Giuseppe G.A. Celano

The Dependency Treebanks for Ancient Greek and Latin

Abstract: The article aims to be an introduction to the dependency treebanks currently available for Ancient Greek and Latin, i.e., the Ancient Greek and Latin Dependency Treebank (AGLDT), the Index Thomisticus Treebank (IT-TB), the PROIEL Treebank, and the SEMATIA Treebank. Their pipelines for creation of morphosyntactic annotations are presented so as to highlight major commonalities and differences. All treebanks share the same basic underlying formalism, whereby syntactic words are connected to each other to form labeled directed acyclic graphs, and their annotation schemes, although different, are comparable to a very large extent.

1 An introduction to the dependency treebank formalism

A dependency treebank is a corpus containing a symbolic representation of the syntax of one or more texts. It can be defined as a set of sentences parsed according to the linguistic formalism of dependency grammar. Most treebanks for Ancient Greek and Latin, i.e., the Ancient Greek and Latin Dependency Treebank (AGLDT), the Index Thomisticus Treebank (IT-TB), the PROIEL Treebank, and the SEMATIA Treebank, are dependency treebanks. Even if, in the present article, I deal only with dependency treebanks, most of what follows in the present section could also be applied, mutatis mutandis, to describe constituency treebanks, such as the Nestle 1904 and SBNLGT Treebanks,[1] the major difference being that in dependency treebanks all nodes except the ROOT node are paired with tokens,[2] while in constituency treebanks non-terminal nodes, which represent phrases, such as VPs or PPs, are also licensed.

The parsed sentences in a treebank are formally represented as labeled directed acyclic graphs, where each token, excluding the ROOT node, is annotated

1 The treebanks and relative documentation can be accessed at https://github.com/biblicalhumanities/greek-new-testament/tree/master/syntax-trees (last access 2019.01.31).

2 The term is here used to mean a syntactic word, i.e., the unit for syntactic analysis.

Giuseppe G.A. Celano, Universität Leipzig

https://doi.org/10.1515/9783110599572-016

for its linguistic head and syntactic function. The graphs for each sentence can be visualized as trees (hence the name of "tree-bank"), whose vertices/nodes, excluding the ROOT node, correspond to a sentence's tokens and whose directed edges depict syntactic dependencies (i.e., head-dependent relationships) specified for syntactic functions.

Figure 1 shows an example of a parse tree. Tokens are connected via labeled arrows (i.e., edges with direction from heads to dependents): for example, αὐτόν is a dependent of ὀνομάζουσι and its syntactic function is OBJ (i.e., it is the object of its head ὀνομάζουσι). Notably, the original sentence in Figure 1 contains the graphic word ἐγῷμαι, where two syntactic words, ἐγὼ and οἶμαι, are merged together by crasis. This phenomenon well illustrates the necessity of keeping the concept of graphic word and token/syntactic word distinct in treebanking: even if, in Ancient Greek and Latin, the ratio between tokens and graphic words is very close to one (i.e., one graphic word usually corresponds to one token), there exist cases where graphic words clearly need to be split (e.g., Latin enclitic *que* also requires tokenization).

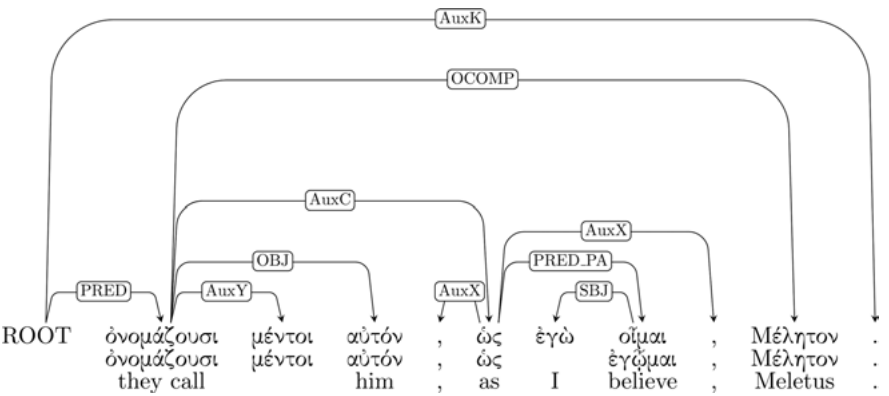

Figure 1: A dependency tree from Plato's *Euthyphro* (2b).

Each token in a parse tree can only have one head (also called "governor"), while one head can have more than one dependent. This translates, graphically, into one or more arrows originating from a token (e.g., ὀνομάζουσι has four dependents), but only one arrowhead per token. In the constituency/dependency grammar parlance, it is common to also describe relationships between tokens as kinship relationships: for example, ὀνομάζουσι is the parent of μέντοι, αὐτόν, ὡς, and Μέλητον, which are in turn its children. These latter tokens are, with respect to each other, siblings, in that they have a common parent.

In this regard, it is noteworthy that the dependency formalism fits perfectly into the XML/XPath/XQuery data model,[3] in that each sentence graph could be serialized as an element node whose child and descendant elements represent the linguistic tree structure (see Figure 2). The XML Path language 3.1 provides a tool to smoothly query such elements, by allowing traverse along a rich set of axes, such as `parent::`, `descendant::`, or `following-sibling::`.[4]

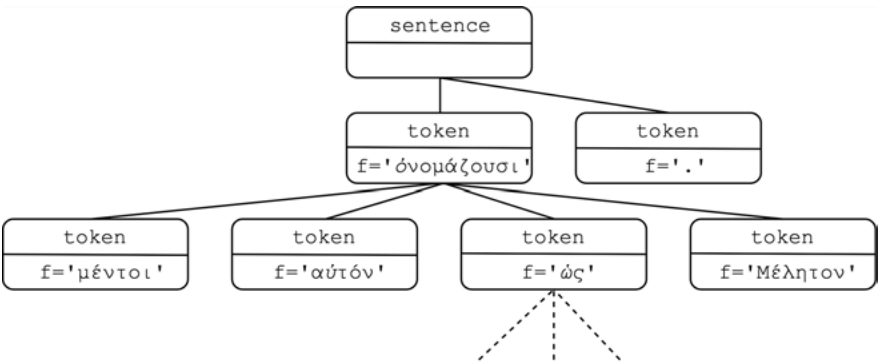

Figure 2: An XML serialization of the sentence in Figure 1.

The dependency formalism used for treebanks offers a model for syntactic annotation, which is, like any model, a trade-off between description completeness/accuracy and simplicity, the latter being required to make the entire process of annotation of large data sets doable. One of the clearest limitations of the dependency model is that all relationships are formally represented as dependency relationships (i.e., subordinate relationships). This poses a challenge when it comes to representing non-subordinate relationships, such as constituents coordinated by conjunctions such as καί or *et* ("and"), appositions, or, as in Figure 1, parenthetical clauses.

3 https://www.w3.org/TR/xml/; https://www.w3.org/TR/xpath-datamodel-31/; https://www.w3.org/TR/xpath-functions-31/; https://www.w3.org/TR/xquery-31/ (last access 2019.01.31).

4 Interestingly, the XML structures of the serializations of both the AGLDT and the PROEIL Treebank privilege readability by keeping unaltered the token order of original texts: tokens are serialized as sibling elements and the linguistic dependency structure is expressed via internal links. On the contrary, the Prague Markup Language (Pajas 2010) used for the IT-TB shows a closer mapping between the parent-child relationships of the linguistic structure and those of the XML structure.

In the dependency formalism, such non-subordinate structures are formally represented, like any other, as subordinate relationships in order to preserve the simplicity of the model. Notwithstanding, they have to be correctly interpreted as *technical* dependencies,[5] their linguistic reality being different. This is even clearer with punctuation marks, which are commonly part of a syntactic tree, although their syntactic relevance is often questionable (especially for ancient texts, where punctuation is usually added by modern editors). The ROOT node can also be interpreted as a technical node: in the AGLDT, it typically governs the verb of the main clause, which receives the syntactic label PRED, and the final punctuation mark, which always receives the syntactic label AuxK.

Differently from constituency treebanks, constituents such as noun phrases or subordinate clauses are not explicitly annotated in a dependency treebank. However, most of them can be indirectly identified combining morphological and syntactic information. In the AGLDT, for example, a noun and its dependents can be taken to form an NP; similarly, a node labeled with AuxC, used to annotate a subordinate conjunction, and its dependents form a subordinate clause; a finite verb form with syntactic label ATR plus its dependents is a relative clause.

A treebank typically contains morphological annotation and lemmatization, which are layers of annotation preliminary to syntactic annotation. Further annotation layers, such as, for example, pragmatics or semantics, can also be added. All dependency treebanks for Ancient Greek and Latin contain morphological annotation and lemmatization. Only a few Ancient Greek and Latin texts have also been enriched with some semantic and/or pragmatic annotation (see following sections).

The pipeline for the creation of a treebanked text typically includes the following steps:

- automatic tokenization of an original text,
- automatic morphological (and syntactic) annotation,
- manual correction of the tokenization/morphosyntactic annotation.

Even though original texts can be in any format, they are usually available as TEI/EpiDoc XML, which is the standard for text encoding. Each treebank has developed its own algorithms to perform both tokenization and morphosyntactic annotation. The automatic morphosyntactic annotation is now usually performed via statistical POS taggers/parsers which have been previously trained on gold data annotated manually or by rule-based algorithms.

5 https://www.cil19.org/cc/en/abstract/contribution/754/index.html (last access 2019.01.31).

Manual annotation for a given text is usually performed by one, two, or three annotators. The one-annotator model is the "scholarly model": annotation reliability is assumed because of the expertise of the annotator, who is a trained advanced scholar. In the two-annotator model, one expert annotator's annotation is reviewed by another expert annotator (whose judgment therefore prevails). The three-annotator model requires that the annotations of two annotators are adjudicated by a third expert annotator, who resolves discrepancies.

Layers of annotations are not provided as pure stand-off markup.[6] Morphosyntactic (and pragmatic) annotation is usually attached to tokens within the same file. Even when the layers of annotations are kept separate, as in the Prague Markup Language (PML) used for the IT-TB, references to offsets in the (unannotated) original texts are not given – the tokenization of an original text therefore becomes the new base text. The original physical format for all treebanks is some flavor of XML. Since the schemas adopted are rather simple, it is also possible to easily convert the XML formats to other formats, such as CoNLL.

In the following sections, I will describe the Ancient Greek and Latin dependency treebanks in more detail, trying to offer an overview whose aim is to describe the main features, commonalities, and differences of the treebanks. I introduce the AGLDT in Section 2, while in Section 3 the IT-TB is presented. I outline the PROEIL Treebank in Section 4 and the SEMATIA Treebank in Section 5. Section 6 contains some conclusive remarks.

2 The Ancient Greek and Latin Dependency Treebank

The Ancient Greek and Latin Dependency Treebank is the oldest treebank for Ancient Greek and Latin (Bamman and Crane 2011). The project started at Tufts University in 2006 and is currently continued mostly at Leipzig University. The treebank contains entire texts or parts of texts belonging to classical antiquity, the choice of which reflect research interests of the annotators. The current release is 2.1. Most of the annotations have been performed by single scholars or university students under the supervision of a teacher.

6 A recently approved DFG-project ("Revising, standardizing, and expanding the Ancient Greek and Latin Dependency Treebank") aims, among other things, to provide stand-off annotation for the AGLDT using PAULA XML: http://gepris.dfg.de/gepris/projekt/408121292?language=en (last access 2019.01.31).

The Ancient Greek part of the treebank currently contains 557,922 tokens. It includes the annotations of (parts of) the following works: *Ilias, Odysseia, Hymnus in Demetrem*, Aeschylus' and Sophocles' tragedies, Hesiod's *Theogonia, Operae et dies*, and *Scutum*, Plato's *Euthyphro*, Lysias' *De caede Eratosthenis, In Alcibiadem I, In Alcibiadem II, In Pancleonem*, Plutarch's *Alcibiades* and *Lycurgus*, Aesop's *Fabulae*, Athenaeus' *Deipnosophistae*, Diodorus Siculus' *Bibliotheca Historica*, Herodotus' *Historiae*, Polybius' *Historiae*, Pseudo Apollodorus' *Bibliotheca*, and Thucydides' *Historiae*.

The Latin part of the treebank contains 79,697 tokens. The following works have been (partly) annotated: Augustus' *Res Gestae*, Jerome's *Vulgata*, Ovid's *Metamorphoses*, Sallust's *Bellum Catilinae*, Caesar's *Commentarii de Bello Gallico*, Cicero's *In Catilinam*, Vergil's *Aeneid*, Petronius's *Satyricon*, Phaedrus' *Fabulae*, Propertius' *Elegiae*, Suetonius' *Vita Divi Augusti*, and Tacitus' *Historiae*.

As Ancient Greek and Latin are morphologically rich languages, the morphological annotation of each token in the AGLDT treebank is based on tagsets[7] identifying both parts of speech and morphological features.[8] It is to be noted that, differently from syntactic annotation, most of the Ancient Greek and Latin texts have been annotated without specific morphological guidelines. Guidelines for the annotation of Ancient Greek morphology have been added from release 2.0.[9]

The lack of morphological guidelines is a common feature of all the Ancient Greek and Latin dependency treebanks. Since writing guidelines is a labour-intensive task, all projects have given priority to syntactic guidelines, syntax being arguably more complex to annotate. It is however acknowledged that morphological guidelines are needed. While morphological annotation for most tokens may seem uncontroversial, there are a number of known phenomena requiring rules.

These include, for example, distinction of the category noun/adjective (e.g., is *Athenienses* always an adjective?) or definition of the category "pronoun," the latter being able to be used to cover examples such as ἐγώ and ἐμός (= possessive

7 They are documented at https://github.com/PerseusDL/treebank_data/tree/master/v1/greek (Ancient Greek) and https://github.com/PerseusDL/treebank_data/tree/master/v1/latin (Latin) (last access 2019.01.31).

8 The annotations have been performed using the full-fledged Arethusa annotation tool, which also allows automatic tokenization and sentence split – which, as any other piece of annotation, can then be manually corrected. It is accessible at http://sosol.perseids.org/sosol/, while its code, including the one for the tagsets, is documented at https://github.com/alpheios-project/arethusa (last access 2019.01.31).

9 (Celano 2014).

adjective). Similarly, rules are needed to consistently annotate, for example, ὁ and τις, which could be articles (definite and indefinite, respectively) or pronouns. Another example is the distinction between relative adverbs and conjunctions: when Latin *ubi* means "when", it tends to be annotated as a subordinate conjunction, but when it means "where" as a relative adverb. Many of such morphological issues are treated in the Guidelines for the Ancient Greek Dependency Treebank 2.0.[10] Guidelines for Latin morphology are missing.

Technically, morphological annotation has been performed semi-automatically, a morphological analyzer suggesting an annotation, which is then validated by an annotator.[11] It is physically encoded as a 9-character long string, whose first position represents the POS and the following ones the morphological features. For example, the morphology of βασιλέα corresponds to "n-s—ma-," which stands for noun ("n"), singular ("s"), masculine ("m"), and accusative ("a"). Each position in the string always corresponds to a definite morphological category having definite values represented by letters. For example, the third position always encodes "number" with three possible values: singular ("s"), plural ("p"), and dual ("d"). Similarly, the seventh position always encodes "gender" and can take three values: masculine ("m"), feminine ("f"), and neuter ("n"). When a category is not relevant for a certain word form, a hyphen is used to mean lack of that morphological feature. For example, the feature "person" is always encoded in second position: as nouns are not specified for "person", the string for βασιλέα shows a hyphen.

Syntactic annotation for both Ancient Greek and Latin has been performed following an annotation scheme informed by the guidelines for the analytical layer (i.e., syntax) of the Prague Dependency Treebank 2.0.[12] The initial guidelines for Ancient Greek and Latin comprise 19 and 20 syntactic labels, respectively. In general, dependency rules, as well as the meaning of the labels, used to annotate Ancient Greek and Latin are very similar.

The initial guidelines for Ancient Greek and Latin, as most other guidelines, present syntactic descriptions which cannot necessarily be complete, but represent a compromise between annotation feasibility and description completeness. The guidelines for Ancient Greek have been further extended by the

10 (Celano 2014).

11 The morphological analyzer used is Morpheus (Crane 1991), which is integrated into the Arethusa annotation framework. A POS tagger has been more recently used for Ancient Greek (Celano et al. 2016).

12 (Hajič et al. 1999).

annotation guidelines for the AGDT 2.0.[13] Their novelty consists in making the annotation rules more precise by incorporating H.W. Smyth's *Greek Grammar for Colleges* (henceforth SG; Smyth 1920) both notionally and formally via hyperlinks to the relevant sections of the Perseus digital edition.[14] Many definitions are also provided with treebanked examples to make them clearer. On the contrary, the *Guidelines for the Syntactic Annotation of Latin* remain the only documentation for syntactic annotation of Latin (but see Section 3 for the *Rules of Annotation for the Analytical Layer of the Index Thomisticus Treebank*).[15]

The basic unit for syntactic annotation is the sentence, whose end is formally defined by the presence of a strong punctuation mark, such as a full stop or a colon. Prototypically, a sentence has a main verb, which is annotated as the linguistic root of the dependency tree and gets the label PRED (= "predicate"). Problematic is the case where a main verb is missing: in this case, an elliptical node functioning as the main verb is usually added. If a sentence starts with a coordinating conjunction, as is often the case in both Ancient Greek and Latin, the conjunction is chosen as the linguistic root (COORD), while its associated verb is made dependent on it with the label PRED_CO.

The suffix _CO is added to any node depending on a coordinate conjunction. Appositions are similarly encoded: the appositive nodes get labels depending on their syntactic function within a clause, which are terminated by the suffix _AP.

The sentence in Figure 3 shows the annotation style for appositions in Latin: *Fulvius* and *filius* are both annotated as subjects and get the suffix _AP. Notably, this annotation style differs from how apposition is treated in Ancient Greek and Latin traditional grammars, where apposition is conceptualized of as an independent syntactic function and not as a coordinate structure. According to this latter, *Fulvius* should be annotated as SBJ depending on *erat* and *filius* as APOS depending on *Fulvius*. This annotation style is favored for Ancient Greek starting from the guidelines 2.0.

The label OBJ is meant to captures all arguments not being SBJ, PNOM, or OCOMP. This means that it is not only used for direct objects, but also for a great variety of other complements "required" by a given verb. These include, for example, indirect objects and prepositional objects, such as those governed by verbs of motion. Admittedly, the notion of argument is notoriously difficult

13 (Celano 2014).

14 http://www.perseus.tufts.edu/hopper/text?doc=Perseus%3Atext%3A1999.04.0007 (last access 2019.01.31).

15 (Bamman et al. 2007). In the present and following sections, I will focus on the most noteworthy annotation phenomena/questions; for more details the reader is referred to the guidelines and data of each treebank.

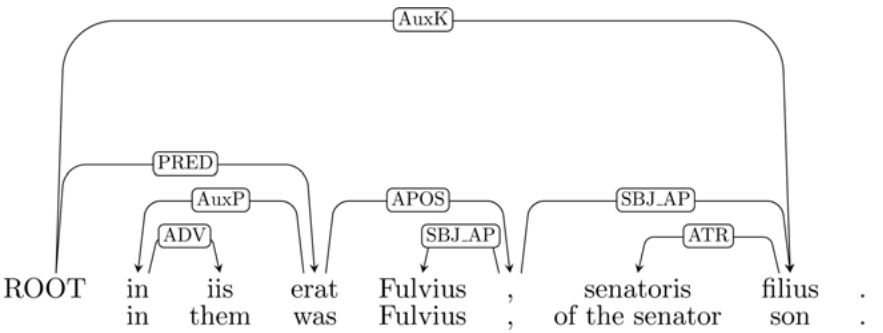

Figure 3: A dependency tree showing apposition.
Note: The example is drawn from Bamman et al. 2007, p. 28. The syntactic label of *iis* is however questionable: I would rather consider it as OBJ.

to define and both the original guidelines for both Ancient Greek and Latin do not attempt to define it precisely[16], therefore leaving to the annotator the task of identifying them.

The syntactic function "predicate nominal" (PNOM) is prototypically meant to capture complements depending on the copula. The Ancient Greek guidelines 2.0 further specify that this same label should also be used for complements, including supplementary participles not in indirect discourse, which depend on copulative verbs, a list of which is given in SG 917. Similarly, the label OCOMP is used for object complements and, according to the Ancient Greek guidelines 2.0, for supplementary participles not in the indirect discourse depending on verbs of perceiving and finding (SG 2110–2115).

The label ADV ("adverbial") is used to tag dependents being adjuncts. Contrary to OBJ nodes, adverbials are those dependents which are not verb specific and could therefore potentially modify any verb. Typical adjuncts are, for example, temporal modifications. The annotation schemes for Ancient Greek and Latin also include a number of Aux- labels to annotate dependents whose function is "auxiliary", which usually correspond to function words, such as prepositions and conjunctions.

The Ancient Greek guidelines 2.0 also provide rules for the annotation of a third annotation layer, which is called "advanced syntax layer", whose nature is at the interface between syntax and semantics.[17] More precisely, it can

16 See, however, https://github.com/PerseusDL/treebank_data/blob/master/AGDT2/guidelines/Greek_guidelines.md#obj (last access 2019.01.31).
17 (Celano 2015).

be defined as a syntax-driven semantic layer corresponding to categories such as "genitive of possession" or "purpose clause" elaborated by traditional grammars and summarized in Smyth (1920). This layer of annotation is currently available only for the annotated passages of Aesop's fables and Diodorus Siculus' *Bibliotheca Historica*.

The advanced syntax layer has been designed as an algorithm which, starting from the morphological annotation of any given token, guides the annotator, through successive steps, towards a more specific, semantic annotation. For example, a noun morphologically tagged as a genitive, can be further annotated – depending on its function within the clause – as "genitive proper > genitive of possession" or as "ablatival genitive > genitive of cause". An underlying assumption for the creation of this layer is that classicists are familiar with its categories, which are considered useful for the study and description of the language.

The Ancient Greek and Latin Dependency Treebank is released in a native XML format. The XML schema is very simple and intuitive. Each text is contained in a `treebank` element, which is the outermost element of the XML file. The `treebank` element has `sentence` elements as children, which contain `word` elements, each of which conveys the morphosyntactic information relative to a sentence token. The `word` elements are specified for at least six attributes: the `id` attribute content is an integer signaling the position of a given token in the sentence; the `form` attribute contains the actual word form for the given token, while the `lemma` attribute its lemma; the `postag` attribute content is the 9-character long string standing for the morphological analysis; the `relation` attribute shows the syntactic function the given token bears with respect to its head, which is referred to, in the `head` attribute, via the integer of the ID of its corresponding `word` element.[18]

3 The Index Thomisticus Treebank

The Index Thomisticus Treebank (IT-TB)[19] is, together with the Latin Dependency Treebank (LDT) (see Section 2), the oldest treebank for Latin.[20] It has been run, since its inception, at the Catholic University of Milan. It contains

18 Many texts also contain cite attributes containing the corresponding cts:urn identifier: https://www.homermultitext.org/hmt-doc/index.html (last access 2019.01.31).

19 The IT-TB is described in more detail, in this volume, in the contribution by Marco Passarotti.

20 (McGillivray et al. 2009); (Passarotti 2011).

parts of Thomas Aquinas' *Summa Theologica* which have been annotated for morphology ("morphological layer") and syntax ("analytical layer"). It currently consists of 277,547 tokens.[21] Part of the data (28,886 tokens) has also been annotated for pragmatics and semantics ("tectogrammatical layer"). The annotation has been performed by a single scholar and then reviewed by another scholar.

The IT-TB relies on the data provided by the *Index Thomisticus* (IT),[22] which is one of the first projects in what we would now call Digital Humanities/ Computational Linguistics: it aimed to digitize all works of Thomas Aquinas and POS tag and lemmatize them.

The morphological annotation of the IT-TB[23] is, due to its historical connection with the IT, the most peculiar one when compared to the morphological layers of the other treebanks. While morphological annotation in the LDT (see Section 2) and the PROIEL Treebank (see Section 4) reflects the categorization elaborated by traditional grammar, the IT-TB is freer in this respect: parts of speech are, for example, identified by a combination of two values: the first describes "flexional categories", most of which specify declension/conjugation type, while the second gives information for "flectional types" ("Nominal", "Participial", "Verbal", "Invariable", and "Pseudo-lemma"). For example, *multitudinis* is tagged as "C1", where "C" stands for "third declension" and "1" for "Nominal". Similarly, *pigmentaria* is annotated as "A1", i.e., "first declension" and "Nominal", while *et* is tagged as "04", i.e., "invariable" and "invariable".

These examples show that a mapping from the IT-TB morphological categories to the more widely known of the LDT is not always possible (without recourse to external resources such as dictionaries or morphological analyzers). The "Nominal" category, for example, can correspond to either a noun or an adjective. Similarly, a token labeled as "04" could be a conjunction, such as "et", or an adverb, such as "bene".

The IT-TB shares the same annotation guidelines for syntax with the LDT (Bamman et al. 2007; see Section 2 for more details). They have also been complemented by the *Rules of Annotation for the Analytical Layer of the Index Thomisticus Treebank* (henceforth RALIT).[24]

21 The calculations on the PML format have been performed on the release available at https://itreebank.marginalia.it/view/download.php on 2018.10.13 (excluding the more recent Golden Age texts).

22 http://www.corpusthomisticum.org/it/index.age (last access 2019.01.31).

23 Documentation for the tagset is available at https://itreebank.marginalia.it/doc/Tagset_IT.pdf and https://itreebank.marginalia.it/doc/Tagset_IT_README.txt (last access 2019.01.31)

24 (Passarotti 2016).

In RALIT annotation issues are addressed, which are particularly (but not exclusively) relevant to Thomas Aquinas' Latin. These include sentences starting with a conjunction, such as *sed* or *quia*. The case of sentences introduced by *quia* is particularly interesting in that it has to do with the phenomenon of ellipsis, which is notoriously difficult to deal with theoretically and therefore also regulate. While the introduction of elliptical nodes in a sentence raises a number of theoretical issues (especially when it comes to comparing different annotations of the same text, elliptical nodes altering the initial token number), it is clear that building of a syntactic tree often requires them. As to the specific case of *quia*, it is stated in RALIT that it is treated as the linguistic root of the tree and the verb depending on it receives the label PRED. In this case, therefore, no elliptical node is added. In general, adequate treatment of ellipsis currently remains one of the major challenges for treebanking.

Interestingly, in RALIT an annotation rule for the construction "ita ... sicut" is also given. The adverb *ita* is made the head of the subordinate clause introduced by *ut*. This kind of syntactic construction is very frequent in both Ancient Greek and Latin. Typically, a pronoun or an adverb anticipates a following subordinate clause, which is usually introduced by a conjunction. Such a clause is of an explicative nature with respect to the anaphor/cataphor in the superordinate clause. English shows a similar construction with clauses introduced by, for example, "to such an extent that" or "to the point that", where the subordinate conjunction "that" introduces the clause explaining what the extent/point is. This construction can take different forms: the kind of the anaphor/cataphor can be a pronoun, an adverb, or a noun, while the subordinate clause can be introduced by different conjunctions (or even be an infinitive clause). The subordinate clause is in traditional grammar described as being in apposition to the anaphor/cataphor (SG 991).

The IT-TB also provides tectogrammatical annotation for a subset of its morphosyntactically annotated sentences.[25] The guidelines for this layer of annotation are based on the annotation manual for the tectogrammatical layer of the Prague Dependency Treebank[26] complemented by the *Guidelines for*

25 The tectogrammatical layer is also, together with the morphological and analytical layers, made available for some Latin texts of the Golden Age: https://itreebank.marginalia.it/doc/15-01-2018_all_resources_all_formats.zip (last access 2019.01.31). A syntactically-based valency lexicon and semantically-based valency lexicon have also been created: https://itreebank.marginalia.it/view/resources.php (last access 2019.01.31).

26 (Mikulová 2006).

Tectogrammatical Annotation of Latin Treebanks: The Treatment of some Specific Constructions, which provide rules for specific Latin constructions.[27]

The tectogrammatical layer provides a description for the pragmatics and semantics of a given sentence. Particularly noteworthy are the formalisms for semantic (macro)roles (called "functors"), coreference, and topic-focus articulation. Since a tectogrammatical tree is meant to represent the semantics and pragmatics of a sentence, its tree structure is different from the corresponding syntactic tree, which has to do with surface syntax: prepositions, for example, are not assigned separate nodes in a tectogrammatical tree, but are merged with nodes of the nouns they govern, in that they enable functor identification (and therefore semantic roles).

The IT-TB is released in different formats. All three annotation layers are available only in the PML (Prague Markup Language) language, which is an XML format specifically designed to encode the linguistic annotation layers for the Prague Dependency Treebank. The PML format keeps the three annotation layers physically separate and links them via IDs. Notably, the XML syntax encoding linguistic trees is informed by the parent-child relationships between linguistic nodes (see Pajas 2010 for the full specification). The morphosyntactic annotation is also available in two other common formats: CoNLL and Tiger XML.

4 The PROIEL Treebank

The PROIEL Treebank originates from the PROIEL (Pragmatic Resources in Old Indo-European Languages) project, which was run at Oslo University between 2008 and 2013. The project produced morphosyntactic (and partly pragmatic) annotation for the Greek New Testament and its translations into the following Old Indo-European languages: Latin, Gothic, Armenian, and Old Church Slavonic.[28]

After the end of the PROIEL project, the PROIEL Treebank has continued to be augmented with new Ancient Greek and Latin texts, mainly belonging to classical antiquity (see also Eckhoff (2017) for the PROEIL Treebank family). Currently, it comprises, besides the Greek and Latin New Testaments, (parts of) the following works: Herodotus' *Histories*, Sphrantzes' *Chronicles*, Caesar's *De*

27 (Passarotti and Gonzáles Saavedra 2015).
28 (Haug et al. 2008); (Haug et al. 2009).

Bello Gallico, Cicero's *Epistulae ad Atticum* and *De Officiis*, *Peregrinatio Aetheriae* and Palladius' *Opus Agriculturae*. The Ancient Greek part of the treebank consists of 250,455 tokens, while the Latin part of 225,064 tokens (release 20180408).[29] Annotations are stored in files containing detailed metadata for each text, such as the source of the text, tagset abbreviations and their expansions, and names of annotators and reviewers.[30]

The morphological annotation[31] for both Latin and Ancient Greek is very similar to that of the Ancient Greek and Latin Dependency Treebank. The parts of speech acknowledged are however more numerous (i.e., 27). This is mainly due to subcategorizations of more general parts of speech: for example, pronouns are classified as demonstrative, indefinite, interrogative, personal, personal reflexive, possessive, possessive reflexive, reciprocal, and relative. Similarly, nouns are of two types: common and proper. On the contrary, the morphological features are essentially the same as those of the AGLDT, with minor deviations. Remarkably, some of them are defined on a purely morphological basis, in that specific values for morphologically ambiguous terminations are allowed: for example, the gender for an adjective such as *cotidianis* is annotated as "masculine, feminine or neuter", even though the gender of its governing noun can disambiguate it.

The syntactic annotation is based on an annotation scheme[32] similar to those for the AGLDT, in that it also relies on the annotation guidelines for the analytical layer of the Prague Dependency Treebank. Also in the PROEIL Treebank argument structure is annotated. Besides subjects and objects (the latter corresponding mostly, but not exclusively, to arguments in the accusative case), arguments conveyed by oblique cases or prepositional phrases are distinguished through the label OBL. In the case of prepositional phrases, the label is used for both the preposition and its dependent noun. If the preposition introduces an adjunct, it receives the label ADV, but the dependent noun is still annotated as OBL, being always considered an oblique argument – on the contrary, in the AGLDT prepositions are always annotated as AuxP and their

29 The numbers do not consider punctuation marks, which are not encoded as token elements.

30 The treebank can be downloaded at https://github.com/proiel/proiel-treebank/releases (last access 2019.01.31).

31 See https://proiel.github.io/handbook/developer/#apis-and-libraries for the code documenting, among other things, text preprocessing (last access 2019.01.31).

32 (Haug 2010).

dependent nouns can get the label OBJ or ADV depending on whether they are or are not arguments.

Differently from the AGLDT, a specific label (XOBJ) is used to mark the syntactic function of predicative complements depending on verbs such as *esse*, *videari*, or *creare*. The distinction between subject and object complements is marked via a system called *slash notation*, which consists in the addition of a dependency relation of the subject or object on the predicative complement. The XOBJ label is also used to annotate infinitives depending on auxiliary verbs such as *posse* and *velle* or on the passive forms of verbs such as *putare* and *dicere* (e.g., *dicitur ad urbem venisse*, with *venisse* receiving the XOBJ label). The PROIEL Treebank also has specific labels for partitives (PART), conjunct participles (XADV), and complement sentences (COMP).[33]

Remarkably, in the PROEIL Treebank function words such as prepositions and subordinate conjunctions take the labels describing the syntactic function of the phrases they heads, while in the AGLDT it is the governed nouns or verbs that receive such labels, prepositions and conjunctions being tagged with function labels (AuxP and AuxC, respectively).

Parts of the Ancient Greek and Latin texts have also been annotated for information structure.[34] This includes addition of pro-drop subjects and of a few labels mostly pertaining to the information status of referents. In particular, the "new" and "old" labels are the ones which prototypically help identification of foci and topics in a sentence.[35]

The PROIEL Treebank is released in a native XML format, which includes all annotation layers. The XML structure is very similar to that of the AGLDT (See Section 2), with `sentence` elements containing `token` elements, each of which has at least 9 attributes. Among these two are peculiar: `citation-part`, which contains the passage reference and `presentation-after`, which contains any punctuation mark following the given token. Optionally, the information-status attribute can be present. The slash notation, which adds further dependency relations, is encoded in slash elements within `token` elements. The morphosyntactic information is also made available in a CoNLL format.

33 A precise description of these labels is outside the scope of the present article. The reader is referred to the guidelines for a detailed account of all their uses.

34 (Haug et al. 2014).

35 All possible information-structure labels are listed in each treebank file. The information structure-annotation is currently best readable when accessing the texts at http://foni.uio.no:3000/, where a visualization for each single annotation layer is provided (last access 2019.01.31).

5 The SEMATIA Treebank

The SEMATIA Treebank[36] aims to provide morphosyntactic annotations for papyri using the formalism of the AGLDT.[37] It is currently maintained at Helsinki University and contains 313 papyrological texts[38] coming from the Duke Databank of Documentary Papyri.[39]

The fragmentary nature of papyri represents one of the biggest challenges for annotation. As is known, more or less extensive parts of a text may be missing, which the papyrologist tries to interpret and integrate. Moreover, documentary papyri usually do not have word division, which often, as one can imagine, raises ambiguities. The complexity of papyri is mirrored in the heavy markup contained in the original TEI XML documents, which challenges automatic extraction of the text itself.[40]

For this reason, two annotated versions of a given papyrus are provided in the SEMATIA Treebank: one for the original text (also called "original layer") and one for the same text after editorial intervention. Moreover, sections of a papyrus written by different authors also receive separate annotations.

Texts are preprocessed and annotated using the Arethusa annotation framework (see note 8). The morphological annotation is facilitated by the Morpheus morphological analyzer, whose outputs however require to be frequently corrected because it is based on Classical Greek and Classical Latin and the lexicon of papyri differs from them in many respects.[41] The SEMATIA Treebank is released in the same XML format as the AGLDT.

6 Conclusion

In the present paper, I have overviewed the existing dependency treebanks for Ancient Greek and Latin, with the aim not to describe them exactly but to present their most relevant features and how they relate to each other.

36 The SEMATIA Treebank is available at https://sematia.hum.helsinki.fi and https://github.com/ezhenrik/sematia-tb (last access 2019.01.31).

37 (Vierros 2018).

38 There are 19,340 tokens for Ancient Greek and 1,400 tokens for Latin in the data sets available at https://github.com/ezhenrik/sematia-tb on 2018.10.13.

39 http://papyri.info/ (last access 2019.01.31).

40 (Vierros and Henriksson 2017).

41 See also Celano (2018).

There currently exist four dependency treebanks for Ancient Greek and Latin: the AGLDT contains texts from classical antiquity, while the IT-TB (mainly) contains Thomas Aquinas' *Summa Theologica*; the PROIEL Treebank originally contained the New Testament and some of its translations into Old Indo-European languages, but texts from classical antiquity have subsequently been added; the SEMATIA Treebank contains documentary papyri.

The AGLDT and the IT-TB share the same Latin syntactic guidelines, while the AGLDT and the SEMATIA treebank the same Ancient Greek and Latin syntactic guidelines. The PROEIL treebank has developed its own syntactic guidelines for both Ancient Greek and Latin, even though they are similar to the ones adopted by the AGLDT, all relying on the annotation guidelines for the analytical layer of the Prague Dependency Treebank. Some treebanks also provide a few texts annotated for semantics/pragmatics.

Despite some differences, it is safe to say that the morphosyntactic annotation of one treebank can be converted into that of another treebank to a large extent: this is also evidenced by the ongoing work to convert the original annotation schemes into the Universal Dependencies[42] one. The original format of all treebanks is XML, but none of them has so far adopted pure stand-off annotation (i.e., where tokens are referenced to the offsets of the original unannotated text).

Acknowledgements: This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation: project number 408121292).

Abbreviations

AGLDT = Ancient Greek and Latin Dependency Treebank
AGDT = Ancient Greek Dependency Treebank
IT = Index Thomisticus
IT-TB = Index Thomisticus Treebank
LDT = Latin Dependency Treebank
SG = H. W. Smyth's *Greek Grammar for Colleges,* 1920

42 http://www.universaldependencies.org (last access 2019.01.31).

Bibliography

Bamman, D.; Passarotti, M.; Crane, G.; Raynaud, R. (2007): Guidelines for the Syntactic Annotation of Latin Treebanks. https://github.com/PerseusDL/treebank_data/blob/master/v1/latin/docs/guidelines.pdf (last access 2019.01.31).

Bamman, D.; Crane, G. (2011): "The Ancient Greek and Latin Dependency Treebank". In: C. Sporleder; A. van Den Bosch; K. Zervanou (eds.): Language Technology for Cultural Heritage. Berlin and Heidelberg: Springer, 79–89.

Celano, G.G.A. (2014): Guidelines for the Annotation of the Ancient Greek Dependency Treebank 2.0. https://github.com/PerseusDL/treebank_data/edit/master/AGDT2/guidelines (last access 2019.01.31).

Celano, G.G.A.; Crane, G. (2015): "Semantic Role Annotation in the Ancient Greek Dependency Treebank". In: D. Markus; E. Hinrichs; A. Patejuk; A. Przepiórkowski (eds.): Proceedings of the Fourteenth International Workshop on Treebanks and linguistic Theories (TLT14). Warszawa: Institute of Computer Science, 26–34.

Celano, G.G.A; Crane, G.; Majidi, S. (2016): "Part of Speech Tagging for Ancient Greek". Open Linguistics 2:1, 393–399. https://doi.org/10.1515/opli-2016-0020.

Celano, G.G.A. (2018): "An Automatic Morphological Annotation and Lemmatization for the IDP Papyri". In: N. Reggiani (ed.): Digital Papyrology II. Berlin and Boston: De Gruyter, 139–148.

Crane, G. (1991): "Generating and Parsing Classical Greek". Literary and Linguistic Computing 6, 243–245.

Eckhoff, H.; Bech, K.; Bouma, G.; Eide, K.; Haug, D.; Haugen, O.E.; Jøhndal, M. (2017): "The PROIEL Treebank Family: A Standard for Early Attestations of Indo-European Languages". Language Resources and Evaluation 52, 29–65.

Hajič, J.; Panevová, J.; Buráňová, E.; UrešOvá, Z.; Bémová, A.; (in cooperation with) Kárník, J.; Štěpánek, J.; Pajas, P. (1999): Annotations at Analytical Level: Instructions for Annotators. https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/index.html (last access 2019.01.31).

Haug, D.T.T.; Jøhndal, M.L. (2008): "Creating a Parallel Treebank of the Old Indo-European Bible Translations". In: C. Sporleder; K. Ribarov (eds.): Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH). Association for Computational Linguistics, 27–34.

Haug, D.T.T.; Jøhndal, M.L.; Eckhoff, H.M.; Hertzenberg, M.J.; Müth, A. (2009): "Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages". Traitement automatique des langues 50:2, 17–45.

Haug, D.T.T. (2010): PROEIL Guidelines for Annotation. http://folk.uio.no/daghaug/syntactic_guidelines.pdf (last access 2019.01.31).

Haug, D.T.T.; Eckhoff, H.M.; Welo, E. (2014): "The Theoretical Foundations of Givenness Annotation". In: K. Bech; K.G. Eide (eds.): Information Structure and Syntactic Change in Germanic and Romance languages. Amsterdam: John Benjamins, 17–52.

McGillivray, B.; Passarotti, M.; Ruffolo, P. (2009): "The Index Thomisticus Treebank Project: Annotation, Parsing and Valency Lexicon". Traitement Automatique des Langues 2009, 103–127.

Mikulová, M.; Bémová, A.; Hajič, J.; Hajičová, E.; Havelka, J.; Kolárová, V.; Kučová, L.; Lopatková, M.; Pajas, P.; Panevová, J.; Razímová, M.; Sgall, P.; Štěpánek, J.;

Urešová, Z.; Veselá, K.; Žabokrtský, Z.; (translation) Součková, K.; Böhmová, A.; Čermáková, K.; Havelka, J.; Corness, P. (2006): Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank: Annotation Manual. http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html (last access 2019.01.31).

Pajas, P. (2010): The Prague Markup Language (version 1.1). http://ufal.mff.cuni.cz/jazz/pml/doc/pml_doc.pdf (last access 2019.01.31).

Passarotti, M. (2011): "The State of the Art of Latin and the Index Thomisticus Treebank Project". In: M. Ortola (ed.): Corpus Anciens et Bases de Données, ALIENTO. Échanges Sapientiels en Méditerranée 2, 301–320.

Passarotti, M.; Gonzáles Saavedra, B. (2015): Guidelines for Tectogrammatical Annotation of Latin Treebanks: The Treatment of some Specific Constructions. https://itreebank.marginalia.it/doc/Guidelines_tectogrammatical_Latin.pdf (last access 2019.01.31).

Passarotti, M. (2016): Rules of Annotation for the Analytical Layer of the Index Thomisticus Treebank.https://itreebank.marginalia.it/doc/Guidelines_analytical_latin_specific-constructions.pdf (last access 2019.01.31).

Smyth, H.W. (1920): Greek Grammar for Colleges. New York: American Book Company.

Treebank: Annotation Manual. http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html (last access 2019.01.31).

Vierros, M.; Henriksson, E.; (2017): "Preprocessing Greek Papyri for Linguistic Annotation". In: M. BüChler; L. Mellerin (eds.): Journal of Data Mining and Digital Humanities. Special Issue on Computer-Aided Processing of Inter-textuality in Ancient Languages. https://jdmdh.episciences.org/paper/view?id=1385 (last access 2019.01.31).

Vierros, M. (2018): "Linguistic Annotation of the Digital Papyrological Corpus: Sematia". In: N. Reggiani (ed.): Digital Papyrology II. Berlin and Boston: De Gruyter, 105–118.