Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09706-8

PROJECT NOTES

Multi-layered semantic annotation and the formalisation of annotation schemas for the investigation of modality in a Latin corpus

Helena Bermúdez-Sabel[1] · **Francesca Dell'Oro**[1] · **Paola Marongiu**[1]

Accepted: 14 November 2023

Abstract

This paper stems from the project *A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language* (WoPoss) which involves a corpus-based approach to the study of modality in the history of the Latin language. Linguistic annotation and, in particular, the semantic annotation of modality is a keystone of the project. Besides the difficulties intrinsic to any annotation task dealing with semantics, our annotation scheme involves multiple layers of annotation that are interconnected, adding complexity to the task. Considering the intricacies of our fine-grained semantic annotation, we needed to develop well-documented schemas in order to control the consistency of the annotation, but also to enable an efficient reuse of our annotated corpus. This paper presents the different elements involved in the annotation task, and how the description and the relations between the different linguistic components were formalised and documented, combining schema languages with XML documentation.

Keywords Semantic annotation · Feature structures · Schema languages · Modality · Latin

✉ Helena Bermúdez-Sabel
helena.bermudez@jinntec.de

Francesca Dell'Oro
delloro.fr@gmail.com

Paola Marongiu
paola.marongiu@unine.ch

[1] Institut des sciences du langage, University of Neuchâtel, Neuchâtel, Switzerland

Published online: 06 January 2024

1 Introduction

The development of semantically annotated corpora is a time consuming endeavour and yet of vital importance in the development of many Natural Language Processing (NLP) tasks. In particular, modality—here broadly defined as the qualification of a state of affairs as non-factual based on the use of the notions of necessity, possibility and volition—, is a challenging domain for both theory and annotation, and with multiple applications. For instance, it can be exploited in sentiment analysis and opinion mining, or event extraction, among others. Beside its application as a training dataset for a number of NLP tools, an annotated corpus can also be used to explore and test linguistic theories. However, this reusability is conditioned by the implementation of well-documented annotation schemas. In the framework of the project *A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language* (WoPoss), we are setting up a corpus annotated with several layers of linguistic annotation and with a focus on modality. This paper presents the WoPoss annotation scheme focusing on how its formalisation is meant to facilitate the reuse of both the dataset and the annotation scheme itself. In Sect. 2, we present the state of the art concerning linguistically annotated resources in Latin, in particular those dealing with semantics, along with other projects aiming to annotate modality in other languages. We carry out a brief presentation of the informational requirements on which our complex annotation scheme is based in Sect. 3. Section 4 describes the annotation task (Sect. 4.1) with a detailed annotation example (Sect. 4.2). The formalisation of the annotation scheme is presented in Sect. 5 and further discussed in Sect. 7. The exploitation functionalities of our dataset will be illustrated through an example in Sect. 6. Finally, the major points are summed up in the conclusions (Sect. 8).

2 State of the art

In addition to the WoPoss corpus, the realm of digital resources for Latin benefits from several corpora and datasets that offer various forms of semantic and/or pragmatic annotation.

One noteworthy resource is the Index Thomisticus Treebank (IT-TB) (Passarotti, 2019), which comprises texts by Thomas Aquinas encompassing a total of 350,000 tokens. A portion of the IT-TB (28,000 tokens) also has a tectogrammatical layer, annotated according to the Prague Dependency Treebank (PDT) style. This layer is built upon the morphological layer, containing information about lemmas and morphology, as well as the analytical layer, encoding syntactic information through dependency trees. The tectogrammatical layer presents additional semantic and pragmatic information. In this layer, the underlying structure of the sentence is represented by tectogrammatical trees, where the nodes

correspond to content words exclusively. Each node is assigned a semantic role label called 'functor,' which specifies the semantic role of a mandatory argument (e.g., 'Actor' or 'Patient') or the type of adverbial (e.g., 'Place' or 'Manner'). Moreover, information about the illocutionary force of the sentence is encoded by assigning relevant attributes (e.g., 'enunciative') to the main verb. Ellipsis resolution is provided through the addition of new empty nodes (e.g., for omitted subjects in pro-drop constructions), while coreference analysis is supported through the use of arrows connecting the nodes involved in coreference relations. Lastly, the horizontal arrangement of the nodes reflects the topic-focus order of the sentence, thereby conveying the sentence's information structure. For the annotation of modality, cf. below.

Another notable resource is the PROIEL treebank (Haug & Jøhndal, 2008) for Latin, which encompasses various texts such as the New Testament and additional works that have been included in subsequent releases of the treebank, e.g., Caesar's *De bello Gallico*, Cicero's *Epistulae ad Atticus* and Palladius' *Opus Agriculturae*. The Latin section of the PROIEL treebank counts approximately 225,000 tokens. All texts in the treebank are annotated with morphosyntactic information. Moreover, certain texts have an additional layer of annotation that captures semantic and pragmatic information, specifically animacy and information structure. The latter is aimed at recording the givenness status of the referents in the sentence i.e., if and how they relate to other referents previously introduced in the text (Celano, 2019, p. 293). There are specific tags e.g., for anaphora (givenness tag 'OLD', for referencing previously introduced referents), for newly introduced content without contextual reference (givenness tag 'NEW'), for referents accessible through inference based on encyclopaedic knowledge (givenness tag 'ACC-GEN' that is, accessible through general knowledge), and others (Haug et al., 2014, pp. 28–37). Both the PROIEL treebank and the IT-TB were converted into the Universal Dependencies annotation style. However, it is important to note that the semantic annotation can only be accessed in their original releases.

In a recent work by McGillivray et al. (2022), a different type of semantic annotation focusing on lexical semantics and word senses was introduced. A team of annotators manually annotated forty lemmas. These were selected based on their known semantic changes due to, for instance, political and social transformations (e.g., *dux* transforming from 'leader' to 'duke') (Clackson, 2011; Clackson & Horrocks, 2007), including the advent of Christianity (e.g., *beatus* changing from 'happy, fortunate' to 'blessed'). The sense inventory for each target word was built based on information retrieved from the Latin dictionaries Du Cange et al. (1883–1887 [1678]), Lewis (1890) and Lewis and Short (1879). Each annotator was assigned sixty passages extracted from the LatinISE corpus (McGillivray & Kilgarriff, 2013) for each target word. The annotation process followed a graded approach based on the DuRel annotation framework (Schlechtweg et al., 2020). Annotators could assign a value from 1 to 4 to each sense in the sense inventory, to indicate how closely the meaning of the target word in a specific

passage expresses each sense provided in the inventory (with 1 corresponding to 'Unrelated' and 4 to 'Identical').[1]

The textual annotation of modality types (such as 'deontic', 'epistemic' and so on, cf. below) is at least as old as the study of modality based on corpora (cf., e.g., Coates, 1983). In the last two decades, the growing interest for modality in NLP along with the development of annotation tools has enabled the development of complex annotation schemas. Without any pretence of exhaustiveness, we mention here some projects focusing on the textual annotation of modality and some projects aiming to set up databases of modal meanings.

Rubinstein et al. (2013), after providing a useful overview of previous projects (Rubinstein et al., 2013, p. 39), outline their own annotation schema. It features 'lemma', 'modality type' (based on the formal framework—cf., in particular, Portner, 2009—and divided into a coarse and a fine subdivision), 'environmental polarity' (taking into account not only negation, but also items creating a semantically negative environment such as *to reject*), 'propositional arguments' (this is the "textual span corresponding to the proposition a modal applies to" (Rubinstein et al., 2013, p. 42), corresponding to the scope of the modal marker in the WoPoss annotation (cf. below). The schema also includes, when they are pertinent, 'source' (e.g., the entity placing the obligation in the case of deontic modality), 'background' (textual elements providing information such as the circumstances on which the modal claim is based), 'modified element' (in the case of modal items used to modify another item, as in *the probable answer*), 'degree indicator' (an item indicating the degree of modal necessity or possibility, as in *very high likelihood*), 'outscoping quantifier' (quantifying elements taking scope over the modal marker) and 'additional notes'.

Inputs for the development of taggers for the automatic annotation of modality are provided by Baker et al. (2014), based on the setting up of a dedicated annotation scheme and lexicon. Baker et al. (2014) develop previous results on the annotation of the factuality of events by Saurí et al. (2006). Based on their definition of modality in terms of pure non-factuality, their annotation includes items such as effort (as expressed, e.g., by the verb *to try*) and success (as expressed, e.g., by the verb *to reach*). It is worth specifying that in many other modality frameworks there is a clear focus on the notions of necessity and possibility alongside non-factuality (cf. Narrog, 2012; Portner, 2009; van der Auwera & Plungian, 1998 among others). Another important point of divergence from the mainstream treatment of modality is the fact that Baker et al. (2014) do not distinguish between lexical (cf. *Tents are needed*, p. 1405) and properly modal meaning. In English, this is the case when the verb is used as an auxiliary (cf. *He need not go*, p. 1405) that modalises the propositional content. The original annotation scheme is based on the identification of a trigger (the modal marker), a target (the scope in the WoPoss project or the focus in other projects) and a holder (which partially overlaps with the WoPoss notion of participant, cf. below). With respect to other projects aiming to annotate modality, it is peculiar that the modality type is annotated on the target and not on the marker or

[1] It is worth specifying that the WoPoss team has contributed to this project by annotating some words and then discussing the annotation of modal markers. The annotation carried out in the framework of McGillivray et al. (2022) does not follow the WoPoss annotation schema.

on the whole passage. For example, in the sentence *He might be able to go to NY*, *go* is annotated as ability (p. 1404).

With respect to the multi-layered annotation of modality set up in the framework of the WoPoss project, the distinction between the layers 'marker', 'scope' and 'relation' derives from the project *MODAL—Modèles de l'annotation de la Modalité à l'Oral* (Ghia et al., 2016). Their schema allows annotators to describe marker and scope independently from the modal meaning, which is described separately in the layer 'relation' Laboratoire Ligérien de Linguistique, 2017.

The *Proceedings of the Workshop on Models for Modality Annotation* (Nissim & Pietrandrea, 2015) gathers some other annotation experiences, including annotation of modality in Portuguese (Ávila et al., 2015, with further bibliography) and multi-level annotation of modality with the tool GraphAnno (Gast et al., 2015).

With reference to Latin, the only—to our knowledge—attempt to annotate modality before the WoPoss project was carried out in the framework of the IT-TB project. Based on the Prague Dependency Treebank guidelines for the annotation of modality in Czech, the tectogrammatical layer (covering semantics and pragmatics) of the IT-TB and of a portion of the Latin Dependency Treebank[2] provides a basic annotation of Latin modal verbs in terms of modal types, including necessity, possibility, obligation, permission, volition/intention and ability.

Alongside projects focusing on the textual annotation of modality, it is worth mentioning here also some other projects aimed at setting up databases of modal markers. Projects aimed at describing a list of modal markers by associating modal markers to modal values—without textual contextualisation—are EUROEVIDMOD[3] and *A Database for Modal Typology*.[4] The first project had the purpose of annotating modal and evidential[5] markers and some of their semantic features in some languages of Europe. Data and results are not publicly accessible. Following the formal approach to modality (Kratzer, 1981; Matthewson, 2016; Portner, 2009), the second project, which is still under development, has the aim of gathering the analysis of modal markers in different world languages in terms of their flavour (epistemic, deontic) and force (weak, strong).

With respect to the features annotated in the WoPoss project and beyond the specific nomenclature used to identify types and sub-types of modality, one of the main differences consists in the fact that WoPoss does not annotate sources—the 'source of the modality' and the 'source of the event mention' (cf. Ávila et al., 2015, p. 3). However, WoPoss annotates features that are not taken into consideration in other annotation projects, such as the description of the modalised state of affairs (target, scope or focus) in terms of control and dynamicity and the animacy of the main participant in the modalised state of affairs. Another important difference is that, while in the other annotation projects the focus is on how annotators annotate modality, in the case of WoPoss the focus is on publishing a corpus annotated in a consistent way.

[2] Cf. https://itreebank.marginalia.it/view/download.php.

[3] See https://www.ucm.es/euroevidmod/ (accessed 20 June 2023).

[4] See https://clmbr.shane.st/modal-typology/ (accessed 20 June 2023).

[5] The relationship between modality and evidentiality is debated. A proper discussion of this point is beyond the scope of this paper.

3 Informational requirements

The main research question of the WoPoss project addresses the issue of how the semantics of Latin modal markers evolved over a very long period of time, such as the millennium between the third century BCE and the seventh century CE. The devised methodology is based on the setting up of a corpus of Latin texts and the annotation of the modal passages based on the presence of a predefined list of potentially modal markers, such as the verb *debeo* 'owe (not modal)/must (modal)', the adverb *certe* 'certainly' or the adjective *necessarius* 'necessary'. The complete list is available in Dell'Oro (2023, pp. 8–9). Texts are selected based on two main criteria: diachronic representativity and sociolinguistic representativity (in terms of diatopic, diastratic and diaphasic variation), inasmuch as this goal is compatible with the extant Latin texts from Antiquity. Therefore, the corpus includes both literary and documentary texts, belonging to diverse textual genres, transmitted through different types of support (codices, inscriptions, and papyri). In addition, we also include authors from different Latin-speaking regions of the ancient world. The corpus planned size is about 500,000 tokens.

The concept of 'modality' is notoriously a complex one and its borders are not clearly defined (cf. e.g., Nuyts, 2005). Moreover, some modal markers are polyfunctional and can be used to express several (sub)types of modality. This can imply ambiguity, when a modal marker can be associated with more than one type of modality in the same passage. As ambiguity can be important diachronically and be one of the triggers of semantic change, it is mandatory for us to take into account double or triple possible readings of the same passage. However, this is only one of the aspects which can be relevant in the annotation of modality. Based on the construction-centred approach of the project *MODAL*, we extended their annotation of one type of modality—epistemic modality—to the other types and sub-types (cf. also Sect. 2). Moreover, we added other information, e.g., about the state of affairs and the type of participant in it, making the annotation of the three elements of a modal construction (i.e. marker, scope and relation) more fine-grained. Based on the state-of-the-art outlined in Sect. 2, it is possible to state that the WoPoss corpus stands out as the first diachronic Latin corpus providing a fine-grained annotation of modality along with other relevant features. It is worth stressing that the focus of the project was not on annotating modality and the analysis of the annotation, but on annotating a corpus to provide the community with a consistently annotated resource.

At the moment only some samples of the annotated corpus are searchable through a dedicated interface.[6] The corresponding dataset is freely available (see WoPoss Project, 2022b). We plan to make other samples available little by little in the next future.

[6] See https://woposs.unine.ch/search (accessed 20 May 2023).

4 The annotation task

4.1 Overview

The annotation task is based on the annotation guidelines specifically devised in the framework of the project in order to annotate the texts of the corpus. The WoPoss guidelines for the annotation of modality (Dell'Oro, 2023) develop the above-mentioned annotation model of the project *MODAL* (cf. Sect. 2 and Sect. 3) by adding additional features and criteria. Some additional features are the same as those annotated in the framework of the modality projects led by Jan Nuyts (Nuyts, 2019). Some other features and criteria were introduced to take into account specific markers, such as adjectival modal suffixes, which are very important in the Latin modal system (cf. the annotation example under Sect. 4.2).

Texts are automatically annotated with regard to tokenisation, sentence division, lemmatisation, PoS-tagging, morphological analysis and syntactic dependencies. This automatic annotation was carried out by implementing the NLP library for Python Stanza, developed by the StanfordNLP research group (Qi et al., 2020), with the model trained with the Perseus Treebank (Universal Dependencies, 2021). This model was selected after carrying out a number of tests with the other available models for Latin.

The CONLL-U files derived from the automatic annotation are imported in the annotation platform INCEpTION (Klie et al., 2018). We have created custom annotation layers with the pertinent features and tagsets in order to carry out the semantic annotation. The different layers with their features and controlled vocabularies, formalised in JSON, are available in WoPoss Project (2022a). The annotators always read the whole text uploaded to INCEpTION and look for the modal markers listed in the guidelines. It is worth mentioning that the annotators check the imported text with a reference edition in order to ensure the philological exactitude and quality of the text. The presence of a modal marker allows them to identify a passage as potentially modal. Then they need to decide whether the passage is modal or not according to the theoretical framework of the project. If the passage is considered non-modal, the annotators check the option 'not pertinent' and introduce some specifications as needed (e.g., whether the meaning is premodal, that is, a meaning from which the modal sense evolved). If the passage is modal, the annotation continues. With regard to the modal passages, the annotators also check and correct the following automatically annotated layers: lemmata, parts of speech and the morpho-syntactic analysis.

After this preliminary work, annotators characterise the modal passage with regard to semantics. They describe the modal marker, its scope and the resulting modal relation according to modal types, subtypes and degrees, if relevant (see Table 1). The marker and its scope are annotated with a span layer (named 'modal unit'), which enables the creation of annotations over spans of text. The modal relation is defined through a relation layer which allows for drawing arcs between span annotations (in our case, between the marker and its scope).

Table 1 Features annotated to describe the semantics of a modal relation

Modality	Type	Degree/subtype	Other features
Dynamic	Possibility	Participant-inherent	
		Participant-imposed	
	Necessity	Situational	(If applicable) Inevitability Prospective
Deontic	Acceptability	Absolutely necessary	
		Desirable	
		Acceptable	
		Undesirable	
		Unacceptable	
	Authority	Obligation	Type of source (moral/ethical norms, religious norms, unspecified norms) Type of context (official or non-official)
		Permission	
		Recommendation	
	Volition		
	Intention		
Epistemic		Absolutely certain	
		Probable	
		Possible	
		Improbable	
		Impossible	

Both the marker and the scope are described in terms of polarity and presence/absence of an interrogative context. If the marker or the scope has negative polarity, the negation marker is annotated and connected to its target. The modal scope is also described in terms of the features '± control' and '± dynamic' with regard to the modalised event. The annotators identify the main participant in the modal event and annotate it based on criteria of animacy and agentivity. The option 'animate' or 'inanimate' is checked for both agent-like and patient-like participants. The WoPoss annotation does not take into consideration semantic roles. Due to a necessary operational simplification, the option 'patient' is checked by default for the syntactic subject of a passive verbal form. This allows the user to distinguish the subject of a passive verb from that of a deponent verb, as in the morpho-syntactic annotation both are annotated as passive. Usually only the main first argument participant is annotated, but in the case of passive verbs the agent can also be annotated. The participant(s) are then connected to their scope. In some cases, the presence of a participant is implicit or there is no participant at all (for example with atmospheric verbs). In such cases, the annotators indicate that the participant is implicit or that there is no participant in the corresponding feature of the scope layer. This thorough description of the modalised event through the dynamicity and control features and with the identification of the participant may provide an important insight on how modality works by examining whether

certain types of modality and/or modal markers are preferred (or avoided) with certain types of events.

It is worth mentioning that some elements, typically the modal scope, can be discontinuous. In such cases, the annotators combine the use of the span layer 'modal unit' with a chain layer which enables the connection between the different discontinuous elements.

Finally, some values can be ambiguous. If the modal passage can be interpreted according to more than one modal reading, the annotators annotate the passage for both interpretations. If in the scope the presence of dynamicity and control is ambiguous, the annotators choose the option '±'.

Each work is annotated by at least one annotator. The annotation is then checked and, if necessary, corrected, by the Principal Investigator of the WoPoss project. Thus, each text is usually annotated by one person and then reviewed by a different person. A double annotation was carried out for one of the texts in order to evaluate the inter-annotator agreement. The text contained 251 potential modal markers. With regards to the retrieval of these potential modal markers, the precision is 1 for both annotators and the average recall is 0.96. Considering pertinence evaluation (this is, whether a potential marker is modal and pertinent according to the project criteria), when both annotators retrieved the marker, their agreement has a Krippendor's Alpha (1980)[7] score (α) of 0.676. When both annotators agreed that a marker was modal, their agreement concerning the type of modality (i.e. dynamic, deontic or epistemic) has a α score of 0.758. Within this evaluation, if an annotator considers that a passage is ambiguous and provides a double annotation but the other annotator does not, it is measured as a disagreement (interestingly, there are no modal passages that both annotators interpret as ambiguous). If in those eleven passages annotated by one of the annotators with a double reading we consider that there is agreement if at least one of the two ambiguous readings matches the interpretation of the other annotator, then the agreement concerning the type of modality increases to a α score of 0.862. Agreement related to semantic annotation tasks can be notoriously low (see Véronis, 1998). In the case of modality annotation in English, Rubinstein et al. (2013) had a α score of 0.49 measuring the agreement of their possible 10 modal types. By collapsing these types into two categories (priority vs non-priority) the agreement score increases to 0.89. In McGillivray et al. (2022) which also worked with Latin in a task related to polysemy, the average pairwise agreement calculated as Spearman correlation coefficient was 0.69. While the inter-annotator calculation presented may give an indication of the annotations' reliability, it is not representative of the reliability of the final dataset, since in it, the work of the annotators goes through a curation process: firstly the PI reviews all the annotations, and secondly, additional corrections take place during post-processing (see Sect. 7) thanks to the schema association which flags any inconsistencies in the annotation.

[7] The scores were calculated using the NLTK Python library. For further details about the formula implemented see https://www.nltk.org/_modules/nltk/metrics/agreement.html.

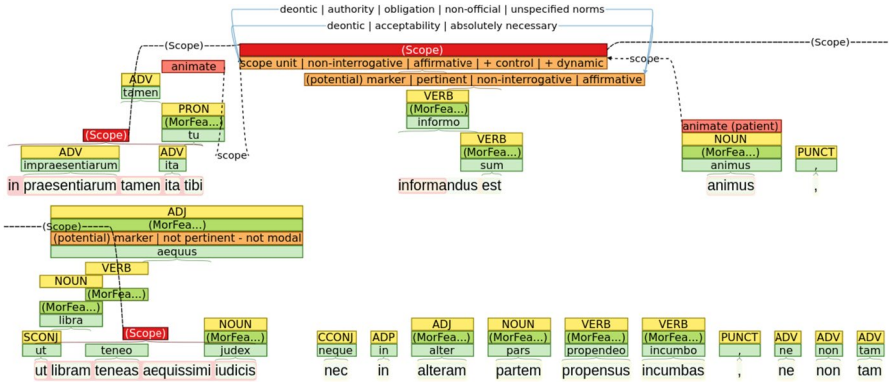

Fig. 1 Example of the annotation of a modal passage in INCEpTION. (colour figure online)

4.2 Annotation example

The annotation of a modal passage in the WoPoss framework can reach a high level of complexity. In example 1, we show the annotation of a modal passage taken from the WoPoss corpus.

(1) Minucius Felix, *Octavius*, 5 [...] *in praesentiarum tamen ita tibi informandus est animus, ut libram teneas aequissimi iudicis* [...]
[...] yet for the time being you must deliberately hold the balance of impartial justice [...][8]

Figure 1 illustrates how this passage was annotated in the INCEpTION platform.

As mentioned in Sect. 4.1, the text uploaded in INCEpTION is already tokenised, lemmatised, morphologically tagged and syntactically analysed. The lemmas appear in the first layer of annotation, which corresponds to the light green boxes in Fig. 1. The second layer of information is the morphological annotation, which appears in green boxes and displays morphological features in the format adopted by Universal Dependencies (UD) (de Marneffe et al., 2021). Finally, each token is given a PoS. The results of PoS-tagging appear in yellow boxes above the morphological annotation. The PoS tags also adhere to UD standards. Therefore, the form *tibi* in the passage, which is a personal pronoun in dative singular, will display the following annotation: 'tu' as lemma; 'Case = Dat|Gender = Masc|Number = Sing' as morphological features; 'PRON' as PoS. If the automatic annotation is incorrect, the annotator clicks on the box corresponding to the layer of annotation that needs to be corrected and changes the values of that field. In the passage illustrated in Fig. 1, the automatic annotation for *in praesentiarum* has been corrected. This phrase was automatically annotated as two separate tokens with the following description: *in* as a preposition with lemma 'in', and *praesentiarum* as a verb in gerundive with

[8] Translation by T.R. Glover and Gerald H. Rendall (Tertullian & Minucius Felix, 1931, p. 321).

lemma ‘praesentia’. The annotation of *praesentiarum* as a verb is obviously wrong. Moreover, there is also a problem with the lemma. As explained in the guidelines, in order to check the lemmatisation the annotator should use the Oxford Latin Dictionary (OLD) (Glare, 2012) as reference. In the description for the noun *praesentia* the OLD mentions the phrase *in praesentiarum* and refers back to the contracted form *impraesentiarum* ‘for the time being’. Therefore, it was necessary to correct the automatic annotation in order to adhere to the OLD. However, in this specific case it was also necessary to keep the original text, which does not present the adverb *impraesentiarum* but reads *in praesentiarum*. In order to solve this issue, the annotator deleted the annotation for the preposition *in*, and only annotated *praesentiarum* as an adverb with lemma *impraesentiarum*, adding a note to justify the changes.

Once the automatic annotation has been checked and corrected, the annotator proceeds with the manual semantic annotation. In Fig. 1 the modal marker is the morphological element *-ndus* with the copula *est*. The marker scopes over the following portion of text: *in praesentiarum tamen ita tibi informa- […] ut libram teneas aequissimi iudicis*. The layer ‘modal unit’ is used in order to annotate both the modal marker and its scope. The annotation of these two elements appears in orange boxes above the automatic annotation and displays the annotated features for each of them. The modal marker *-ndus est* is annotated as a ‘(potential) marker’: this option is defined as such because each lemma listed in the guidelines can be a modal marker, but the presence or absence of a modal value is defined by the context in which the lemma appears. If the occurrence does not have a modal value in that specific context, it will not be defined as a modal marker. This is the case in example 1 for *aequissimi*, a superlative form of the adjective *aequus*. This adjective is listed in the *Guidelines* as one of the lemmas to be annotated in the WoPoss corpus. In this specific context, however, the adjective is not modal: it is used as a modifier to the noun *iudicis* (they agree in gender and number) and does not have a scope to modalise. Therefore, it was annotated as a ‘(potential) marker’, but ‘not pertinent – not modal’. On the contrary, *-ndus est* is a modal marker, and as such is annotated as ‘pertinent’. The sentence function for both the marker and the scope is annotated as non-interrogative, and their polarity is affirmative. The layer of the scope also shows the features that describe the modalised event (‘+control’ and ‘+dynamic’): these are defined based on the main verb of the scope, which in example 1 is the lexical base of the gerundive, *informa-*. The main verb is annotated with the layer ‘modal unit’, and the rest of the scope with the chain layer ‘scope’, which is then linked back to the main verb. This option allows the annotator to deal with discontinuous scopes: sometimes the portion of text to which the marker refers is interrupted by other elements that are not part of its scope. In example 1 this is the case for the subordinate clause *ut libram teneas aequissimi iudicis,* which—since the WoPoss annotation takes into account subordinate clauses depending on the immediate target of the modal marker (here *ita tibi informa-*)—is part of the scope but is separated from the rest of it by the modal marker and the participant *animus*. Other elements of a modal passage can be discontinuous, such as modal markers and participants. In order to tackle this, the annotator annotates each discontinuous element with the corresponding chain layer (‘scope’ for the situation illustrated in Fig. 1, ‘marker’ or ‘participant’ in the other cases) and links them together. Figure 1 also shows two

annotated participants. This is because the modal passage contains a passive periphrastic, instantiated by the gerundive with the copula *est*. In this case there are two expressed participants: a patient and an agent. The annotator chooses the layer 'participant' and annotates *animus* as 'animate – patient' and *tibi* as 'animate'. The two participants are then linked to the scope by an arrow. The presence of an annotated marker and its scope enables the annotation of the modal relation between them. This type of information appears above the two modal units, between blue arrows. As shown in Fig. 1, two possible modal relations were created for the same marker and scope: the first one is 'deontic – acceptability – absolutely necessary' and the second one is 'deontic authority – obligation – non-official – unspecified norms'. As the modal reading of this passage was ambiguous between two different subtypes of deontic modality, both of them were annotated as possible readings: acceptability (which implies an evaluation of the modalised event based on ethical/moral norms) and authority (which implies the presence of a source of authority).

5 Formalisation of the annotation scheme

The corpus is encoded in TEI-XML, a standard with a historic and extensive use in linguistic corpus development—e.g., the *Corpus Encoding Standard* (Ide, 1998). TEI was particularly adequate to combine both the linguistic annotation and the philological and editorial information available for the source text, since it is crucial in our corpus to encode certain editorial decisions (such as additions). The advantages of this particular standard when combining different levels of annotations within linguistic projects have been highlighted in previous literature (e.g., Przepiórkowski & Bański, 2011). In addition, the TEI offered encoding strategies to define our particular theoretical framework in regards to modality, facilitating the formalisation of the semantic annotation.

The encoding strategy we implemented combines inline and stand-off annotation, this is, the insertion of annotation outside the text flow in opposition, thus, to inline annotation (see Figs. 2, 3, cf. below for the explanation). The annotation schema was formalised using the ODD specification format which enables the customisation of the TEI schema in a literate programming fashion (TEI Consortium, 2018). The ODD file is then processed to generate a file in a schema language by using the TEI Stylesheets. The ODD files and the resulting Relax-NG schemas[9] are available under a CC-BY licence (WoPoss Project, 2022a). In addition, the annotated texts released so far are freely available in GitHub.[10]

The TEI Guidelines are a flexible scheme that can be customised to answer the modelisation needs of very heterogeneous and diverse projects. A consequence of this flexibility is the multiple ways in which the same information can be encoded. In the case of linguistic information, the TEI offers both main strategies mentioned

[9] REgular LAnguage for XML Next Generation (RELAX NG) is a schema language for XML (Murata, 2014).

[10] https://github.com/WoPoss-project/WoPoss-corpus.

```
<s n="380">
    <w msd="Case=Acc|Gender=Fem|Number=Sing" pos="ADJ" lemma="omnis">Omnem</w>
    <w pos="ADV" lemma="adeo">adeo</w>
    <w pos="NOUN" lemma="mundus" msd="Case=Acc|Gender=Masc|Number=Sing">
        <seg function="participant" type="inanimate" corresp="#oct2349230">mundum</seg></w>
    <pc>,</pc>
    <w pos="SCONJ" lemma="si">si</w>
    <w msd="Case=Acc|Gender=Masc|Number=Sing" pos="NOUN" lemma="sol">solem</w>
    <w msd="Case=Acc|Gender=Fem|Number=Sing" pos="NOUN" lemma="lunam">lunam</w>
    <w msd="Case=Acc|Gender=Neut|Number=Plur" pos="ADJ" lemma="reliquus">reliqua</w>
    <w msd="Case=Acc|Gender=Masc|Number=Plur" pos="NOUN" lemma="aster">astra</w>
    <w msd="Aspect=Perf|Mood=Sub|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act"
        pos="VERB" lemma="desio">desierit</w>
    <w msd="Case=Gen|Gender=Masc|Number=Plur" pos="ADJ" lemma="fontium">fontium</w>
    <w msd="Case=Acc|Gender=Fem|Number=Plur" pos="ADJ" lemma="dulcis">dulcis</w>
    <w msd="Case=Abl|Gender=Fem|Number=Sing" pos="NOUN" lemma="aqua">aqua</w>
    <w pos="CCONJ" lemma="et">et</w>
    <w msd="Case=Abl|Gender=Fem|Number=Sing" pos="NOUN" lemma="aqua">aqua</w>
    <w msd="Case=Abl|Gender=Fem|Number=Sing" pos="ADJ" lemma="marinum">marina</w>
    <w msd="Tense=Pres|VerbForm=Inf|Voice=Act" pos="VERB" lemma="nutrio">nutrire</w>
    <pc>,</pc>
    <seg function="scope" next="abi" part="Y" ana="#oct2349230">
        <w pos="ADP" lemma="in">in</w>
        <w pos="NOUN" lemma="vis" msd="Case=Acc|Gender=Fem|Number=Sing">vim</w>
        <w pos="NOUN" lemma="ignis" msd="Case=Acc|Gender=Masc|Number=Plur">ignis</w>
    </seg>
    <w pos="VERB" lemma="abeo"
        msd="Case=Acc|Gender=Masc|Number=Sing|Tense=Fut|VerbForm=Part|Voice=Act">
        <seg function="scope" ana="#oct2349230" part="Y">abi</seg>
        <seg function="marker" ana="#oct2248029">turum</seg>
    </w>
    <pc>,</pc>
    <w msd="Case=Dat|Gender=Masc|Number=Plur" pos="NOUN" lemma="Stoicus">Stoicis</w>
    <w msd="Case=Nom|Gender=Fem|Number=Sing|Tense=Pres|VerbForm=Part" pos="VERB"
        lemma="consto">constans</w>
    <w msd="Case=Nom|Gender=Fem|Number=Sing" pos="NOUN" lemma="opinium">opinio</w>
    <w msd="Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act"
        pos="VERB" lemma="sum">est</w>
    <pc>,</pc>
    <w pos="SCONJ" lemma="quod">quod</w>
    <w msd="Case=Abl|Gender=Masc|Number=Sing" pos="VERB" lemma="consumo">consumto</w>
    <w msd="Case=Abl|Gender=Masc|Number=Sing" pos="NOUN" lemma="umor">umore</w>
    <w msd="Case=Nom|Gender=Masc|Number=Sing" pos="ADJ" lemma="mundus">mundus</w>
    <w msd="Case=Nom|Gender=Masc|Number=Sing" pos="PRON" lemma="hic">hic</w>
    <w msd="Case=Acc|Gender=Masc|Number=Plur" pos="ADJ" lemma="omnis">omnis</w>
    <w msd="Mood=Ind|Number=Sing|Person=3|Tense=Fut|VerbForm=Fin|Voice=Act" pos="VERB"
        lemma="ignesco">ignescet</w>
    <pc>.</pc>
</s>
```

Fig. 2 TEI-XML annotation of the sentence *'Omnem adeo mundum, si solem lunam reliqua astra desierit fontium dulcis aqua et aqua marina nutrire, in vim ignis abiturum, Stoicis constans opinio est, quod consumto umore mundus hic omnis ignescet.'* (Minucius Felix, *Octavius*, 34) which contains a modal passage

above: inline annotation and stand-off annotation, the latter generally modelised through the implementation of Feature Structures (FS, cf. below) (TEI Consortium, 2021a). The WoPoss approach combines both strategies in order to increase the legibility of the source files.

5.1 Metadata

A TEI document has at least two parts: a header containing metadata describing the document, and the text itself (see Burnard, 2014). In our corpus, the element <teiHeader> (TEI header) supplies the main metadata associated with the digital file, such as the different agents responsible for its contents (annotators and data curators) and the reference to the electronic source that we used. In addition, it also contains the identification of the work and its author (if any) pointing to an authority

```
<fs xml:id="oct2349314" type="relation">
   <f name="marker" fVal="oct2248029"/>
   <f name="scope" fVal="oct2349230"/>
   <f name="modality">
      <symbol value="dynamic"/>
   </f>
   <f name="meaning">
      <symbol value="necessity"/>
   </f>
   <f name="type">
      <symbol value="situational"/>
   </f>
   <f name="subtype">
      <symbol value="inevitability"/>
   </f>
   <f name="note">
      <string>The necessity meaning is probably
         bound to the conditional sentence.</string>
   </f>
</fs>
```

Fig. 3 Description of a modal relation through feature structures

file also encoded in TEI. This file contains descriptive metadata that provides the sociolinguistic description of the author (period, gender, geographical provenance) and the description of the work according to the parameters defined in the WoPoss guidelines (Dell'Oro, 2023, pp. 24–26): date of composition, type of transmission, basic textual features (whether the text is in verse, dialogued and/or a translation) and the textual genre.

5.2 Structural and editorial markup

Each work of the corpus contains a basic division into major structural elements (elements) to delimit the chapters or sections. These sections are made up of one or more <ab> elements (anonymous block). Due to the well-known challenges of modelling texts as ordered hierarchies of content objects (OHCO) imposed by the XML format (Renear et al., 1996), we also resort to the implementation of boundary marking with empty elements of certain structural features. For example, instances of direct speech or verse passages within a prose work are delimited with the empty element <anchor/> with a @type attribute to categorise it and a @subtype attribute with the possible values of 'start' and 'end' to establish the boundary.

As mentioned above (Sect. 4.1) the philological quality of the annotated passages as presented in the source texts is verified by consulting the best available editions. It is thus common that there is editorial information that we want to formalise in our

annotated version. For example, we include textual passages that are considered in the edition of reference as superfluous within the element <surplus>. Editorial additions are enclosed in the element <supplied> with a @source attribute pointing to the bibliographical description of the reference edition.

5.3 Linguistic markup

The results of the automatic annotation concerning sentence division and tokenisation are formalised using the structural elements <s> (sentence), <w> (word),[11] and <pc> (punctuation character). Each word contains as well the attributes @lemma and @pos (part of speech), and, if pertinent, @msd (morphosyntactic description): the values of these attributes resulted from the automatic annotation. Note that the contents of these attributes were manually reviewed and corrected for words that are part of an annotated passage.

As explained in Sect. 4, we decompose a modal passage in different constituent units. Each one of these components is enclosed in a <seg> element (arbitrary segment) with a @function attribute to define the type of constituent: modal marker, its scope, negation or participant of the state of affairs (see in Fig. 2, lines 27–31 which contain a modal marker, the suffix *-turum*, a part of its scope, the stem of the verb). In the case of the major components, marker and its scope, an @ana attribute (analysis) points to a <fs> element that contains the detailed description of this component by including the features outlined in Sect. 4 (e.g., type of utterance, polarity, dynamicity, control). Although the theoretical model behind FS stems from the structuralist framework of the Prague school and later developments, FS can be considered a general type of data structure (Stegmann & Witt, 2009) with successful implementations not limited to the linguistic field.[12] A feature structure is a group of *attribute:value* pairs, in which the values may be either atomic or composed by complex units in the form of nested feature structures (Stegmann & Witt, 2009). In the case of a feature structure containing the description of a modal marker, we have a <f> (feature) element for each individual feature (e.g., pertinence, whether the item is modal, type of utterance, polarity). This element may contain a <symbol> (symbolic value) for controlled vocabularies, this is, when we have a finite list of symbols, or a <binary> element for Boolean features (e.g., pertinence).

The <seg> elements may contain an attribute @part to indicate whether any of the modal units is fragmented and, by having the same @ana value, we are able to reconstruct discontinuous elements. Figure 2 shows the annotation of a sentence that

[11] The element <w> is not exactly equivalent to the notion of token: for example in the case of *in praesentiarum* mentioned in Sect. 4.2 in relation to example 1, the adverb *impraesentiarum* is taken as unit of reference, thus we have one <w> element (with the corresponding @pos and @lemma attributes) with the content 'in praesentiarum'.

[12] For examples of linguistic annotation implemented through FS see Bański and Przepiórkowski (2009); for applications in other domains, see Bermúdez Sabel (2020) for a philological implementation, or Triplette et al. (2018) for a literary one. For a discussion of the advantages of FS see Langendoen and Simons (1995), and Stegmann and Witt (2009).

contains a modal passage formalised in TEI-XML (Minucius Felix, *Octavius*, 34).[13] The complete annotated file is available under a CC-BY licence (WoPoss Project, 2022b).

Within the scope of the modal marker, the <w> element containing its main verb displays a @function attribute with the value 'main'. In the case of analytic passive forms, the auxiliary verb also presents the @function attribute with the value 'aux'. The <seg> elements concerning the participants and the negation elements have a @corresp attribute (corresponds) that points to the scope of a marker, in the case of the participant, and to a marker or, in certain cases,[14] to its scope, in the case of the negation (see Fig. 2, line 5 for the annotation of a participant).

There are three types of feature structures concerning the description of the main components of a modal passage: the marker, its scope and the modal relation (see Fig. 3 with the description of the modal reading of the passage from *Octavius* shown in Fig. 2).

The three types of feature structures are heavily constrained.

6 Data exploitation

This section will present a brief use case to illustrate the type of information that can be extracted thanks to our comprehensive annotation scheme.[15] The use case focuses on one of the annotated markers: the adverb *certe* 'certainly' which is a polysemic marker and thus is relevant to examine its different functions. This adverb has (1) an epistemic use, stating absolute certainty and therefore being modal; (2) a reinforcing use by which a firm assertion that something is true is made; (3) a scalar use to establish that something is at least the case in comparison to something else; and (4) a pragmatic use in interrogative questions.[16] According to our theoretical framework, only the first function is modal, and functions 2 and 3 are what we consider postmodal meanings (see van der Auwera & Plungian, 1998).

One of the first things that can be explored is the comparison of frequencies of this potential modal marker with other markers from the same word family (Fig. 4),

[13] *Omnem adeo mundum, si solem lunam reliqua astra desierit fontium dulcis aqua et aqua marina nutrire, in vim ignis abiturum, Stoicis constans opinio est, quod consumto umore mundus hic omnis ignescet*. 'So too the universe, if sun, moon and stars are deprived of the fountains of fresh water and the water of the seas, will disappear in a blaze of fire. The Stoics firmly maintain that when the moisture is dried out, the universe must all take fire', translation by T.R. Glover and Gerald H. Rendall (Tertullian & Minucius Felix, 1931, p. 418).

[14] The negation of the modal scope is annotated as an independent component only when it is not inside the scope itself, that is, when there is a negation which is semantically affecting the state of affairs, but it is not syntactically in its scope.

[15] These results are based on the analysis of WoPoss Project (2022b), which, at the time of writing, included the following works: the epigraphic document known as the *Senatus consultum* (2nd c. BCE), the first book of the *Metamorphoses* by Ovid (1st c. BCE), *Satyricon* by Petronius (1st c.), *De spectaculis* by Tertullian (2nd c.), and *Octavius* by Minucius Felix (2nd c.).

[16] For the different functions see Schrickx (2011, p. 215); for an overview of the meanings of *certe* see Marongiu and Dell'Oro (2021).

like the adjective *certus* 'certain' or the adverb *certo* 'certainly' along with the compounds with the negative prefix: the adverb *incerte* 'uncertainly', the adjective *incertus* 'uncertain' and the noun *incertum* 'uncertainty'. This type of query is easily implemented by retrieving the values within the @lemma and @pos attributes and then the modal meaning (if any). In this sense, we can see that (1) *certe* is the most frequent form of this word family, (2) some words within the family were not attested (*certo*) and we can clearly see that there are some lexical items more likely to be modal than others.

Previous works that have addressed this adverb or class of adverbs (that is, potentially epistemic modal adverbs according to our theoretical framework), have described it as an adverb that cannot be in the scope of a negation (but that may have negations in its scope) and that hardly ever occur in interrogative sentences (Pinkster, 2014, pp. 192–193; Schrickx, 2011, p. 212). We can verify this by interrogating the features 'type of utterance' and 'polarity' for both the marker and its scope and the results confirm these previous expectations.

In addition, we can examine other linguistic features annotated to attempt to see if there is any linguistic evidence that would allow us to predict if *certe* will function as a modal epistemic marker or as a post-modal discourse marker, i.e. whether it organises the flow and structure of the discourse. By analysing the state of affairs (described through the features 'dynamicity' and 'control') no clear trend emerges (*certe* may be modal in any of the possible combinations). What we can see is that in all the cases in which *certe* is immediately preceded by a conjunction it functions as a discourse marker: this is the expected behaviour in constructions like *aut … aut certe* or *uel.. uel certe* (which could be translated as 'and in addition'), but we find the same behaviour with *aut* without the repetition and also with the coordinating conjunction *et* ('and').

7 Discussion

Stand-off annotation presents a number of advantages (Bański, 2010). We can create complex structures of annotations using as reference any linguistic unit (from the phoneme to the sentence, for instance) and it is possible to combine multiple types of analysis without the concern of elements overlap, the main syntax restriction of XML. Reusing the datasets is also facilitated by separating the primary text from its annotation. Thanks to this separation, the primary text remains more legible and neutral, and the analytical and interpretative annotation is kept in a different file or files. Therefore, other community members can formalise their own interpretations by easily modifying the existing annotation or by developing a complete new set of annotations from the same 'clean' source document. Latin has resources that implement this annotation method, like the corpus *Opera Latina Adnotata* (Celano, 2021).

After carrying out the semantic annotation in the INCEpTION platform, we transform the UIMA CAS XMI results into TEI. We decided to change a purely stand-off annotation format into a hybrid one because the UIMA CAS XMI system is based on character offsets relations, this is, it is based on specific character positions counting from the beginning of the text: this strategy was suboptimal

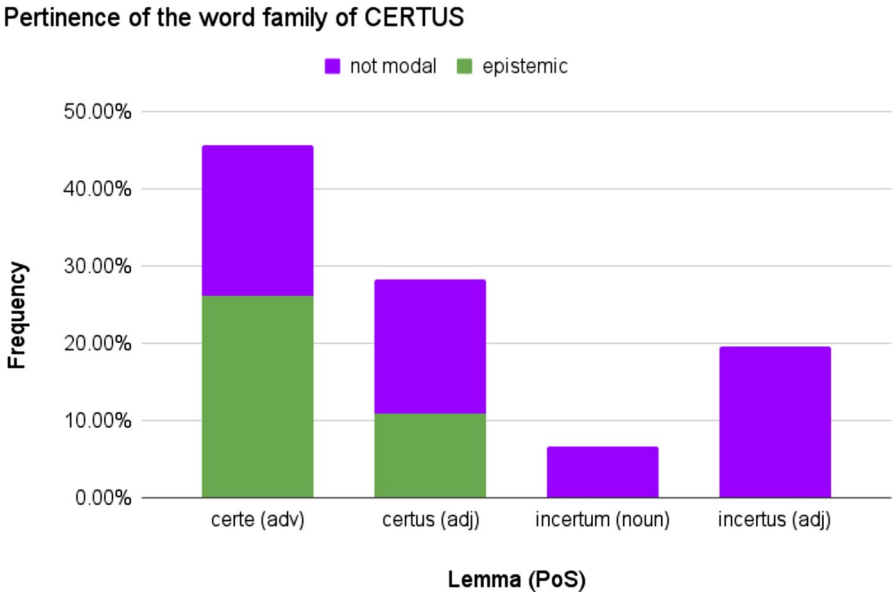

Fig. 4 Relative frequency of each attested member of the word family of *certus* with its modal and non-modal meanings

in our case for several reasons. Firstly, during the manual annotation process, the annotators detect different types of textual problems that need to be corrected during post-processing, which would mean that all character offsets need to be recalculated. Then, a second pass of corrections of the linguistic annotation is required because the constraints that can be formalised in the annotation platform are insufficient. INCEpTION allows us to define certain constraints that are extremely useful. For example, we can declare controlled vocabularies, and we can also design an annotation workflow by displaying certain fields conditioned by previously selected values in another field. However, due to the complexity of the WoPoss annotation scheme, we need to implement very restrictive rules and constraints that can only be properly formalised in a schema language. After the curation of the annotation within INCEpTION, a second correction process takes place during post-processing after the TEI conversion. During this correction certain inconsistencies are detected and potential modal markers that have not been annotated are also flagged. To implement all these corrections we must work within an annotation model that facilitates the manual edition. A TEI file combining both inline annotation and attribute-based stand-off annotation provided us with a legible format, without losing interoperability nor reusability advantages. For example, as annotators also modify the results of the automatic annotation concerning the morphological description that are formalised in the TEI file through the attribute @msd, we defined several rules into our schema to verify the conformance of this attribute. This means that the feature has always the proper name, and that its value corresponds to the controlled list of values defined by UD. In addition, we have further rules that check the consistency

between the values of @pos and the value of @msd (for example, if the part-of-speech has the value "VERB" then @msd must contain at least, the features 'VerbForm', 'Tense' and 'Voice'). If an inconsistency within the @msd attribute is found in a word that belongs to a modal passage, then it is flagged as an error by the schema; if the word does not belong to a modal passage, it is flagged as a warning and can be used as a cue to improve the linguistic annotation that has not been manually reviewed.

The semantic annotation is modelled through the use of FS and we argue that this formalisation increases the reusability of our dataset. The TEI recommendations for feature structures have been adopted as ISO Standard,[17] thus providing stability to the model (Romary, 2015). In addition, the *attribute:value* syntax of FS (which does not allow mixed content)[18] enables a straightforward transformation to other data models, like semantic triples, which facilitates a possible publication of our dataset as linked data.

The detailed description of the feature structure model is formalised in a Feature System Declaration (FSD) (TEI Consortium, 2021b). The FSD is a way to define a well-formed and valid feature structure, formalising its necessary components together with other types of constraints, and declaring all the feature names and values. The FSD is a means to document the implementation of a feature-structure-based schema. Figure 5 shows two conditions formalised in the FSD: if the feature 'modality' has as its value 'deontic', then it must contain a feature 'type' with one of the declared four possible values (authority, acceptability, volition, intention); a second condition establishes the only possible values of a mandatory feature 'subtype' when the 'type' is 'authority'. Thanks to this model, we are able to process this declaration and automatically transform it into a series of constraints formalised in Schematron[19] and later inserted in the ODD file (see Fig. 6). This transformation was written in XSLT by modifying previously developed code (Bermúdez Sabel, 2022). The FSD (and the code to process it) is available under a CC-BY licence (WoPoss Project, 2022a).

8 Conclusions

The WoPoss scheme combines multiple layers of annotation which are particularly complex due to the granularity of the semantic description and to the interrelations between the different components—which may be discontinuous and thus creating the need for additional linking between the parts. If we aim to provide the community with a useful annotated dataset, having both a well-documented annotation

[17] See the standard *24610-1 Language Resource Management—Feature Structures—Part One: Feature Structure Representation* (ISO/TC 37/SC 4, 2007).

[18] Mixed content refers to the presence of both elements and text nodes as children of a given element.

[19] Schematron is a language for making assertions about the presence or absence of patterns in linked XML documents (Jelliffe, 2021).

```
<cond>
   <fs>
      <f name="modality">
         <symbol value="deontic"/>
      </f>
   </fs>
   <then/>
   <fs>
      <f name="type">
         <vAlt>
            <symbol value="authority"/>
            <symbol value="acceptability"/>
            <symbol value="volition"/>
            <symbol value="intention"/>
         </vAlt></f>
   </fs>
</cond>
<cond>
   <fs>
      <f name="type">
         <symbol value="authority"/>
      </f>
   </fs>
   <then/>
   <fs>
      <f name="subtype">
         <vAlt>
            <symbol value="obligation"/>
            <symbol value="recommendation"/>
            <symbol value="permission"/>
         </vAlt></f>
   </fs>
</cond>
```

Fig. 5 Excerpt of the Feature System Declaration containing two constraints

scheme and restrictive schemas is crucial. The *WoPoss annotation guidelines* are a fundamental part of our documentation: beside their internal function as a reference for the annotators, they present our theoretical framework and exemplify how to tackle specific challenges. For its part, the formalised annotation schema is very specific and detailed. It also fulfils the double role of having (i) a practical internal function that ensures the consistency of our annotation and (ii) an external function that facilitates the reusability of our dataset. With this goal in mind, we have developed a FSD. This FSD (i) presents the linguistic components used to define each constituent of the modal reading, (ii) ensures the correct implementation of the UD tagset in the morphosyntactic description, and (iii) declares certain constraints, thus providing additional metadata about the annotation scheme. In addition, the use of ODD to customise the TEI schema facilitates the understanding of our annotation, which is

```
<sch:rule
    context="tei:fs[@type eq 'relation'][tei:f[@name eq
    'modality'][tei:symbol[@value eq 'deontic']]]">
    <sch:assert test="tei:f[@name eq 'type']">Missing
        tei:f[@name eq 'type']</sch:assert>
    <sch:assert
        test="tei:f[@name eq 'type']/tei:symbol/@value =
        ('authority', 'acceptability', 'volition',
        'intention')" >Possible values of tei:f[@name eq
        'type'] are 'authority' 'acceptability' 'volition'
        'intention'</sch:assert>
</sch:rule>
<sch:rule
    context="tei:fs[@type eq 'relation'][tei:f[@name eq
    'type'][tei:symbol[@value eq 'authority']]]">
    <sch:assert test="tei:f[@name eq 'subtype']">Missing
        tei:f[@name eq 'subtype']</sch:assert>
    <sch:assert
        test="tei:f[@name eq 'subtype']/tei:symbol/@value =
        ('obligation', 'recommendation', 'permission')">
        Possible values of tei:f[@name eq 'subtype'] are
        'obligation' 'recommendation' 'permission'
    </sch:assert>
</sch:rule>
```

Fig. 6 Automatically-generated Schematron rules based on the constraints shown in Fig. 5

indispensable for reusing our dataset. Thanks to both the FSD and the ODD we are also providing a well-documented annotation schema that could be implemented by other projects interested in carrying out the semantic annotation of modality within a similar theoretical framework (either in Latin or in other languages).

Acknowledgements This work was funded by the Swiss National Science Foundation (SNSF N° PP00P1 176778 and N° PP00P1 214102) and is led by Francesca Dell'Oro at the University of Neuchâtel. We wish to thank Jan Nuyts and Paola Pietrandrea for providing us support in the elaboration of our annotation model. We would like to thank the reviewers for the effort and expertise they contributed in reviewing the article.

Author contributions This paper was written collaboratively: H. Bermúdez Sabel is mainly responsible for Sects. 1, 4.1 (second part), 5, 6, 7 and 8; F. Dell'Oro is mainly responsible for Sects. 2 (second part), 3 and 4.1 (first part) and the general supervision; P. Marongiu is mainly responsible for Sects. 2 (first part) and 4.2.

Funding The research leading to these results is funded by the Swiss National Science Foundation (SNSF N° PP00P1 176778 and N° PP00P1 214102).

Declarations

Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Ávila, L. B., Mendes, A., & Hendrickx, I. (2015). Towards a unified approach to modality annotation in Portuguese. In *Proceedings of the workshop on models for modality annotation*. Retrieved April 7, 2022, from Association for Computational Linguistics. https://aclanthology.org/W15-0301

Baker, K., Bloodgood, M., Dorr, B. J., Filardo, N. W., Levin, L., & Piatko, C. (2014). *A modality lexicon and its use in automatic tagging* (arXiv:1410.4868). arXiv. https://doi.org/10.48550/arXiv.1410.4868

Bański, P. (2010). Why TEI stand-off annotation doesn't quite work: And why you might want to use it nevertheless. In *Proceedings of Balisage: The markup conference 2010* (Vol. 5). Presented at the Balisage: The markup conference 2010, Montréal, Canada. https://doi.org/10.4242/BalisageVol5.Banski01

Bański, P., & Przepiórkowski, A. (2009). Stand-off TEI annotation: The case of the National Corpus of Polish. In *ACL-IJCNLP '09: Proceedings of the third linguistic annotation workshop* (pp. 64–67). Presented at the third linguistic annotation workshop, Suntec, Singapore: Association for Computational Linguistics. https://doi.org/10.3115/1698381.1698392

Bermúdez Sabel, H. (2020). Encoding of variant taxonomies in TEI. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.2676

Bermúdez Sabel, H. (2022). *FS-validator*. XSLT. Retrieved April 13, 2022, from https://github.com/HelenaSabel/FS-Validator

Burnard, L. (2014). The structural organization of a TEI document. In *What is the text encoding initiative?: How to add intelligent markup to digital resources*. OpenEdition Press. Retrieved June 9, 2022, from http://books.openedition.org/oep/681

Clackson, J. & Horrocks, G. (2007). The Blackwell history of the Latin language. Oxford: Wiley-Blackwell.

Clackson, J. (ed.). (2011). A companion to the Latin language. Oxford: Wiley-Blackwell.

Celano, G. (2019). The dependency treebanks for ancient Greek and Latin. In *Digital classical philology* (pp. 279–298). https://doi.org/10.1515/9783110599572-016

Celano, G. G. A. (2021). *Opera Latina Adnotata (OLA)*. Retrieved April 6, 2022, from http://ola.informatik.uni-leipzig.de/en/index.html

Coates, J. (1983). *The semantics of the modal auxiliaries*. Croom Helm.

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal dependencies. *Computational Linguistics, 47*(2), 255–308. https://doi.org/10.1162/coli_a_00402

Dell'Oro, F. (2023). WoPoss guidelines for the annotation of modality. Revised version. *Zenodo*. https://doi.org/10.5281/zenodo.10427053

Du Cange, C. du F., Carpenter, P., Henschel, L. G. A., & Favre, L. (1883–1887 [1678]). *Glossarium mediæ et infimæ latinitatis* [Glossary of Middle and Low Latin]. Favre.

Gast, V., Bierkandt, L., & Rzymski, C. (2015). Annotating modals with GraphAnno, a configurable lightweight tool for multi-level annotation. In *Proceedings of the workshop on models for modality annotation*. Association for Computational Linguistics. https://aclanthology.org/W15-0303

Ghia, E., Kloppenburg, L., Nissim, M., Pietrandrea, P., & Cervoni, V. (2016). A construction-centered approach to the annotation of modality. In H. Bunt (Ed.), *Proceedings of the 12th joint ACL-ISO workshop on interoperable semantic annotation* (pp. 67–74). ACL, ISO.

Glare, P. G. W. (Ed.). (2012). *Oxford Latin dictionary* (2nd ed., Vols. 1–2). Oxford University Press.

Haug, D. T. T., Eckhoff, H. M., & Welo, E. (2014). The theoretical foundations of givenness annotation. In K. Bech & K. G. Eide (Eds.), *Information structure and syntactic change in Germanic and Romance languages* (pp. 17–52). John Benjamins.

Haug, D. T. T., & Jøhndal, M. L. (2008). Creating a parallel treebank of the old Indo-European Bible translations. In C. Sporleder, & K. Ribarov (Eds.), *Proceedings of the second workshop on language technology for cultural heritage data (LaTeCH 2008)* (pp. 27–34).

Ide, N. (1998). Corpus encoding standard: SGML guidelines for encoding linguistic Corpora. In *Proceedings of the first international language resources and evaluation conference* (pp. 463–470).

ISO/TC 37/SC 4. (2007). *ISO 24610–1:2006, language resource management—Feature structures—Part 1: Feature structure representation*. Distributed through American National Standards Institute.

Jelliffe, R. (2021). Schematron. Retrieved April 6, 2022, from https://www.schematron.com/home.html

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R., & Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the*

27th international conference on computational linguistics: System demonstrations (pp. 5–9). Association for Computational Linguistics. http://tubiblio.ulb.tu-darmstadt.de/106270/

Kratzer, A. (1981). The notional category of modality. In *The notional category of modality* (pp. 38–74). De Gruyter. https://doi.org/10.1515/9783110842524-004

Krippendorff, D. K. (1980). *Content analysis: An introduction to its methodology*. Sage Publications Inc.

Laboratoire Ligérien de Linguistique. (2017). Modal—modèles de l'annotation de la modalité à l'Oral. https://hdl.handle.net/11403/modal/v1

Langendoen, D. T., & Simons, G. F. (1995). A rationale for the TEI recommendations for feature-structure markup. *Computers and the Humanities, 29*(3), 191–209. https://doi.org/10.1007/BF01830616

Lewis, C. T. (1890). *An elementary Latin dictionary*. Oxford University Press.

Lewis, C. T., & Short, C. (1879). *A Latin dictionary, founded on Andrews' edition of Freund's Latin Dictionary. Revised, enlarged and in great part rewritten by Charlton T. Lewis, PhD. and Charles Short*. Clarendon Press.

Marongiu, P., & Dell'Oro, F. (2021). "certe". v.1.0. WoPoss. https://woposs.unine.ch/maps/map-certe.html

Matthewson, L. (2016). Modality. In M. Aloni & P. Dekker (Eds.), *The Cambridge handbook of formal semantics* (pp. 525–559). Cambridge University Press. https://doi.org/10.1017/CBO9781139236157.019

McGillivray, B., & Kilgarriff, A. (2013). Tools for historical corpus research, and a corpus of Latin. In P. Bennett, M. Durrell, S. Scheible, & R. J. Whitt (Eds.), *New methods in historical corpus linguistics* (pp. 247–257). Narr.

McGillivray, B., Kondakova, D., Burman, A., Dell'Oro, F., Bermúdez Sabel, H., Marongiu, P., & Márquez Cruz, M. (2022). A new corpus annotation framework for Latin diachronic lexical semantics. *Journal of Latin Linguistics, 21*(1), 47–105. https://doi.org/10.1515/joll-2022-2007

Murata, M. (2014). RELAX NG. Retrieved April 6, 2022, from https://relaxng.org/

Narrog, H. (2012). *Modality, subjectivity, and semantic change: A cross-linguistic perspective*. Oxford University Press.

Nissim, M., & Pietrandrea, P. (Eds.). (2015). *Proceedings of the workshop on models for modality annotation*. Association for Computational Linguistics. https://aclanthology.org/W15-03

Nuyts, J. (2005). The modal confusion: On terminology and the concepts behind it. In A. Klinge & H. H. Müller (Eds.), *Modality. Studies in form and function* (pp. 5–38). Equinox Publishing.

Nuyts, J. (2019). Things to keep in mind when investigating the diachrony of modal expressions. In *Workshop on modality: From theory to encoding*. University of Lausanne.

Passarotti, M. (2019). The project of the Index Thomisticus Treebank. In M. Berti (Ed.), *Digital classical philology. Ancient Greek and Latin in the digital revolution* (pp. 299–319). Walter De Gruyter GmbH.

Pinkster, H. (2014). Attitudinal and illocutionary satellites in Latin. In H. Aertsen, M. Hannay, & R. J. Lyall (Eds.), *Words in their places. A festschrift for J. Lachlan Mackenzie* (pp. 191–198). Vrije Universiteit Amsterdam.

Portner, P. (2009). *Modality*. Oxford University Press.

Przepiórkowski, A., & Bański, P. (2011). Which XML standards for multilevel corpus annotation? In Z. Vetulani (Ed.), *Human language technology. Challenges for computer science and linguistics* (pp. 400–411). Springer. https://doi.org/10.1007/978-3-642-20095-3_37

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations*. https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Renear, A., Mylonas, E., & Durand, D. (1996). Refining our notion of what text really is: The problem of overlapping hierarchies. In N. Ide & S. Hockey (Eds.), *Research in humanities computing* (Vol. 4, pp. 263–280). Clarendon Press.

Romary, L. (2015). Standards for language resources in ISO—Looking back at 13 fruitful years. arXiv:1510.07851 [*cs*]. Retrieved April 7, 2022, from http://arxiv.org/abs/1510.07851

Rubinstein, A., Harner, H., Krawczyk, E., Simonson, D., Katz, G., & Portner, P. (2013). Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 workshop on annotation of modal meanings in natural language (WAMM)* (pp. 38–46). Association for Computational Linguistics. Retrieved April 7, 2022, from https://aclanthology.org/W13-0306

Saurí, R., Verhagen, M. & Pustejovsky, J. (2006). Annotating and recognizing event modality in text. In G. Sutcliffe& R. Goebel (Eds.), FLAIRS Conference (pp. 333–339). AAAI Press.

Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., & Tahmasebi, N. 2020. SemEval-2020 Task 1: Unsupervised lexical semantic change detection. In *Proceedings of the fourteenth workshop on semantic evaluation* (pp. 1–23). Barcelona (online). International Committee for Computational Linguistics.

Schrickx, J. (2011). *Lateinische Modalpartikeln: "Nempe", "Quippe", "Scilicet", "Videlicet" Und "Nimirum."* Brill Academic Pub.

Stegmann, J., & Witt, A. (2009). TEI feature structures as a representation format for multiple annotation and generic XML documents. In *Proceedings of Balisage: The markup conference 2009* (Vol. 3). Presented at the Balisage: The markup conference 2009, Montréal, Canada. https://doi.org/10.4242/BalisageVol3.Stegmann01

TEI Consortium. (2018). ODD. *TEI Wiki*. Retrieved March 22, 2022, from https://wiki.tei-c.org/index.php/ODD

TEI Consortium. (2021a). Feature structures. *TEI P5: Guidelines for electronic text encoding and interchange*. Version 4.3.0. Retrieved April 6, 2022, from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html

TEI Consortium. (2021b). Feature system declaration. In *TEI P5: Guidelines for electronic text encoding and interchange* (Vol. Version 4.3.0). Retrieved April 6, 2022, from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html#FD

Tertullian & Minucius Felix. (1931). *Apology. De Spectaculis. Minucius Felix: Octavius* (T. R. Glover, & G. H. Rendall, Trans.). Harvard University Press.

Triplette, S., Beshero-Bondar, E., & Bermúdez Sabel, H. (2018). A digital humanities approach to cultural translation in Robert Southey's Amadis of Gaul. *Journal of Translation Studies, 2*(1), 35–58.

Universal Dependencies. (2021). *UD Latin Perseus*. Universal Dependencies. Retrieved February 1, 2022, from https://github.com/UniversalDependencies/UD_Latin-Perseus

van der Auwera, J., & Plungian, V. A. (1998). Modality's semantic map. *Linguistic Typology, 2*, 79–124.

Véronis, J. (1998). A study of polysemy judgements and inter-annotator agreement. In *Advanced papers of the SENSEVAL workshop*, Sussex, UK.

WoPoss Project. (2022a). *Annotation schemes of the WoPoss Project*. XSLT, WoPoss. Retrieved April 13, 2022, from https://github.com/WoPoss-project/annotation-schemes

WoPoss Project. (2022b). *The WoPoss modality corpus*. WoPoss. Retrieved May 20, 2022, from https://github.com/WoPoss-project/WoPoss-corpus