Challenges in Converting the *Index Thomisticus* Treebank into Universal Dependencies

Flavio Massimiliano Cecchini and **Marco Passarotti**
Università Cattolica del Sacro Cuore, CIRCSE Research Centre
Largo Gemelli 1, 20123 - Milan, Italy
{flavio.cecchini}{marco.passarotti}@unicatt.it

Paola Marongiu
Università degli Studi di Pavia
Corso Strada Nuova 65, 27100 - Pavia, Italy
paola.marongiu01@universitadipavia.it

Daniel Zeman
Charles University in Prague, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 - Prague, Czech Republic
zeman@ufal.mff.cuni.cz

Abstract

This paper describes the changes applied to the original process used to convert the *Index Thomisticus* Treebank, a corpus including texts in Medieval Latin by Thomas Aquinas, into the annotation style of Universal Dependencies. The changes are made both to harmonise the Universal Dependencies version of the *Index Thomisticus* Treebank with the two other available Latin treebanks and to fix errors and inconsistencies resulting from the original process. The paper details the treatment of different issues in PoS tagging, lemmatisation and assignment of dependency relations. Finally, it assesses the quality of the new conversion process by providing an evaluation against a gold standard.

1 Introduction

Since release 1.2, Universal Dependencies (UD) (Nivre et al., 2016)[1] has been including treebanks for ancient languages or historical phases of modern ones. In the current release of UD (2.2), there are treebanks for Ancient Greek, Gothic, Latin, Old Church Slavonic, Old French and Sanskrit.

Among these languages, Latin is not only the one provided with most data in UD 2.2 (520K tokens), but also the one with the most treebanks (3). These are PROIEL (Haug and Jøhndal, 2008), which includes the entire New Testament in Latin (the so called *Vulgata* by Jerome) and texts from the Classical era (199K tokens), the Latin Dependency Treebank (LDT) by the Perseus Digital Library (Bamman and Crane, 2006), which collects a small selection of texts by Classical authors (29K tokens), and the *Index Thomisticus* Treebank (IT-TB) (Passarotti, 2011), based on works written in the XIIIth century by Thomas Aquinas (291K tokens).

The greater number of treebanks available for Latin than for other ancient languages reflects the large diachronic (as well as diatopic) span of Latin texts, which are spread across a time frame of more than two millennia and in most areas of what is called Europe today. This aspect is peculiar to Latin, which has represented for a long time a kind of *lingua franca* in Europe. The variety of textual typologies in Latin is thus wide: to name just a few, scientific treaties, literary works, philosophical texts and official documents were mostly written in Latin for centuries all around Europe. Today, this makes it impossible to build a textual corpus that can be sufficiently representative of "Latin", just because there are too many varieties of Latin, which can be even very different from each other.[2]

The three Latin treebanks were all developed before UD came into use and thus have been following a different annotation style. Although they are all dependency-based, only the IT-TB and the LDT have been sharing the same annotation

[1] http://universaldependencies.org/

[2] For instance, Ponti and Passarotti (2016) show the dramatic decrease of accuracy rates provided by a dependency parsing pipeline trained on the IT-TB when applied on texts of the Classical era taken from the LDT.

Proceedings of the Second Workshop on Universal Dependencies (UDW 2018), pages 27–36
Brussels, Belgium, November 1, 2018.

guidelines since the beginning of their respective projects (Bamman et al., 2007), while PROIEL has adopted a slightly different style.[3] The treebanks had been originally converted into the UD style by means of different and independent processes, which led to a number of inconsistencies in treating syntactic constructions as well as in part-of-speech (PoS) tagging and lemmatisation. In order to overcome such situation, a consensus has been achieved between the three projects with the aim of bringing the Latin treebanks closer to each other, establishing fundamental common criteria for both syntactic and morphological annotation.

In particular, so far the IT-TB has been always converted into UD through the same process used for the Prague Dependency Treebank for Czech (PDT) (Hajič et al., 2017), since both treebanks follow the same annotation style; just few modifications were made to cope with issues in PoS tagging.

This paper describes the changes applied to the original process of conversion from the IT-TB into the UD style, both to harmonise the IT-TB with the other Latin treebanks and to fix errors and inconsistencies during conversion. The result of the new conversion process is the UD version of the IT-TB that will be made available in the release 2.3 of UD, scheduled to be published in November 2018.

The paper is organised as follows. Section 2 describes the conversion process, by detailing its two phases, i. e. the so called *harmonisation*, which mostly deals with issues in PoS tagging and lemmatisation (Section 2.1), and the UD *conversion proper*, which is responsible for assigning dependency relations and rearranging the nodes in the syntactic trees to fit the UD annotation style (Section 2.2). Section 3 provides an evaluation of the conversion process. Finally, Section 4 concludes the paper and sketches some future work.

2 Conversion Process

The conversion process is performed via two sets of scripts, both written in Perl language[4] and embedded as modules in TREEX's[5] architecture. They consist of a preparatory *harmonisation* phase (Section 2.1), followed by the UD *conversion proper* (Section 2.2).

2.1 Harmonisation

Here, with *harmonisation* we mean adjusting a treebank to the PDT annotation style, with regard to the notation of lemmas, PoS and dependency relations. This is the starting point for the current UD conversion proper script (developed as part of the HamleDT project (Rosa et al., 2014)), which in a second phase infers morphological features and intervenes on the structure of the syntactic trees. In our case, harmonisation also includes making the IT-TB adhere to the agreed-upon annotation criteria for the three Latin treebanks, by means of a number of interdependent harmonisation scripts.

In what follows, we describe the most relevant issues that are dealt with during harmonisation of the IT-TB and their treatment in the script.

2.1.1 PoS Tagging of Inflectable Words

The syntactic annotation style of the IT-TB already substantially coincides with the PDT one (with the exception of one *afun*;[6] see Section 2.1.6). Hence, no substantial changes have to be carried out during harmonisation in this respect. However, the IT-TB does not distinguish PoS: instead, it applies a tripartite classification on a morphological basis between (a) nominal inflection, including nouns, adjectives, pronouns and numerals, (b) verbal inflection, including verbs, and (c) no inflection, including conjunctions, prepositions, adverbs and interjections.[7]

This means that, while words belonging to the class of verbal inflection (including also their nominal forms; see Section 2.1.2) can be readily assigned PoS VERB,[8] assigning a PoS to words of the other classes is not straightforward. To this end, we take advantage of the finer morphological classification provided by LEMLAT (Passarotti, 2004), where each inflectable nominal, adjectival and pronominal paradigm is treated differently. This gives us a PoS tagging for inflectable word classes, but not for uninflectable ones. From LEM-

[3] http://folk.uio.no/daghaug/syntactic_guidelines.pdf

[4] https://www.perl.org/

[5] TREEX is a modular software system in Perl for Natural Language Processing. It is described in (Popel and Žabokrtský, 2010) and available online at http://ufal.mff.cuni.cz/treex.

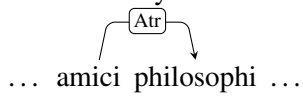
[6] *afun* means "*a*nalytical *fun*ction", which is the term used for syntactic labels in the surface syntax ("analytical") layer of annotation in the PDT. The corresponding term in UD is *deprel*, standing for "*dep*endency *rel*ation".

[7] Actually, the IT-TB also considers a fourth inflectional class to acknowledge the nominal inflections in verbal paradigms, like for instance for participles and gerunds.

[8] UD makes use of the Universal PoS tagset by (Petrov et al.).

LAT we thus obtain three lists of lemmas, respectively for nouns, adjectives and pronouns, which are hard-coded into the Perl script as look-up tables for PoS assignment (lemmas are already provided by the IT-TB annotation).

These lists are manually checked and partly corrected; indeed, some terms that are new to Thomistic Latin, or that have changed PoS or gained a new one in the passage from the Classical to the Medieval era[9] need to be added to their respective list or moved to a different one. This procedure does not resolve lexical ambiguity: for example, *philosophus* 'philosopher; philosophical' can function both as a noun and as an adjective. This ambivalence between noun and adjective can not be solved by look-up tables alone, but requires taking into account the syntactic behaviour of the word in the dependency tree. More precisely, if in the IT-TB the node in question is found to be dependent on another node and has *afun* `Atr` (attribute)[10] and they agree by case, number and gender, we will label it as an adjective; otherwise, as a noun. The genitive case needs to be excluded from this procedure, as one of its functions is to make a noun the attribute of another noun; e. g., a phrase like *amici philosophi*, where both words are in the genitive case, might be interpreted as 'of the philosophical friend' (noun *amicus* and adjective *philosophus*), 'of the philosopher's friend' (two nouns), or 'of the philosopher friend' (noun and nominal apposition). This ambiguity can not be solved *a priori*, as in the IT-TB all these three constructions yield the same annotation:



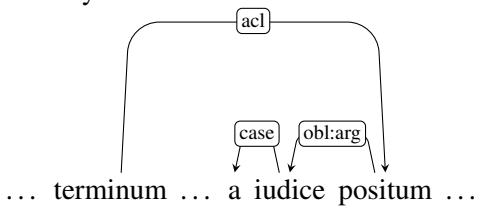
In general, the boundaries between adjectives and nouns are blurred. Thus, in those occurrences where an adjective is not assigned *afun* `Atr` in the IT-TB, we give it PoS NOUN in UD.

2.1.2 PoS of Verbal Nouns

Words belonging to the verbal inflectional class are always assigned PoS tag VERB, also when nominal forms are concerned (participles, gerunds, gerundives, supines).[11] Since a verbal noun is still able to take complements, the strongest argument in favour of this decision is that in the current version of UD nominals can not govern the same syntactic relations as verbs (e. g. no core/oblique distinction between complements is made). For example, in the sentence from *Summa contra gentiles*, Lib. III, Cap. CXXIX[12]

> *Transgredi autem terminum hunc a iudice positum, non est secundum se malum...*
> 'But to pass over a boundary line set up by a judge is not essentially evil...'

we have *positum* 'set up' (perfect participle of *pono*) acting as a modifier of *terminum* 'boundary line', whose Agent is represented by the prepositional phrase *a iudice* 'by a judge'. Our UD conversion yields:[13]



Here, if we were to use `amod` (adjectival modifier) instead of `acl`, we would not be able to identify *a iudice* as an agent, and for the corresponding node we should then choose between `nmod` (noun modifier) and *amod*, both however unsuitable to this context.[14]

In the IT-TB, the only possible identification of a verbal noun as an adjective or another nominal is made at the level of lemmatisation: some occurrences of e. g. *abstractus* 'abstract' (adjective), perfect participle of *abstraho* 'to drag away', are assigned their own adjectival lemma (reported in the look-up table) instead of the verbal one, on the basis of their lexicalisation.

2.1.3 PoS Tagging of Uninflectable Words

Words belonging to uninflectable classes (prepositions, conjunctions, adverbs, interjections) are all

[9] E. g. *sanctus* 'saint' was originally only a participial form of the verb *sancio* 'to ratify', but subsequently it was perceived and used also as an independent noun or adjective.

[10] For further details about *afun*s, see the annotation guidelines for PDT's analytical layer at `https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/`.

[11] A practical reference for Latin grammar is (Greenough and Allen, 2006).

[12] Here and thereafter, English translations of excerpts from *Summa contra gentiles* are taken from (Aquinas, 1955–1957). Those from *Scriptum super sententiis* are based on the Italian translation provided by (d'Aquino, 2001).

[13] `acl`: adjectival clause; `obl:arg`: oblique argument; `case`: case-marking element.

[14] The IT-TB is not the only treebank following this approach, another one being the Sanskrit treebank: `http://universaldependencies.org/treebanks/sa_ufal/index.html`.

labeled with a common PoS tag I (for "invariable") in LEMLAT.

To assign a Universal PoS tag to such words, we use a number of *ad hoc* rules relying on the original IT-TB syntactic annotation, where such words are assigned specific *afun*s: `AuxC` for subordinating conjunctions, `AuxZ` and `AuxY` for a closed subset of non-derived adverbs, and `Coord` for coordinating conjunctions.[15] All those uninflectable words that are not assigned a PoS by these *ad hoc* rules are considered to be non-derived adverbs.

2.1.4 PoS and Lemmas of Derived Adverbs

In the IT-TB, the lemma of a derived adverb is the adjective or the verb from which it is regularly formed. For example, *continue* 'continuously', *continuius* 'more continuously' (comparative) and *continuissime* 'most continuously' (absolute superlative) are all lemmatised under the adjective *continuus*, while the lemma for *abundanter* 'abundantly', *abundantius* 'more abundantly' and *abundantissime* 'most abundantly' is the verb *abundo*, on whose present participle (*abundans*) the adverb is formed. However, in UD Latin treebanks, the lemma of an adverb is defined to be its positive degree. In the examples above, we will thus have lemmas *continue* and *abundanter*.

To assign a PoS to derived adverbs, we exploit the original tagging of the IT-TB, which features a specific morphological tag for the "adverbial case", as this is considered to be part of the nominal inflection (so that e. g. *continue* is the adverbial case of *continuus*).

2.1.5 The Article

Latin does not feature the lexical category of the article, but all modern Romance languages descended from it, like Italian, have developed one. Remarkably, in the IT-TB we find 8 occurrences of the otherwise unattested word *ly*, as in *ly homo* 'the human being'. This is clearly an ancestor of the Italian definite article making its way in the XIIIth-century Latin of Thomas, whose mother tongue was a southern Italian variety. In the IT-TB, *ly* is then the only word receiving PoS DET (determiner); it does not show any inflection.

2.1.6 Verbal Complements

For what concerns the *afun* tagset, the only innovation of the IT-TB with respect to the PDT standard is the *afun* `OComp` for predicative complements (or secondary predicates), precisely for object complements (the *afun* `Pnom` being used for subject complements). For example, see *Summa contra gentiles*, Lib. II, Cap. XXXVIII (`OComp` highlighted):

> . . . *posuerunt mundum* ***aeternum***.
> '. . . (they) asserted the world's eternity.'
> lit. '. . . (they) supposed the world eternal.'

In UD this syntactic relation is represented by assigning the *deprel* `xcomp` (open clausal complement) to object complements. However, in the original version of the conversion script, `OComp` was equated to *afun* `Obj` (direct or indirect object) and as such erroneously translated into UD as *deprel* `obj`.[16] Since the harmonisation to the PDT style does not accept the `OComp` *afun*, we have to mark the affected nodes by using a "miscellaneous" field in the XML TREEX file, so that we will be able to treat `OComp` as a subcase of `Obj` later during conversion proper. A similar approach is also pursued for appositions (cf. Section 2.2.3).

2.2 UD Conversion Proper

The UD conversion script manages the relabeling of *afun*s into *deprel*s and, most importantly, rearranges the dependencies in the tree according to the UD style.

After describing the main differences between the IT-TB and UD annotation styles (2.2.1), in this Section we will focus on two syntactic constructions that we deem to be particularly challenging to tackle while adapting the conversion script to the IT-TB: namely, ellipsis (2.2.2) and apposition (2.2.3).

2.2.1 Differences between IT-TB and UD

The main difference between the IT-TB and UD styles is that in the IT-TB conjunctions, prepositions and copulas govern their phrases, while UD favours dependencies between content words, with function words tending to end up as leaves of the tree.[17] To illustrate this with an example, we consider the following excerpt from *Scriptum super sententiis* (Lib. I, Dist. III, Qu. II, Art. II):

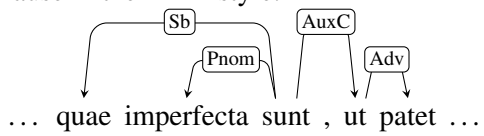
[15] `AuxY` is also assigned to coordinating conjunctions occurring in multiple coordinations (like . . . *et*. . . *et*. . . '. . . and. . . and. . . ').

[16] In the IT-TB, the *afun* `Obj` is also used for annotating oblique nominals expressing Result, Origin and Target (mostly) with motion verbs. As these are considered to be (non-core) arguments, they are assigned *deprel* `obl` (oblique nominals) with a specific subtype `arg` (argument).

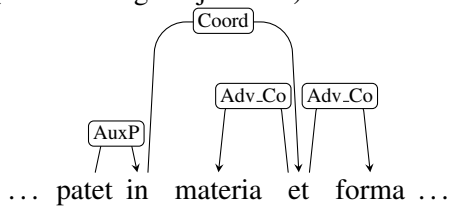
[17] The basic principles of UD are explained at `http://universaldependencies.org/u/overview/syntax.html`.

> . . . *quae imperfecta sunt, ut patet in materia et forma* . . .
> '. . . which are imperfect, as it clearly appears in matter and form. . . '

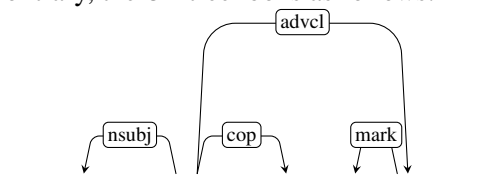
Here, *sunt* '(they) are' is a copula and *ut* 'as' is a conjunction introducing a subordinate clause. They both govern the predicate of their respective clause in the IT-TB style:[18]



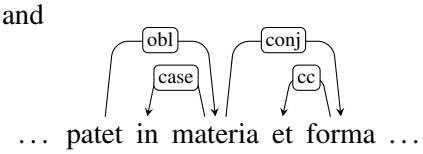
The same goes for *in* 'in' (preposition) and *et* 'and' (coordinating conjunction):[19]



Here, *in* and *et* govern the two conjuncts of the coordinated phrase *in materia et forma*. On the contrary, the UD tree looks as follows:[20]



and



Once a treebank is harmonised into a standard PDT-style form, the UD conversion script acts in two ways: (a) it translates all *afun*s into UD *deprel*s. This translation is not always biunivocal and is handled through a set of rules exploiting both morphological and syntactic annotation: e. g., *afun* `Adv` can correspond to different *deprel*s, like `advcl` or `advmod` (adverbial modifier); (b) it rearranges the nodes in the tree. TREEX features a number of specific modules to manage different kinds of constructions, such as coordinations and subordinate clauses. These Perl subroutines are language-independent and make use of the PoS, the morphological features and the *afun*s found in the source data. Thus, even after harmonisation, the basic conversion script is still inadequate to properly handle a language-specific treebank. Therefore, we have to tune the script to better address the specific needs of the IT-TB.

We will illustrate this point with the aid of two constructions: ellipsis and apposition.[21]

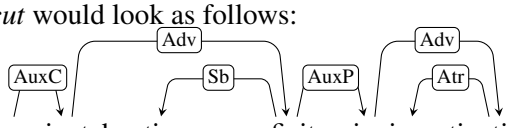
2.2.2 Ellipsis

The IT-TB and UD styles treat ellipsis quite differently, in a way that is not directly related to the UD primacy of content words. To clarify this point, we will use the following excerpt from the IT-TB (*Scriptum super sententiis*, Lib. IV, Dist. VII, Qu. I, Art. III):

> *In illis autem sacramentis quae perficiuntur in usu materiae, sicut baptismus [perficitur] in ipsa tinctione. . .*
> 'In those sacraments, however, which are accomplished through the use of matter, like baptism [is accomplished] through the submersion itself. . . '

The text in square brackets (a verb) is the elided part of the sentence. In the IT-TB, the only recorded ellipses, i. e. constructions for which the *afun* `ExD` (external dependency) is used, are those of verbal elements. On the contrary, nominal ellipses are not explicitly marked in the annotation. Therefore, in the following we will consider verbal ellipses only.

In the IT-TB style, if ellipsis resolution were applied, the comparative clause introduced by *sicut* would look as follows:



Since the node for *perficitur* is missing, the nodes for *baptismus* and *in* (head of *tinctione*), lacking their governor, become children of their closest

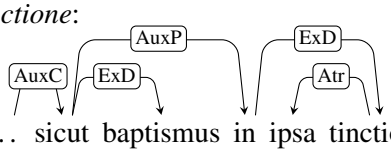
[18] `Sb`: subject; `Pnom`: nominal predicate; `AuxC`: subordinating conjunction; `Adv`: adverbial.

[19] `AuxP`: adposition; `Coord`: coordinating element; `Co` adscript: member of a coordination.

[20] `nsubj`: nominal subject; `cop`: copula; `advcl`: adverbial clause; `mark`: marker introducing a finite subordinate clause; `obl`: oblique nominal (see footnote 17); `conj`: conjunct; `cc`: coordinating conjunction. The complete list of *deprel*s and their explanations can be found at `http://universaldependencies.org/u/dep/index.html`.

[21] Ellipsis and apposition are challenging constructions where different UD teams have faced similar problems and sometimes found different, yet compatible, solutions. Discussion about the treatment of such constructions in different languages can be found in (Aranzabe et al., 2014), (Dobrovoljc and Nivre, 2016), (Pyysalo et al., 2015), (Tandon et al., 2016) and (Zeman, 2015).

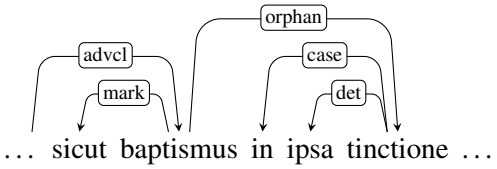
ancestor (in this case *sicut* for both) and are assigned *afun* `ExD`. Since nodes labeled with `AuxP`, `AuxC` or `Coord` can never take the *afun* `ExD`, this percolates down the tree to the first content word. Here, this happens from *in* to *tinctione*:



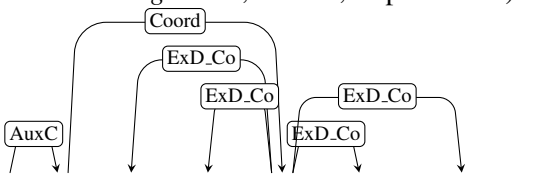
In UD a member of the elliptical clause is promoted to clause's head on the basis of its *coreness* value[22] and receives the *deprel* that would have been otherwise assigned to the elided predicate. The remaining nodes of the clause become its children and are assigned the special *deprel* `orphan` to avoid misleading dependencies.[23]

For elliptical constructions, the task of our conversion script is then to identify one of the `ExD` siblings in the IT-TB source data as the node to promote to head of the elliptical clause in UD. Following the UD guidelines, we consider a coreness hierarchy that gives precedence to a subject over an object, to an object over an indirect object, to an indirect object over an oblique one, and generally to core complements over peripheral ones. Now, the *afun* `ExD` obscures such relations. However, we can retrieve this information heuristically, by exploiting the rich Latin morphology (word order being much less meaningful) and cross-checking it with the PoS assigned during harmonisation.

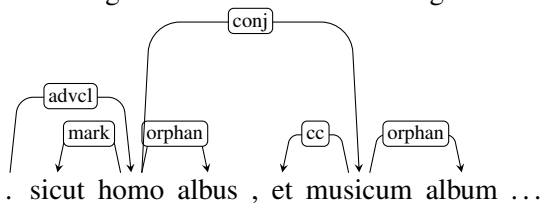
In the example above, the conversion script has to choose the head of the elliptical clause between *baptismus* and *tinctione* (*tinctione* being the content word, and thus the UD head, in its prepositional phrase). Both are nominals (with PoS NOUN assigned by harmonisation), but the fact that *baptismus* is in the nominative case, while *tinctione* is in the ablative (lemma *tinctio*) tells us that the former is most probably the subject of the elliptical clause, while the latter is an oblique complement. Hence, the script promotes *baptismus* and restructures the subtree as follows:



Such approach shows some limitations, especially when dealing with coordinating constructions, which are quite tricky when paired with elliptical constructions. Indeed, *a priori* it is not possible to set a hierarchy of the `ExD` siblings occurring in a coordination, since they all equally depend on one common coordinating element. For example (*Summa contra gentiles*, Lib. III, Cap. LXXIV):



This clause means "just like a man [is] white and a musical being [is] white". First, we know that the `ExD` siblings need to be distributed among the (at least) two members of the coordination, but, in principle, we do not know this distribution: e. g., both *homo albus*/*musicum album* and *homo*/*albus musicum album* might be valid splits.[24] To address this issue, we implement a heuristic approach that takes into account both frequently used separators (like commas and conjunctions) and word order to identify the most probable boundaries between coordination members; in the example above, the two members *homo albus* and *musicum album* are separated by the coordinating conjunction *et*. Second, head promotion for elliptical constructions takes place according to the PoS hierarchy described above: in our example, the nouns *homo* and *musicum* become governors of the adjectives *albus* and *album* respectively, via *deprel* `orphan`. The resulting UD subtree is the following:[25]



As it clearly stands out from the UD subtree above, in such a case our conversion fails. Here, the adjectives are nominal predicates and only their

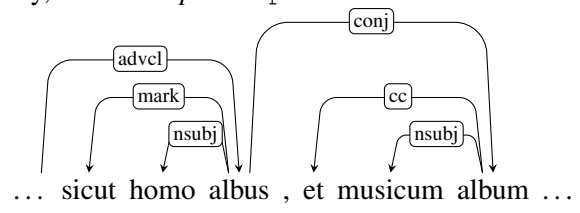
[22] See the UD guidelines at http://universaldependencies.org/u/overview/specific-syntax.html#ellipsis.

[23] Again, this does not apply to function words like conjunctions and prepositions, which keep their *deprel*.

[24] The latter is probably not grammatical, but we are working at a very shallow level here.

[25] `cc`: coordinating conjunction; `conj`: conjunct.

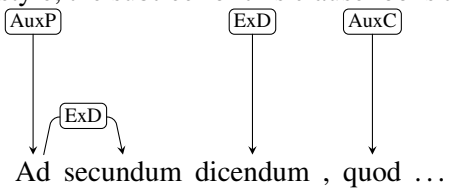
copulas (*est*) are missing, so that the correct dependencies should be assigned in the opposite way, with no *deprel* `orphan` involved:



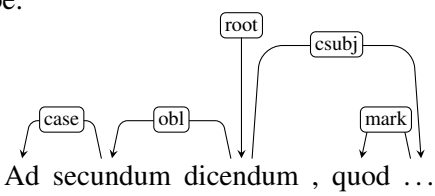
Being aware of such limitations, when treating specific elliptical constructions we print different kinds of warnings at the end of the conversion process to support a subsequent manual revision.

In the previous version of the IT-TB conversion script, ellipsis was not dealt with at all, providing the TREEX modules with no clues about how to interpret such constructions correctly. The way we treat elliptical constructions exemplifies how to take advantage of properties of a language like Latin to address linguistic issues that impact the UD conversion.

A particular case of ellipsis is the omission of the auxiliary verb *sum* 'to be' in the gerundive construction when occurring at the beginning of a sentence, e. g. in a frequent formula of the type *Ad secundum dicendum [est], quod...* 'Secondly, [it has] to be said that...'. According to the IT-TB style, the subtree for this clause looks as follows:



The nodes for *ad*, *dicendum* and *quod* directly depend on the root as a consequence of the missing root node for *est*. The conversion script promotes *dicendum* to the head, as verbs have priority over nominals. The children of *dicendum* are then assigned the correct *deprel* (instead of `orphan`), by using heuristics similar to those to establish coreness hierarchy. In the end, the UD subtree will be:[26]



In the UD subtree, the elided node for *est* would be a child of the node for *dicendum* with *deprel* `aux:pass`. This is a case where an elliptical construction represented in the IT-TB style is not apparent anymore in UD, because the primacy of content words obscures the ellipsis of *est* in the UD subtree.

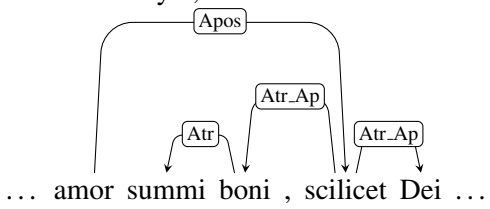
[26] `csubj`: clausal subject.

2.2.3 Apposition

Just like in the PDT style, in the IT-TB an apposition is defined as a binary relation where one phrase or clause is reworded or specified in some way by another following phrase or clause, which is separated from the first one by punctuation or a grammatical element.[27] In the IT-TB, this element is in most cases *scilicet* 'that is, namely', less frequently *sicut* 'as', like in *Summa contra gentiles*, Lib. III, Cap. CXVI:

> *... amor summi boni, scilicet Dei...*
> '...the love of the highest good, namely, God...'

In the IT-TB style, we have:[28]



Apposition in this sense can take place for any noun, verb or adverb phrase. However, the definition of the UD *deprel* `appos` is stricter[29] and limited to a noun immediately following another one and specifying it, like in *Moyses, propheta iudaeorum* 'Moses, prophet of the Jews', where *propheta* is assigned *deprel* `appos` and is made dependent on the node for *Moyses*.

This means that we can not translate the IT-TB *afun* `Apos` directly into the UD *deprel* `appos`, but have to resort to other *deprel*s expressing modifiers, according to their appropriateness. These include `acl`, `nmod`, `amod`, `advmod` (adverbial modifier) and `advcl`. Anyway, according to the definitions of such *deprel*s in the current UD guidelines, none of them is suitable to express (and thus convert) the joint, coordination-like relationship holding between the two members of an apposition as meant in the IT-TB. In particular, the status of *scilicet* remains unclear, as it can neither

[27] https://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/html/ch03s04x12.html

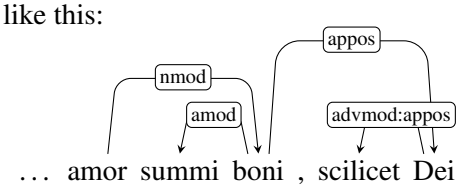
[28] `Apos`: Apposition (assigned to the connecting element); `Ap` adscript: member of an apposition.

[29] http://universaldependencies.org/u/dep/appos.html

be considered an adverbial modifier (it introduces, but does not modify the apposition), nor a coordinating conjunction (*deprel* `cc`).

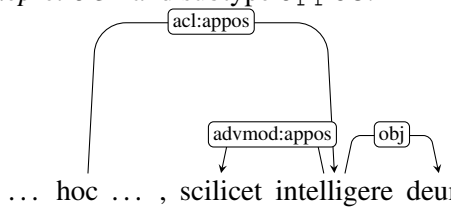
We address this issue by assigning a specific subtype `appos` (a) to appositive adverbial modifiers like *scilicet*, (b) to non-nominal appositions and (c) to appositions whose second member does not immediately follow the first one.

Our UD subtree for the example above will look like this:



Here we can use the *deprel* `appos` since the apposition is made of two nominals (*boni*, lemma *bonum*, and *Dei*, lemma *Deus*).

A case of two non-nominals involved in an apposition is the following (*Summa contra gentiles*, Lib. III, Cap. XXV), where the second member of the apposition (*scilicet intelligere deum* 'namely, to understand God') is an attributive clause modifying the pronoun *hoc* 'this' and it is thus assigned *deprel* `acl` and subtype `appos`:



Treating appositions also requires a quite substantial rearrangement of the nodes in an IT-TB subtree prior to the UD conversion proper, including a complex system of cross-references in the Perl script to reconstruct all considered syntactic dependencies, that was completely absent from the original conversion script.

3 Evaluation

We perform an evaluation to assess to what degree our modifications to the IT-TB–UD conversion process impact the quality of the conversion. To this aim, we first build a gold standard that we use as a benchmark for our data.

The 2.2 UD version of the IT-TB includes 21 011 sentences (291K tokens), 17 721 of which pertain to the first three books of *Summa contra gentiles*, the remaining 3 290 being the concordances of the lemma *forma* 'form' from a selection of works of Thomas Aquinas. We randomly extract 994 sentences out of the IT-TB and check that they are balanced and representative of the whole treebank according to a number of topological and annotation parameters.[30] Then, the gold standard is built by manually checking the output of the automatic conversion of these 994 sentences into the UD style and fixing the mistakes.

	LAS	LA	UAS	PoS	Lemma
Orig.	84.8	87.9	94.2	95.5	95.2
New	97.0	98.0	98.3	98.5	99.8

Table 1: Evaluation of original and new conversion.

Finally, we compare the gold standard with (a) the output of our new conversion process and (b) the output of the original conversion process. We compute the rates for the usual evaluation metrics of dependency parsers: LAS (Labeled Attachment Score), LA (Label Accuracy) and UAS (Unlabeled Attachment Score) (Buchholz and Marsi, 2006). Table 1 shows the results together with the accuracy rates for PoS tagging and lemmatisation, as a way to evaluate the harmonisation phase too.

Results reveal a general improvement of the quality of conversion. In particular, there is a substantial increase in LAS, while this is smaller for what concerns UAS. This shows that, while the basic TREEX conversion modules are already capable of addressing well the rearrangement of some subtrees required by the conversion to UD, they nonetheless need and greatly benefit from a language-specific fine-tuning, mainly but not only for what concerns the assignment of *deprel*s.

4 Conclusion

We presented the new conversion process of the *Index Thomisticus* Treebank of Medieval Latin into the Universal Dependencies annotation style. We detailed the changes applied not only to make the IT-TB consistent with the other UD treebanks, but also to harmonise it with the other Latin treebanks available in the UD dataset. This aspect is particularly relevant, because the wide diachronic and diatopic span of Latin language requires to collect (and annotate) several sets of textual data to represent its different varieties. These corpora need to follow a common set of guidelines for annotation so as to enable users to run queries pro-

[30] Length of the sentence; depth of trees; cases of elliptical constructions (`ExD`) and of coordination chains (a `Coord` governing another `Coord`); distribution of PoS and *afun*s.

viding results that support research in comparative linguistics, as well as to train stochastic NLP tools.

Beside harmonisation, refining the original conversion process has opened questions concerning the annotation of specific constructions. This is e. g. the case of appositions, where our decision is to use the subtype `appos` to address structures that are not yet considered in the current UD guidelines. We hope that our solution will be helpful also for other treebanks getting through similar problems.

Given the good quality of the conversion, as shown by our evaluation, after publishing the new version of the IT-TB in the release 2.3 of UD, we plan to start working on enriching the treebank with enhanced dependencies.

The current harmonisation and conversion scripts can be downloaded from the Github pages of the TREEX and HamleDT projects.[31]

Acknowledgments

Marco Passarotti gratefully acknowledges the support of the project LiLa (Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin). This project has received funding from the European Research Council (ERC) European Union's Horizon 2020 research and innovation programme under grant agreement No 769994.

References

Thomas Aquinas. 1955–1957. *Summa contra gentiles*. Hanover House, New York, NY, USA. Accessible at https://dhspriory.org/thomas/ContraGentiles.htm.

S. Tommaso d'Aquino. 2001. *Commento alle sentenze di Pietro Lombardo*. Edizioni Studio Domenicano, Bologna, Italy.

Maria Jesus Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz, Iakes Goenaga de Ilarraza, Koldo Gojenola, and Larraitz Uria. 2014. Automatic conversion of the basque dependency treebank to universal dependencies. In *Proceedings of the Fourteenth Workshop on Treebanks an Linguistic Theories (TLT 14)*, pages 233–241, Warszawa, Poland. Polish Academy of Sciences.

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 67–78, Prague, Czech Republic. Univerzita Karlova.

[31] https://github.com/ufal/treex; https://github.com/ufal/hamledt.

David Bamman, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of Latin treebanks. *Tufts University Digital Library*.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York, USA. Association for Computational Linguistics.

Kaja Dobrovoljc and Joakim Nivre. 2016. The universal dependencies treebank of spoken slovenian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1566–1573, Paris, France. European Language Resources Association (ELRA).

James B Greenough and JH Allen. 2006. *Allen and Greenough's new Latin grammar*. Dover publications, Mineola, NY, USA.

Jan Hajič, Eva Hajičová, Marie Mikulová, and Jiří Mírovský. 2017. Prague Dependency Treebank. In Nancy Ide and James Pustejovsky, editors, *Handbook of Linguistic Annotation*, pages 555–594. Springer, Dordrecht, Netherlands.

Dag TT Haug and Marius Jøhndal. 2008. Creating a parallel treebank of the old Indo-European Bible translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Marco Passarotti. 2004. Development and perspectives of the Latin morphological analyser LEMLAT. *Linguistica computazionale*, XX-XXI:397–414.

Marco Passarotti. 2011. Language resources. The state of the art of Latin and the *Index Thomisticus* treebank project. In *Corpus anciens et Bases de données*, number 2 in ALIENTO. Échanges sapientiels en Méditerranée, pages 301–320, Nancy, France. Presses universitaires de Nancy.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *ArXiv e-prints*. arXiv:1104.2086 at https://arxiv.org/abs/1104.2086.

Edoardo Maria Ponti and Marco Passarotti. 2016. Differentia compositionem facit. A slower-paced and reliable parser for Latin. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 683–688, Portorož, Slovenia. European Language Resources Association (ELRA).

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: modular NLP framework. In *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin - Heidelberg, Germany. Springer.

Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. 2015. Universal dependencies for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (Nodalida 2015)*, pages 163–172, Linköping, Sweden. Linköping University Press.

Rudolf Rosa, Jan Mašek, David Mareček, Martin Popel, Daniel Zeman, and Zdeněk Žabokrtský. 2014. HamleDT 2.0: Thirty dependency treebanks Stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2334–2341, Reykjavík, Iceland. European Language Resources Association (ELRA).

Juhi Tandon, Himani Chaudhry, Riyaz Ahmad Bhat, and Dipti Sharma. 2016. Conversion from paninian karakas to universal dependencies for hindi dependency treebank. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 141–150, Berlin, Germany. Association for Computational Linguistics.

Daniel Zeman. 2015. Slavic languages in universal dependencies. In *Natural Language Processing, Corpus Linguistics, E-learning (proceedings of SLOVKO 2015)*, pages 151–163, Bratislava, Slovakia. RAM-Verlag.