Bulletin de l'Institut de linguistique

30

Université de Lausanne

ISSN 1023-134X

Digital Tools for Semantic Annotation: the WoPoss Use Case

Helena Bermúdez Sabel

&

The Semantics of Irish Determiner Phrases

Benjamin Storme

Université de Lausanne

Digital Tools for Semantic Annotation: the WoPoss Use Case – Helena Bermúdez Sabel

Abstract

This paper examines the use of annotation platforms to perform semantic annotation of textual contents. It focuses on a specific tool called INCEpTION. This review stems from a project that studies modality in Latin from a diachronic perspective; thus, the analysis emanates from the development of an annotation pipeline for this particular use case. I briefly overview the role of semantic annotation in the project so as to delve into the specific requirements of the annotation process and how a customized tool assists in this procedure. After justifying the selection of INCEpTION over other annotation environments, a description of the functionalities of the tool is presented. The paper continues with a discussion of the tool's customization that was undertaken in order to meet the requirements of the project. This part draws attention to how the annotation challenges were tackled. To conclude, a general reflection on the use of annotation platforms is presented.

1. Rationale

This paper is developed in the framework of the FNS project *A World of Possibilities. Modal pathways over an extra-long period of time: the diachrony of modality in the Latin language* (WoPoss), led by Francesca Dell'Oro and whose members are Paola Marongiu and the present author.[1] This project studies the evolution of modal meanings in Latin, analysing modality mainly from a

1 See <http://woposs.unil.ch> (accessed on 01/11/2019).

semantic perspective, although not exclusively. In the WoPoss project, modality is understood as the expression of possibility, necessity and probability Modal meanings are empirically elicited by annotating modal passages in a diachronic corpus.

As pointed out by Nissim et al. (2013), many projects dealing with the annotation of modality entail a mere classification task in which annotators assign modality values to pre-selected markers or expressions (Nissim et al. 2013, 8). With regard to WoPoss, we have developed a complex annotation scheme (Dell'Oro 2019) that dissects a modal expression into its different components, that is, the modal marker, its scope, the state of affairs and the modal relation between marker and scope. These units are later described using various linguistic features.

The theoretical framework of WoPoss is largely based on Nuyts (2016). With respect to the annotation scheme, it was influenced by the work developed under the umbrella of the project *Modal – Modèles de l'annotation de la modalité à l'oral* (Ghia et al. 2016).[2] We also drew inspiration from the annotation parameters used by Jan Nuyts in his projects on the diachrony of the Dutch modal verbs.[3]

The complexity of our schema requires, on the one hand, a tool able to formalize the intricacy of this multifaceted linguistic phenomenon. On the other hand, it demands an annotation environment that makes it possible for people with different profiles to work collaboratively. We need a space for the inexperienced annotators to learn and practice and the experienced ones to guide them.

2 The complete annotated corpus is available online (Pietrandrea et al. 2016). For more information about this project see <https://modal.msh-vdl.fr/> (accessed on 01/11/2019).
3 For a list of projects by Jan Nuyts concerning modality, please see the list available at <http://woposs.unil.ch/credits.php> (accessed on 01/11/2019).

The following section outlines the workflow of the project in order to contextualize the role of semantic annotations as part of the development (Section 2.1). In Section 2.2, I detail the specific requirements that an annotation platform should meet for the correct formalization of our annotation scheme. After analysing different annotation platforms and workflows (Section 3), the members of the WoPoss project concluded that INCEpTION[4] was the most suitable annotation tool for our necessities. A brief description of this platform will be introduced in Section 4 and this review will focus on the functionalities that make this tool different from other annotation platforms. Section 5 sketches the customization of the tool that was developed in order to make of this platform a functional resource for our project. This section will pay special attention to the challenges presented in section 2.2 explaining how the customization works around them. The paper will conclude with some general remarks about the use of tools for semantic annotation.

2. The semantic annotation of modality

2.1. Project workflow

WoPoss has a corpus-based approach for the study of modality. Diversity was used as a determining factor for the selection of the texts to be included: we aim at a representative corpus in terms of diachronic, diatopic, diastratic and diaphasic parameters. The corpus spans the period from the 3rd BCE to the 7th century CE, and besides the different textual types and genres, we also took into consideration the various sources of transmission of ancient texts.[5]

4 <https://inception-project.github.io> (accessed on 01/11/2019).
5 About the importance of including both documentary and literary texts for the study of ancient languages, see Dell'Oro (2015).

Firstly, the selected works are gathered by retrieving them from different online resources that contain the texts under a free licence.[6] These sources present the texts in different formats, so they are first converted into plain text. Pseudo-markup is added to the textual content in order to preserve the pertinent semantic information previously conveyed in the XML or HTML tags: this affects the chapter and book divisions, verse lines and foreign words, among other pieces of information that were contained in the source files.

Afterwards, these documents are automatically annotated using the StanfordNLP library for Python.[7] Thus, lemmas, part of speech categorization, morphological features, and syntactic dependencies are added. The resulting CONLL-U[8] files are uploaded to the annotation platform INCEpTION.

Through this platform, the annotators add the relevant semantic information. They read the complete text paying closer attention to any occurrence of the pre-selected modal markers (Dell'Oro 2019, 9-10). Every time a potential modal passage is found, they correct the automatic annotation (if needed) and then they annotate the passage according to the WoPoss scheme and guidelines. Inter-annotator agreement is frequently checked, especially when it involves less experienced annotators. Disagreements are discussed in order to evaluate whether the passage is ambiguous or if one of the annotators misinterpreted either the text or the (sub-)type of modality.

6 See <http://woposs.unil.ch/credits.php> (accessed on 13/11/2019) for a list of the digital libraries and sources employed.
7 <https://stanfordnlp.github.io/stanfordnlp> (accessed on 13/11/2019).
8 In the CONLL-U format, annotations are encoded in plain text files. Blank lines mark sentence boundaries, each line concerns the analysis of a word and each value of this analysis is separated by a single tab character. For a detailed explanation of this format see <https://universaldependencies.org/format.html> (accessed on 20/11/2019).

The revised annotated texts are then exported to the XMI format, one of the output formats available in INCEpTION, and one that is easy to transform into XML-TEI. Our annotated dataset will be preserved in TEI with the linguistic information encoded through stand-off annotation.

After this first transformation, a series of steps are implemented to cure the annotated documents and add more information.

The first one entails the automatic addition of linguistic features concerning the most ancient meaning of each modal marker. This meaning is elicited by reviewing and synthesizing lexicographical resources.[9]

The second step affects the pseudo mark-up that was added when the sources were first converted to plain text. These graphical conventions are then transformed into TEI elements. In a similar manner, miscellaneous information that was kept (unstructured) during the annotation process in a field labelled “note” is analysed and disambiguated, adding the pertinent XML elements when necessary.

The parameters that are relevant for the selection of the corpus – textual genre, type of transmission, chronology and origin of the author – are part of the metadata that will be added automatically using the Digital Humanities Toolkit (DHTK) (Picca and Egloff 2017).

At this point, the dataset is ready to be stored in a no-SQL database and to be published and exploited through a user-friendly interface.

9 The bases of this work are the entries of the markers in the *Thesaurus Linguae Latinae* (Thesaurusbüro München Internationale Thesaurus-Kommission, *n.d.*), and when this resource does not yet provide the description of the lemma, the *Oxford Latin Dictionary* is consulted (Glare 2012). In addition, current etymological dictionaries have also been consulted (Ernout and Meillet 2001; Meiser 2010; Vaan 2008).

As the description of the workflow suggests, WoPoss needed an annotation platform that, in the first place, could import the output format of natural language processing tools. It was important for the annotators to be able to access (and edit) the results of the linguistic automatic annotation so they could implement the pertinent corrections when necessary. In the second place, the support of rich semantic annotation schemes was required. It was important to have the ability to formalize large tagsets and create restrictions around them to facilitate the annotation process and to ensure the accuracy of the annotation. Finally, we needed to be able to export the annotated dataset to a format that allowed its transformation to different output formats. This guarantees the sustainability of the dataset and its efficient exploration and exploitation.

2.2. Specific requirements of the fine-grained annotation

In this section, I will detail the elements that needed to be formalized through the annotation platform from a generic point of view: the rationale of this section is to expound the technical functionalities that an annotation platform must have for the correct modelling of modality as understood in the WoPoss project.

- Interaction between multiple layers of annotation. Semantic interpretation is conditioned by other levels of linguistic analysis. Therefore, it is critical to discern between different layers of linguistic annotation.
- Annotation of relations. As briefly mentioned in Section 1, the theoretical approach to modality of WoPoss discriminates the different components of a modal passage. As understood in

this project, modality concerns the expression of the notions of possibility, necessity and probability. We identify the lexical elements that articulate these notions, that is, the modal markers. Modality concerns the stance of a speaker on a specific representation. This representation is the state of affairs.[10] To analyse the state of affairs, we identify the scope, that it, the part of the clause to which the marker refers, and the participant or participants in the state of affairs, when pertinent. Finally, for each modal passage we examine the abstract relation between the marker and its scope. Therefore, a network of relations needs to be established between different linguistic components: the relation of the marker with its scope and, when relevant, the role of the participant with the scope.

- Annotation of linguistic contents below the word level. A word-based tokenization would not be granular enough to identify the linguistic units that comprise a modal expression. For instance, some of the modal markers selected for annotation concern morphological units smaller than the words, such as the adjectival suffixes *-bilis* or *-turus*. This means that the presence of these adjectives with a modal meaning required a segmentation of the word in which the suffix must be analysed as a modal marker, and the root as part of the scope (Dell'Oro 2020).
- Annotation of discontinuous elements. As the relevant linguistic elements for the annotation may not be contiguous, a method to identify tokens belonging to the same structure is needed.

10 There are special cases when, for example, the state of affairs is not explicit.

- Overlapping and stacking. Again, syntactic structures may determine the discontinuity and overlap of the segments that form a modal expression. In addition, various combinations of the modal units are possible so a flexible annotation system must be implemented: a marker might affect multiple scopes, or the same scope could be conditioned by more than one marker.
- Annotation of ambiguity. In order to understand modal shift, annotators take care to annotate the possibility of two (or more) modal readings.[11]

3. Testing phase: an overview of annotation tools

Numerous benchmarks for the evaluation of software are available, including benchmarks that were specifically created for the recommendation of XML editors (van den Broek, Wiering, and van Zwol 2005) which could have been a starting point for the procedure of selecting an annotation tool. However, considering the specificities of the WoPoss project (briefly presented in the Introduction), it was decided to perform a hands-on experience with different annotation tools. Therefore, a mock-up for each evaluated tool was developed as a proof of concept.

For the selection of tools to be reviewed, we took into consideration the resources used by other projects.[12] I will briefly present the reasons why Analec and <oXygen/> XML Editor were

11 In addtion, the discrimination of the meaning conveyed by natural language expressions requires a large amount and wide range of contextual information which is not always available in a project that analyses textual contents created thousands of years ago. Therefore, ambiguity is inevitable (Bunt 2017).
12 After the proof of concept was finished, it came to my attention that the project *Portuguese Corpus Annotated for Modality - MODAL* (Hendrickx, Mendes, and Mencarelli 2012) used the tool MMAX which was not evaluated by the WoPoss team. This tool is very versatile but it does not support an installation as a service, which is especially useful for working collaboratively in a production environment. Moreover, it has not been updated since 2013. For a description of this tool see Müller and Strube (2006).

rejected in favour of WebAnno. Then, INCEpTION was used instead of WebAnno because the development of the later project merged with that of INCEpTION.

Analec[13] is a specific tool for textual annotation that has a desktop version and also a plug-in as part of the modular platform for textometry, TXM.[14] The main advantage of Analec is its very intuitive interface (see Figure 1). It also provides additional functionalities thanks to the built-in analytical tools through which different calculations can be made: computation of frequencies, search of correlations or the establishment of the inter-annotator agreement. There is also the possibility to perform advanced queries of the annotations. Although the modifications of the scheme are easy to implement (and update), the definition of restrictions offers few possibilities. For instance, it is not possible to specify the cardinality of a feature, that is, to define whether a feature is optional or mandatory and whether it can be repeated. Moreover, restrictions conditioned by the value of a specific feature cannot be established, which is an important handicap considering, for instance, how much the description of an epistemic modal passage differs from that of a dynamic one in terms of pertinent features. Therefore, an annotator would have to read over non-pertinent features instead of having more guided annotation choices. The last disadvantage is that the same tokens cannot be analysed more than once. This makes the encoding of ambiguity especially convoluted, since the most straightforward annotation will entail the analysis of the same passage with the different meanings that are the source of the ambiguity.

13 For information about Analec, see <http://explorationdecorpus.corpusecrits.huma-num.fr/analec-2> (accessed on 04/11/2019) and Landragin et al. (2012).
14 <http://textometrie.ens-lyon.fr> (accessed on 04/11/2019).

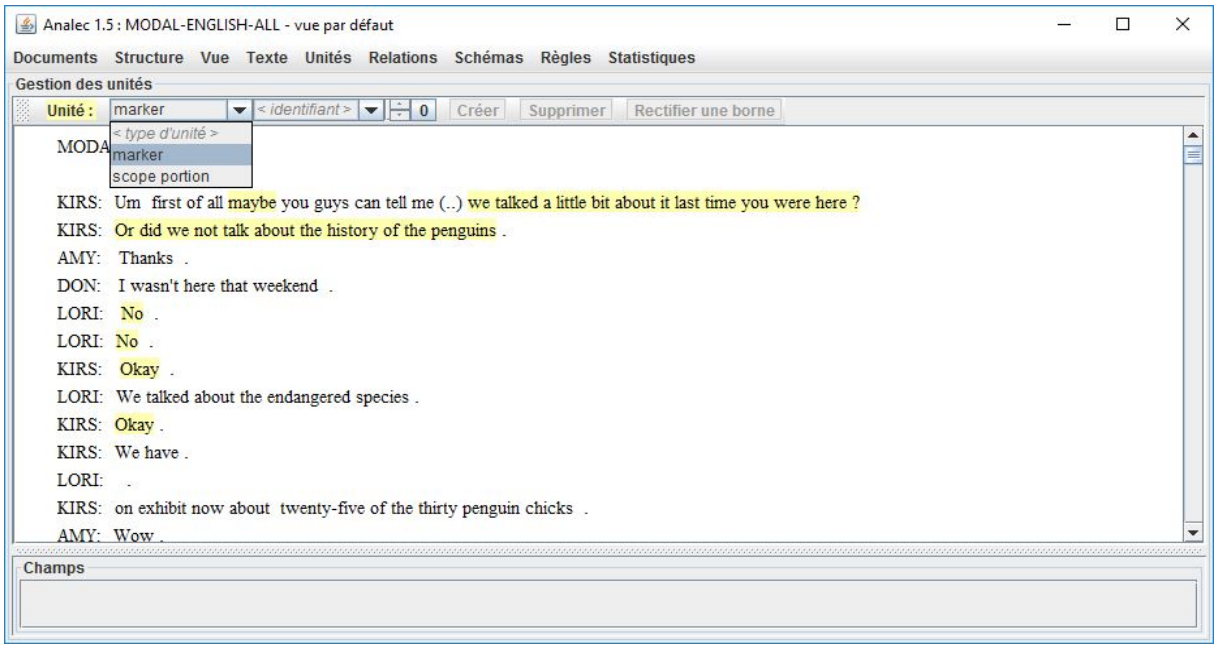

Figure 1. Screenshot of the annotation interface of Analec (*Modal* project)[15]

We also performed a proof of concept with the XML editor <oXygen/>[16] using CSS and its Author Mode to customize the annotation experience and make it more user-friendly (Figure 2). The encoding strategy for this test implemented TEI (TEI Consortium 2019b) encoding in which manual annotations were added through stand-off methods.[17] In contrast to Analec, the great advantage of directly editing the XML is the possibility to define a very complex scheme formalized through feature structures. A feature structure is a group of *attribute:value* pairs, where the values may either be atomic or nested feature structures (Witt and Stegmann 2009) so that complex hierarchies can be created, achieving a great level of granularity by describing a linguistic phenomenon as an accumulation of feature structures.

15 Taken from <https://modal.msh-vdl.fr/index.php/2016/12/10/english-using-the-analec-tool> (accessed on 19/11/2019).
16 <https://www.oxygenxml.com> (accessed on 04/11/2019).
17 More specifically, the mark-up technique detailed in Bermúdez Sabel (2018) built upon the TEI feature structures module (TEI Consortium 2019a).

By using this environment, no changes of format need to be done throughout the workflow since we would be working with XML and XML technologies from the source retrieval to the publication of the annotated dataset. The main disadvantage of this tool is that the annotation process comes off as tedious, especially for less experienced annotators, and specifically when dealing with discontinuous elements for which the boundaries of each segment need to be made explicit.

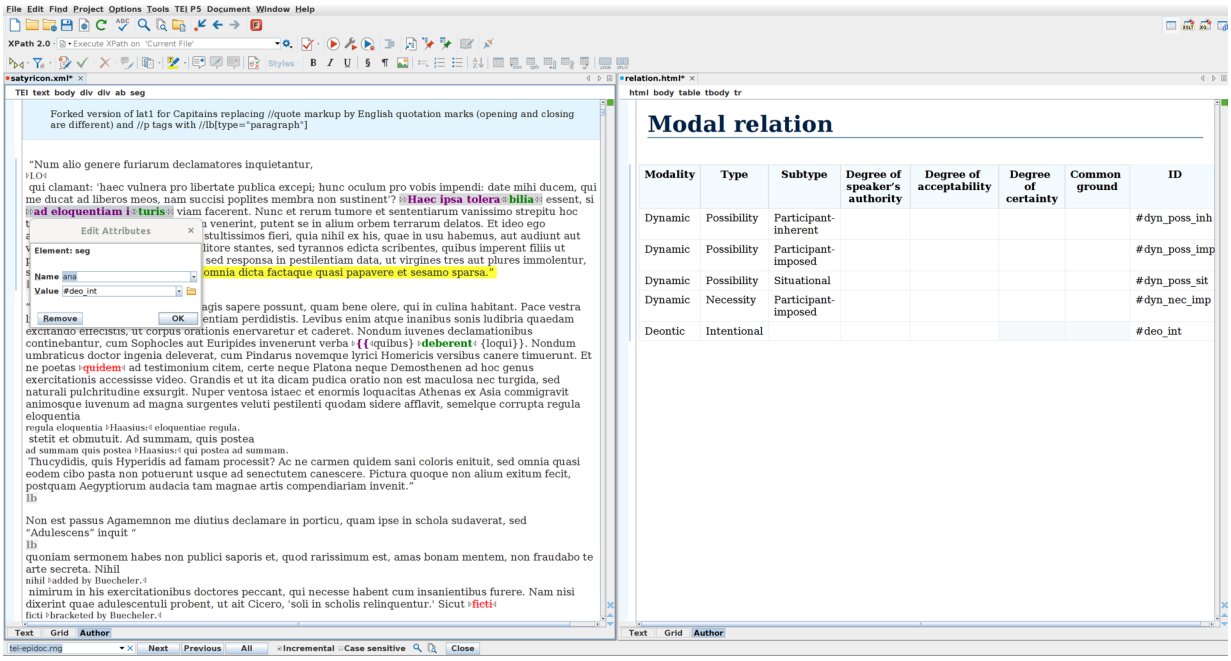

Figure 2. Screenshot of <oXygen/>. In the left panel, green and magenta indicate the different type of modal units already annotated, grey highlight defines the modal relation. A dialog prompts the introduction of the code to describe the segment being analysed. Theese codes are searchable in the table dispalyed on the right panel.

WebAnno is a general purpose web-based annotation tool. Although conceived for linguistic annotation, it allows the

customization of any layer of annotation thus enabling its use even for non-linguistic annotation.

Since 2018, WebAnno entered a phase of development that is mainly prompted by the specifications of the INCEpTION project. Thus, any updates of WebAnno are done on the basis of how certain parts of this tool can be reused by the INCEpTION project. Therefore, instead of describing WebAnno, I proceed to the examination of INCEpTION in the following section.

4. Description of the annotation platform

INCEpTION is a multi-functional and multi-modular platform that enables the creation of corpora, the annotation of texts, and the management of knowledge.

INCEpTION is presented as a tool particularly adept at handling semantic annotations (Klie et al. 2018). One of the reasons behind this statement is the flexible multi-layer annotation support: different layers can be combined, all of them being implemented with freely configurable annotation schemes. In our case, the interaction and conditioning of semantics with other aspects of linguistic analysis, especially morphosyntactic features, needed to be made explicit so it was crucial to work with an annotation tool that supported multiple levels of annotation.

Besides providing the framework to develop tangential tasks directly related to text annotation, such as corpus management, INCEpTION also includes different features to improve the efficiency of the tasks themselves. Concerning the annotation procedure, it provides intelligent annotation assistance under the form of machine learning recommenders. These recommenders can be used during the annotation process to generate predictions

that the annotators may accept or reject. Through an active learning process, the evaluations by the users are employed to further improve the quality of the predictions (Klie 2018). Besides the built-in recommenders, users can train their own recommenders as their dataset is progressively annotated and validated. This means that the manual annotations can later be used for training and implementing an automatic annotation.

INCEpTION facilitates knowledge management thanks to the knowledge base module. This feature allows users to create their own knowledge base, to import one or to connect to remote knowledge bases, like DBpedia[18] or YAGO.[19] A knowledge base can be used, for instance, for linking entities. Besides adding more information about a particular entity, this step is especially useful for disambiguating mentions. Through a knowledge base, cross-document co-references can be introduced. Adding this type of references not only enriches the annotation of concepts or named entities, but it is also helpful for the addition of complex semantic information like, for example, taxonomic or meronymic relations (Eckart de Castilho et al. 2018).

In regard to the supervision of an annotation project, INCEpTION provides different functionalities to manage collaboration. Thanks to the various types of users available, members of a project with different profiles have a specific environment to perform their particular tasks: corpus management, customization of schemas, annotation, monitoring and curation (see Figure 3). All these functionalities are available depending on the type of user, so while an annotator may only access the annotation interface (see Figure 4), curators have access to a monitoring environment where

18 <https://wiki.dbpedia.org/> (accessed on 18/11/2019).
19 <https://datahub.io/collections/yago> (accessed on 18/11/2019).

they can check the progress of the annotation. In addition, inter-annotator agreement, that is, the degree of agreement between the annotators of the same text, can be automatically calculated according to three different types of measure.[20] Besides the statistical approach, the curator can easily compare the results of different annotators and validate (or reject) their annotations.

As a final remark about INCEpTION, it should be noted that the development of the tool is open: not only is the code freely available in a public repository,[21] but the discussions and development tasks are also publicly managed via GitHub. It is also worth mentioning that the community of users of this resource has at its disposal an active mailing-list in which the developers are quick to offer their support by answering back to any problems or doubts posed by users.

20 The available measurements are the Cohen's kappa, Fleiss' kappa and Krippendorff's alpha. For more information about the differences between this type of measures see Gwet (2014).
21 See <https://github.com/inception-project/inception> (accessed on 18/11/2019).

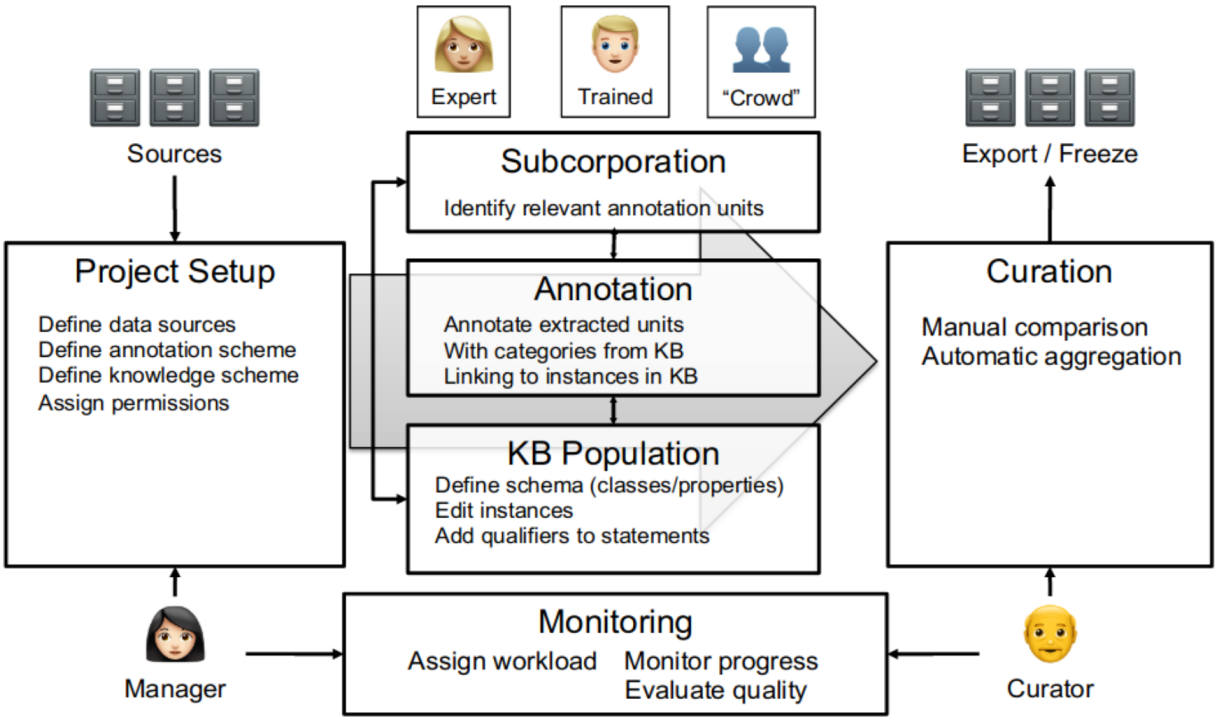

Figure 3. Workflow of a project managed through INCEpTION (Eckart de Castilho et al. 2019)

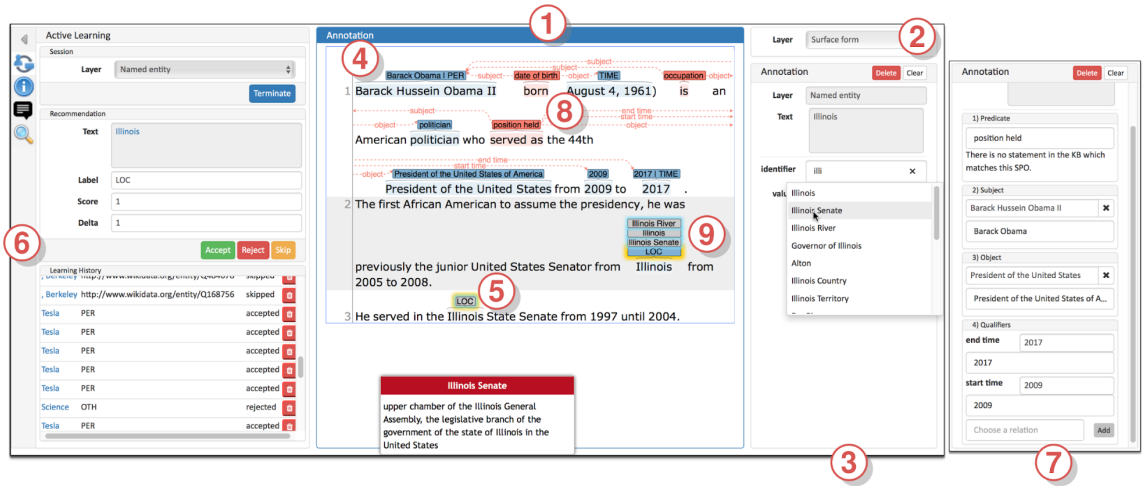

Figure 4. INCEpTION annotation editor: 1) annotation area, 2) annotation layer selection, 3) entity linking feature editor, 4) named entity linked to Wikidata, 5) entity mention suggestion, 6) active learning sidebar, 7) fact linking editor, 8) annotated fact, 9) entity linking recommendations (Klie et al. 2018, 8).

5. Customization of the annotation platform

This section will not delve into the details of user administration or document management. However, it will explain how the theoretical framework and the informational needs elicited during a first annotation test[22] were formalized in INCEpTION. Thus, I will present the steps followed to create an annotation scheme in this tool.

The formalization of the annotation scheme entails the definition of different components, viz. the layers, their features and the tagsets that can control the values of those features. In addition, constraints can be declared in order to define the optional or mandatory nature of these elements as well as the relations between them.

The results of the automatic linguistic analysis are conveyed in fourth layers: *Lemma*, *Morphological Features*, *Part of speech*, and *Dependency*.

To encode the different elements that constitute a modal expression, the layer *Modal unit* was created. This is a layer of type "span": it enables the annotation over a span of text. Span annotations can have any length, can overlap, can stack, can nest, and can cross sentence boundaries, but all these behaviours need to be configured. As explained in Section 2.2, both the marker and the scope do not respond to word boundaries, thus, this layer requires a configuration that stated that the level of granularity of the span is the character. Also, any type of overlap should be allowed. In regard to the features of this layer, they concern the

22 The WoPoss annotation guidelines (Dell'Oro 2019) were designed after the annotation of one text, the *Satyricon* by Petronius, by the three members of the project. The annotations were done independently and later, the results for each modal passage were put in common and throughly discussed.

definition of either the marker or the scope. There is a first feature to discriminate between the type of modal unit that is being analysed, and the value of this feature conditions the following elements that the annotator needs to define.

It was coherent to include both the marker and the scope in the same layer because, on the one hand, they share some linguistic features like the type of utterance or the polarity; and on the other hand, one can create a layer of type "relation" which enables the description of the relationship between spans that belong to the same layer. This last aspect was suitable for the creation of the layer *Modal relation*. Therefore, this layer is attached to the *Modal unit* one and it is used to define the abstract relationship between a marker and its scope (or scopes). Among other features, the different types and subtypes of modality are defined in this layer. Since the features to describe an epistemic passage are not pertinent to define, for instance, a dynamic one, different conditions are put into place so features appear in the annotation interface when they are really pertinent.

There are other linguistic elements that are relevant for the study of modality, therefore additional layers are created.

The state of affairs is the representation that is modalized in a modal passage. In general terms, it can be equivalent to the scope of the marker, but other contextual elements might be needed to reconstruct the state of affairs. This is the reason behind the annotation of a third layer named *Participant*. With this layer we identify the participant of the state of affairs even when it is made explicit at a large distance of the scope. Every participant needs to be linked to at least one scope.

Inspired by studies focused on the negation of modal expressions – e.g. van der Auwera (2001) – we decided to explicitly annotate the lexical element that provides a negative meaning to the marker, hence the existence of the layer *Negation*. In the same way that a *Participant* must be linked to a scope, it is mandatory to relate a negation with the relevant marker.

As exposed in section 2.2, both the marker and the scope can be discontinuous. To deal with this circumstance, a chain layer for each one of those elements was created. A chain layer includes both span and relation annotations into a single structural layer. This is an efficient way to deal with discontinuous elements. Of course, to avoid an annotation abuse, these layers are only pertinent when tackling segmented modal units.

Without entering into the functionalities related to the use of knowledge bases, INCEpTION presents two different types of features: link features and primitive ones.

Link features can be used to link one annotation to others. A link feature is the one which allows us to connect a participant with its scope or, for instance, a negative particle with the marker that it is affected by it.

The primitive feature types supported by INCEpTION are string, boolean, integer, and float. Boolean features are displayed in the user interface as a checkbox that can either be marked or unmarked. Integer and float features are displayed using a number field (although for short ranges radio buttons can be displayed instead). String features are filled in using a text field and they can be displayed as a single field or as a text area with multiple rows. However, if a string feature has a tagset associated with it, a drop-down menu appears instead.

Although there are elements in the annotation scheme of WoPoss that could be formalized as a boolean feature, they are defined instead as a string feature whose possible values are “true” or “false”. This is due to the fact that, at the time of writing,[23] the value of a boolean feature cannot be used in the second part of a conditional statement when defining restrictions.

In the WoPoss project, tagsets are created for any feature whose value is a string, except for the element “note”. As mentioned above, features defined with a tagset are displayed as a drop-down list which only allows the choices declared in the tagset to be selected (users cannot type a value).

The element “note” is an open feature in which the annotator can add any relevant information for the annotation that is not formalized in the other features. For instance, they can record any textual problems here for a latter use. During the curation of files done after the annotation these contents can be reviewed and the pertinent editorial modifications can be implemented.

In the previous paragraphs, the notion of constraints was mentioned when explaining the functionalities of the scheme declaration in INCEpTION. Constraints are used to establish conditional features, that is, features that only become available in the annotation interface if another feature has a specific value. Figure 5 shows the constraints that affect the layer *Modal unit*. The statements before the arrow are the conditions and the elements after the arrow are the features and values that appear if the conditions are met. As we can see, the syntax of the constraints is very straightforward.

23 New versions of INCEpTION are released very frequently and they usually provide new functionalities.

Constraints may also be used for reordering the tags or restricting certain values in a given context.

To sum up, the use of constraints not only speeds up the annotation process, but it enables an annotation less prone to error.

```
Unit {

      Typeofmodalunit="(potential) marker" -> Pertinence="pertinent" |
                     Pertinence="not pertinent -  modal" | Pertinence="not pertinent - not
                     modal";
      Pertinence="not pertinent - not modal" -> Diachrony="post-modal" | Diachrony="pre-modal";
      Pertinence="pertinent" -> Typeofutterance="interrogative" |
                     Typeofutterance="non-interrogative" | Polarity="affirmative" |
                     Polarity="negative";
      Typeofmodalunit="scope unit" -> Typeofutterance="interrogative" |
                     Typeofutterance="non-interrogative" | Polarity="possitive" |
                     Polarity="negative" | SoAcontrol="+control" | SoAcontrol="-control" |
                     SoAcontrol="+/-control" |  SoAdynamicity ="+dynamic" |
                     SoAdynamicity ="-dynamic" |  SoAdynamicity ="+/-dynamic";

}
```

Figure 5. Snippet of the constraint file

Besides the elaboration of the annotation scheme with the definition of constraints to aid (and validate) the annotation process, INCEpTION supports the display of additional documentation. The members and collaborators of WoPoss can access the annotation guidelines through the annotation interface at any moment. In addition, the tagsets employed in the layers created by the automatic analysis[24] are also available so these annotations can easily be reviewed (and corrected when needed).

In this section, all the challenges previously expounded in section 2.2 were addressed. Thereby, I showed that INCEpTION provides a suitable environment for the development of our project. This,

24 The automatic annotation is done using a model trained with the Perseus treebank: <https://github.com/UniversalDependencies/UD_Latin-Perseus> (accessed on 21/11/2019). For more information about the annotation of this resource see <https://universaldependencies.org/treebanks/la_perseus/index.html> (accessed on 21/11/2019).

however, does not mean that some improvements would not be welcomed.[25]

6. Final remarks

To conclude this paper, I underline the importance of working with the appropriate tools when tackling such a complex phenomenon as the semantic analysis of modality.

A tabular formalization, that is, the type of description that can be made in a spreadsheet, is hardly suitable for the definition of notions that are so intrinsic to the context. A correct semantic interpretation requires contextual information that includes the complete morphological and syntactical structure of the linguistic expression containing the modal passage. In this sense, a platform that supports the annotation directly on the text seems to be imperative. The WoPoss approach to modality requires the overview of the interaction between multiple linguistic elements, so a tool that enables the implementation of relations and links between those elements seems to be the most convenient resource. Moreover, it is very practical to use platforms that provide different standards as output formats. This guarantees the sustainability and interoperability of our data as well as their exploitation in different ways without depending on the tool in which the dataset was annotated.

Annotators need an environment in which they can work collaboratively. When evaluating the utility of a given tool, we must consider the learning curve. In general terms, great efforts are made to ensure annotation platforms are as intuitive as possible and made usable by people without a technical background.

25 For instance, it would be useful if, through the constraint rules, one could define the order in which the list of features should appear, so features closely related would be displayed one after the other.

Although using a tool requires some practice, more time should be invested in the issues arisen from the complexity of the annotation scheme than in using the tool itself.

In this paper, I presented a review of the annotation procedure of a specific use case, the diachronic study of modality in the Latin language. Attention was paid to the particular challenges of this project and how a specific annotation platform, INCEpTION, was suitable for the formalization and implementation of a complex annotation scheme.

This annotation platform provides functionalities that we have not explored yet. Future steps of the project envision, on the one hand, the testing of the machine-assisted annotation, and on the other hand, the creation of a knowledge base that would formalize the theoretical framework in an ontology.

References

Auwera, Johan van der. 2001. 'On the Typology of Negative Modals'. In *Perspectives on Negation and Polarity Items*, edited by J. Hoeksema, H. Rullmann, V. Sánchez-Valencia, and T. van der Wouden, 23–48. Amsterdam and Philadelphia: Benjamins.

Bermúdez Sabel, Helena. 2018. 'Anotación Multicamada Externa e o Enriquecemento de Edicións Dixitais/Multi-Layered Stand-off Annotation and the Enrichment of Digital Scholarly Editions'. In *Humanidades Digitales*, edited by Déborah González and Helena Bermúdez Sabel, 4–17. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110585421-002.

Broek, Thijs van den, Frans Wiering, and Roelof van Zwol. 2005.

‘Backing the Right Horse: Benchmarking XML Editors for Text-Encoding’. *Humanities, Computers and Cultural Heritage*, 78.

Bunt, Harry. 2017. ‘The Semantics of Semantic Annotation’. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*, 13–29. Korean Society for Language and Information.

Dell’Oro, Francesca. 2015. ‘What Role for Inscriptions in the Study of Syntax and Syntactic Change in the Old Indo-European Languages? The Pros and Cons of an Integration of Epigraphic Corpora’. In *Perspectives on Historical Syntax*, edited by Carlotta Viti, 271–90. Studies in Language Companion 169. Amsterdam: Benjamins.

———. 2019. ‘WoPoss Guidelines for Annotation’. Lausanne: Université de Lausanne. DOI: 10.5281/zenodo.3560951.

———. 2020. ‘L’expression de la modalité par des adjectifs : une comparaison entre l’adjectif grec ancien –ιμος et l’adjectif latin en –bilis’. *ACME*. Manuscript submitted for publication.

Eckart de Castilho, Richard, Jan-Christoph Klie, Naveen Kumar, Beto Boullosa, and Iryna Gurevych. 2018. ‘Linking Text and Knowledge Using the INCEpTION Annotation Platform’. In *Proceedings of the 14th EScience IEEE International Conference*, 327–28. http://tubiblio.ulb.tu-darmstadt.de/106983/.

Eckart de Castilho, Richard, Jan-Christoph Klie, Ute Wichenbach, and Iryna Gurevych. 2019. ‘Beyond WebAnno: The INCEpTION Text Annotation Platform’. CLARIN Annual Conference, Leipzig.

Ernout, Alfred, and Antoine Meillet. 2001. *Dictionnaire etymologique de la langue latine*. 4th ed. Paris: Klincksieck.

Ghia, Elisa, Lennart Kloppenburg, Malvina Nissim, Paola Pietrandrea, and Valerio Cervoni. 2016. 'A Construction-Centered Approach to the Annotation of Modality'. In *Proceedings of the 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*. Portoroz. https://www.academia.edu/32925215/A_Construction-centered_Approach_to_the_Annotation_of_Modality.

Glare, P. G. W., ed. 2012. *Oxford Latin Dictionary*. Second Edition. 2 vols. Oxford, New York: Oxford University Press.

Gwet, Kilem Li. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*.

Hendrickx, Iris, Amália Mendes, and Silvia Mencarelli. 2012. 'Modality in Text: A Proposal for Corpus Annotation'. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 1805–1812. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/520_Paper.pdf.

Klie, Jan-Christoph. 2018. 'INCEpTION: Interactive Machine-Assisted Annotation'. In *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems*, 105–105. http://tubiblio.ulb.tu-darmstadt.de/106627/.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. 'The INCEpTION

Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation'. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5–9. Association for Computational Linguistics. http://tubiblio.ulb.tu-darmstadt.de/106270/.

Landragin, Frédéric, Thierry Poibeau, and Bernard Victorri. 2012. 'ANALEC: A New Tool for the Dynamic Annotation of Textual Data'. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 357–362. Istanbul, Turkey: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/638_Paper.pdf.

Meiser, Gerhard. 2010. *Historische Laut- und Formenlehre der lateinischen Sprache*. 3rd ed. Darmstadt: wbg academic.

Müller, Christoph, and Michael Strube. 2006. 'Multi-Level Annotation of Linguistic Data with MMAX2'. In *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*, edited by Sabine Braun, Kurt Kohn, and Joybrato Mukherjee, 197–214. Frankfurt a.M., Germany: Peter Lang.

Nissim, Malvina, Paola Pietrandrea, Andrea Sansò, and Caterina Mauri. 2013. 'Cross-Linguistic Annotation of Modality: A Data-Driven Hierarchical Model'. In *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*, 7–14. Potsdam. https://www.aclweb.org/anthology/papers/W/W13/W13-0501/.

Pietrandrea, Paola and *varii auctores*. 2016. '*Modal – Modèles de*

l'annotation de La Modalité à l'Oral [Corpus]'. ORTOLANG (Open Resources and TOols for LANGuage). https://hdl.handle.net/11403/modal.

TEI Consortium. 2019a. 'Feature Structures'. In *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. http://www.tei-c.org/release/doc/tei-p5-doc/en/html/FS.html.

———. 2019b. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Thesaurusbüro München Internationale Thesaurus-Kommission (n.d.). *Thesaurus linguae Latinae. Editus iussu et auctoritate consilii ab academiis societatibusque diversarum nationum electi*. Berlin, Boston: De Gruyter.

Vaan, Michiel de. 2008. *Etymological Dictionary of Latin: And the Other Italic Languages*. Brill.

Witt, Andreas, and Jens Stegmann. 2009. 'TEI Feature Structures as a Representation Format for Multiple Annotation and Generic XML Documents'. In *Proceedings of Balisage: The Markup Conference 2009*. Montréal, Canada: Balisage Series on Markup Technologies. http://www.balisage.net/Proceedings/vol3/html/Stegmann01/BalisageVol3-Stegmann01.html.