1, t2⟩⟩ =

= [T| $t \in T \rightarrow$ AbsInt($I_T(\langle$*inst*, $\langle$t1$\rangle\rangle)(t)$, begin),
AbsInt($I_T(\langle$*inst*, $\langle$t2$\rangle\rangle)(t)$, end)]

Example: *From two to five on January first"*.
$I_T(\langle$*inst*, $\langle$*date*,$\langle$january,1$\rangle$, T14:00$\rangle$,
$\langle$*inst*, $\langle$*date*,$\langle$january,1$\rangle$,T17:00$\rangle\rangle$) =
= [T| $t \in T \rightarrow$ month(t) = january,
datynumber(t)=1, [$Y, Z|y \in Y \rightarrow$
clocktime(y) = 14:00, included(z, x),
begin(t, y)], $z \in Z \leftrightarrow$ clocktime(z) =
17:00, included(z, x), end(t, z)]]

2. A labeled triple ⟨instant, time amount, relation⟩, formed by the beginning or end of the period, its length, and the relation 'before' or 'after' .

   Example:
   "the last two weeks before Christmas"
   $I_T(\langle$*per*, $\langle$christmas, $\langle$*amt*, $\langle$tdistance, 2, week$\rangle\rangle$,
   before$\rangle$) =
   = [T| $t \in T \rightarrow$ before($t1$, christmas),
   distance(t, christmas) $\leq$ 2]

3. Like the previous case, but with an events as begin/end rather than an instant,
   as in *"the first two days after the attack"*. The semantics is analogous.

4.2.4. Time-amount structures

A time-amount structure is a labeled triple ⟨*amt*, ⟨dimension, rational number, temporal unit⟩⟩:
$I_T(\langle$*amt*, ⟨D, k, u⟩⟩ = $\lambda x.D(x, u) = k$.

Example: *" half an hour"*.
$I_T(\langle$*amt*,⟨tdistance, 0.5, hour⟩⟩) =
$\lambda x.$ tdistance(x,hour)= 0.5.

4.2.5. Frequency structures

A frequency is either a labeled natural number ⟨*freq*, n⟩ for *"twice"*, etc.) or a labeled pair ⟨*freq*, ⟨natural number, temporal unit⟩⟩ (*"twice a week"*).

Semantics:
$I_T(\langle$*freq*, $k\rangle$) = [$T|t \in T \rightarrow [E||E| = k, e \in E \rightarrow$
included(e, t)]]

$I_T(\langle$*freq*, $\langle k, u\rangle\rangle$) = [$T|t \in T \leftrightarrow I_T(u)(t), t \in T \rightarrow$
$[E||E| = k, e \in E \rightarrow$ included(e, t)]]

4.2.6. 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 the @quant values "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 either a labeled pair ⟨*qset*, ⟨domain, quantifier⟩⟩ or a labeled triple ⟨*qset*, ⟨domain, quantifier, frequency⟩⟩.

Semantics:
1. $I_T(\langle$*qset*, ⟨D, some⟩⟩) = [$T|t \in T \rightarrow D(t)$]
2. $I_T(\langle$*qset*, ⟨D, all⟩⟩) = [$T|t \in T \leftrightarrow D(t)$]
3. $I_T(\langle$*qset*, ⟨D, some, f⟩⟩) = [$T|$ $t \in T \rightarrow D(t)$,
   $[E|$ $I_T(f)(E), e \in E \rightarrow$ included(e, t)]]
4. $I_T(\langle$*qset*, ⟨D, all, f⟩⟩) = [$T|$ $t \in T \leftrightarrow D(t)$, $t \in T \rightarrow$
   $[E|$ $I_T(f)(E), e \in E \rightarrow$ included(e, t)]]

Examples:
1. *"Friday"* used as an existental or universal quantifier, see examples in (7).
2. Quantification with frequency, as in *"twice a week"*:
   $I_T(\langle$*qset*, ⟨week, all, $\langle$*freq*,2$\rangle\rangle$) = [$T|$ $t \in T \leftrightarrow$
   $week(t), t \in T \rightarrow [E||E|$=2,$e \in E \rightarrow$ included(e, t)]]

4.3. Link structures

The abstract syntax of ISO-TimeML has seven different link structures, distinguished by labels: for (1) anchoring events in time (label *etRel*); (2) temporal ordering of events (*eteRel*, (3) ordering of periods, dates or instants relative to each other (*ttRel*); (4) measuring a time interval (*tmRel*); (5) specifying subordination relations between events (*esRel*); (6) indicating aspectual relations between events (*eaRel*), and (7) quantified temporal anchoring a set of events (*qtRel*). Their semantics is as follows.

a. Temporal anchoring: $I_T(\langle$*etRel*,$\langle e_1, t_1, R\rangle\rangle)$ =
$\cup^+(I_T(t_1), I_T(e_1), [T|t \in T \rightarrow$
$[E|e \in E \rightarrow I_T(R)(e, t)]])$

b. Temporal event relations:
$I_T(\langle$*eteRel*$\langle e_1, e_2, R\rangle\rangle) =$
$\cup^+(I_T(e_1), I_T(e_2), [E1|e \in E1 \rightarrow$
$[E2|e' \in E2 \rightarrow I_T(R)(e, e')]])$

c. Intra-time relations:
$I_T(\langle$*ttRel*, $\langle t_1, t_2, R\rangle\rangle)$ =
$\cup^=(I_T(t_1), I_T(t_2), R)$

e , f. Aspectual and subordination relations:
identical to temporing anchoring:
$I_T(\langle$*eaRel*,$\langle e_1, e_2, R\rangle\rangle)$ =
$I_T(\langle$*esRel*,$\rangle\langle e_1, e_2, R\rangle\rangle) =$
$\cup^+(I_T(e_1), I_T(e_2), [E1|e \in E1 \rightarrow$
$[E2|e' \in E2 \rightarrow I_T(R)(e, e')]])$

g. Quantified temporal anchoring:
$I_T(\langle$*qtRel*,$\langle E_Q, T_Q, R\rangle\rangle)$ =
$\cup^+(I_T(t_1), I_T(e_1), [T|t \in T \rightarrow$
$[E|e \in E \rightarrow I_T(R)(e, t)]])$
Example: *"Carl teaches thrice a week"*:
$\cup^+$ ($I_T(\langle$*qset*, ⟨week, all, $\langle$*freq*, 3$\rangle\rangle$, [$T|$ $t \in T \leftrightarrow$
week(t), $I_T(\langle$*evt*, ⟨teach, occurrence, process⟩⟩,
included) =

= $[T|t \in T \leftrightarrow$ week(t), $t \in T \rightarrow [E||E| = 2$,
$e \in E \rightarrow$ teach(e), occurrence(e), process(e),
included(e, t)]]

5. Interlinked annotation semantics

The implementation of QuantML - ISO-TimeML interlinking includes the construction of an integrated abstract annotation structure from the decodings of the two concrete representations. To this end a function dF_{QT} is defined which combines the results of the decoding functions dF_Q and dF_T, plus the decoding function dF_{IL} applied to the <*idLink*> structures - see (5d).

In QuantML, an annotation structure is a quadruple $\langle \epsilon_{eQ}, S_Q, pL_Q, scL_Q \rangle$ consisting of an event structure ϵ_e, a set of participant structures S_Q, a set of participation link structures pL_Q, and a set of scope link structures scL_Q. An ISO-TimeML annotation structure is a similar quadruple $\langle \epsilon_{eT}, S_T, L_T, \emptyset \rangle$, with temporal, aspectual, subordinate and measure links instead of participation links, and with S_T containing temporal entities and events. The integrated abstract annotation structure has the same overall structure, with the content described in (8).

(8) A_{QT} = $\langle \epsilon_{eQT}, S_{QT}, L_{QT}, scL_{QT} \rangle$, with:

- ϵ_{eQT} = $\langle \epsilon_{eQ}, \epsilon_{eT} \rangle$
- S_{QT} is the set of all pairs of entities from S_Q and S_T that are linked through an <*idLink*> element, plus the individual elements of S_Q and S_T that are not linked.
- L_{QT} is the union of the pL_Q and L_T sets of link structures while replacing any linked entity by the pair consisting of that entity and the one that it is linked with.
- scL_{QT} = scL_Q

The semantics of the integrated annotation structure is calculated by the interpretation function I_{QT}, following the same approach as the QuantML semantics. This approach uses the information in scope link structures for combining the information in participation link structures, which in turn include the information in the event and participant structures. In QuantML the semantics of annotation structures (as defined by the abstract syntax is therefore simply the semantics of the sequence of participation links, ordered by the scope relations. This can be copied for the semantics of integrated annotations since the decoding function dF_{QT} has replaced the elements in QuantML participation links by interlinked pairs (of events and participant/temporal objects).

The interpretation function I_{QT} is thus defined as follows[1], where the order $pL1, pL_2, ..pL_n$ is determined by the scope relations in scL_Q, interpreted by the functions σ_{ij}.

(9) $I_{QT}(A_{QT})$ = $I_{QT}(\langle pL_1, pL_2, ..pL_n \rangle)$ =
= $\sigma_{12}(I_{QT}(pL_1), \sigma_{23} I_{QT}(pL_2), ...$
$\sigma_{n-1,n} I_{QT}(pL_n)$

Formulating the semantics of ISO-TmeML annotations in the form of a translation to DRSs, like the QuantML semantics, makes it possible to use the operations on DRSs defined in DRT and in QuantML. In addition, the following variant of the 'scoped merge' operation defined in QuantML is used.

Linked merge This operation generalizes the 'scoped merge' operation $\cup^*$ to take into account that the identity relations expressed by interlinking statements lead to pairs of entity structures in the integrated annotation structure, instead of single entity structures.

Scoped merge:
$\cup^*([A|a \in A \rightarrow \alpha], [B|b \in B \rightarrow \beta], R)$ =
= $[A|a \in A \rightarrow \alpha, [B|b \in B \rightarrow \beta, R(a, b)]]$

Scoped merge generalized to linked merge:
$\cup^{**}(\langle [A_1|a \in A_1 \rightarrow \alpha_1], [A_2|a \in A_2 \rightarrow \alpha_2] \rangle,$
$[B_1|b \in B_1 \rightarrow \beta_1], [B_2|b \in B_2 \rightarrow \beta_1] \rangle,$
$\langle R_1, R_2 \rangle, \langle \pi_1, \pi_2 \rangle)$ =
= $[A|a \in A \rightarrow \alpha_1, \alpha_2, [B|b \in B \rightarrow \beta_1, \beta_2,$
$R_1(a, b), R_2(a, b)]]$

The use of this operation is illustrated by the semantic interpretation of the interlinked annotation structure for the example sentence *"John called at midnight"* in Appendix A. This example shows step by step how the interlinking works, starting from the concrete representation via the integrated abstract annotation structure and down to the semantic representation computed by the combined interpretation functions.

6. Conclusion

With the introduction of labeled n-tuples in the abstract syntax of both QuantML and ISO-TimeML together with the definition of a DRT-based semantics for ISO-TimeML, we have paved the way for effectively combining QuantML and ISO-TimeML annotations (1) at the concrete syntax level by linking representations by means of identity links, (2) at the abstract syntax level by defining a decoding function that constructs an integrated annotation structure, and (3) at the level of semantics by

[1] For more information see ISO 24617-12:2025, Section B.7.

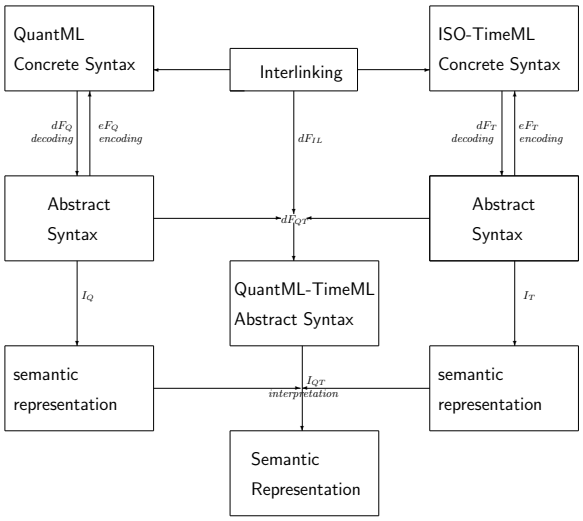

Figure 2: Interlinking of QuantML and ISO-TimeML annotation schemes.

means of operations that combine DRSs into a single integrated DRS. This is visualized in Fig. 2 and illustrated in detail by the example in Appendix A.

The next step in exploring the interlinking of ISO-TimeML and QuantML annotations is to specify the interpretation function I_{QT} systematically and in full detail. Further work also includes the study of some details that have been left out of consideration in this paper, such as the ISO-TimeML attributes @temporalFunction and @functionInDocument.

Another issue to be explored in future work is the interlinking of annotations from other SemAF schemes, such as those for discourse relations and coreference. So far, the development of interlinking as a way to build rich annotations while preserving existing annotated corpora seems promising as the basis for a new initiative to extend the ISO Semantic Annotation Framework.

Bibliographical References

J. Allen. 1984. A General Model of Action and Time. *Artificial Intelligence*, 23-29.

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

H. Bunt. 2020. Annotation of Quantification: The Current State of ISO 24617-12. In *Proceedings of the 16th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-16)*, pages 1 – 13, Marseille, France.

H. Bunt. 2024. Combining annotation schemes through interlinking. In *Proceedings ISA-20, Twentieth Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 111–121, Turin, Italy.

H. Bunt, A. Fang, K. Lee, V. Petukhova, M. Silvano, and J. Pustejpvsky. 2025. Revisiting the Abstract Syntax of ISO-TimeML. In *Proceedings ISA-21, Twenty-first Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, pages 12–21, Duesseldorf, Germany.

L. Champollion. 2015. The interaction of compositional semantics and event semantics. *Linguistics and Philosophy*, 38 (1):31–66.

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.

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.

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

S. Peters and D. Westerståhl. 2006. *Quantifiers in Language and Logic*. Oxford University Press, New York.

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.

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.

Appendix A

In this appendix we show in detail how the use of interlinking combines ISO-TimeML and QuantML annotations at the levels of concrete and abstract syntax, with a decoding function that constructs an integrated annotation structure, and an interpretation that computes the meaning of integrated annotation structures..

(10) *"John called at midnight. "*

Markables:
m0 = "John called at midnight", m1 = "John", m2 = "called", m3 = "at'", m4 = "at midnight", m5 = "midnight"

CONCRETE SYNTAX:

<linkedRep> xml:id="tq1" target="#m0">
<timeml xml:id="crT">
<EVENT xml:id="eT1"
target= "#m2" pred="call"
class="OCCURRENCE"
type="TRANSITION'"/>
<TIMEX3 xml:id="tT1" target= "#m4"
type= TIME" value="XXXX-YY-ZZT00:00"/>
<SIGNAL xml:id="s1" pred="AT"/>
<TLINK eventID="#eT1" relatedToTime="#tT1"
relType="SIMULTANEOUS"/>)
</timeml>
<quantml xml:id="crQ" target="#m0">
<event xml:id="eQ1" target="#m4"
pred="call"/>
<entity xml:id="xQ1" target="#m1" ref-
Domain="#xQ2" individuation="count"
size="1" involvement="all"/>
<refDomain txml:id="#xQ2" target="#m1"
pred="John" determinacy="det"/>
<participation event="eQ1" participant="xQ1"
semRole="agent" distr="individual"
' evScope="narrow" />
<entity xml:id="xQ3" target="#m3 ref-
Domain="#xQ4 individuation="count"
involvement="all" size="1"/>
<refDomain xml:id="#xQ4" arget="#m3"
pred="friday" determinacy="det"/>
<participation event="eQ1" participant="xQ3"
semRole="time" distr="individual"
' evScope="narrow" />
</quantml>
<idl xml:id="crIL" target="#m0">
<idLink xml:id="i1" arg1="#eQ1" arg2="#eT"1/>
<idLink xml:id="i2" arg1="#xQ2" arg2="#tT1"/>
</idl>
</linkedRep>

ABSTRACT SYNTAX:

Using the schematic notation of (3), instantiated as above:

QuantML

$A_Q = dF_Q(crQ) =$
$= \langle \epsilon_{Qe}, \langle \epsilon_{xQ1}, \epsilon_{xQ2} \rangle, \langle pL_{Q1}, pL_{Q2} \rangle, \langle scL_1 \rangle \rangle$
$\epsilon_{Qe} = dF_Q(IdEL(crQ, e_Q)) = \langle$*evt*, ⟨call, past⟩$\rangle$
$\epsilon_{xQ1} = dF_Q(IdEL(crQ, xQ1))$
$\langle$*ent*, ⟨John, determinate, count, 1⟩$\rangle$
$\epsilon_{xQ2} = dF_Q(IdEL(crQ, xQ2))$
$\langle$*ent*, ⟨midnight, determinate, count, 1⟩$\rangle$
$pL_{Q1} = \langle$*pLink*, $\langle \epsilon_{Qe}, \epsilon_{xQ1}$, agent, individual$\rangle\rangle$
$pL_{Q2} = \langle$*pLink*, $\langle \epsilon_{Qe}, \epsilon_{Qxt}$, time, individual$\rangle\rangle$,
$scL_{Q1} = \langle$*scope*, $\langle pL_{Q1}, pL_{Q2}$,wider$\rangle\rangle$

ISO-TimeML

$A_T = dF_T(crT) = \langle \epsilon_{Te}, \langle \epsilon_t \rangle, \langle tL_1 \rangle \langle\rangle\rangle$
$\epsilon_{Te} = dF_T(IdEL(crT, e_t))$
$= \langle$*evt*, ⟨call, occurrence, transition, past⟩$\rangle$
$\epsilon_{Tt} = dF_T(IdEL(crT, tT1) = \langle$*instant,* ⟨T00:00⟩$\rangle$
$tL_{T1} = dF_T(TL(crT) = \langle$*tLink*, $\langle \epsilon_{Te}, \epsilon_{Tt}$, included$\rangle\rangle$

Interlinking:

$A_{IL} = dF_{IL}(crIL) =$
$= \langle dF_{IL}(\text{IdEl}(crIL), i1), dF_{IL}(\text{IdEl}(crIL, i2))\rangle$
$= \langle\langle \epsilon_{Qe}, \epsilon_{Te} \rangle, \langle \epsilon_{xQ2}, \epsilon_{Tt} \rangle\rangle$crIL

Interlinked annotation structure:

$A_{QT} = dF_{QT}(crQ + crT + crIL)$
$\langle \epsilon_e, \epsilon_x, pL, scL \rangle$, where
$\epsilon_e = dF_{IL}(\text{IdEl}(crIL), i1) = \langle \epsilon_{Qe}, \epsilon_{Te} \rangle$
$\epsilon_x = \langle dF_{QT}(\epsilon_{Qx1}), dF_{QT}(\epsilon_{xQ2}), dF_{QT}(\epsilon_{Tt}) \rangle$
$= \langle \epsilon_{Qx1}, \langle \epsilon_{xQ2}, \epsilon_{T1} \rangle \rangle$
$=_D \langle \epsilon_1, \epsilon_2 \rangle$
$pL = \{pL_1, pL_2\}$
$pL_1 = pL_{Q1}$
$pL_2 =$ PLink($\langle pL_{Q2}, tL_{T1} \rangle$) =
$= \langle \epsilon_e, \epsilon_2, \langle$SRole($pL_{Q2}$), SRole($tL_{T1}\rangle$,
Params(pL_{Q2})$\rangle$
$= \langle \epsilon_e, \epsilon_2, \langle$time, simultaneous$\rangle$,
⟨individual, narrow, non-exh, positive⟩$\rangle$
$scL = \{dF_{QT}(scL_{Q1})\}$

SEMANTICS:

$I_{QT}(A_{QT}) = I_{QT}(scL_1)$
$= I_{QT}(pL_1) \cup^* (I_{QT}(pL_2)$
$I_{QT}(pL_1) = I_Q(pL_{Q1}) = [X| \ |X| = 1, \ x \in X \leftrightarrow$ john$_0(x), x \in X \rightarrow$
[E | e ∈ E → [call(e), agent(e,t)]]
$I_{QT}(pL_2) = I_{QT}(\langle \epsilon_e, \epsilon_2,$
⟨time, simultaneous⟩, individual, narrow,
⟨inon-exh, positive)⟩=
$= I_{QT}(\langle\langle \epsilon_{Qe}, \epsilon_{Te} \rangle, \langle \epsilon_{xQ2}, \epsilon_{Tt1} \rangle, \langle$time,
simultaneous⟩, ⟨individual, narrow,
non-exh, positive)⟩⟩
$= \cup^+(I_Q(\epsilon_{Qe}) \cup I_T(\epsilon_{Te}), I_Q(\epsilon_{xQ2}) \cup I_T(\epsilon_{Tt1}),$
$\lambda x, y.$ time$(e, y) \wedge$ simultaneous(x, y)
$= \cup^+([E|e \in E \rightarrow$ call(e), past(e), occurrence(e),
transition(e)],
$[T||T| = 1, t \in T \leftrightarrow$ midnight$_0$, $t \in T \rightarrow$
clocktime(t) = 00:00],
$\lambda x, y.$ time$(e, y) \wedge$ simultaneous(x, y))
$= [T||T| = 1, t \in T \leftrightarrow$ midnight$_0$, $t \in T \rightarrow$
clocktime(t)= 00:00], $[E|e \in E \rightarrow$
call(e), past(e), occurrence(e), transition(e)),
time(e, t), simultaneours(e, t)]]

The interpretation of the integrated annotation is thus:
$I_{QT}(A_{QT}) = I_{QT}(pL_1) \cup^{**} (I_{QT}(pL_2)$
$= [X| \ |X| = 1, x \in X \leftrightarrow$ john$_0(x)$,
$x \in X \rightarrow [T||T| = 1, t \in T \leftrightarrow$ midnight$_0(x)$,
$t \in T \rightarrow$ clocktime(t)= 00:00),
$[E|e \in E \rightarrow$ call(e), past(e), occurrence(e),
transition(e)), agent(e,x), time(e, t),
simultaneous(e, t)]]

From Categories to Decisions: A Framework for Attitudinal Analysis of Evaluative Language

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, alex.fang@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-Comte, France ({francois_claude.rey, sylviane.cardey}@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

The study reported in this paper aims to contribute to the development of an annotation scheme for evaluative language, based on Appraisal Theory, that addresses key sources of classification problems. In particular, it aims to develop a unified annotation scheme that proposes (1) a three-component annotation model comprising Appraiser, Appraised and Appraisal Element, (2) the operationalised distinction between Affect and Appreciation governed by a criterion of experiencer salience and a criterion distinguishing personal emotions from evaluations of conduct, and (3) a decision framework for the Judgement-Appreciation distinction structured on the target and lexis types operating through override conditions and substitution tests. The revised framework is illustrated with examples selected from a corpus of news discourse in English and is designed to be replicable across future Appraisal-based studies of evaluative language.

Keywords: evaluative language, appraisal theory, affect, judgement, appreciation, corpus

1. Introduction

Appraisal Theory (Martin & White, 2005) provides one of the most widely adopted frameworks for analysing evaluative language in discourse, and its Attitude system, comprising Affect, Judgement, and Appreciation, has been applied extensively in corpus-based studies of media, political, and academic discourse (Fuoli, 2018; Read & Carroll, 2012). Despite this broad uptake, the annotation of Attitude in practice remains a source of persistent difficulty. In particular, three problems have proven resistant to resolutions within existing annotation methodologies. First, most annotation schemes mark up only the evaluative expression itself without formally specifying its source (the Appraiser) and target (the Appraised), leaving the relational structure of evaluation implicit and creating conditions for undetected disagreement (Read & Carroll, 2012; Fuoli, 2018). Secondly, while theoretically motivated by the concept of 'institutionalised Affect' (Martin, 2000; Martin & White, 2005), the boundary between Affect and the other two subsystems lacks operationally explicit criteria for deciding when an expression has crossed from felt emotions into Judgement or Appreciation. Thirdly, the boundary between Judgement and Appreciation, particularly for abstract human-derived entities such as policies, plans and reports, depends on the relationship between the evaluative lexis and its target in ways that existing accounts describe but do not fully proceduralise (Bednarek, 2009; Martin & White, 2005; White, n.d.).

The present study seeks to address these three problems by developing a unified annotation framework grounded in Appraisal Theory. In an earlier companion study (Zeng et al., 2025), we reported the results of an annotation experiment applying this framework, including inter-annotator agreement measures and quantitative evaluation. The present article addresses the annotation principles and decision procedures for distinguishing among the three major Attitude categories at the category level. In what follows, the framework is described and illustrated with

[*] Equal contribution

[**] Corresponding author

Proceedings of the 22nd Joint ACL - ISO Workshop on Interoperable Semantic Annotation and Representation (ISA-22), pages 134–143

examples drawn from a corpus of English-language news discourse on the COVID-19 pandemic. The aim is not to report corpus findings but to demonstrate the annotation procedures themselves, providing a practical and replicable framework for attitudinal analysis in the study of evaluative language.

2. The Annotation Framework

2.1 Why annotate Appraiser, Appraised, and Appraisal Element together

Most existing corpus annotation work in the Appraisal tradition centres on the identification and classification of evaluative expressions, what we term the Appraisal Element (AE), without formally indicating or marking up who is performing the evaluating or the target of the evaluation. This reduced approach has been the practical default in large-scale annotation studies (Fuoli, 2018; Read & Carroll, 2012). However, this reduction has caused theoretical and practical problems. Theoretically, evaluative meaning is inherently relational: the classification of an evaluative expression depends on the source (who is evaluating) and the target (what/who is being evaluated), not on the lexical item alone (Martin & White, 2005). Practically, omitting the source and target of evaluation makes it impossible to recover who is evaluating whom in multi-voice discourse (White, 2012).

The present study addresses these interconnected problems through a proposal to identify three relational components of appraisal A: AS (appraisal source), AT (appraisal target), and AE (appraisal element, i.e. the linguistic realisation), thus $A = \{AS, AT, AE\}$. The idea of identifying AS and AT alongside AE is not new in itself. It is already implicit in the theoretical literature (Martin & White, 2005). What has been lacking is a schematic requirement that all three components be considered and marked up where possible for every evaluative instance, together with a principled procedure for resolving the classificatory ambiguities that arise when multiple readings become available. In the rest of this article, each example of evaluation is annotated with triples in the form of ⟨AS⟩, ⟨AE⟩ and ⟨AT⟩. Unless otherwise indicated, all the examples in the article are selected from a corpus of COVID-19 news reports (Zeng et al, 2025). The source of each example is provided in round backets and coded in the format year-month-area-media.[1]

2.2 What each component captures

The Appraiser is the semantic source of the evaluative act (Martin & White, 2005, pp. 71-81). It can be explicitly realised, as when a named individual or entity is grammatically present, or implicitly realised, as when the evaluation is attributable to the authorial voice. In the Affect subsystem, AS corresponds to the experiencer of emotion. AT is the semantic target of evaluation. It can be a person, behaviour, thing, process, or proposition. In the Affect subsystem, AT functions as the stimulus triggering the emotional response. AE is the linguistic realisation of evaluation, articulated by AS about AT. Its classification carries information about the attitudinal categories including Affect, Judgement, and Appreciation, and their subcategories. Their relational structure can be represented as ⟨AS⟩-⟨AE⟩-⟨AT⟩, that is, AS evaluates AT as possessing a particular attitudinal quality linguistically encoded through AE.

Consider [1] and [2], which illustrate how the components work together and how the same formal structure applies across attitudinal categories.

[1] "I think it was just a <u>really bad</u> idea," Ms. Ratley said. (20-08-US-NT)

⟨AS: *Ms Ratley*⟩
⟨AE: Appreciation: Reaction: *really bad*⟩
⟨AT: *idea*⟩

AS is identified with *Ms. Ratley*, to whom the evaluative proposition is attributed through the reporting clause. The attribution frame simultaneously carries Engagement significance but the present study focuses exclusively on the Attitude layer. AT is *the idea*, a non-human entity, and AE *really bad* realises negative Appreciation in the Reaction subcategory. The evaluative triple can be represented as a tree structure in Figure 1, where the linguistic encodings of evaluation are anchored in terminal nodes and annotated by the non-terminal nodes for appraisal types and categories.

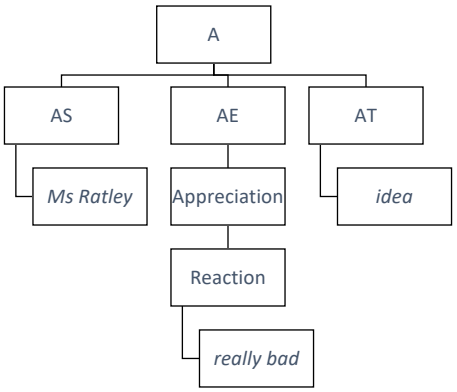

Figure 1: A tree representation of the evaluative structure of Example [1]

[1] The area codes include CN (China), HK (Hong Kong), SG (Singapore), UK and US. Media codes include CD (China Daily), NT (New York Times), SM (South China Morning Post), ST (Strait Times), and TG (The Guardian).

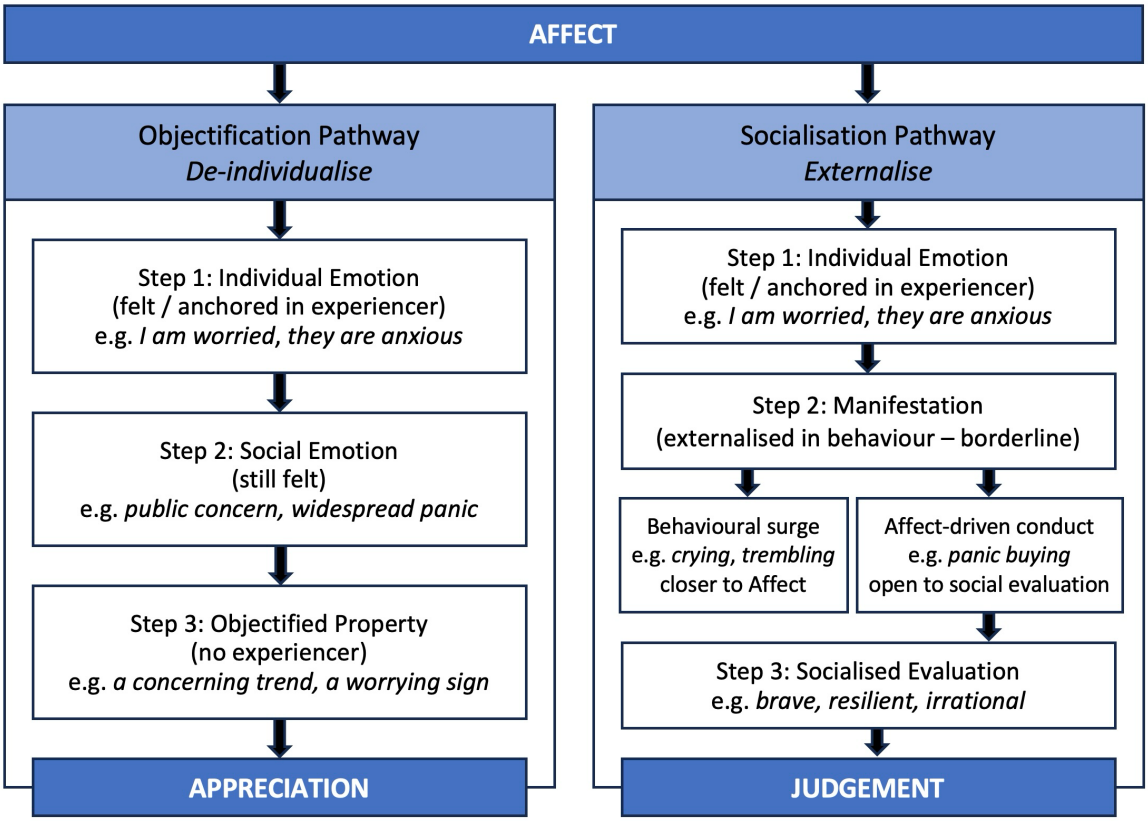

Figure 2: Pathways of Affect towards Appreciation and Judgement

Differently, AE in [2] below indicates an attitudinal evaluation in the Affect sub-system.

[2] "They fear side effects. They don't trust vaccines in principle, or they want to wait and see what happens to other people first," said Denis Volkov, deputy director of the Levada Center. (21-06-US-NT)

⟨AS: *they*⟩
⟨AE: Affect: In/security: *fear*⟩
⟨AT: *side effects*⟩

AS *they* is the experiencer, AT *side effects* is the trigger, the stimulus triggering the emotional response, and AE *fear* realises negative Affect in the In/security subcategory.

3. Distinguishing Affect from Judgement and Appreciation

3.1 Theoretical foundation: Two pathways from Affect

The three Attitude subsystems, Affect, Judgement, and Appreciation, share a common evaluative root rather than occupying wholly separate domains. As Martin (2000) argues, 'AFFECT can perhaps be taken as the basic system, which is then institutionalized in two major realms of uncommon sense discourse': Judgement represents Affect recontextualised as an evaluation matrix for behaviour while Appreciation represents Affect recontextualised as an evaluation matrix for the products of behaviour and the natural world. White (2004) clarifies the practical distinction: evaluative language can be divided into emotion, in which attitudinal assessments are indicated through descriptions of emotional reactions or states of human subjects, and opinion, under which a positive or negative quality is said to be an inherent property of the phenomenon being evaluated. Martin and White (2005) formalise this relationship as institutionalised Affect.

This insight explains why the boundaries between subsystems are not clear-cut: if Judgement and Appreciation are derived from Affect, expressions at the boundary will inevitably retain traces of both. However, 'institutionalised Affect' as a formulation remains too schematic to resolve the specific boundary decisions that arise in practical annotation. We therefore aim to refine this schema by describing the relationship between Affect and the other two subsystems as two directional pathways, each governed by a distinct mechanism. Both pathways begin from the same point of departure, in which feeling moves beyond the private experience of an individual. What distinguishes the two pathways is the direction of that movement. One pathway proceeds toward the objective world while the other toward the social world. Figure 2 illustrates the two pathways and their respective stages.

The first pathway from Affect to Appreciation can be described as individual emotion → social emotion → objectification. At the initial stage, Affect is anchored in a specific experiencer and the evaluative meaning remains at the level of private emotional feelings. In the next stage, emotions expand beyond individual experience into a collectively circulating condition: expressions such as public concern no longer refer to a single experiencer and yet their semantic core still lies in the existence and circulation of emotions within a social field. The emotion has become de-individualised but it has not yet been fully detached from human feelings. In the final stage, affective meaning undergoes objectification: what was originally felt by people is reconfigured as a quality of an entity, trend, situation, atmosphere, or event, as in a worrying sign. Discourse no longer foregrounds who is feeling worried but the object as possessing a worrying quality in itself. The emotional response has been encoded as an attribute of the evaluated entity.

The second pathway, from Affect to Judgement, can be described as individual emotion → individual manifestation → socialised evaluation. Unlike the movement toward Appreciation, which proceeds by objectification, the movement toward Judgement proceeds by socialisation. It begins with felt emotions, which may then become externalised in the form of individual manifestations. Such a manifestation can be divided into two subtypes. The first consists of pure manifestation, corresponding broadly to what is termed 'behavioural surge': crying, trembling, or freezing (Martin and White 2005, p. 47). These are bodily or behavioural manifestations of emotion that start to be publicly visible displays rather than remaining purely internal states. The second subtype consists in an affect-driven social conduct, that is, emotion-driven action that has an impact on others or on the social environment, such as expressions like *panic buying*. Unlike crying or trembling, such conduct is not only expressive but also socially impactive, triggering social or collective response, thus readily open to public evaluation. In the final stage, these manifestations are interpreted as signs of what is classified as social esteem or social sanction (Martin and White 2005, pp. 52-56), that is, as indicators of a person's normality, capacity, tenacity, veracity, or propriety, thereby entering the domain of Judgement. Judgement does not arise directly from emotion itself but from the social reading of emotional manifestations. What moves Affect toward Judgement is not the emotional feeling alone but the fact that feeling becomes publicly visible, thus interpretable and impactive according to social norms.

Following from here, we formulate two operational criteria in Sections 3.2 and 3.3 below.

3.2 Criterion 1: Felt Experience or Objectified Property (Affect vs. Appreciation)

The first annotation criterion follows directly from the objectification pathway, that is, an expression realises Affect when an identifiable experiencer, whether an individual or a collective human entity, is present in whose emotional state the evaluation is grounded. When the emotional quality has been fully detached from any experiencer and attributed to an entity, situation, or text as an inherent property, the expression realises Appreciation. This criterion operates across a gradient of experiencer salience, from individuals to collectivity to absence. The clearest cases are those in which the experiencer occupies subject position and AE directly predicates an emotive state. Consider

[3] Now, we are grieving, afraid and confused. (20-05-US-NT)

⟨AS: *we*⟩
⟨AE: Affect: Un/happiness: *grieving*⟩
⟨AT: unspecified⟩

⟨AS: *we*⟩
⟨AE: Affect: In/security: *afraid*⟩
⟨AT: unspecified⟩

⟨AS: *we*⟩
⟨AE: Affect: Dis/satisfaction: *confused*⟩
⟨AT: unspecified⟩

AS *we* is the explicit experiencer and *grieving, afraid and confused* each directly attribute an emotive state to it, yielding three co-occurring AEs belonging to the Affect category. No trigger is syntactically realised; this illustrates what Martin and White (2005, p. 47) call undirected mood, an emotional state reported without a specific stimulus. The absence of any identifiable AT is itself a diagnostic indicator: Judgement and Appreciation by definition require an entity being evaluated, and its absence strongly favours an Affect classification.

Between the implicit experiencer and full objectification lies a middle zone: social emotion, in which an emotional state is no longer anchored in a specific individual but still circulates as a felt condition within a collectivity.

[4] Criticism and worry continue to plague other colleges hoping to offer in-person learning this fall. (20-08-US-NT)

⟨AS: *other colleges*⟩
⟨AE: Affect: In/security: *worry*⟩
⟨AT: *in-person learning this fall*⟩

Worry is an emotion noun. While no individual experiencer is named, worry here is not presented as a property of the colleges or of the learning environment. It is presented as a social-emotional condition actively afflicting the colleges, as the verb plague makes clear. The experiencer is a

recoverable collective (university communities, administrators, faculty), making this a case of social emotion rather than fully objectified Appreciation. The absence of an explicit individual does not disqualify Affect; what matters is whether the emotional meaning remains grounded in someone experiencing it, even if that someone is a collectivity.

When the emotional quality has been fully detached from any experiencer and functions to characterise an entity or text, the expression realises Appreciation.

[5] This "nightmare scenario" may come about. (20-04-SG-ST)

⟨AS: implicit author⟩
⟨AE: Appreciation/Reaction: *nightmare*⟩
⟨AT: *scenario*⟩

AE *nightmare* is originally an Affect-laden term, but here it functions as a modifier attributing a quality to the scenario itself. This parallels the contrast that White (n.d.) draws between *the building bores me*, where the evaluative quality is anchored in a human experiencer, and *the building is boring*, where the emotion has been detached from any human experiencer and reattached to the entity as an intrinsic property. In [5], *nightmare* functions in precisely this second way: no specific or collective experiencer is construed, and the clause foregrounds the object (*scenario*) and its inherent affective charge rather than anyone's experience of it.

The contrast between Examples [4] and [5] is instructive. In [4], *worry* is something that the colleges have or suffer; in [5], *nightmare* is a quality that the scenario is said to possess. The former foregrounds an experiential state and the latter foregrounds an entity property. This distinction marks the boundary at which Affect gives way to Appreciation along the objectification pathway.

3.3 Criterion 2: Personal Emotion or Evaluation of Conduct (Affect VS. Judgement)

The second annotation criterion follows from the socialisation pathway: an expression realises Affect when it foregrounds a personal emotional state, whether a mood, disposition, or surge, and realises Judgement when the primary evaluative content concerns how someone behaved, assessed against the dimensions of social esteem or social sanction. The boundary between the two is particularly subtle when an expression can be read as either an individual's feeling or an assessment of conduct, a difficulty documented empirically in annotation research. Read and Carroll (2012) provide an instructive case, where two annotators independently coded [6]:

[6] Like him, Vermeer, or so he chose to believe, was an artist neglected and wronged by critics and who had died an almost unknown.

One annotator tagged *neglected* and *wronged* as Affect: Satisfaction, reading the passage as expressing Vermeer's dissatisfaction with his treatment while another annotator tagged them as Judgement: Propriety, reading it as moral reproach directed at the critics, both readings semantically available. Read and Carroll (2012) acknowledge that the ambiguity reflects an insufficiently specified annotation guideline rather than an error by either annotator.

In the present annotation scheme, the Vermeer case is classified as Judgement: Propriety. AEs *neglected* and *wronged* do not foreground Vermeer's felt emotional reaction; there is no explicit mood or surge, and he is not described as feeling distressed, angry, or sorrowful. Instead, the passage foregrounds the ethical status of the critics' conduct: being wronged implies a violation of norms of fair treatment that holds independently of any individual's emotional response. The following example from the present corpus provides a clear contrast:

[7] Some critics were frustrated that university officials opted to take the chance at all, given the state's ongoing struggle to contain the virus. (20-08-US-NT)

⟨AS: *some critics*⟩
⟨AE: Affect: Dis/satisfaction: *frustrated*⟩
⟨AT: *university officials opting to take the chance*⟩

AE *frustrated* directly names an emotional state experienced by AS *critics*. The evaluation is explicitly anchored in their subjective experience; the frustration is the evaluative content, not an inference about anyone's moral conduct. There is no assessment against the dimensions of social esteem or social sanction; the clause simply reports that a group of individuals experienced a negative emotional feeling.

The evaluative weight of *wronged* in [6] above lies in the ethical but not emotive domain. The contrast between Examples [6] and [7] yields the annotation guideline: when the clause foregrounds what or how someone feels as its primary evaluative content, annotate AE as Affect; when AE foregrounds how someone behaves, assessed against the dimensions of social esteem or social sanction, annotate AE as Judgement, even if the expression carries emotional resonance.

4. Decision Framework for Judgement versus Appreciation Classification

The Judgement-Appreciation boundary has attracted sustained attention in the Appraisal literature, but existing proposals have not converged on a resolution. Bednarek (2009) observes that grammatical patterns alone cannot reliably distinguish Judgement from Appreciation,

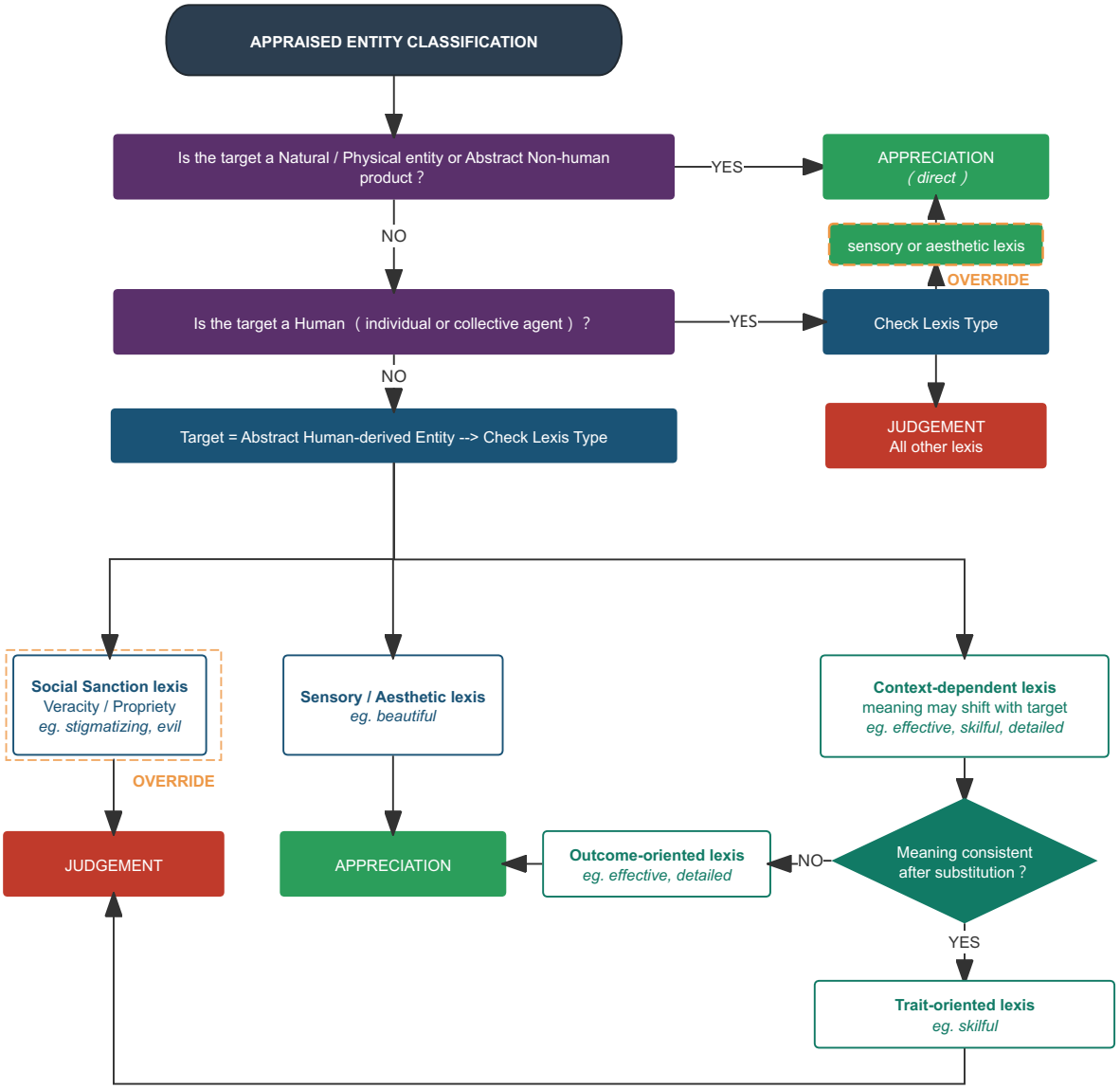

Figure 3: The complete decision framework for Judgement-Appreciation classification

proposing two key classification criteria, lexical type and target entity, that jointly determine category assignment without establishing clear precedence between them. Thompson (2014) prioritises target characteristics, suggesting that all evaluations of non-human targets should be classified as Appreciation, even when employing judging lexis, including nominalisations as non-human entities. However, Fuoli (2018) points out practical limitations: expressions like *industry-leading* and *disciplined* resist classification under traditional Appreciation subcategories while clearly evaluating human-related qualities. Starfield et al. (2015) address this issue by relocating Judgement subcategories into Appreciation: Valuation, accommodating judging lexis applied to non-human targets by expanding the Appreciation taxonomy. While this approach has resolved some difficult classification issues, it has unnecessarily over-broadened the Valuation category beyond its original conceptual boundaries. Taboada & Carretero (2012) propose a different strategy, prioritising lexical properties through an ethics-aesthetics distinction: ethics-based evaluations belong to Judgement while aesthetics-based evaluations belong to Appreciation, regardless of target type, with abstract nouns classified as Judgement when representing nominalisations of human actions. In the present study, the framework described below adopts target type as the primary determinant of classification while refining the target taxonomy into three operationally distinct categories. Lexis type is retained as a secondary mechanism but its role is restricted to two well-defined override conditions and a substitution test, rather than operating as an independent co-equal

criterion. The decision framework for Judgement-Appreciation classification is visually represented in Figure 3.

4.1 Foundational principles

The framework rests on two interacting dimensions: the nature of the AT and the semantic core of lexis in AE. These two dimensions do not carry equal weight. AT type is the primary determinant of classification while AE lexis operates as a secondary mechanism that intervenes selectively through override conditions or through a substitution test. The central theoretical question underlying this priority order is the extent to which a human agent is implicated in the evaluation: Judgement evaluates human behaviour, character, and social conduct while Appreciation evaluates the properties of objects, phenomena, and artefacts.

AE lexis enters the framework as an override mechanism under exactly two conditions. First, AEs whose semantic core is inherently oriented towards social sanction dimensions, specifically Veracity or Propriety, trigger Judgement classification regardless of whether AT is human or nonhuman: the evaluative meaning presupposes human volitional agency and cannot be coherently detached from it. Second, AEs whose semantic core is inherently sensory or aesthetic in nature trigger Appreciation classification regardless of AT types since their evaluative meaning pertains to perceptual qualities rather than normative assessments of conduct or character. Outside these two override conditions, AT type governs classification throughout.

AEs whose semantic core does not pre-determine classification are designated in the present framework as context-dependent, in the sense that their evaluative meaning shifts according to the nature of AT. A substitution test is required when AEs are directed at abstract human-derived entities: The appraisal is transferred from the human-derived entity to its recoverable human agent, and the evaluative meaning of the resulting expression is examined for consistency. If the meaning remains essentially the same, that is, the evaluation retains the same qualitative dimension after substitution, AE is annotated as Judgement and retrospectively identified as trait-oriented AE, i.e. elements whose semantic core describes inherently human qualities that transfer coherently from non-human entities to human agents. If the meaning changes, that is, the evaluation shifts to a qualitatively different dimension after substitution, AE is classified as Appreciation and identified as outcome-oriented, i.e. elements whose semantic core describes functional effectiveness or compositional properties that cannot transfer to a human agent without altering the evaluative meaning. The substitution test therefore operationalises a distinction that is latent within the context-dependent AE category itself.

4.2 Classification Procedures

Category 1: Natural and physical entities, and abstract non-human products → Appreciation

Where AT refers to a natural or physical object, or an abstract entity not traceable to deliberate human action or decision, AE is classified as Appreciation without any further test. Human agency is either entirely absent or insufficiently implicated to anchor an ethical or social evaluation.

[8] Two hotels in the exclusive Swiss Alpine resort of St Moritz were quarantined and ski schools closed to try to curb an outbreak of the highly infectious new coronavirus variant. (21-01-UK-TG)

⟨AS: implicit author⟩
⟨AE: Appreciation: *infectious*⟩
⟨AT: *coronavirus variant*⟩

The Delta variant is a natural entity. No human agency is implicated in its properties. AE *infectious* describes a biological characteristic of the variant itself, and the element is annotated as Appreciation.

[9] Five days later, the outbreak was out of control, with nearly 400 virus cases among a campus student population that is usually around 6,000. (20-09-US-NT)

⟨AS: implicit author⟩
⟨AE: Appreciation: *out of control*⟩
⟨AT: *outbreak*⟩

AT *outbreak* is an abstract non-human entity, a developing epidemiological situation not directly traceable to a single deliberate human decision. AE *out of control* characterises the state of the outbreak itself and receives straightforward Appreciation classification.

Category 2: Human and collective human targets → Judgement

Where AT is human or functional as human, such as organisations, institutions, and government bodies, AE receives Judgement classification as the default outcome. Human agency is directly and fully implicated, satisfying the foundational condition for Judgement.

[10] The island has struggled to secure vaccine supplies but the United States delivered 2.5 million Moderna doses last weekend. (21-06-HK-SM)

⟨AS: implicit author⟩
⟨AE: Judgement: Capacity: *struggled*⟩
⟨AT: *island*⟩

In [10], AT functions as a collective human agent responsible for its own public health governance. AE is therefore classified as Judgement: Capacity, reflecting a negative assessment of the collective

agent's ability to fulfil its responsibilities. AS is implicit, assumed to be the writer-speaker.

The sole exception to the human-target default arises when AE carries an inherently sensory or aesthetic semantic core. In such cases, the override condition takes precedence over the AT-primary rule and AE receives Appreciation classification regardless of the human AT. For instance, an evaluation such as *she is beautiful* describes a perceptual quality of the human AT's appearance whose semantic core cannot be coherently transferred to an evaluation of behaviour, competence, or social conduct; the Judgement reading is overridden and Appreciation assigned to AE. The human AT does not trigger the default Judgement outcome.

Category 3: Abstract human-derived entities

The operationally most complex category involves entities produced through or traceable to deliberate human action or decision such as reports, figures, plans, accusations, measures, and similar products of human agency. Classification in this category is determined by AE type, proceeding through three distinct steps described below.

Step 1: Social sanction AE override → Judgement

AE receives Judgement analysis if its semantic core is inherently oriented toward Veracity or Propriety. This override applies because the evaluative meaning presupposes a human agent whose conduct is being assessed against normative standards, regardless of whether AT is human. Consider the following example:

[11] The WHO chief stressed that having a name matters to prevent the use of other names that can be inaccurate or stigmatizing. (20-02-CN-CD)

⟨AS: *WHO chief*⟩
⟨AE: Judgement: Veracity: *inaccurate*⟩
⟨AT: *other names*⟩

⟨AS: *WHO chief*⟩
⟨AE: Judgement: Propriety: *stigmatizing*⟩
⟨AT: *other names*⟩

In [11], AT is *other names*, a human-derived entity. However, AE *inaccurate* evaluates the truthfulness of the naming practice: to call a name inaccurate is to assert that whoever produced it failed to represent reality truthfully, an evaluation that presupposes a human agent capable of truth or falsehood. AE *stigmatizing* evaluates the ethical conduct behind the naming: to stigmatise is to cause harm through social labelling, an evaluation that presupposes volitional human action subject to ethical norms. Neither evaluation can be coherently detached from human agency. The social sanction override applies to both items, which are classified as Judgement without proceeding to the substitution test.

Step 2: Context-dependent AE → Substitution test

The substitution test is applied in cases of AEs not covered by the social sanction or sensory-aesthetic overrides, where AT is transferred from the human-derived entity to its recoverable human agent, and the evaluative meaning of the resulting expression is examined for consistency.

The logic of the test can be illustrated through a pair of contrasting cases before turning to corpus examples. Consider

[12] It was a skilful innings.

The example is taken from Martin and White (2005, p. 59), where AE *skilful* is annotated as Appreciation. However, when the substitution test developed in the present framework is applied, substituting *a skilful innings* with *he is skilful* yields an evaluation with essentially the same meaning: both formulations assess the technical mastery. The consistency of evaluative meaning indicates that *skilful* is trait-oriented and therefore classified differently as Judgement in the present framework. Now consider AE *important*, which is noted in Bednarek (2009) as Appreciation when modifying a thing, issue, or proposition but as Judgement when modifying a person for the social standing and influence. The shift in evaluative meaning identifies *important* as context-dependent whose classification depends on the nature of the target.

These two contrasting cases illustrate the distinction that the substitution test is designed to operationalise. If the evaluative meaning remains essentially the same after substitution, the item is trait-oriented and receives Judgement classification, thus

Meaning consistent → Judgement (trait-oriented AE)

Consider [13], an example from the corpus of news discourse.

[13] One, by the name Kou, said: "With such a perfunctory state of emergency, holding the Olympics will be impossible." (21-01-SG-ST)

⟨AS: *Kou*⟩
⟨AE: Judgement: Tenacity: *perfunctory*⟩
⟨AT: *state of emergency*⟩

AT *state of emergency* is a human-derived entity traceable to a governmental decision. Substituting *a perfunctory state of emergency* with *the government was perfunctory* yields an evaluation of essentially the same meaning: both formulations assess diligence and commitment. AE *perfunctory* describes a disposition, a lack of seriousness or effort that transfers coherently from the product to its human agent. The evaluative meaning is consistent, and so AE is classified as Judgement.

If the evaluative meaning changes, the item is outcome-oriented and receives Appreciation classification, thus

Meaning inconsistent → Appreciation (outcome-oriented AE)

Now consider [14] as a final example in this study.

[14] A detailed study in California found that the variant easily spread from an unvaccinated teacher to children and, in a few cases, their families. (21-08-US-NT)

⟨AS: implicit author⟩
⟨AE: Appreciation: Composition: *detailed*⟩
⟨AT: *study*⟩

AT *study* is traceable to human action. However, substituting *a detailed study* with *the researchers are detailed* changes the evaluative meaning: the evaluation shifts from assessing the study's structural properties to a characterological assessment of the researchers' disposition. The change in meaning revealed through the substitution test classifies *detailed* as Appreciation.

5. Conclusion

This study has developed a methodological framework for the attitudinal annotation of evaluative language. It is grounded in Appraisal Theory and aims to address three persistent problems in the classification of evaluative language.

The first contribution is the three-component annotation model incorporating AS, AT, and AE for every evaluative instance. The model restores the relational structure of evaluation that is theoretically central to Appraisal Theory but has been routinely omitted in annotation practice. The selection principle developed alongside this model provides a principled procedure for resolving cases in which a single expression participates in more than one evaluative reading, replacing the ad hoc treatment of what Thompson (2014) termed the 'Russian doll' syndrome with a transparent decision mechanism grounded in the realisation and ontological status of AT.

The second contribution is the two-pathway model of the Affect boundary. By considering the relationship between Affect and the other two subsystems as two directional processes, where objectification tends toward Appreciation and socialisation toward Judgement, the model transforms the schematic notion of 'institutionalised Affect' discussed in Martin (2000) into two operational criteria. The first criterion distinguishes Affect from Appreciation by locating expressions along a gradient of experiencer salience, from individual felt emotion through social emotion to fully objectified property. The second criterion distinguishes Affect from Judgement by considering whether the expression foregrounds a personal emotive state or an evaluation of conduct against the dimensions of social esteem or social sanction.

The third contribution is the decision framework for the Judgement-Appreciation boundary. The framework establishes the target type as the primary determinant of classification and defines two lexis-based override conditions (Social Sanction lexis and sensory or aesthetic lexis) that operate independently of the target type. For context-dependent lexis directed at abstract human-derived entities, the substitution test provides a falsifiable diagnostic: consistency of evaluative meaning after substitution indicates trait-oriented lexis and Judgement classification, while a shift in meaning indicates outcome-oriented lexis and Appreciation classification.

Several limitations should be acknowledged. The methodology has been developed and illustrated using English-language media discourse, and its applicability to other languages, registers, and genres remains to be tested. The selection principle addresses the most common configurations of multiple evaluative readings but does not claim to exhaust all possible cases; further corpus work may reveal configurations that require refinement of the priority ordering and related procedures. The substitution test, while designed to be replicable, involves an element of analytical judgement in determining whether the evaluative meaning has changed, and further empirical validation across corpora and annotation settings would be a valuable next step.

Despite these limitations, the methodology that we outline in this study offers a coherent and operationally explicit framework for Attitude annotation. By integrating the three-component model, the two-pathway criteria, and the target-lexis decision framework into a single analytical workflow, it aims to provide annotators with a sequence of principled decisions rather than a set of isolated category definitions. The objective has been to make the reasoning behind classification decisions transparent, replicable, and open to empirical testing so that disagreements in attitudinal analysis can be traced to specific points in the decision process rather than attributed to the inherent fuzziness of the categories themselves.

6. Acknowledgement

Research described in this article was partially supported by grants received from City University of Hong Kong (Project No 9360115), China National Social Science Fund (Project No 24&ZD28), and Beijing Social Sciences Foundation (Project Nos 18JDYYA005 and 19YYA001).

7. Bibliographical References

Bednarek, M. (2009). Language patterns and

ATTITUDE. *Functions of Language*, 16(2):165–192.

Fuoli, M. (2018). A stepwise method for annotating APPRAISAL. *Functions of Language*, 25(2):229–258.

Martin, J.R. (2000). Beyond exchange: APPRAISAL systems in English. In S. Hunston & G. Thompson (Eds.), *Evaluation in text: Authorial stance and the construction of discourse*. Oxford University Press, pp. 142–175.

Martin, J.R. & White, P.R.R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Read, J. & Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46(3):421–447.

Starfield, S., Paltridge, B., McMurtrie, R., Holbrook, A., Bourke, S., Fairbairn, H., Kiley, M., & Lovat, T. (2015). Understanding the language of evaluation in examiners' reports on doctoral theses. *Linguistics and Education*, 31:130–144.

Taboada, M. & Carretero, M. (2012). Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish. *Linguistics and the Human Sciences*, 6(1–3):275–295.

Thompson, G. (2014). AFFECT and emotion, target-value mismatches, and Russian dolls: Refining the APPRAISAL model. In G. Thompson & L. Alba-Juez (Eds.), *Evaluation in context*. John Benjamins, pp. 47–66.

White, P.R.R. (n.d.). An introductory tour through appraisal theory. Retrieved from http://www.grammatics.com/appraisal/

White, P.R.R. (2004). Subjectivity, evaluation and point of view in media discourse. In C. Coffin, A. Hewings, & K. O'Halloran (Eds.), *Applying English grammar: Corpus and functional approaches*. Hodder Arnold, pp. 229–246.

White, P.R.R. (2012). Exploring the axiological workings of 'reporter voice' news stories: Attribution and attitudinal positioning. *Discourse, Context & Media*, 1(2–3):57–67.

Zeng, J., Wang, H., Bunt, H., Cao, X., Cardey, S., Dong, M., Hao, T., Jia, Y., Lee, K., Liao, S., Pustejovsky, J., Rey, F.C., Romary, L., Zong, J., & Fang, A.C. (2025). Enhanced evaluative language annotation through refined theoretical framework and workflow. In H. Bunt (Ed.), *Proceedings of the 21st Joint ACL–ISO Workshop on Interoperable Semantic Annotation (ISA-21)*, pages 76–84. Association for Computational Linguistics.

Author Index