Marco Passarotti

The Project of the Index Thomisticus Treebank

Abstract: The paper introduces the project of the Index Thomisticus Treebank (IT-TB). The IT-TB is a dependency-based treebank based on the corpus of the Index Thomisticus by father Roberto Busa (IT), which includes the *opera omnia* of Thomas Aquinas, for a total of approximately 11 million words. Currently, the IT-TB is the largest Latin treebank available, with more than 350,000 nodes in around 17,000 sentences. The annotation covers the entire books 1, 2 and 3 of *Summa contra Gentiles*, plus excerpts from *Scriptum super Sententiis Magistri Petri Lombardi* and *Summa Theologiae*. The paper details the multi-layer annotation style of the IT-TB and its background theoretical motivations. The conversion process to the now widely used Universal Dependencies style is described as well. Across more than a decade, the project has developed a number of linguistic resources and NLP tools for Latin connected to the IT-TB. As for the resources, the paper presents the syntax-based subcategorization lexicon IT-VaLex and the valency lexicon Latin Vallex. As for the tools, the automatic dependency parsing process is described, highlighting the core issue of portability of NLP tools across the wide diachronic and diatopic span of Latin texts. A section is dedicated to automatic morphological analysis of Latin, introducing the analyzer Lemlat and its recent enhancement with information on derivational morphology and a new set of lexical entries covering a large *Onomasticon* (from Forcellini dictionary) and Medieval Latin (from Du Cange glossary).

1 Introduction

The name of the Italian Jesuit Roberto Busa is quoted in almost every introduction to Computational Linguistics or Digital Humanities. His often recounted

Note: The author 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.

Marco Passarotti, Università Cattolica del Sacro Cuore, Milano

https://doi.org/10.1515/9783110599572-017

meeting in New York with the founder of IBM, Thomas Watson Sr., in 1949 is considered one of the funding moments of the discipline.[1]

Similarly, the Index Thomisticus (IT), the most important outcome of that meeting, is usually mentioned among the first annotated textual corpora available in machine-readable format.[2] The result of thirty years of work and funding from IBM, the IT contains the *opera omnia* of Thomas Aquinas (118 texts) as well as 61 texts by other authors related to Thomas, for a total of approximately 11 million tokens. The corpus is morphologically tagged and lemmatized and it is available on paper, CD-ROM and on-line (http://www.corpusthomisticum.org).

Already at the time when the IT was just published, Busa planned to enhance the corpus with syntactic metadata. After a number of pilot attempts since the Nineties, the process of syntactic annotation of the IT started in 2006 with the so-called Index Thomisticus Treebank (IT-TB; http://itreebank.marginalia.it), which today represents the largest syntactically annotated corpus for Latin available.

Father Busa, who died in 2011, had the opportunity to see the start of the project and followed its first steps. In December 2009, he gave his last speech at a scientific event, the eighth edition of the international workshop on *Treebanks and Linguistic Theories* (http://tlt8.unicatt.it). The talk of Busa was entitled *From Punched Cards to Treebanks: 60 Years of Computational Linguistics*. The following excerpt from the unpublished transcription of that talk epitomizes both the objective and the motivation of the IT-TB:

> The [. . .] aim is to construct a summa of the entire syntax of Aquinas with statistics and percentages of each grammatical element, including punctuation marks (this is the Index Thomisticus Treebank project): this will then serve as a yardstick to compare or contrast the Latin grammar of St Thomas with that of others in other languages as well.

The objective of Busa was huge: to perform the syntactic annotation of the entire corpus of Thomas Aquinas' works not only to get a deep knowledge of his language and, thus, philosophy, but also to be able to compare Latin with other languages. This sounds like a plan perfectly fitting the needs of current research in the area of linguistic resources. The Universal Dependencies project (http://universaldependencies.org), which the IT-TB takes part of, represents

1 (Passarotti 2013, 17).
2 (Busa 1974–1980).

today the most rising effort from the research community to build a common annotation style for an ever growing number of languages. Starting from the empirical description of the syntactic constructions of a single language, this can be compared with those of other languages thanks to shared formats, schemes and tools. The IT-TB today contributes to such common effort, providing evidence about the specific variety of Latin represented by the works of Thomas Aquinas.

Across more than a decade, the IT-TB has grown into a larger project, which has gone beyond the construction of the treebank of Thomas Aquinas' texts. Starting from the IT-TB, the project has built a number of other linguistic resources and tools for automatic processing of Latin, making the CIRCSE research center in Milan (where the project is run since its beginning) an internationally known hub in the field and contributing to lead Latin out of its status of under-resourced language, which was still the case in mid 2000s when the IT-TB was started.

This paper wants to provide an overview of the IT-TB project, by detailing both the theoretical and the practical aspects connected to the building and the use of its resources and Natural Language Processing (NLP) tools for Latin.

The paper is organized as follows. Section 2 describes the main linguistic resource of the project, namely the IT-TB, presenting the theoretical framework supporting its annotation style, and its recent conversion into the Universal Dependencies style. Section 3 details two lexical resources strictly related to the IT-TB: the syntactic subcategorization lexicon IT-VaLex and the valency lexicon Latin Vallex. Section 4 deals with NLP tools for Latin. First, it introduces the version 3.0 of the Latin morphological analyzer Lemlat, particularly focusing on its enhancement with information about derivational morphology. Second, it presents the state of the art of automatic dependency parsing of the IT-TB, sketching the problem of portability of NLP tools for Latin across time and space. Finally, Section 5 concludes the paper by discussing a number of open challenges in the field and by looking at the near future of the IT-TB project as well as of the several linguistic resources and NLP tools for Latin built so far, presenting the objectives of the new ERC-Consolidator Grant LiLa, which is run at CIRCSE.

2 The Index Thomisticus Treebank

2.1 Theoretical background

The IT-TB is a dependency treebank based on a subset of the IT. The project is carried out at the CIRCSE research center of the Università Cattolica del Sacro Cuore in Milan, Italy (http://centridiricerca.unicatt.it/circse).[3]

The dependency-based annotation style of the IT-TB is grounded on Functional Generative Description (FGD),[4] a theoretical framework developed in Prague and intensively applied and tested while building the Prague Dependency Treebank of Czech (PDT).

FGD is based on the assumption that language must be considered as a form-meaning composite. Consistently and like the PDT, the IT-TB features three layers of annotation ordered as follows:[5]

(1) a morphological layer: disambiguated morphological annotation and lemmatization;
(2) an "analytical" layer: annotation of surface syntax (the "form");
(3) a "tectogrammatical" layer: annotation of underlying syntax (the "meaning").

Both analytical and tectogrammatical layers describe the sentence structure with dependency tree-graphs, respectively named "analytical tree structures" (ATSs) and "tectogrammatical tree-structures" (TGTSs).

In ATSs every word and punctuation mark of the sentence is represented by a node of a rooted dependency tree. The edges of the tree correspond to dependency relations labeled with (surface) syntactic functions called "analytical functions" (like Subject, Object, etc.).

3 (Passarotti 2010). The IT-TB is freely available from the IT-TB website under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License. Data can be queried by using PML Tree Query (PML-TQ), a highly portable query language and search engine (Pajas and Štěpánek 2009). PML-TQ is available both as a local extension of the tree editor TrEd (http://ufal.mff.cuni.cz/tred/) and as an on-line implementation which, in the case of the IT-TB, enables users to run queries on the linguistic resources of the IT-TB project (http://itreebank.marginalia.it/view/resources.php). The portion of the IT-TB annotated at the analytical layer is accessible also through the web-based treebank search and visualization application TüNDRA (Martens and Passarotti 2014) as part of the web infrastructure of linguistic resources and tools CLARIN (https://www.clarin.eu: last access 2019.01.31).

4 (Sgall et al. 1986).

5 (Hajič et al. 2000).

TGTSs describe the underlying structure of the sentence, conceived as the semantically relevant counterpart of the grammatical means of expression (described by ATSs). The nodes of TGTSs represent content words only, while function words and punctuation marks are left out. The nodes are labeled with semantic role tags called "functors", which are divided into two classes according to valency: (a) arguments, called "inner participants", i.e. obligatory complements of verbs, nouns, adjectives and adverbs: Actor, Patient, Addressee, Effect and Origin; (b) adjuncts, called "free modifications": different kinds of adverbials, like Place, Time, Manner etc. TGTSs feature two dimensions that represent respectively the syntactic structure of the sentence (the vertical dimension) and its information structure ("topic-focus articulation", TFA), based on the underlying word order (the horizontal dimension). Also ellipsis resolution and coreference analysis are performed at the tectogrammatical layer and are represented in TGTSs through newly added nodes (ellipsis) and arrows (coreference).

2.2 Analytical layer

During the first three years of the project, the analytical annotation of the IT-TB was performed fully manually. Since 2009, analytical data are annotated in semi-automatic fashion by using various combinations of stochastic parsers trained on different subsets of the IT-TB (see Section 4.1), whose output is manually checked by two human annotators.

Currently the number of analytically annotated nodes in the IT-TB is around 370,000, corresponding to approximately 23,000 sentences excerpted from three works of Thomas Aquinas: *Scriptum super Sententiis Magistri Petri Lombardi* (*Sent.*), *Summa contra Gentiles* (*ScG*) and *Summa Theologiae* (*ST*). In particular, the IT-TB includes the following texts annotated at the analytical layer:

A. concordances of the lemma *forma* in *Sent.*, *ScG* and in the first 76 *quaestiones* of *ST*;
B. entire first, second and third books and chapters 1–11 of the fourth book of *ScG*.

Analytical annotation is performed according to a specific manual for the syntactic annotation of Latin treebanks,[6] which was developed on the basis of the PDT guidelines for analytical annotation.[7]

6 (Bamman et al. 2007).
7 (Hajič et al. 1999).

Figure 1 reports the ATS of the following sentence from the IT-TB: “tunc enim unaquaeque res optime disponitur cum ad finem suum convenienter ordinatur;” (‘So, each thing is excellently arranged when it is properly directed to its purpose;’) (*ScG* I, ch. 1, no. 2).

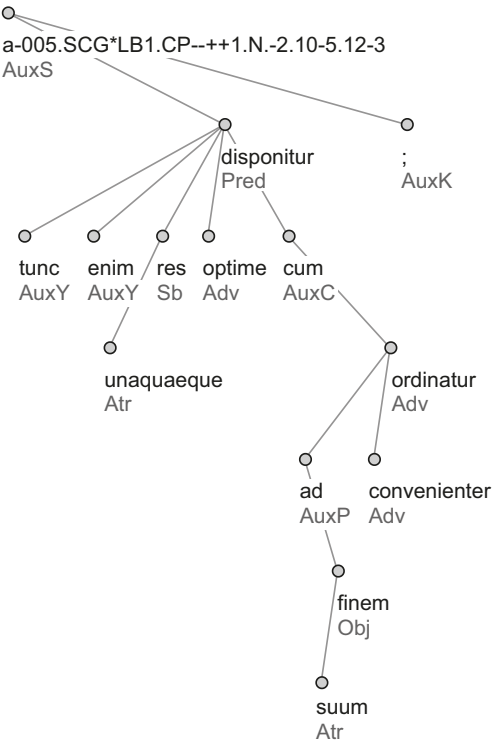

Figure 1: An analytical tree structure.

Except for the technical root of the tree (which reports the textual reference of the sentence), each node in the ATS corresponds to either one word or punctuation mark in the sentence. Nodes are arranged from left to right according to surface word order; they are connected in governor-dependent fashion and each relation is labeled with an analytical function. For instance, the relation between the word *res* and its governor *disponitur* is labeled with the analytical function Sb (Subject), i.e. *res* is the subject of *disponitur*. Four kinds of analytical functions that occur in the tree are assigned to auxiliary sentence members, namely AuxC (subordinating conjunctions: *cum*), AuxK

(terminal punctuation marks), AuxP (prepositions: *ad*) and AuxY (sentence adverbs: *enim, tunc*).[8]

2.3 Tectogrammatical layer

The tectogrammatical annotation workflow of the IT-TB is based on TGTSs automatically converted from ATSs.[9] Conversion is performed by adapting to Latin a number of ATS-to-TGTS scripts provided by the NLP framework Treex.[10] The TGTSs that result from conversion are then checked and refined manually by two independent annotators. The annotation guidelines are those for the tectogrammatical layer of the PDT.[11]

So far, the first 2,000 sentences of *ScG* have been fully annotated at tectogrammatical level (corresponding to approximately 28,000 nodes).[12] Figure 2 shows the TGTS corresponding to the ATS of the sentence reported in Figure 1.

Since only nodes for content words can occur in TGTSs, auxiliary sentence members labeled with analytical functions AuxC, AuxK and AuxP are collapsed. Analytical functions are replaced with functors. The nodes for the lemmas *enim* and *tunc* are both assigned the functor PREC, since they represent expressions linking the clause to the preceding context; they are given node-type "atom" (atomic nodes), which is used for adverbs of attitude, intensifying or modal expressions, rhematizers and text connectives.[13] *Res* is the Patient (PAT) of *dispono*, as it is the syntactic subject of a passive verbal form (*disponitur*).[14] Both the adverbial forms of *bonus* (*optime*) and *convenio* (*convenienter*) are labeled with functor MANN, which expresses manner by specifying an

8 The other analytical functions occurring in this sentences are the following: Adv (adverbs and adverbial modifications, i.e. adjuncts), Atr (attributes), AuxS (root of the tree), Obj (direct and indirect objects), Pred (main predicate of the sentence).

9 (González Saavedra and Passarotti 2014).

10 (Popel and Žabokrtský 2010).

11 (Mikulová et al. 2006).

12 Also some texts excerpted from the Latin Dependency Treebank of Classical Latin (LDT; Bamman and Crane 2007) were annotated at the tectogrammatical layer in the context of the IT-TB project. In particular, these are 100 sentences from Caesar and Cicero, and the entire text of *Bellum Catilinae* by Sallust (Passarotti and González Saavedra 2018).

13 (Mikulová et al. 2006, 17).

14 Conversely, syntactic subjects of active verbal forms are usually labeled with the functor ACT (Actor). However, this does not always hold true, since the functor of the subject depends on the semantic features of the verb.

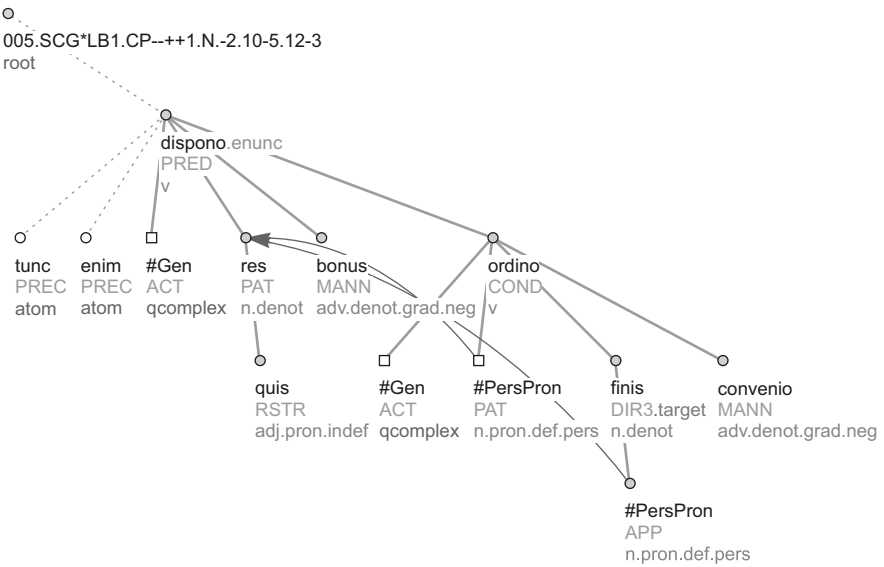

Figure 2: A tectogrammatical tree structure.[15]

evaluating characteristic of the event, or a property. *Unusquisque* is a pronominal restrictive adnominal modification (RSTR) that further specifies the governing noun *res*. The clause headed by *ordinatur* (lemma: *ordino*) is assigned the functor COND, as it reports the condition on which the event expressed by the governing verb (*disponitur*; lemma: *dispono*) can happen. The lemma *finis* is assigned the functor DIR3 (Directional: to), which expresses the target point of the event. *Finis* is then specified by an adnominal modification of appurtenance (APP).

Three newly added nodes occur in the tree (square nodes), to provide ellipsis resolution of those arguments of the verbs *dispono* and *ordino* that are missing in the surface structure. *Dispono* is a two-argument verb, the two arguments being respectively the Actor and the Patient, but only the Patient is explicitly expressed in the sentence, i.e. the syntactic subject *res*. The missing argument, i.e. the Actor (ACT), is thus replaced with a "general argument" (#Gen), because the coreferred element of the omitted modification cannot be clearly identified

15 In the default visualization of TGTSs, word forms are replaced with lemmas.

with the help of the context. The same holds also for the Actor of the verb *ordino* (#Gen), whose Patient (#PersPron, PAT) is coreferential with the noun *res*, as well as the possessive adjective *suus* (#PersPron, APP). In the TGTS, these coreferential relations are shown by the blue arrows linking the two #PersPron nodes with the node for *res*.[16]

The nodes in the TGTS are arranged from left to right according to TFA, which is signaled by the color of the nodes (white nodes: topic; yellow nodes: focus). A so-called "semantic part of speech" is assigned to every node: for instance, "denotational noun" is assigned to *finis*.[17] Finally, the illocutionary force class informing about the sentential modality is assigned to the main predicate of the sentence *dispono* ("enunciative").

2.4 The Index Thomisticus Treebank in Universal Dependencies

Universal Dependencies (UD)[18] is one of the most notable projects currently ongoing in computational linguistics. The project, run by contributors from the research community, aims at creating a collection of dependency treebanks for different languages built according to a cross-linguistically consistent annotation style meant to complement (but not to replace) the single language/treebank-specific schemes.

Started in 2014 with the first set of guidelines, the project has published a new release of the collection of the treebanks roughly every six months. Version 2 (v2), which introduces a new set of guidelines, was released in March 2017. The current version is 2.2 (July 2018). It includes 122 treebanks and 71 languages.

The IT-TB is part of UD since version 1.2 (November 2015), thanks to an automatic conversion procedure from ATSs to UD.[19] The UD annotation guidelines show a number of differences from those of the IT-TB original scheme for ATSs. Figure 3 presents the UD v2 compliant tree of the sentences whose ATS is shown in Figure 1.

From Figure 3 it stands out clearly that one of the basic annotation principles of UD is that fundamental dependencies do hold between content words, while function words depend on the content word they modify. For instance,

16 #PersPron is a "t-lemma" (tectogrammatical lemma) assigned to nodes representing possessive and personal pronouns (including reflexives).

17 (Mikulová et al. 2006, 47).

18 http://universaldependencies.org (last access 2019.01.31); (Nivre 2015).

19 (Cecchini et al. forthcoming).

one can see that in the UD tree of Figure 3 there is a direct dependency relation between the main predicate of the sentence (*disponitur*) and that of the subordinate clause *ordinatur*, while this is not the case in the ATS of Figure 1, where such relation is mediated by the the subordinating conjunction (*cum*). Consistently, in UD trees prepositions depend on the head of the prepositional phrase. In Figure 3, the node for the preposition *ad* depends on *finem* (thus creating a direct relation between the content words *ordinatur* and *finem*), while the opposite holds in the ATS of Figure 1, resulting in an indirect relation between *ordinatur* and *finem*.

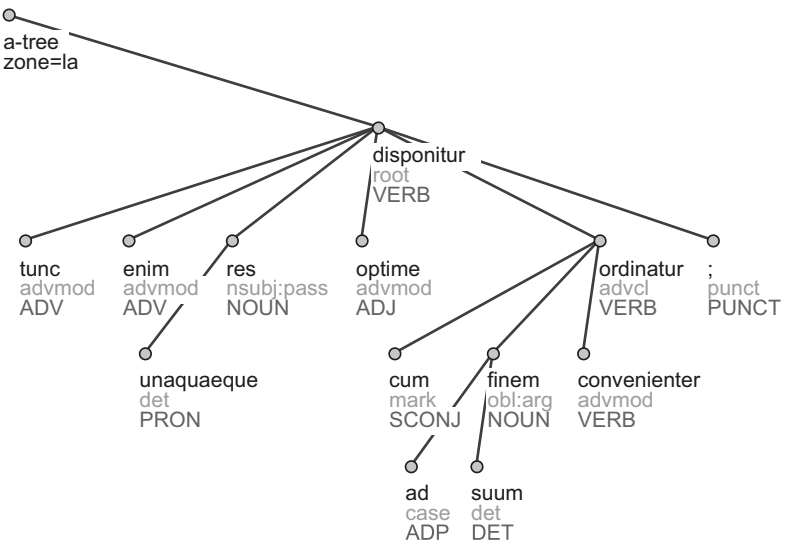

Figure 3: A UD v2 tree.

3 Subcategorization and valency lexica

Following the basic assumption of frame semantics,[20] according to which the meaning of some words can be fully understood only by knowing the frame elements that are evoked by that word, valency lexica for several languages are today available.[21] These lexica play an important role in NLP thanks to their

20 (Fillmore 1982).

21 See, for instance, PropBank (Kingsbury and Palmer 2002), FrameNet (Ruppenhofer et al. 2006) and PDT-Vallex (Hajič et al. 2003), which were first created in intuition-based fashion

large applicability in tasks like semantic role labeling, word sense disambiguation, automatic verb classification, selectional preference acquisition and also treebanking.[22]

As for Latin, the IT-TB project has developed two lexica for Latin based on the notion of valency: IT-VaLex and Latin Vallex.

3.1 IT-VaLex

IT-VaLex is a corpus-driven syntactic subcategorization lexicon whose entries (verbs only) are automatically induced from the analytical layer of annotation of the IT-TB.[23]

Being developed in corpus-driven fashion, IT-VaLex fully reflects the empirical evidence shown by corpus data and can always be rebuilt using a new version of the source treebank. The lexicon provides a full account of the syntactic subcategorization behavior of the verbs in the IT-TB. This means that only those arguments that are explicitly realized by a lexical item in the text are reported in IT-VaLex, thus resulting in cases where, for instance, typically three-argument verbs (like *do* 'to give') are assigned a subcategorization frame featuring only one argument (e.g. the subject), reflecting the fact that, among the three possible arguments, only one is realized by a lexical item in the occurrences of the verb represented by that frame.

Each entry in IT-VaLex corresponds to a verbal token in the treebank. All those tokens that share a common lemma are then collected together, to build the lexical entry of that lemma in the lexicon.

Subcategorization frames are enhanced with a number of properties concerning their occurrences in the IT-TB. These are the voice of the verb, the morpho-syntactic and syntactic features of its arguments and the order of the verb and its arguments in the sentence.

For example, one of the patterns referring to the active instances of the verb *compono* 'to join' in the lexicon is "A_Sb[nom]+V+Obj[acc]+(cum)Obj[abl]". "A" stands for "active" and the sign "+" links the elements in the linear order in

and then checked and refined by using data taken from corpora. Examples of valency lexica automatically acquired from annotated corpora are VALEX (Korhonen et al. 2006) and LexShem (Messiant et al. 2008).

22 (Urešová 2004).

23 (McGillivray and Passarotti 2009). The same structure of IT-VaLex is resembled by a lexicon created from the Latin Dependency Treebank and described by McGillivray (2013, 31–60).

which they appear in the sentence. Sb and Obj are analytical functions. The case of the arguments is enclosed in square brackets and the preposition *cum* introducing the ablative argument is in round brackets. This pattern thus corresponds to those active occurrences of *compono* preceded by a nominative subject and followed by an accusative argument and an ablative argument introduced by the preposition *cum*, like in the following sentence of Thomas Aquinas "intellectus componit privationem cum subiecto" ('The intellect links privation to the subject') (*Sent.* III, Dist. 6, Q. 2, Art. 1).

Currently IT-VaLex includes 1,276 lexical entries, corresponding to 65,535 verbal occurrences in the IT-TB. The lexicon is downloadable from the IT-TB website and can be queried through a dedicated web graphical interface (http://itreebank.marginalia.it/itvalex). Complex queries can be run by merging different search criteria, namely the number of arguments, their order, their morpho-syntactic labels and their lemma.

3.2 Latin Vallex

Latin Vallex is a valency lexicon built in conjunction with the tectogrammatical annotation of the IT-TB and the LDT performed by the IT-TB project.[24]

Each valency-capable word occurring in the semantically annotated portion of the two treebanks is assigned one frame entry in Latin Vallex. These can be verbs (*do* 'to give'), adjectives (*contrarius* 'opposite'), nouns (*descriptio* 'representation') and adverbs (*similiter* 'similarly').

The structure of the lexicon resembles that of the valency lexicon for Czech PDT-Vallex in the theoretical context of FGD. On the topmost level, the lexicon is divided into word entries. A word entry consists of a non-empty sequence of frame entries relevant for the lemma in question, where each different frame entry usually corresponds to one of the lemma's senses. Each frame entry contains a description of the valency frame itself and of the frame attributes. A valency frame is a sequence of frame slots. Each frame slot represents one complement of the given lemma. The surface morphological features of the frame slots are recorded, coming from the the textual evidence provided by the tectogrammatical annotation of the two Latin treebanks Latin Vallex is built on. Attributes are functors used to express types of relations between lemmas and their complements. The functors reported in the frame entries of Latin Vallex are those for inner participants ('arguments'). Also some free

24 (Passarotti et al. 2016).

modifications ('adjuncts') can enter the frame entries and are recorded as optional slots. The most frequent functors for adjuncts appearing in Latin Vallex are the locative and directional ones, which are mostly used in the frame entries for motion verbs.[25] For instance, the prototypical frame entry for the verb *venio* features three slots, whose functors are ACT, DIR1 (Direction-From) and DIR3 (Direction-To).

Presently, Latin Vallex includes 1,373 lexical entries and 3,406 frame entries. Like the treebanks which is based on, it is downloadable from the website of the IT-TB and can be queried either locally via TrEd or online through a PML-TQ implementation (http://itreebank.marginalia.it/view/resources.php). Users can move between a specific frame entry in the lexicon and its occurrences in the source treebanks.

4 Natural Language Processing tools

4.1 Morphological analysis; Lemlat and word formation Latin

Lemlat is a morphological analyzer for Latin whose version 3.0 was recently released.[26]

Among the available morphological analyzers for Latin,[27] Lemlat has proved to be the best performing together with LatMor[28] and the one provided with the largest lexical basis. In versions 1.0 and 2.0, this consists in the collation of three Latin dictionaries[29] for a total of 40,014 lexical entries and 43,432 lemmas. In version 3.0, the lexical basis of Lemlat was further enlarged at CIRCSE by adding the *Onomasticon* provided by the fifth edition of the Forcellini Dictionary.[30]

25 (Mikulová et al. 2006, 503–514).

26 (Passarotti et al. 2017). For details about credits of the different versions of Lemlat see http://www.lemlat3.eu/about/credits/ (last access 2019.01.31).

27 The main ones are *Words* (http://archives.nd.edu/words.html), Lemlat (http://www.lemlat3.eu), *Morpheus* (https://github.com/tmallon/morpheus), reimplemented in 2013 as *Parsley* (https://github.com/goldibex/parsley-core), the PROIEL Latin morphology system (https://github.com/mlj/proiel-webapp/tree/master/lib/morphology) and *LatMor* (http://cistern.cis.lmu.de) (last access 2019.01.31). Morpheus, Parsley and LatMor are all capable of analyzing word forms into their morphological representations including vowel quantity.

28 For the results of a comparison between the morphological analyzers for Latin see Springmann et al. (2016, 389) and Passarotti et al. (2017, 28).

29 GGG: (Georges and Georges 1913–1918); (Glare 1982); (Gradenwitz 1904).

30 (Budassi and Passarotti 2016).

Most recently, in the context of the IT-TB project the lexical basis of Lemlat was enhanced by CIRCSE also with *Glossarium Mediae et Infimae Latinitatis*, a reference dictionary for Medieval Latin comprising approximately 86,000 lemmas.[31] This makes Lemlat able to analyze the inflected forms of more than 150,000 Latin lemmas spread over a large diachronic span.

4.1.1 Word form analysis

Given an input word form recognized by Lemlat, the tool produces in output the corresponding lemma(s) and a number of tags conveying (a) the part of speech of the lemma(s) and (b) the morphological features of the input word form. The analysis is run on types rather than on tokens, which means that no contextual disambiguation is performed.

If the analyzed word is morphologically derived, its derivation process is provided by reporting the base lemma and the word formation rule applied (see Section 4.1.2). For instance, the word form *amabilem* is analyzed by Lemlat as singular masculine/feminine accusative of the adjective *amabilis* 'lovable', which is derived from the verb *amo* 'to love' via a word formation rule that builds second class deverbal adjectives with suffix *-bil-*.

The lexical database of Lemlat 3.0 is available at https://github.com/CIRCSE/LEMLAT3, where also a Command Line Interface (CLI) implementation of the tool for Linux, OSX and Windows can be downloaded.

4.1.2 Derivational morphology

The information on derivational morphology provided by Lemlat is taken from Word Formation Latin (WFL; Litta et al. 2016), a derivational morphology resource for Latin built by CIRCSE in the context of a project funded by the EU Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie Individual Fellowship.

WFL connects the lemmas of the GGG lexical basis of Lemlat by word formation rules (WFRs). Each morphologically derived lemma is assigned a WFR and is paired with its base lemma. All those lemmas that share a common (not derived) ancestor belong to the same "morphological family". For instance, nouns *amator*

31 (Du Fresne Du Cange et al. 1883-1887).

'lover' and *amor* 'love', and adjective *amabilis* all belong to the morphological family whose ancestor is the verb *amo*.

WFL can be accessed via a web application (http://wfl.marginalia.it), where WFR-based relations between the lemmas of a morphological family are represented in a tree graph. In such graph, a node is a lemma, and an edge is the WFR applied to derive the output lemma from the input one (or two, in the case of compounds), along with any affix used. For example, Figure 4 shows a part of the derivation tree for the lemma *amo*. One can see that *amabilis* derives from *amo* and it is in turn the input for two other derived lemmas: *amabilitas* 'loveliness' and *inamabilis* 'repugnant'. Clicking on an edge shows the lemmas built by the WFR concerned in that edge. Lemmas are provided both as a tree graph and as an alphabetical list.

4.2 Dependency parsing

So far, the IT-TB is the treebank providing the training set that allowed to achieve the best accuracy rates for dependency parsing of Latin.[32] This is not surprising, not only because the IT-TB is the largest Latin treebank available, but also because its texts are written in quite a formal variety of Medieval Latin and are very consistent, as they are written by one author only.

The parser developed by Ponti and Passarotti achieves a Labeled Attachment Score (LAS) of 86.5 and an Unlabeled Attachment Score (UAS) of 90.97 and it is the one currently used in the IT-TB project to process automatically the sentences of the IT before double manual checking.[33] The parser was trained on a version of the IT-TB including around 250,000 nodes. Six different stochastic dependency parsers were first trained and tested. The best performing one was then provided with an ad-hoc feature model for Medieval Latin and its settings were tuned. Then, a combination of the outputs of two shift-reduce parsers and one graph-based parser was performed.

The quite high accuracy rates for syntactic parsing achieved on the IT-TB data must be considered carefully when generalizing about the automatic processing of Latin. Indeed, performances of stochastic NLP tools depend heavily on the training set which their models are built on. This problem is particularly hard when Latin is concerned, because Latin texts show a high degree of variation resulting from (a) a wide time span (covering more than two millennia),

32 (Ponti and Passarotti 2016).
33 (Buchholz and Marsi 2006).

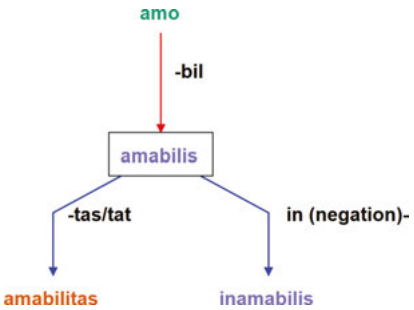

Figure 4: Derivation tree for *amo* (part).

(b) a large variety of genre (ranging from literary to philosophical, historical and documentary texts) and (c) a big diatopic diversity (spread all over Europe and beyond). As a matter of fact, Ponti and Passarotti show that when the best performing IT-TB-based dependency parser is applied on texts from the Classical era taken from the LDT, results drop dramatically: e.g. 28.2 on Caesar and 23.9 on Ovid. This is strictly related to the remarkable incongruity between the varieties of Latin represented in the training set (IT-TB) and in the test data (LDT).

5 Conclusion and future work

Building a linguistic resource is a labor-intensive work, which today goes beyond the simple development of a new collection of (annotated) linguistic data. In a virtuous circle, several different kinds of actors are concerned: textual resources are made of words, which are described in lexical resources and represent the main object of analysis of NLP tools, which in turn tend to achieve better accuracy rates when trained on larger empirical evidence provided by textual data. This is why, in more than a decade the IT-TB project has developed a number of lexical resources and NLP tools connected with the annotated data of the treebank.

The annotation work is also diverse. Beside continuing the analytical annotation of the IT-TB, a core task of the project is to enlarge the available set of sentences annotated at the tectogrammatical layer, to address the current need of semantic annotation in textual resources. The task is time-consuming because the portion of work that can be performed automatically is still very limited and annotators must have a deep understanding of the text both at intra- and inter-sentential level.

Beside linguistic annotation of textual data, there are three other open issues.

First, lexical resources must be enlarged and refined to be able to cover and process a larger (and more diverse) set of Latin data.

Second, the three Latin treebanks available in UD, namely the IT-TB, the LDT and PROIEL,[34] must be harmonized, as they still show differences in tokenization, lemmatization, PoS-tagging and syntactic analysis.

Third is assessing the degree of portability of NLP tools for Latin. As shown in Section 4.2, the sociolinguistic aspects connected to Latin texts open new challenges for the NLP world. Indeed we do not deal with one Latin only, but with several varieties of Latin, which can even heavily differ one from the other. Building sets of annotated empirical data to train stochastic NLP tools to process all such varieties is out of reach of current research. Instead, trying to make NLP processes more dynamic, enabling them to automatically adapt to the specific variety of language they deal with, would represent a major advance not only in the field of resources for Latin but overall in computational linguistics. In this respect, Latin is a perfect case study language, where developing and evaluating techniques, methods and tools for dynamic domain-adaptation in NLP. The harmonization of the three Latin treebanks in UD is a mandatory step also towards such objective, providing a set of texts annotated with a common scheme which can be used as a test bed for different NLP tasks.

This paper focuses on the resources and tools for Latin built by the IT-TB project. They represent just an example of those currently available, as there exists a huge number of digitized Latin texts (and lexical resources as well) built by various projects around the world, spread in different repositories and recorded in various data formats.[35] This is a limit, because linguistic resources become even more useful when linked with each other, which makes it possible to exploit the contribution each of them gives to linguistic analysis. The increasing complexity and diversity of linguistic resources and NLP tools that have become available throughout the last decades have led to a growing interest in their sustainability and interoperability.[36] This was partially approached by building large infrastructures of linguistic resources, like CLARIN (https://www.clarin.eu), DARIAH (http://www.dariah.eu) and META-SHARE (http://www.meta-share.org). However, these represent collections of resources and tools, which can be used and queried from one common place on the web, more than interconnections between them to make the whole greater than the sum of its parts.

34 (Haug and Jøhndal 2008).
35 (Bagnall and Heath 2018).
36 (Ide and Pustejovsky 2010).

Instead, making linguistic resources interoperable requires that all types of information related to a particular word/text get integrated into a common representation. Currently, the most rising approach to make linguistic resources interoperable (and potentially enhanced with NLP web-services) is to apply to them the principles of Linked Data and thus to build a Linguistic Linked Open Data cloud.[37]

The ERC-Consolidator Grant LiLa (Linking Latin. Building a Knowledge Base of Linguistic Resources for Latin: https://lila-erc.eu), recently started at CIRCSE, wants to connect and, ultimately, to exploit the wealth of linguistic resources and NLP tools for Latin assembled so far, in order to bridge the gap between raw language data, NLP and knowledge descriptions. To this aim, the project will build a Knowledge Base for Latin by using the Linked Data paradigm to combine data from disparate linguistic resources, provide NLP web-services and ultimately include also Latin into the multilingual Linguistic Linked Open Data cloud.

Bibliography

Bagnall, R.S.; Heath, S. (2018): "Roman Studies and Digital Resources". The Journal of Roman Studies 108, 1–19.

Bamman, D.; Crane, G. (2007): "The Latin Dependency Treebank in a Cultural Heritage Digital Library". In: Proceedings of The Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007). Prague, Czech Republic: Association for Computational Linguistics, 33–40.

Bamman, D.; Passarotti, M.; Crane, G.; Raynaud, S. (2007): Guidelines for the Syntactic Annotation of Latin Treebanks. Boston, MA: Tufts University Digital Library. http://hdl.handle.net/10427/42683 (last access 2019.01.31).

Buchholz, S.; Marsi, E. (2006): "CoNLL-X Shared Task on Multilingual Dependency Parsing". In: Proceedings of the Tenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 149–164.

Budassi, M.; Passarotti, M. (2016): "Nomen Omen. Enhancing the Latin Morphological Analyser Lemlat with an Onomasticon". In: Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH 2016). Stroudsburg, PA: Association for Computational Linguistics, 90–94.

Busa, R. (1974-1980): Index Thomisticus. Stuttgart-Bad Cannstatt: Frommann-Holzboog.

Cecchini, F.M.; Passarotti, M.; Marongiu, P.; Zeman, D. (forthcoming): "Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies". In: Proceedings of the Universal Dependencies Workshop 2018 (UDW 2018).

37 (Chiarcos et al. 2013).

Chiarcos, C.; Cimiano, P.; Declerck, T.; McCrae, J.P. (2013): "Linguistic Linked Open Data (LLOD). Introduction and Overview". In: Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013). Pisa, Italy: Association for Computational Linguistics, i–xi.

Du Fresne Du Cange, C. (1883-1887): Glossarium Mediae et Infimae Latinitatis. Niort: L. Favre.

Fillmore, C. (1982): "Frame Semantics". In: Linguistics in the Morning Calm. Selected Papers from SICOL-1981. Seoul: Hanshin Publishing Co., 111–137.

Forcellini, A. (1940): Lexicon Totius Latinitatis. Padova: Typis Seminarii.

Georges, K.E.; Georges H. (1913-1918): Ausführliches Lateinisch-Deutsches Handwörterbuch. Hannover: Hahn.

Glare, P.G.W. (1982): Oxford Latin Dictionary. Oxford: Oxford University Press.

González Saavedra, B.; Passarotti, M. (2014): "Challenges in Enhancing the Index Thomisticus Treebank with Semantic and Pragmatic Annotation". In: Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT-13). Tübingen, Germany: Department of Linguistics, University of Tübingen, 265–270.

Gradenwitz, O. (1904): Laterculi Vocum Latinarum. Leipzig: Hirzel.

Hajič, J.; Panevová, J.; Buránová, E.; Urešová, Z.; Bémová, A. (1999): Annotations at Analytical Level. Instructions for annotators. Prague: Institute of Formal and Applied Linguistics. http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf (last access 2019.01.31).

Hajič, J.; Böhmová, A.; Hajičová, E.; Vidová Hladká, B. (2000): "The Prague Dependency Treebank: A Three-Level Annotation Scenario". In: A. Abeillé (ed.): Treebanks: Building and Using Parsed Corpora. Amsterdam: Kluwer, 103–127.

Hajič, J.; Panevová, J.; Urešová, Z.; Bémová, A.; Kolárová-Reznícková, V.; Pajas, P. (2003): "PDT-VALLEX: Creating a Large Coverage Valency Lexicon for Treebank Annotation". In: J. Nivre; E. Hinrichs (eds.): TLT 2003. Proceedings of the Second Workshop on Treebanks and Linguistic Theories. Volume 9 of Mathematical Modelling in Physics, Engineering and Cognitive Sciences. Växjö, Sweden: Växjö University Press, 57–68.

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

Ide, N.; Pustejovsky, J. (2010): "What does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability for Language Technology". In: Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL). Hong Kong.

Kingsbury, P.; Palmer, P. (2002): "From Treebank to Propbank". In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas, Gran Canaria: ELRA.

Korhonen, A.; Krymolowski, Y.; Briscoe, T. (2006): "A Large Subcategorization Lexicon for Natural Language Processing Applications". In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006). Genoa: ELRA, 1015–1020.

Litta, E.; Passarotti, M.; Culy, C. (2016): "Formatio formosa est. Building a Word Formation Lexicon for Latin". In: A. Corazza; S. Montemagni; G. Semeraro (eds.): Proceedings of the Third Italian Conference on Computational Linguistics (CLiC–it 2016). Naples: Accademia

University Press, Collana dell'Associazione Italiana di Linguistica Computazionale. Vol. 2, 185–189.

Martens, S.; Passarotti, M. (2014): "Thomas Aquinas in the TüNDRA: Integrating the Index Thomisticus Treebank into CLARIN-D". In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik: ELRA, 767–774.

McGillivray, B.; Passarotti, M. (2009): "The Development of the Index Thomisticus Treebank Valency Lexicon". In: Proceedings of LaTeCH-SHELT&R Workshop 2009. Athens: ACL, 43–50.

McGillivray, B. (2013): Methods in Latin Computational Linguistics. Leiden and Boston: Brill.

Messiant, C.; Korhonen, A.; Poibeau, T. (2008): "LexSchem: A Large Subcategorization Lexicon for French Verbs". In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). Marrakech: ELRA, 533–538.

Mikulová, M.; Bémová, A.; Hajič, J.; Hajičová, E.; Havelka, J.; Kolářová, V.; Kučová, L.; Lopatková, M.; Pajas, P.; Panevová, J.; Razímová, M.; Sgall, P.; Štěpánek, J.; Urešová, Z.; Veselá, K.; Žabokrtský, Z.; Součková, K.; Böhmová, A.; Čermáková, K.; Havelka, J.; Corness, P. (2006): "Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank Institute of Formal and Applied Linguistics". Prague: Institute of Formal and Applied Linguistics. http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/html/index.html (last access 2019.01.31).

Nivre, J. (2015): "Towards a Universal Grammar for Natural Language Processing". In: Gelbukh A. (ed.): Computational Linguistics and Intelligent Text Processing. CICLing 2015. Cham: Springer, 3–16.

Pajas, P.; Štěpánek, J. (2009): "System for Querying Syntactically Annotated Corpora". In: G. Geunbae Lee; S. Schulte Im Walde (eds.): Proceedings of the ACL-IJCNLP 2009 Software Demonstrations. Singapore: World Scientific Publishing Co Pte Ltd, 33–36.

Passarotti, M. (2010): "Leaving Behind the Less-Resourced Status. The Case of Latin through the Experience of the Index Thomisticus Treebank". In: 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages. La Valletta: ELRA, 27–32.

Passarotti, M. (2013): "One Hundred Years Ago. In Memory of Father Roberto Busa SJ". In: F. Mambrini; M. Passarotti; C. Sporleder (eds.): Proceedings of the Third Workshop on Annotation of Corpora for Research in the Humanities (ACRH-3). Sofia: Bulgarian Academy of Sciences, 15–24.

Passarotti, M.; González Saavedra, B.; Onambele, C. (2016): "Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin". In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Portorož: ELRA, 2599–2606.

Passarotti, M.; Budassi, M.; Litta, E.; Ruffolo, P. (2017): "The Lemlat 3.0 Package for Morphological Analysis of Latin". In: G. Bouma; Y. Adesam (eds.): Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language. Gothenburg: Northern European Association for Language Technology Proceedings Series 32, 24–31.

Passarotti, M; González Saavedra, B. (2018): "The Treebanked Conspiracy. Actors and Actions in Bellum Catilinae". In: J. Hajič (ed.): Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16). Prague: Institute of Formal and Applied Linguistics, 18–26.

Ponti, E.M.; Passarotti, M. (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). Portorož: ELRA, 683–688.

Popel, M.; Žabokrtský, Z. (2010): "TectoMT: Modular NLP Framework". In: H. Loftsson; E. Rögnvaldsson; S. Helgadóttir (eds.): Proceedings of IceTAL, 7th International Conference on Natural Language Processing. Berlin, Heidelberg and New York: Springer, 293–304.

Ruppenhofer, J.; Ellsworth, M.; Petruck, M.R.L.; Johnson, C.R.; Scheffczyk, J. (2006): FrameNet II. Extendend Theory and Practice. https://framenet.icsi.berkeley.edu/fndrupal/node/5400 (last access 2019.01.31).

Sgall, P.; Hajičová, E.; Panevová, J. (1986): The Meaning of the Sentence in its Semantic and Pragmatic Aspects. Dordrecht: D. Reidel.

Springmann, U.; Schmid, H.; Najock, D. (2016): "LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity". In: G. Celano; G. Crane (eds.): Treebanking and Ancient Languages: Current and Prospective Research (Topical Issue). Open Linguistics 2:1, 386–392.

Urešová, Z. (2004): The Verbal Valency in the Prague Dependency Treebank from the Annotator's Point of View. Bratislava: Jazykovedný ústav Ľ. Štúra, SAV.