DHQ: Digital Humanities Quarterly
Preview
2024
Volume 18 Number 1

An Annotated Multilingual Dataset to Study Modality in the Gospels

Helena Bermúdez-Sabel <helena_dot_bermudez_at_jinntec_dot_de>, University of Neuchâtel https://orcid.org/0000-0002-8627-1367

Francesca Dell'Oro <delloro_dot_fr_at_gmail_dot_com>, University of Neuchâtel; Swiss National Science Foundation https://orcid.org/0000-0002-8343-356X

Abstract

This paper presents a number of resources for examining the expression of modality in the Gospels. The main resource is an XML-TEI dataset that contains the linguistic annotation of a predefined list of potentially modal markers in both Ancient Greek and Latin. When one of these markers conveys a modal meaning, each constituent of the modal passage (i.e., the marker, its scope, and the modal relation between them) is annotated with a great level of detail through several linguistic features. One of the original features of our dataset is the implementation of a cross-referencing system that enables the alignment of the potentially modal markers of both languages. To facilitate the exploitation of our data by those unfamiliar with XML technologies, we also provide summary tables with the most relevant features of the annotation. In addition, a program written in Apache Ant allows any user to generate the summary sheets and to align modal passages in both Ancient Greek and Latin with any other language available in the *Multilingual Bible Parallel Corpus* [Christodouloupoulos and Steedman 2015]. This contribution presents the details of the semantic annotation and its formalization, and how our resources may be exploited within semantics and translation studies. In addition, the encoding strategies implemented are relevant for other projects dealing with the combination of multiple layers of (linguistic) annotation and/or tackling the development of parallel corpora.

1 Introduction

1 Parallel corpora are valuable resources for many linguistic fields — such as semantics, translation studies, and foreign language teaching — as well as for the development of natural language processing (NLP) tasks. Moreover, in instances where languages share a certain degree of similarity, semantic annotation carried out on the source text can often be “transferred” to the target languages with minor adjustments. This holds in particular when the translations into a natural variety of the target language tend to be literal.[1] In this paper we present a new resource for studying modality — a notoriously challenging field of semantics — by using the Latin translation of the Gospels by Jerome of Stridon (c. 347–420) and its source text in Ancient Greek. It consists of a richly annotated dataset of modal passages in both languages. The dataset allows for precise answers to specific research questions, such as whether modal markers with the same modal meaning have been consistently translated by the same modal markers throughout the four Gospels (cf. Section 5). As translations of the Bible usually keep the same subdivision into books, chapters, and verses (with small variations) and the verses are numbered, our annotation can easily be reused for analyzing modality in other languages, in particular when languages are similar from a structural point of view. Therefore, we have also set up tools to align the modal passages in both Ancient Greek and Latin with any other language available in the *Multilingual Bible Parallel Corpus* [Christodouloupoulos and Steedman 2015].

2 The four Ancient Greek Gospels were composed at different times during the second half of the first century CE (cf. [Schröter 2010, 277–278] for Mark's Gospel; [Duling 2010, 298] for Matthew's Gospel; [Thompson 2010, 331] for Luke's Gospel; [Painter 2010, 345] for John's Gospel). As the original texts are lost, we need to rely on the manuscript tradition and on philological work. We used the best available online editions, the 28th edition of the *Novum Testamentum*

Graece (Nestle-Aland) for Ancient Greek and the 5th edition of the *Biblia Sacra Vulgata* (Weber-Gryson) for Latin, though there were use restrictions. It is important to stress, however, that we do not have the exact source text from which Jerome translated the Gospels, nor do we know exactly on which of the previous Latin translations he based his revision work. Though we define the Ancient Greek text as the source text here, it must be specified that many translations are not based directly on the Ancient Greek text, but on other translations from Ancient Greek. For example, especially in the past, the Latin translation of the Bible by Jerome, the *Vulgate*, has often been used as the source text. In fact, the first Italian translation of the Bible, the *Malermi Bible* (1471), was based on the *Vulgate*.

After having introduced the notion of modality and briefly described the annotation of the modal passages (Section 2), we outline the workflow (Section 3) and the annotation scheme (Section 4). Then we show the results of our annotation of both the Latin text and the Ancient Greek text. In doing so, we demonstrate how and to what extent annotation created in one language can be used as a basis to annotate modal passages in another language. We also showcase the types of queries that can be carried out thanks to our fine-grained annotation (Section 5), as well as the limitations of our dataset (Section 6). We conclude by suggesting directions for future research (Section 7). 3

2 Modality and Its Annotation

The Ancient Greek/Latin dataset of the Gospels and its annotation was set up following 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). The project's main aim is to study the evolution of modality in the Latin language (3rd c. BCE-7th c. CE) through setting up an annotated diachronic corpus. The corpus is annotated automatically with regard to tokenization, lemmatization, PoS-tagging, morphological analysis, and syntactic dependencies. It is also annotated manually with regard to the semantics of the modal passages (see sections 3 and 4). The *Vulgate* has been selected to be part of the WoPoss corpus because of its cultural and linguistic influence over the centuries beyond the Latin language. However, this choice has forced us to also take the Ancient Greek source text into account. The sub-project described here does not focus on diachrony, but on the relationship between the Ancient Greek Gospels and the Vulgate Gospels. 4

In simplified terms, we define *modality* as the expression of necessity, possibility, and volition [van der Auwera and Plungian 1998]. We follow a basic division of the modal domain into three main types: dynamic, deontic, and epistemic modality. The first refers to possibilities and necessities internal to the main participant in the state of affairs (henceforth SoA) or inherent in the SoA. The second type refers not only to permissions, obligations, and similar notions, but also to the stance of the speaker towards the moral acceptability of an SoA. Finally, the third type refers to the stance of the speaker towards a propositional content in terms of likelihood. Each type is subdivided into subtypes, as outlined in the *WoPoss Guidelines for the Annotation of Modality* [Dell'Oro 2023a]. 5

In order to study its synchronic uses and its evolution over time, guidelines for a fine-grained semantic annotation of the relevant passages (i.e., the passages featuring modal markers) were devised [Dell'Oro 2023a]. It is worth mentioning that for Ancient Greek and Latin we focused on lexical and morphological markers of modality, not on verbal moods. The choice could have been different for other languages, but this issue is beyond the scope of this paper. With reference to Ancient Greek and Latin, we annotated modal verbs, adjectives, nouns, adverbs, and some modal suffixes, such as Ancient Greek -*tos* and Latin -*bilis*, expressing possibility. 6

The annotation of a modal passage starts with the identification of the modal marker (e.g., Ancient Greek *dúnamai* “can”), its syntactic scope, and the type of modality, or “modal relation”, as illustrated by the following example.[2] 7

(1) MT 7.18: 8

ou	*dúnatai*	*déndron*	*agathòn*	*karpoùs*	*poneroùs*	*poieîn* [...]
not	can.3SG	tree	good	fruit.PL	bad.PL	do.INF

Table 1. “A good tree cannot bring forth evil fruit [...]” (English translation from *Multilingual Bible Parallel Corpus*).

9 In this case, *dúnatai* “can.3sg” is the modal marker, its scope is *karpoùs poneroùs poieîn* “bring forth evil fruit”, and the modal type is dynamic situational possibility (i.e., that type of possibility referring to a situation, not to a specific participant in the SoA).[3] For both the marker and its scope, the annotator specifies whether there is a negation scoping over them (negative/positive polarity)[4] and identifies the sentence type as either non-interrogative or interrogative. The scope of the marker is also annotated with respect to the semantics of the featured event (i.e., the SoA). The annotator describes whether or not the event is dynamic (i.e., it involves a change) and whether or not there is control (i.e., someone or something has some control over the unfolding of the event). The annotator identifies the main participant in the SoA and specifies whether they are animate or inanimate. Only in the case of passive verbs does the annotator indicate whether there is also an undergoer (annotated with the label “patient” by default). As modal markers are often ambiguous, the annotator can provide more than one annotation for the same passage or feature. For examples of annotated modal passages, see [Dell'Oro, Bermúdez Sabel, and Marongiu 2020].

10 Though our list of modal markers covers all relevant parts of speech (verbs, adjectives, nouns, adverbs, and even morphological suffixes), it is worth specifying that other words could have been considered to be modal markers in other frameworks. It must be stressed that such a predefined list was not prepared for Ancient Greek. We first looked for the correspondences with the annotated Latin modal passages and annotated those passages. This provided us with a list of potentially modal markers in Ancient Greek as per our theoretical framework. We then searched for the occurrences of these Ancient Greek markers throughout the whole text, and we annotated the missing modal passages.

3 Workflow

11 Plain text versions of the Gospels in Latin and Ancient Greek were retrieved from the resource https://www.academic-bible.com/ [Deutsche Bibelgesellschaft (n.d.)], maintained by the German Bible Society, which provides free access to online editions of the Bible in different languages and which gave us their agreement. This source was selected for the high philological quality of the editions, as explained above. These files were then automatically annotated using the NLP library for Python Stanza[5], developed by the StanfordNLP research group [Qi 2020]. The Latin version was annotated with the model trained with the Perseus treebank [Universal Dependencies 2021b] and the Ancient Greek Gospels were annotated by implementing the model trained with the PROIEL treebank [Universal Dependencies 2021b]. The result was CONLL-U files that include the annotation of the lemmas, part-of-speech categorization, morphological features, and syntactic dependencies, as noted in Section 2. The linguistically annotated files were uploaded to the annotation platform INCEpTION[6] [Klie et al. 2018] in order to proceed with the semantic manual annotation as explained in Section 2. In addition, the automatic linguistic annotation was manually corrected, but only with respect to the tokens inside a modal passage.

12 An annotator carried out the linguistic annotation of a Gospel in each of the two languages. The Latin version was annotated according to the annotation guidelines of the WoPoss project [Dell'Oro 2023a], and then the annotator carried out the linguistic annotation of the equivalent Greek passages. The annotation was then reviewed by the PI of the project. Missing modal passages in the Greek text were retrieved (as explained in Section 2) and annotated.

13 After the revision process for the manual annotations was over, the annotated texts were exported to the UIMA CAS XMI format.[7] The files were then transformed into XML-TEI using a specifically tailored XSLT transformation developed by us. Typographical conventions were transformed into the semantically pertinent TEI elements. For example, square brackets used in the source edition to indicate superfluous contents, likely due to interpolations, were enclosed in the TEI element `<surplus>`.[8] During the manual annotation process, contributors could leave different kinds of comments in a field labeled “note”. These comments might concern tokenization issues (as the annotation platform does not allow

modification of the textual content of each token) or the annotation itself (e.g., implicit linguistic contents that are crucial to the understanding of the modal passage). The contents of this field were manually evaluated one by one and the required modifications were implemented.

14 The chapter structure and verse segmentation of the Greek and Latin versions are not exactly the same, but the differences are not numerous. This is why we were able to implement a semiautomatic alignment of the modal markers by using their location as the main point of reference. This alignment consisted of creating a correspondence between markers in both languages by a system of unique identifiers and URIs.

15 After carrying out the alignment of markers between Ancient Greek and Latin, there were some cases of non-aligned elements in one of the two texts. This means that for one of the languages no marker was annotated at that same location (verse number). This lack of a counterpart can be due to either 1) the nonexistence of that verse in the other version due to the differences in the Gospels tradition for that language (see Section 1); or, more frequently, 2) an inexact correspondence between the two languages.

16 We then performed an automatic identification of the verses using the same system as in the *Multilingual Bible Parallel Corpus Project*. We manually corrected the errors caused by differences in chapter division and verse segmentation between our reference edition and the *Multilingual Bible Parallel Corpus*. Our source texts are not open access, and we therefore have publication constraints. We have reached an agreement with the German Bible Society to publish only annotated verses, together with no more than the three preceding verses and the three following verses. It is worth stressing that for this reason the verses which were not pertinent were deleted before publication. Table 1 presents the percentage of each work (using the number of words as reference unit) made available in our dataset.

	Ancient Greek (GRC)		**Latin (LA)**	
	Published	**Semantically Annotated**	**Published**	**Semantically Annotated**
Gospel of Matthew	74.91%	3.95%	71.52%	3.93%
Gospel of Mark	67.02%	4.61%	64.86%	4.53%
Gospel of Luke	75.55%	4.27%	72.05%	4.23%
Gospel of John	66.11%	4.37%	68.24%	4.55%

Table 2. Published contents and presence of semantic annotation (% of words of the complete work).

4 Description of the Dataset

17 The dataset consists of an XML-TEI file for each Gospel and each language containing the linguistic annotation of modal passages (and potentially modal passages that have been analysed as non-modal). We also provide two summary tables with the most relevant features of the semantic annotation and two processing tools: one to generate the summary tables and one to align the semantic analysis with verses in any other language, if the source file is downloaded from the *Multilingual Bible Parallel Corpus* (or if it follows the same identification system for the verses). Further details about these materials are given at the end of this section. The work is made available through a GitHub and a Zenodo repository [Dell'Oro and Bermúdez Sabel 2023].

18 A customization of the TEI schema was developed in order to account for the issues specific to the project. This customization was formalized using ODD, “a TEI XML-conformant specification format that allows one to customize TEI P5 in a literate programming fashion” [TEI Consortium 2018]. The ODD file and its resulting Relax-NG schema (with embedded schematron rules) are freely available [WoPoss Project 2022].

19 The TEI P5 Guidelines present different methods of annotating linguistic information, but they can be summed up in two ways: 1) inline annotation, by using structural elements like `<s>` (sentence), `<cl>` (clause), or `<w>` (word), and attributes such as `@lemma` (lemma) and `@pos` (part of speech) to linguistically describe each word; and 2) stand-off annotation, generally formalized by implementing the Feature Structures module [TEI Consortium 2022a]. We have

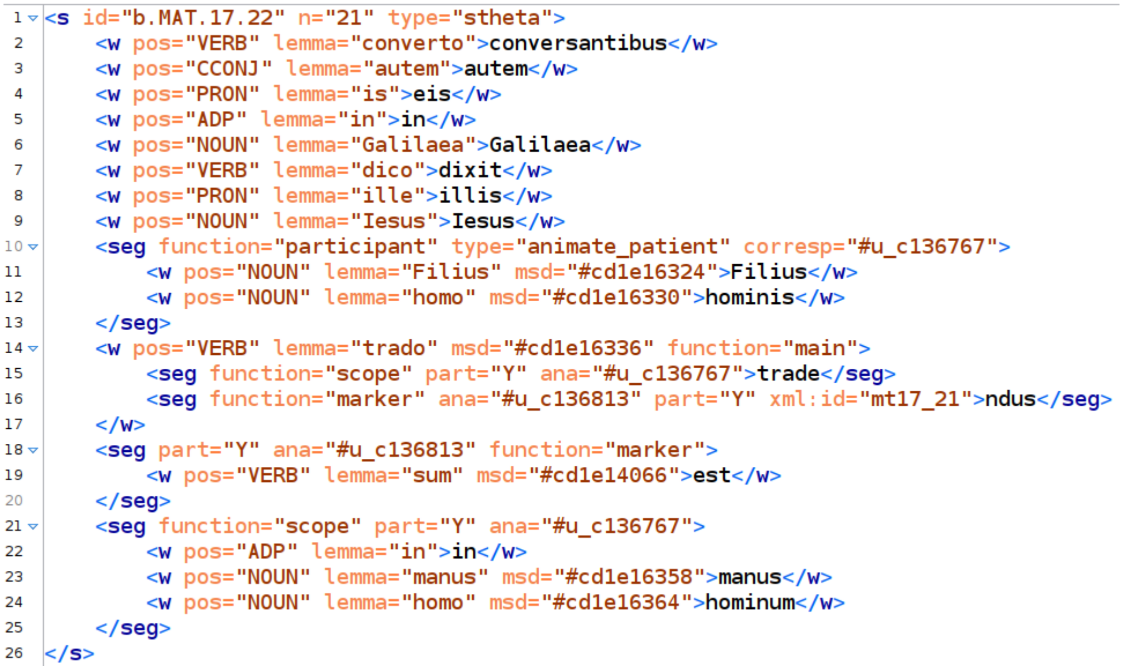
implemented a mixed approach, as is described below.

The <teiHeader> (TEI header) contains the main metadata of each file. Besides the reference to the source text, it contains the names of the annotators and the reviewer of the linguistic annotation of modal passages. In addition, it contains the element <extent> (extent) to describe the total number of words that the text contained before we deleted the contents we were not allowed to publish. Each chapter is contained in a <div> (text division) element. 20

The Latin edition of the Gospels does not contain any punctuation marks, so the results of the automatic sentence segmentation were erroneous. In the Greek text, certain punctuation marks generated errors as well (for example, the raised dot was never properly tokenized). Considering these issues, we decided to enclose each verse in an <s> element, even though the exact correspondence between a sentence and a verse was not always one-to-one. An @n (number / label) attribute allows us to know the verse number[9] according to the source edition (see line 1 in Figure 1). In addition, an @id attribute implements the same identification system as the *Multilingual Bible Parallel Corpus*, which enables the alignment between the two projects (see line 1 in Figure 1). 21

Each word is enclosed in a <w> element and each punctuation mark in a <pc> (punctuation character) element. Every word has a @lemma and a @pos attribute. The annotation of words that belong to modal passages has been manually corrected, but in the other case, the contents of @lemma and @pos are the uncorrected results of automatic annotation. In addition to those two attributes, words inside modal passages contain an @msd (morpho-syntactic description) attribute that points to a feature structure (<fs> element) with a complete morpho-syntactic description (e.g., gender, number, tense, mood; see line 14 in Figure 1, with resolved key in Figure 2). 22

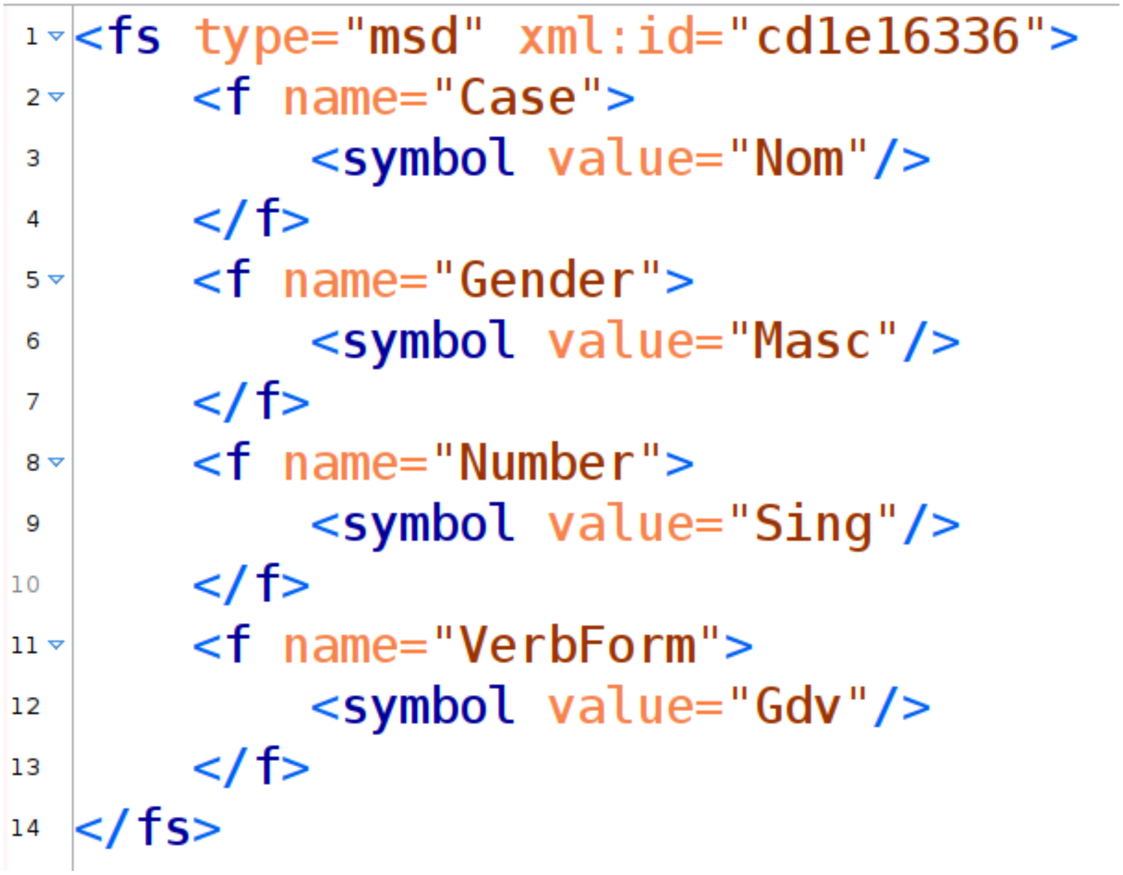
Every component of a modal passage is enclosed in the element <seg> (arbitrary segment; see lines 10-13 in Figure 1). An attribute @function (function) describes the role of the segment with four possible values: modal marker, its scope, its negation (if any), or participant in the state of affairs. The segments concerning the participants and the negation elements have a @corresp (corresponds) attribute that points to the scope of the modal marker, in the case of the participant (see line 10 in Figure 1), and to the modal marker, or, in specific cases, to its scope, in the case of negation.[10] The <seg> elements concerning potentially modal markers and their scope (in the case of modal markers) contain an @ana (analysis) attribute that points to an <fs> element with a complete description according to the features explained in Section 2. The following aspects are described: whether the potential marker is indeed modal and pertinent and, if that is the case, its polarity and type of utterance. In the case of the scope of a modal marker, besides specifying its polarity and type of utterance, the state of affairs is described in terms of dynamicity, control, and type of participant (whether there is a participant or not, and, if there is, whether this is explicit or implicit and then its type). Markers and their scopes may be discontinuous. In these cases, the @ana attribute points to the same description, and a @part attribute with the value “Y” (yes) indicates that the element is fragmented. In Figure 1, we can see a discontinuous scope (*trade-... in manus hominum*; lines 15, 21-25) and a marker annotated as discontinuous to avoid overlap. The marker is *-ndus est*, and due to the word boundary of *tradendus*, we cannot annotate it with only one <seg> element. 23



```xml
<s id="b.MAT.17.22" n="21" type="stheta">
    <w pos="VERB" lemma="converto">conversantibus</w>
    <w pos="CCONJ" lemma="autem">autem</w>
    <w pos="PRON" lemma="is">eis</w>
    <w pos="ADP" lemma="in">in</w>
    <w pos="NOUN" lemma="Galilaea">Galilaea</w>
    <w pos="VERB" lemma="dico">dixit</w>
    <w pos="PRON" lemma="ille">illis</w>
    <w pos="NOUN" lemma="Iesus">Iesus</w>
    <seg function="participant" type="animate_patient" corresp="#u_c136767">
        <w pos="NOUN" lemma="Filius" msd="#cd1e16324">Filius</w>
        <w pos="NOUN" lemma="homo" msd="#cd1e16330">hominis</w>
    </seg>
    <w pos="VERB" lemma="trado" msd="#cd1e16336" function="main">
        <seg function="scope" part="Y" ana="#u_c136767">trade</seg>
        <seg function="marker" ana="#u_c136813" part="Y" xml:id="mt17_21">ndus</seg>
    </w>
    <seg part="Y" ana="#u_c136813" function="marker">
        <w pos="VERB" lemma="sum" msd="#cd1e14066">est</w>
    </seg>
    <seg function="scope" part="Y" ana="#u_c136767">
        <w pos="ADP" lemma="in">in</w>
        <w pos="NOUN" lemma="manus" msd="#cd1e16358">manus</w>
        <w pos="NOUN" lemma="homo" msd="#cd1e16364">hominum</w>
    </seg>
</s>
```

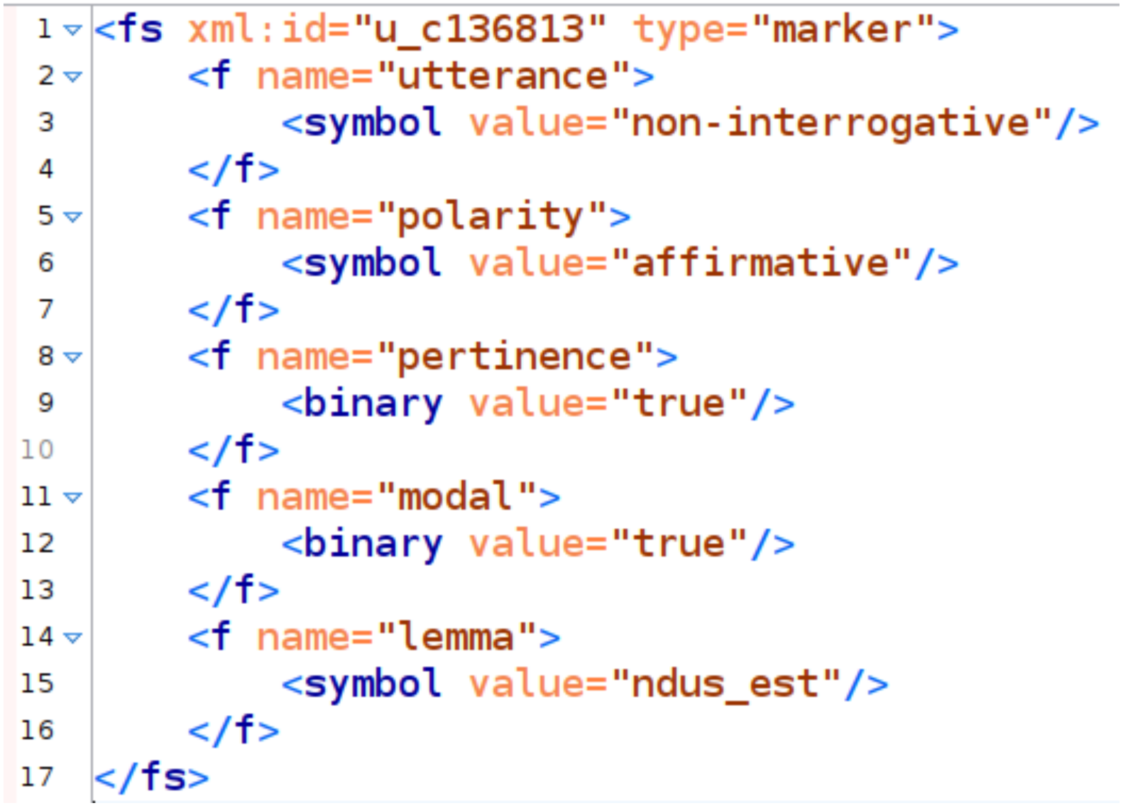
Figure 1. Inline annotation of a verse in the Latin text (Matthew 17.22).

At the end of each file, there is the list of all the `<fs>` elements with four different types: morphological descriptions, markers, scopes, and modal relations. The first three have already been tackled in the above description because there are internal references to them from the text (see figures 2 and 3 for examples). The modal relations contain the features that describe the type of modality. Besides describing a relation to a major type of modality (i.e., deontic, dynamic, or epistemic), it is mandatory to indicate the identifier of both the modal marker and its scope. There is an `<fs>` element for each modal relation (see Figure 4 for an example). In the case of a modal passage where multiple modal readings are possible, there is an `<fs>` element for each different interpretation. 24



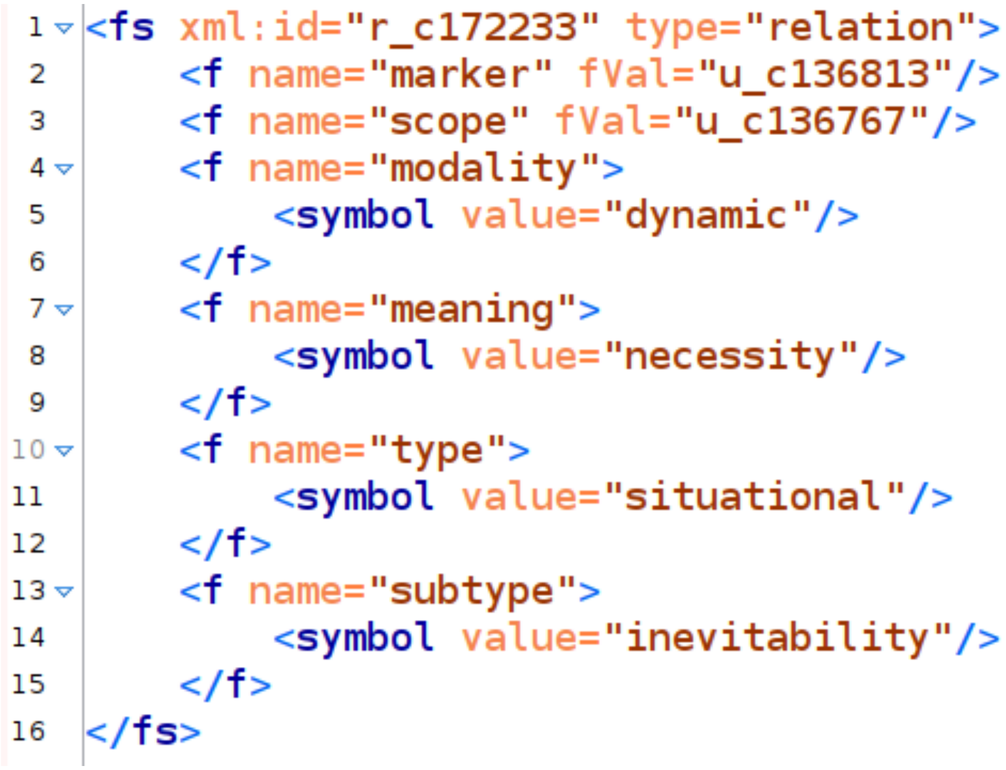
```
<fs type="msd" xml:id="cd1e16336">
    <f name="Case">
        <symbol value="Nom"/>
    </f>
    <f name="Gender">
        <symbol value="Masc"/>
    </f>
    <f name="Number">
        <symbol value="Sing"/>
    </f>
    <f name="VerbForm">
        <symbol value="Gdv"/>
    </f>
</fs>
```

Figure 2. Feature structure that describes the morpho-syntactic description of the word *tradendus* (Figure 1, lines 14-17).



```
<fs xml:id="u_c136813" type="marker">
    <f name="utterance">
        <symbol value="non-interrogative"/>
    </f>
    <f name="polarity">
        <symbol value="affirmative"/>
    </f>
    <f name="pertinence">
        <binary value="true"/>
    </f>
    <f name="modal">
        <binary value="true"/>
    </f>
    <f name="lemma">
        <symbol value="ndus_est"/>
    </f>
</fs>
```

Figure 3. Feature structure that describes the modal marker from Figure 1, lines 16-20.



```
<fs xml:id="r_c172233" type="relation">
    <f name="marker" fVal="u_c136813"/>
    <f name="scope" fVal="u_c136767"/>
    <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>
</fs>
```

Figure 4. Feature structure that describes the modal reading of the verse presented in Figure 1.

25 The model of the different feature structures is described in a Feature System Declaration (FSD) [TEI Consortium 2022b]. This declaration presents the different features and values that are possible depending on the type of feature structure, together with specific constraints [WoPoss Project 2022]. It has helped us model the linguistic annotation within a very restrictive schema. For example, we can declare that, if a modal relation is of the "epistemic" type, then it must include the "degree" feature defined with a controlled vocabulary. The FSD is processed using XSLT to create a series of Schematron constraints that are then added to the ODD file.[11]

26 The relationship between the Ancient Greek and Latin texts is established at the level of the modal marker (`<seg>` element). An identifier is given to the marker in one of the languages (`@xml:id`) and then the corresponding reference is added to its counterpart in the other language through the attribute `@synch`.[12] In the few cases in which there is no correspondence, then the `@synch` attribute is added to the `<s>` element, so we can easily retrieve the complete verse for a close reading of these cases.

27 The Gospels of Luke, Matthew, and Mark are known as the synoptic Gospels because they contain several of the same stories, sometimes transmitted with similar or even identical wording. Kurt Aland enumerates the synoptic passages, and using this work as a reference, we provided each group of synoptic passages with an identifier [Aland 1985]. In this way, the affected verses are recognized across the different Gospels thanks to the attribute `@type` that points to these relations of parallelism (see line 1 of Figure 1).

28 The folder "summary" contains two tabular sheets. One of them, "modal_passages.tsv", contains all the modal passages in each Gospel, aligning the Ancient Greek and the Latin text. It includes 1) the modal marker in both languages (when available); 2) the contents of the verse that contains the modal markers in both languages; 3) features to describe the modal marker (whether it is implicit, the type of utterance, and its polarity); 4) features to describe the scope of the marker (the type of utterance and its polarity, whether there is an SoA and, if there is, its description in terms of dynamicity and control); 5) features to describe the modal reading (whether it is ambiguous and thus receives more than one modal reading, the modal type and its different subcategorizations according to the theoretical framework, and whether it has a rhetoric or pragmatic use); 6) the identification of the synoptic passage; 7) the ID of the modal marker in the XML file; 8) the Gospel in which is located; and 9) the number of the verse following the *Multilingual*

Bible Parallel Corpus. The other file, “potentially_modal_markers.tsv” presents the list of all potentially modal markers, with both languages aligned, including, as in the previously described file, features 1, 2, 7, 8, and 9. Concerning the description of potentially modal markers, the file also contains two pieces of information: 1) whether the marker is pertinent or not (that is, whether it is a modal marker); and 2) whether the marker is implicit.

29 The XQuery code that produces both tabular sheets is available in the folder scripts, and it can be run through an Apache Ant program by running the command `ant analysis`. In addition, we also provide an option to create tabular sheets aligning our semantic analysis of Greek and Latin with any other language. The source files can be downloaded from the *Multilingual Parallel Bible Project*, and they need to be included in the folder “to-be-aligned”. Then the program is launched by running the command `ant align`. The result is a table with a simplification of “modal_passages.tsv” in which the verses in Ancient Greek and Latin are presented along with the verse that contains the modal marker in any of the selected languages.

5 Data Exploration

30 One of the studies that can be carried out thanks to our corpus is the examination of the translation of potentially modal markers from Ancient Greek into Latin. For instance, we can retrieve one-to-one equivalents in order to create simple glossaries (Table 2) or, for example, focus on the markers with a modal meaning that have more than one different translation in Latin (Table 3) and thus explore lexical diversity within the modal context.

Lemma of the Potentially Modal Marker (GRC)	**Lemma of the Translation (LA)**
adunatéō	*impossibilis est*
adúnatos	*impossibilis est*
adúnatos eimí	*impossibilis est*
anéndektos eimí	*impossibilis est*
anánkēn ékhō	*necesse habeo*
ára	*forte*
boúlomai	*volo*
dúnamai	*possum*
dunástēs	*possum*
emós eimí	*meus est*
ékson eimí	*licet*
exousía	*potestas*
thélēma	*voluntas*
ísōs	*forsitan*
pepeisménos	*certus*
prépō	*decet*
-sómenos	*-turus*
-téos	*-ndus*

Table 3. Potentially modal markers in Ancient Greek with a single translation in Latin that matches the predefined list of potentially modal markers for Latin.

Modal Marker in GRC	**LA Translation**
ádikos	*iniquus*
	Not preselected[13]
anánkē	*necesse est*

	Not preselected
deî	*debeo*
	necesse est
	oportet
dúnamis	*-bilis*
	potestas
	Not preselected
dunatós eimí	*possibilis est*
	possum
	volo
éxesti	*licet*
	Not preselected
exousían ékhō	*potestas*
	potestatem habeo
ékhō	*habeo* + inf.
	possum
	Not preselected
thélō	*nolo*
	volo
	Not preselected
iskhúō	*possum*
	valeo
	Not preselected
méllō	*-ndus est*
	-turus
	-turus est
	Not preselected
mḗpote	*forte*
	Not preselected
opheílō	*debeo*
	Not preselected
khreían ékhō	*debeo*
	necesse habeo
	opus est
	Not preselected
-tos	*-bilis*
	possum
	Not preselected

Table 4. Potentially modal markers in Ancient Greek with more than one translation in Latin.

31 As we saw above, there is not always a one-to-one correspondence between the source text and the target text. There can be various reasons for this (see Section 3). In some cases, we find a marker in Latin, where this is implicit in Greek. In the following example we see that Ancient Greek does not use any verb to express the idea that new wine must be

put into new bottles, while Latin and English use modal verbs (*debeo* and *must*, respectively) and passive infinitives (*mitti* and *be put*, respectively).

(2) Mark 2.22 32

all'	*oînon*	*néon*	*eis*	*askoùs*	*kainoús*.		
sed	*vinum*	*novum*	*in*	*utres*	*novos*	*mitti*	*debet*
But	wine	new	in	bottle.PL	new.PL	be.put	must

Table 5. "But new wine must be put into new bottles" (English translation from *Multilingual Bible Parallel Corpus*).

33 Note, moreover, that the syntax of both Ancient Greek and Latin is more flexible than English syntax, so the wordings in Greek and Latin are already perfectly aligned.

34 As suggested at the beginning of this paper, the dataset allows for precise answers to complex questions because it is possible to combine lexical and semantic information. With regard to the question of whether modal markers expressing the same type of modality are consistently translated by the same markers, we find that there is a tendency to use the same markers, though not in a rigid way. See Table 4 for an illustration of the situation for the subtype permission, and Table 5, which illustrates the situation for the subtype obligation. Both subtypes belong to deontic modality — authority.

	licet	**Not Preselected**	***oportet***	***possum***	**Total**
deî			1		1
dúnamai				1	1
éxesti	24	1			25
ékson eimí	1				1
iskhúō				1	1
Total	25	1	1	2	29

Table 6. Correspondences between the Ancient Greek source markers and the Latin target markers with regard to the expression of "permission" (deontic modality — authority).

	debeo	***necesse est***	***oportet***	**Total**
deî		1	6	7
opheílō	3			3
Total	3	1	6	10

Table 7. Correspondences between the Ancient Greek source markers and the Latin target markers with regard to the expression of "obligation" (deontic modality — authority).

35 Permission is expressed in the Ancient Greek Gospels by five markers. The translation tends to take into account the different source markers, as only two Ancient Greek modal markers (*dúnamai* and *iskhúō*) are rendered by one Latin modal marker (*possum*). It is worth noticing that *éxesti* and the etymologically related construction *ékson eimí* are consistently rendered by *licet*. The situation for obligation is similar, though in one passage the Latin translation has introduced a third marker. It is beyond the scope of this paper to discuss the reasons for such slight discrepancies. The point to be made here is that the dataset allows the user to highlight specific tendencies of the translation.

36 The examples above show the type of information that can be retrieved from our dataset, through which linguists may gain new knowledge about how modality is linguistically expressed in both languages. In addition to being a resource for scholars, our dataset can also be exploited in a learning context. As we saw in tables 1 and 2, the creation of bilingual glossaries is very straightforward thanks to the marker-to-marker alignment. Furthermore, this type of vocabulary list

can directly be enriched not only with examples of these words in context but also with a modern language translation of the examples thanks to the alignment with the *Multilingual Bible Parallel Corpus*. It is also possible to design exercises to improve the reading skills of the learners.

The semantic annotation may also be exploited as a means to improve learners' linguistic knowledge. The semantic annotation can easily be explored with the aid of translations of the modal passages in modern languages in order to easily interpret the modal meaning of each passage. Moreover, students can examine different semantic phenomena, such as semantic ambiguity (see Figure 5) or polysemy concerning modal markers. For example, learners can compare the occurrences of modal markers that present different modal readings depending on the context, evaluating for each marker which meanings are more or less frequent. Figure 6 shows markers that convey different types of modality, excluding passages that received a double annotation due to ambiguity (cf. also [Dell'Oro 2023b]).

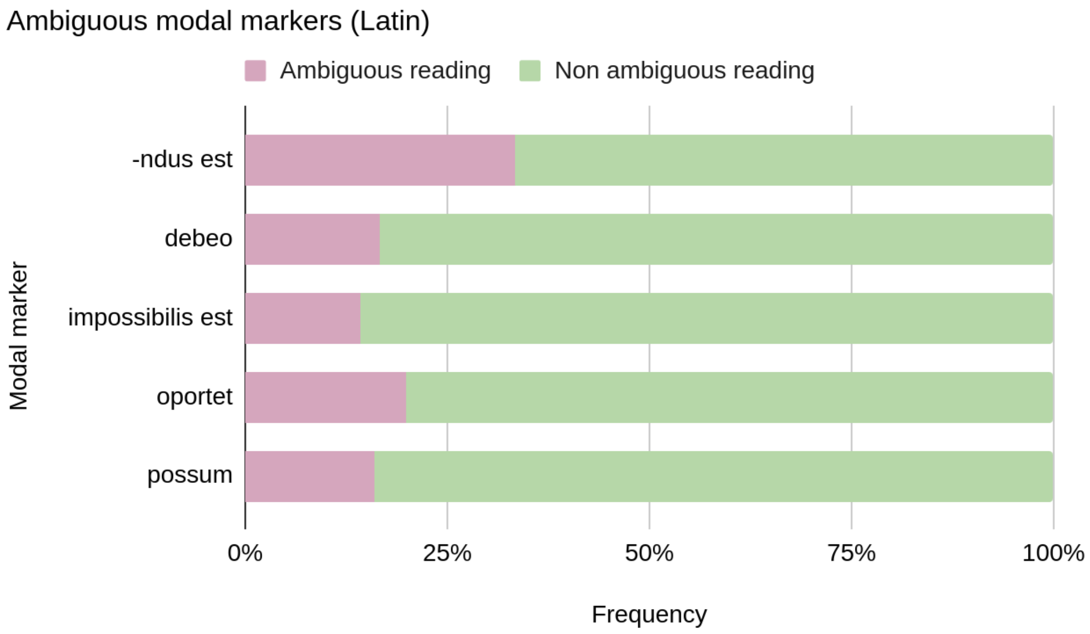


Figure 5. Modal markers in Latin with ambiguous readings and frequency of the phenomenon.

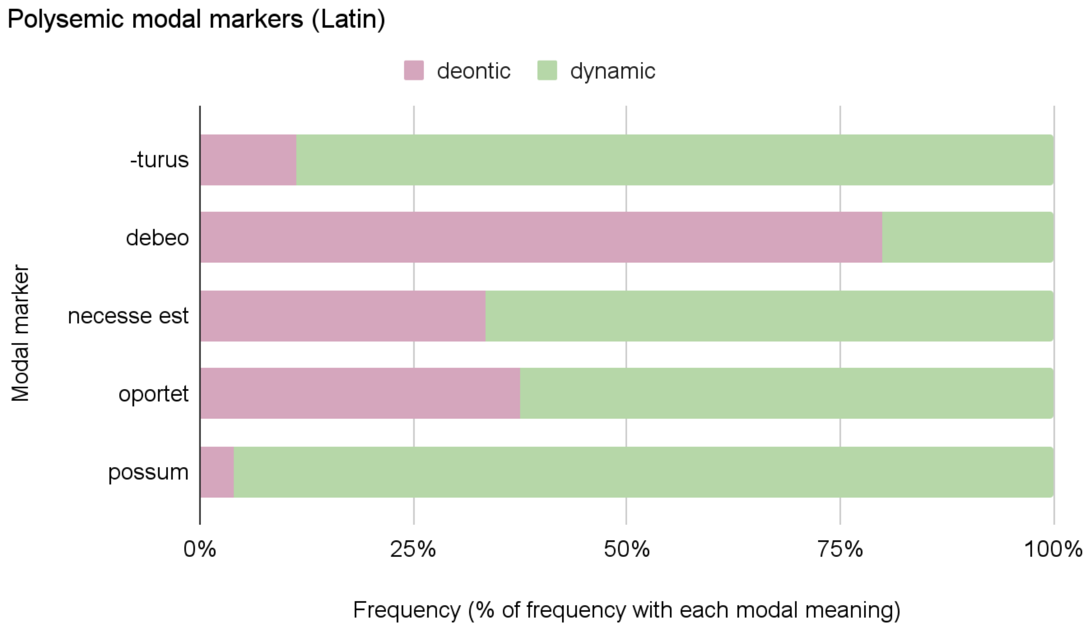


Figure 6. Modal markers in Latin that express more than one type of modality, excluding ambiguous readings.

Our dataset can also be used to explore modality in a typological perspective thanks to the possibility to align the Ancient Greek and Latin texts with the translations contained in the *Multilingual Bible Parallel Corpus* [Christodouloupoulos and Steedman 2015], which includes languages belonging to diverse linguistic families. Therefore, our fine-grained annotation of the modal passages will provide researchers with new insights on the contextual uses of modal markers from a cross-linguistic perspective. For the use of parallel corpora for typological research, see [Cysouw and Wälchli 2007]. 38

6 Limitations

As mentioned in Section 5, each file contains a reference to the complete word count of each Gospel. This allows us to obtain a number of statistical measures, such as the relative frequencies of each marker. However, in order to perform other type of measures, such as the specificity of a collocation, we would need to query one of the resources that provide a lemmatized version of the Gospels, such as *Bible Hub*, although there could be textual differences between editions [Bible Hub 2004]. 39

Another limitation of our annotated dataset is the fact that we only annotate a precise list of markers (for the complete list in Latin, see [Dell'Oro 2023a]; for the list in Ancient Greek, see tables 2 and 3), as outlined above (Section 2). 40

7 Conclusion

The linguistically annotated dataset presented in this paper comprises the four Gospels in both Ancient Greek and Latin. One of the novelties of this project concerns the semantic annotation of modality, since the availability of corpora with this type of annotation is scarce.[14] The annotation pipeline, including the combination of automatic and manual annotation, together with the formalization of different layers of linguistic annotation, can be adapted by other projects within the field of linguistics. Besides the annotation of modal passages, another original aspect of this dataset is the formalization of one-to-one relations between both languages in regard to modal markers. The fine-grained linguistic annotation of this dataset opens up a number of research paths that can be followed, either to carry out monolingual analyses or to tackle contrastive linguistics research. 41

A closely related schema for the semantic annotation of modality has been successfully applied in the complete 42

WoPoss Modality Corpus, available at https://woposs.unine.ch. In addition, the strategies to connect semantic annotation across languages can be implemented in other parallel corpora to align specific linguistic descriptions.

43 Section 5 illustrates several possible applications of this dataset. The dataset is formalized in TEI, and our annotation schemas are well documented, thus increasing the usability of the dataset for the scientific community, as well as within a pedagogical context. To facilitate its exploitation, part of the analytical data is presented in tabular sheets. In addition, the internal references shared with the project *Multilingual Bible Parallel Corpus* enables a straightforward link to this other resource, which can be especially relevant for teaching purposes. A program is provided to facilitate the alignment between our semantic annotation in Ancient Greek and Latin and any of the documents from the *Multilingual Bible Parallel Corpus.*

Funding

44 This work was supported by the Swiss National Science Foundation [PP00P1_176778, PP00P1_214102 to FDO].

Acknowledgements

45 We would like to thank Thomas von Kaenel and Séverine Reymond for their contributions to the annotation task.

Author Contributions

46 This paper was written collaboratively. H. Bermúdez-Sabel is mainly responsible for sections 3, 4, 5 (learning applications), 6, and 7. F. Dell'Oro is mainly responsible for sections 1, 2, 5 (translation, linguistic typology), and the general scientific supervision. H. Bermúdez-Sabel contributed to this paper as a Swiss National Science Foundation post-doctoral researcher.

Notes

[1] It is worth specifying that literal translations can override some of the rules of the target language, giving rise to non-native-like uses. This case represents a special problem which is beyond the scope of this paper.

[2] Linguistic glosses follow the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf).

[3] The labelling of the modal types largely relies on the theoretical framework. We will not go into details here, but a comprehensive overview of modal types and subtypes is described in our annotation guidelines [Dell'Oro 2023a]. It is worth noting that the WoPoss annotation combines the annotation of types of modality with other features of the modal passages. As a result, the analysis does not rely entirely on the intuition of the annotator, but on a combination of different levels of analysis (see below in this section).

[4] For more details about the annotation of polarity, see Section 4 and, in particular, Note 10.

[5] This workflow follows the same main steps as any other text annotated in the framework of the WoPoss project (see [Dell'Oro, Bermúdez Sabel, and Marongiu 2020]), except for the alignment of potentially modal markers.

[6] For a description of the customization of this tool to implement the WoPoss annotation scheme, see [Bermúdez Sabel 2019].

[7] INCEpTION depends on an Unstructured Information Management Architecture (UIMA) engine. The INCEpTION export format we post-process is an XMI serialization form of UIMA Common Analysis Structure (UIMA CAS XMI).

[8] See https://tei-c.org/release/doc/tei-p5-doc/en/html/ref-surplus.html.

[9] This attribute is necessary because we cannot rely on calculated positions since our files do not contain the complete work due to the aforementioned publication restrictions.

[10] The negation particle whose scope is a modal marker is systematically annotated. In a situation where the scope of a modal marker is under the scope of a negation, the negation is annotated only if it is not part of the same syntactic unit as the scope. Let us consider the following passage: “Existimante autem populo et cogitantibus omnibus in cordibus suis de Iohanne ne forte ipse esset Christus” (Luke 3.15), which translates to “And as the people were in expectation, and all men mused in their hearts of John, whether he were the Christ, or not”

(translation provided by the *Multilingual Bible Parallel Corpus*). The subordinate clause that contains the modal marker, the adverb *forte* ("maybe, perhaps") and its scope, *ipse esset Christus*, is introduced by the negative conjunction *ne*. As this negative conjunction introduces the indirect question, it is linked to the scope of the modal marker. When the negation is part of the same syntactic unit as the scope of the marker, we describe the negative polarity of the scope (as a feature structure) without annotating the negation element itself.

[11] For the processing of FSD into Schematron, see [Bermúdez Sabel 2022].

[12] This use might force the semantics of the attribute `@synch`, but other projects have also used it to establish parallel relationships between works; see, as an example, [Triplette, Beshero-Bondar, and Bermúdez Sabel 2018]. The attribute `@corresp` could be a better candidate due to its lax semantics, but, as is explained in this section, it is already being used to establish internal relationships between a negation and its scope, as well as between the participant in a state of affairs and the scope of a modal marker.

[13] This means that the translation does not match a lemma contained in the predefined list of potentially modal markers.

[14] Although notable exceptions do exist, like the project *MODAL: Modèles de l'annotation de la modalité à l'oral*, led by Paola Pietrandrea [Ghia 2016].

Works Cited

Aland 1985 Aland, K. (1985) *Synopsis quattuor evangeliorum: Locis parallelis evangeliorum apocryphorum et patrum adhibitis*. Editio tertia decima revisa. Stuttgart, Germany: Deutsche Bibelgesellschaft.

Bermúdez Sabel 2019 Bermúdez Sabel, H. (2019) "Digital tools for semantic annotation: The WoPoss use case". *Bulletin de Linguistique et Des Sciences Du Langage*, 30, pp. 12–37.

Bermúdez Sabel 2022 Bermúdez Sabel, H. (2022) *FS-Validator*. Available at: https://github.com/HelenaSabel/FS-Validator (Accessed: 3 November 2022).

Bible Hub 2004 Bible Hub (2004) *Bible hub: Search, read, study the Bible in many languages*. Available at: https://biblehub.com/ (Accessed 28 April 2022).

Christodouloupoulos and Steedman 2015 Christodouloupoulos, C. and Steedman, M. (2015) "A massively parallel corpus: The Bible in 100 languages", *Language Resources and Evaluation*, 49(2), pp. 375–95. https://doi.org/10.1007/s10579-014-9287-y.

Cysouw and Wälchli 2007 Cysouw, M. and Wälchli, B. (2007) "Parallel texts: Using translational equivalents in linguistic typology". *Language Typology and Universals*, 60(2), pp. 95–99. https://doi.org/10.1524/stuf.2007.60.2.95.

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

Dell'Oro 2023b Dell'Oro, F. (2023) "Corpus parallèles et apprentissage des langues anciennes: Les Évangiles comme corpus multilingue pour apprendre le grec ancien et le latin (avec un focus sur la modalité)", in Zalesskaya, D. (ed.): *La traduction et son processus didactique*, Lausanne, Switzerland: Cahiers du CLSL 66, pp. 65–84.

Dell'Oro and Bermúdez Sabel 2023 Dell'Oro, F. and Bermúdez Sabel, H. (2023) *The WoPoss dataset on modality in the Gospels*. Available at: https://github.com/WoPoss-project/Gospels.

Dell'Oro, Bermúdez Sabel, and Marongiu 2020 Dell'Oro, F., Bermúdez Sabel, H., and Marongiu, P. (2019). "Implemented to be shared: The WoPoss annotation of semantic modality in a Latin diachronic corpus". *Proceedings of the DARIAH-CH workshop, 2019*. Neuchâtel, Switzerland, 5-6 December. https://doi.org/10.5281/zenodo.3739440.

Deutsche Bibelgesellschaft (n.d.) *Deutsche Bibelgesellschaft* (n.d.) Available at: https://www.academic-bible.com/ (Accessed: 13 January 2022).

Duling 2010 Duling, D.C. (2010) "The Gospel of Matthew", in Aune, D.E. (ed.) *The Blackwell companion to the New Testament*. Chichester, England: John Wiley & Sons, Ltd, pp. 296–318.

Ghia 2016 Ghia, E. et al. (2016) "A construction-centered approach to the annotation of modality". *Proceedings of the 12th joint ACL-ISO workshop on interoperable semantic annotation, ACL, 2016*. Portorož, Slovenia, 28 May. pp. 67–74. Available at: https://sigsem.uvt.nl/isa12/ISA12Proceedings.pdf.

Klie et al. 2018 Klie, J.-C. et al. (2018) "The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation". *Proceedings of the 27th international conference on computational linguistics, ACL, 2018*. Santa Fe, New Mexico, USA, 20-26 July. pp. 5–9. Available at: http://tubiblio.ulb.tu-darmstadt.de/106270/.

Nestle and Aland 2012 Nestle, E. and Aland, K. (eds.) (2012) *Novum Testamentum Graece*. 28th edn. Stuttgart, Germany: Deutsche Bibelgesellschaft.

Painter 2010 Painter, J. (2010) “Johannine literature: The Gospel and letters of John”, in Aune, D.E. (ed.) *The Blackwell companion to the New Testament*. Chichester, England: John Wiley & Sons, Ltd, pp. 344–72.

Qi 2020 Qi, P. et al. (2020) “Stanza: A Python natural language processing toolkit for many human languages”. *Proceedings of the 58th annual meeting of the association for computational linguistics, ACL, 2020*. Online, 6-10 July. Available at: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Schröter 2010 Schröter, J. (2010) “The Gospel of Mark”, in Aune, D.E. (ed.) *The Blackwell companion to the New Testament*. Chichester, England: John Wiley & Sons, Ltd, pp. 272–95. https://doi.org/10.1002/9781444318937.ch17.

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

TEI Consortium 2022a TEI Consortium (2022a) “Feature structures”, in *TEI P5: Guidelines for electronic text encoding and interchange*. Available at: https://tei-c.org/Vault/P5/4.5.0/doc/tei-p5-doc/en/html/FS.html (Accessed: 3 November 2022).

TEI Consortium 2022b TEI Consortium (2022b) “Feature system declaration”, in *TEI P5: Guidelines for electronic text encoding and interchange*. Available at: https://tei-c.org/Vault/P5/4.5.0/doc/tei-p5-doc/en/html/FS.html#FD (accessed 3 November 2022).

Thompson 2010 Thompson, R.P. (2010) “Luke–Acts: The gospel of Luke and the acts of the apostles”, in *The Blackwell companion to the New Testament*. Chichester, England: John Wiley & Sons, Ltd, pp. 319–43. https://doi.org/10.1002/9781444318937.ch19.

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

Universal Dependencies 2021a Universal Dependencies (2021a) *UD Latin Perseus*. Available at: https://github.com/UniversalDependencies/UD_Latin-Perseus (Accessed: 1 February 2022).

Universal Dependencies 2021b Universal Dependencies (2021b) “UD Latin PROIEL”. Available at: https://github.com/UniversalDependencies/UD_Latin-PROIEL (Accessed: 28 April 2022).

Weber and Gryson 2007 Weber, R. and Gryson, R. (eds.) (2007) *Biblia sacra vulgata*. 5th edn. Stuttgart, Germany: Deutsche Bibelgesellschaft.

WoPoss Project 2022 WoPoss Project (2022) *Annotation schemes of the WoPoss project*. Available at: https://github.com/WoPoss-project/annotation-schemes (Accessed: 13 April 2022).

van der Auwera and Plungian 1998 van der Auwera, J. and Plungian, V. A. (1998) “Modality's semantic map”. *Linguistic Typology*, 2(1), pp. 79–124.