From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin

Marco Passarotti
CIRCSE Research Centre
Università Cattolica del Sacro Cuore
Largo Gemelli, 1 – 20123 Milan, Italy
marco.passarotti@unicatt.it

Abstract

Assuming that collaboration between theoretical and computational linguistics is essential in projects aimed at developing language resources like annotated corpora, this paper presents the first steps of the semantic annotation of the *Index Thomisticus* Treebank, a dependency-based treebank of Medieval Latin. The semantic layer of annotation of the treebank is detailed and the theoretical framework supporting the annotation style is explained and motivated.

1 Introduction

Started in 1949 by father Roberto Busa SJ, the *Index Thomisticus* (IT; Busa, 1974-1980) has represented a groundbreaking project that laid the foundations of computational linguistics and literary computing. The IT is a morphologically tagged and lemmatized corpus of Medieval Latin containing the *opera omnia* of Thomas Aquinas (118 texts), as well as 61 texts by other authors related to Thomas, for a total of around 11 million tokens.

The *Index Thomisticus* Treebank (IT-TB: http://itreebank.marginalia.it) is the syntactically annotated portion of the IT. Presently, the IT-TB includes around 220,000 nodes (approximately, 12,000 sentences).

The project of the IT-TB is now entering a new phase aimed at enhancing the available syntactic annotation with semantic metadata. Starting such a task needs to choose a theoretical approach and framework that supports the annotation style. Indeed, performing linguistic annotation of a textual corpus should be strictly connected to fundamental issues in theoretical linguistics in a kind of virtuous circle. On its side, theoretical linguistics serves as the necessary backbone for solid annotation guidelines; no theory-neutral representation of a sentence is possible, since every representation style needs a theory to extract its meaning. On the other hand, applying a theoretical framework to real data makes it possible to empirically test and possibly refine it. According to Eva Hajičová, "corpus annotation serves, among other things, as an invaluable test for the linguistic theories standing behind the annotation schemes, and as such represents an irreplaceable resource of linguistic information for the construction and enrichment of grammars, both formal and theoretical" (Hajičová, 2006: 466).

Further, the task of developing language resources like annotated corpora supports interaction between intuition-based and corpus-based/-driven approaches in theoretical linguistics (Tognini-Bonelli, 2001). No intuition-based grammar is able to manage all the possible

Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014, pages 100–109,
Gothenburg, Sweden, April 26 2014. ©2014 Association for Computational Linguistics

variations in real data, and no induction-based grammar can reflect all the possible well-formed constructions of a language (Aarts, 2002; Sinclair, 2004a).

This paper describes the first steps towards the semantic annotation of the IT-TB, by first presenting and motivating its theoretical background (section 2) and then sampling a number of specific aspects of annotation (section 3). Finally, section 4 reports a discussion and sketches the future work.

2 The Theoretical Background of the *Index Thomisticus* Treebank

Hosted at the CIRCSE research centre of the Università Cattolica del Sacro Cuore in Milan, Italy (http://centridiricerca.unicatt.it/circse), the IT-TB is a dependency-based treebank (McGillivray et al., 2009). The choice of a representation framework alone does not determine the representation for a given sentence, as there can be many (correct) dependency-based (as well as constituency-based) trees for even simple sentences. Thus, a fine-grained linguistic theory must be selected to support the specific aspects raised by a large-scale annotation of real data. In this respect, the annotation style of the IT-TB is based on Functional Generative Description (FGD; Sgall et al., 1986), a dependency-based theoretical framework developed in Prague and intensively applied and tested while building the Prague Dependency Treebank of Czech (PDT).

FGD is rooted in Praguian structuralism-functionalism dating back to the 30s, one assumption of which is the stratificational approach to sentence analysis pursued by Functional Sentence Perspective (FSP), a linguistic theory developed by Jan Firbas in the mid-1950s on the basis of Vilém Mathesius' work (Firbas, 1992). According to FSP, the sentence is conceived as: (a) a singular and individual speech event [utterance-event]; (b) one of the possible different minimal communicative units (means) of the given language [form]; (c) an abstract structure (a pattern) [meaning].

Considering language as a form-meaning composite is a basic assumption also of FGD, which is particularly focused on the last point above, aiming at the description of the so-called 'underlying syntax' of the sentence. Underlying syntax (the meaning) is separated from (but still connected with) surface syntax (the form) and represents the linguistic (literal) meaning of the sentence, which is described through dependency tree-graphs.

This approach is consistent with the functional and pragmatic analysis of language pursued by the Prague Linguistic Circle since its very beginning, along the so-called 'first period' of the Circle (Raynaud, 2008). Language is conceived as "un système de moyens d'expression appropriés à un but" ("a system of purposive means"; Cercle linguistique de Prague, 1929: 7). The "moyens d'expression" correspond to the 'form' (surface syntax), while the fact that they are "appropriés à un but" corresponds to the 'meaning' (underlying syntax).

The description of surface and underlying syntax in FGD is dependency-based mostly because dependency grammars are predicate-focused grammars. This enables FGD to face one of the basic statements of the Prague Linguistic Circle: "l'acte syntagmatique fondamental […] est la prédication" ("the basic syntagmatic act is predication"; Cercle linguistique de Prague, 1929: 13). Further, during the second period of the theory of predication pursued by the Circle, while accounting for the three-level approach to sentence in FSP, Daneš claims that "[t]he kernel syntactic relation is that of dependance" (Daneš, 1964: 227) and stresses the strict connection holding between form and meaning: "we are convinced that the interrelations of both levels, semantic and grammatical must necessarily be stated in order to give a full account of an overall linguistic system" (Daneš, 1964: 226).

Consistently with such a theoretical background, the PDT (as well as the IT-TB) is a dependency-based treebank with a three-layer structure, in which each layer corresponds to one of the three views of sentence mentioned above (Hajič et al., 2000). The layers are ordered as follows:

- a morphological layer: morphological tagging and lemmatization;
- an 'analytical' layer (i.e. the presently available layer of annotation of the IT-TB): annotation of surface syntax;
- a 'tectogrammatical' layer: annotation of underlying syntax.

The development of each layer requires the availability of the previous one(s). Both the analytical and the 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 that are labelled with (surface) syntactic functions called 'analytical functions' (like Subject, Object etc.).

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 autosemantic words only, while function words and punctuation marks are left out. The nodes are labeled with semantic role tags called 'functors'. These are divided into two classes according to valency: (a) arguments, called 'inner participants', i.e. obligatory complementations 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.. The 'dialogue test' by Panevová (1974-1975) is used as the guiding criterion for obligatoriness. 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). In FGD, TFA deals with the opposition between contextual boundness (the 'given' information, on the left) and contextual unboundness (the 'new' information, on the right). Also ellipsis resolution and coreferential analysis are performed at the tectogrammatical layer and are represented in TGTSs through newly added nodes (ellipsis) and arrows (coreference).

Since its beginning, the IT-TB has been following the PDT annotation style for both typological and structural reasons. As far as the former are concerned, Latin and Czech share certain relevant properties, such as being richly inflected, showing discontinuous phrases, and having a moderately free word-order and a high degree of synonymity and ambiguity of the endings. Both languages have three genders (masculine, feminine, neuter), cases with roughly the same meaning and no articles. As for the latter, the tight connection between the three-layer structure of the PDT and a sound background theory like FGD integrates each layer of annotation into a more general framework driven by a functional perspective aimed at understanding the underlying meaning of sentences through its relation with the surface form. Moreover, tectogrammatical annotation includes several pragmatic aspects that, although much present in Latin linguistics research, are still missing from the available treebanks of Latin[1].

The organization of functors into inner participants and free modifications is further exploited by linking textual tectogrammatical annotation with fundamental lexical information

[1] Some semantic-pragmatic annotation of Latin texts is available only in the PROIEL corpus (Haug & Jøndal, 2008). The Latin subset of PROIEL includes Classical texts from the 1st century BC (Caesar, Cicero), the *Peregrinatio Aetheriae* and the *New Testament* by Jerome (both from the 5th century AD).

provided by a valency lexicon that features the valency frame(s) for all those verbs, nouns, adjectives and adverbs capable of valency that occur in the treebank. The valency lexicon of the IT-TB is being built in a corpus-driven fashion, by adding to the lexicon all the valency-capable words that annotators progressively get through[2].

3 Moving From Analytical to Tectogrammatical Tree Structures

As the tectogrammatical annotation of the IT-TB has just started and no Latin texts annotated at the tectogrammatical layer are available yet, we cannot train and use probabilistic NLP tools to build TGTSs. Thus, the annotation workflow is based on TGTSs automatically converted from ATSs. The TGTSs that result from conversion are then checked and refined manually by two independent annotators. Conversion is performed by adapting to Latin a number of ATS-to-TGTS scripts provided by the NLP framework Treex developed in Prague (Popel and Žabokrtský, 2010). Relying on ATSs, the basic functions of these scripts are: (a) to collapse ATSs nodes of function words and punctuation marks, as they no longer receive a node for themselves in TGTSs, but are included into the autosemantic nodes; (b) to assign basic functors (such as Actor and Patient); (c) to assign 'grammatemes', i.e. semantic counterparts of morphological categories (for instance, pluralia tantum are tagged with the number grammateme 'singular').

The annotation guidelines are those for the tectogrammatical layer of the PDT (Mikulová et al., 2006).

In the following, three examples of tectogrammatical annotation of sentences taken from the IT-TB are reported and discussed in detail.

3.1 Example A

Figure 1 reports the ATS of the following sentence of 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;", *Summa contra Gentiles* 1.1).

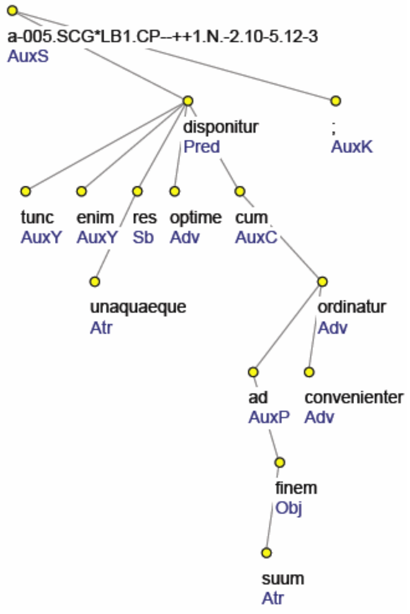

Figure 1. Analytical Tree Structure A

Except for the technical root of the tree (holding the textual reference of the sentence), each node in the ATS corresponds to 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 labelled with an analytical function. For instance, the relation between the word *res* and its governor *disponitur* is labelled 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*). The other analytical functions occurring in this sentences are the following: Atr (attributes), Adv (adverbs and adverbial modifications, i.e. adjuncts), AuxS

[2] A syntactic-based valency lexicon of the IT-TB (named IT-VaLex) was induced automatically from the analytical layer of the treebank and made available on-line (http://itreebank.marginalia.it/itvalex) (McGillivray and Passarotti, 2009).

(root of the tree), Obj (direct and indirect objects), Pred (main predicate of the sentence).

Figure 2 shows the TGTS corresponding to the ATS of this sentence.

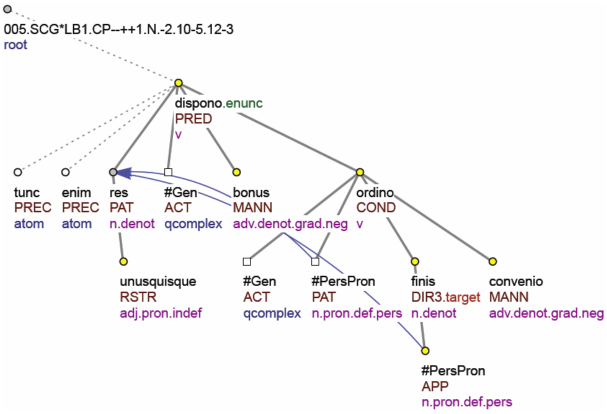

Figure 2. Tectogrammatical Tree Structure A[3]

As only autosemantic nodes can occur in TGTSs, auxiliary sentence members labelled with AuxC, AuxK, or AuxP are collapsed.

Analytical functions are replaced with functors. The nodes of the lemmas *tunc* and *enim* are both assigned the functor PREC, since they represent expressions linking the clause to the preceding context; further, *tunc* and *enim* are given nodetype 'atom' (atomic nodes), which is used for adverbs of attitude, intensifying or modal expressions, rhematizers and text connectives (which is the case of *tunc* and *enim*) (Mikulová et al., 2006: 17). *Res* is the Patient (PAT) of *dispono*, as it is the syntactic subject of a passive verbal form (*disponitur*)[4]. Both the adverbial forms of *bonus* (*optime*) and *convenio* (*convenienter*) are labelled with functor MANN, which expresses manner by specifying an 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*; analytical function: Adv) 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, even 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 that link the two #PersPron nodes with the node of *res*. #PersPron is a 't-lemma' (tectogrammatical lemma) assigned to nodes representing possessive and personal pronouns (including reflexives).

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

[3] In the default visualization of TGTSs, wordforms are replaced with lemmas.

[4] Conversely, syntactic subjects of active verbal forms are usually labelled 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.

3.2 Example B

Figure 3 shows the ATS of this sentence: “unde et earum artifices, qui architectores vocantur, nomen sibi vindicant sapientum.” (“Thus, also the makers of them, who are called architects, claim the title of wise men for themselves”, *Summa contra Gentiles* 1.1).

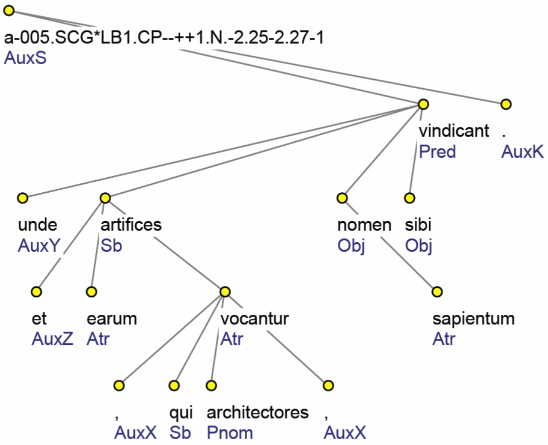

Figure 3. Analytical Tree Structure B

In addition to the analytical functions assigned to auxiliary sentence members in the tree of figure 1, this tree features one occurrence of AuxZ (particles that emphasize a specific sentence member) and two of AuxX (commas).

Figure 4 presents the TGTS of the sentence in question.

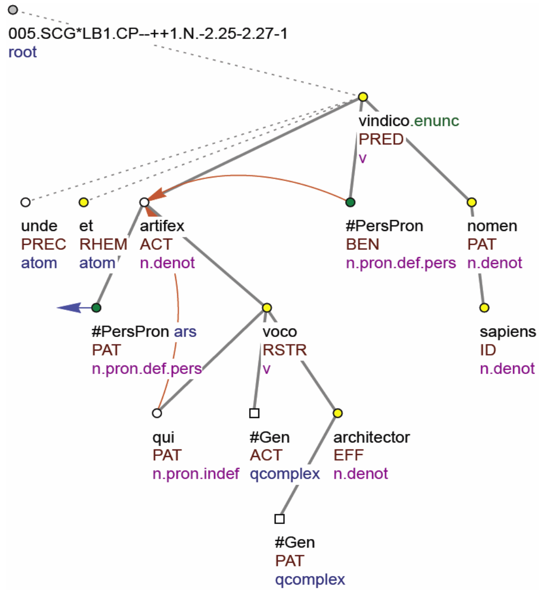

Figure 4. Tectogrammatical Tree Structure B

Sentence members labelled with AuxK, or AuxX are collapsed.

The tree reported in figure 4 features arrows of different colour. The red arrows that link both the relative pronoun *qui* and the reflexive pronoun *sibi* (assigned t-lemma #PersPron) with the noun *artifex* stand for so-called ‘grammatical coreferences’, i.e. coreferences in which it is possible to pinpoint the coreferred expression on the basis of grammatical rules. Instead, the blue arrow represents a ‘textual coreference’, i.e. a coreference realized not only by grammatical means, but also via context (mostly with pronouns) (Mikulová et al., 2006: 998 and 1,100). In figure 4, a blue arrow links *earum* (#PersPron) with the word *ars*, which occurs in the previous sentence in the text.

Sibi (#PersPron) is assigned the functor BEN, because it is the beneficiary of the action carried out by the Actor (*artifex*) of the verb *vindico*. *Sapiens* has functor ID (Identity), which labels explicative genitives. *Earum* (#PersPron) is the Patient (PAT) of the noun *artifex*, because agent nouns are valency-capable nouns; for this reason, a newly added node with functor PAT is made dependent on the agent noun *architector*. This is assigned functor EFF (Effect), which is used for arguments referring to the result of the event, among which are obligatory predicative complements (i.e. the role played by *architector* with respect to *voco*). *Voco* is a RSTR, which is the functor assigned to the main predicates of attributive relative clauses. *Et* is a rhematizer, which has the noun *artifex* in its scope. According to Mikulová et al. (2006: 1,170), in a TGTS the node representing the rhematizer is placed as the closest left sister of the first node of the expression that is in its scope. This is why the node of *et* in the TGTS reported in figure 4 depends on *vindico* instead of *artifex*, while in the ATS of figure 3 it depends on the node of *artifices*. Despite its left position in the TGTS, the node of *et* is marked as focus in TFA and thus the colour of its node is yellow.

3.3 Example C

Figure 5 presents the ATS of the following sentence: “ego in hoc natus sum, et ad hoc veni in mundum, ut testimonium perhibeam veritati.” (“For this I was born and for this I came to the world, to provide the truth with evidence”, *Summa contra Gentiles* 1.1, quoting the Gospel of John 18:37).

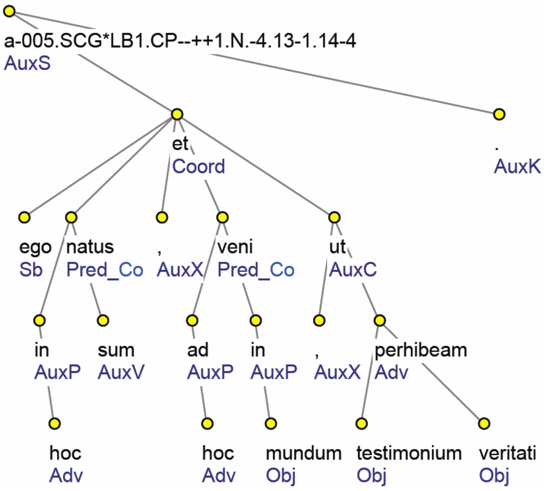

Figure 5. Analytical Tree Structure C

This sentence features two main predicates coordinated by the conjunction *et*: *veni* and *natus sum*, the latter being a complex verb, formed by the perfect participle *natus* and by the auxiliary verb *sum*, which is assigned the analytical function AuxV (collapsed in the corresponding TGTS). The fact that the two predicates are coordinated is signalled by the suffix _Co appendend to their analytical function (Pred). Those nodes that depend on the coordinating conjunction *et* and are not labelled with an analytical function suffixed with _Co are meant to depend on every member of the coordination. Thus, *ego* is the subject of both *natus sum* and *veni*, as well as the subordinate clause headed by *perhibeam* (via the subordinative conjunction *ut*) represents an adverbial modification of both the verbs.

Figure 6 presents the TGTS corresponding to the ATS of figure 5.

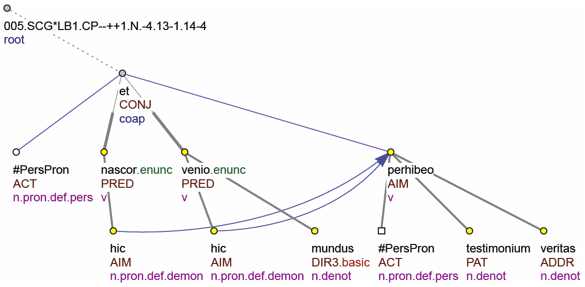

Figure 6. Tectogrammatical Tree Structure C

The conjunction *et* is assigned nodetype ‘coap’ (coordinations and appositions) and functor CONJ (Conjunction), used for the root nodes of paratactic structures.

Veritas is the Addressee (ADDR) of the verb *perhibeo*[5]. *Mundus* is assigned functor DIR3 and subfunctor ‘basic’, the latter specifying that here the meaning of DIR3 is the basic one, i.e. “where to”[6]. The two occurrences of *hic* are respectively the Aim (AIM) of the verb *nascor* and of the verb *venio*, as well as the subordinate clause headed by *perhibeo* represents the Aim of both the coordinated predicates.

The TGTS in figure 6 presents two textual coreferences, linking both the occurrences of *hic* with *perhibeo*. Indeed, the subordinate clause “[…] ut testimonium perhibeam veritati” is coreferent with the two occurrences of *hic* and makes their meaning explicit in a cataphoric manner; this is signalled by the direction of the arrows, which go from left to right (cataphora) instead of from right to left (anaphora), like in figures 2 and 4.

4 Discussion and Future Work

Recently funded by the Italian Ministry of Education, Universities and Research (MIUR), the project aimed at both providing semantic annotation of Latin texts and building a semantic-based valency lexicon of Latin has just

[5] On the bordeline between Beneficiary and Addresse, see Mikulová et al. (2006: 123-126).

[6] Instead, the DIR3 node occurring in the tree of figure 2 is specified by subfunctor ‘target’.

started. So far, only the first 200 sentences of *Summa contra Gentiles* of Thomas Aquinas have been fully annotated at tectogrammatical level (corresponding to 3,112 words and 451 punctuation marks). Such a limited experience on data does not make it possible to provide an evaluation neither of the ATS-to-TGTS conversion scripts nor of the inter-annotator agreement. Presently, the valency lexicon contains 221 verbs; the task of building the lexical entries for nouns, adjectives and adverbs is going to start in the very near future.

Analytical annotation is available not only for Medieval Latin texts, but also for Classical Latin, as the guidelines for the analytical layer of annotation of the IT-TB are shared with the Latin Dependency Treebank (LDT; http://nlp.perseus.tufts.edu/syntax/treebank/), a dependency-based treebank including around 55,000 words from texts of different authors of the Classical era (Bamman et al., 2007). By exploiting the common annotation style of the IT-TB and the LDT, our project will also perform tectogrammatical annotation of the Classical Latin texts available in the LDT and will build the corresponding valency lexicon.

While enhancing a corpus with a new layer of annotation from scratch still remains a labor-intensive and time-consuming task, today this is simplified by the possibility of exploiting the results provided by previous similar experiences in language resources development. Such results can be used for porting background theories, methods and tools from one language to another in a rapid and low-cost fashion. This is the approach pursued by our project, which wants to apply to Latin a treebank scenario originally created for Czech and now used also for other languages (including Arabic and English). Such an application meets and raises a number of issues specifically related to corpora of ancient languages, which make tectogrammatical annotation of such data a particularly difficult task. For instance, while treebanks of modern languages mostly include texts taken from newspapers, this does not hold true for both the IT-TB and the LDT, which contain respectively philosophical (IT-TB) and literary texts (LDT). These textual genres present several specific linguistic features in terms of syntax (quite complex in poetry), semantics (some words undergo a kind of technical shift of meaning in philosophical texts) and lexicon (high register words are pretty frequent). Further, the absence of native speakers often makes different interpretations of texts possible and increases the difficulty of tasks like TFA.

As mentioned above, a large-scale application of a linguistic theory to real data helps to empirically test how much sound the theory is. In our case, the evaluation of the degree of applicability of FGD to Latin is at its very beginning. However, analytical annotation has shown a strong compatibility between the ATS-based description of surface syntax and its application to Latin. As a matter of fact, the PDT manual for analytical annotation was adapted in just a few details for the treatment of specific constructions of Latin (such as the ablative absolute or the passive periphrastic) that could be syntactically annotated in several different ways (Bamman et al., 2008). This experience represents a positive background for a project that wants to build a set of theoretically-motivated advanced language resources for Latin that will provide users with information about morphology, surface syntax and semantics at both textual and lexical level.

Such advanced language resources for Latin will both improve the understanding of Latin language and question the usual research methods pursued by scholars in Classics.

As for the former, research in Latin linguistics dealing with issues like semantic role labelling, valency, ellipsis resolution, coreferential analysis and information structure will finally be able to ground on a relevant amount of empirical evidence not created for the aims of one specific research, thus preventing the vicious circle of building a corpus just for studying a single linguistic phenomenon (Sinclair, 2004b). Also, making available language resources that both feature Latin texts of differents eras and share the same annotation style with language resources of modern languages will impact diachronic research and support studies in comparative linguistics.

As for the latter, building advanced language resources for Latin by connecting a large-scale empirical analysis of Latin data with a modern and broadly evaluated linguistic theory represents a challenging and unconventional approach, which is expected to strongly impact the usual research methods in the field of Classics. Indeed, due to an age-old split holding between linguistic and literary studies, the study of Latin (and of Ancient Greek, as well) has been primarily pursued by focusing on literary, philological and glottological aspects. Further, a large number of classicists is, still today, unwilling both to apply computational methods to textual analysis and to use language resources like annotated corpora and computational lexica. Computational linguists, in turn, are more prone to develop language resources and NLP tools for living languages, which have stronger commercial, media and social impact. Considering collaboration between Classics and computational linguistics to be essential, this project provides an opportunity for innovation of both fields.

Both the treebanks and the valency lexicon will be publicly available datasets with explicit annotation guidelines. This will make the results achieved by using these language resources replicable, which is a not yet consolidated practice in Classics.

Acknowledgments

This research is part of the project “Developing and Integrating Advanced Language Resources for Latin”, supported by the Italian Ministry of Education, Universities and Research (MIUR) as part of the grant FIR-2013.

References

Jan Aarts. 2002. Does corpus linguistics exist? Some old and new issues. Leiv Breivik and Angela Hasselgren (eds.), *From the COLT‘s mouth...and others*. Rodopi, Amsterdam, 1-17.

David Bamman, Marco Passarotti, Gregory Crane and Savina Raynaud. 2007. *Guidelines for the Syntactic Annotation of Latin Treebanks*. «Tufts University Digital Library». Available online from http://hdl.handle.net/10427/42683.

David Bamman, Marco Passarotti, Roberto Busa and Gregory Crane. 2008. The annotation guidelines of the Latin Dependency Treebank and *Index Thomisticus* Treebank. The treatment of some specific syntactic constructions in Latin. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*. ELRA, Marrakech, 71-76.

Roberto Busa. 1974-1980. *Index Thomisticus*. Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Cercle linguistique de Prague. 1929. Thèses présentées au Premier Congrès des philologues slaves. *Travaux du Cercle linguistique de Prague 1: Mélanges linguistiques dédiés au Premier Congrès des philologues slaves*. Jednota Československých matematiků a fysiků, Prague, 5-29.

František Daneš. 1964. A three-level approach to syntax. Josef Vachek (ed.), *Travaux linguistiques de Prague 1: L’École de Prague d’aujourd’hui*. Éditions de l’Académie Tchécoslovaque des Sciences, Prague, 225-240.

Jan Firbas. 1992. *Functional Sentence Perspective in Written and Spoken Communication*. Cambridge University Press, Cambridge, UK.

Jan Hajič, Alena Böhmová, Eva Hajičová and Barbora Vidová Hladká. 2000. The Prague Dependency Treebank: A Three-Level Annotation Scenario. Anne Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora*. Kluwer, Amsterdam, 103-127.

Eva Hajičová. 2006. Old Linguists Never Die, They Only Get Obligatorily Deleted. *Computational Linguistics*, 32(4): 457-469.

Dag Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008)*. ELRA, Marrakech, 27-34.

Barbara McGillivray and Marco Passarotti. 2009. The Development of the *Index Thomisticus* Treebank

Valency Lexicon. *Proceedings of LaTeCH-SHELT&R Workshop 2009, Athens, March 30, 2009*. 43-50.

Barbara McGillivray, Marco Passarotti and Paolo Ruffolo. 2009. The *Index Thomisticus* Treebank Project: Annotation, Parsing and Valency Lexicon. *Traitement Automatique des Langues*, 50(2): 103-127.

Marie Mikulová, et alii. 2006. *Annotation on the Tectogrammatical Layer in the Prague Dependency Treebank*. Institute of Formal and Applied Linguistics, Prague. Available online from http://ufal.mff.cuni. cz/pdt2.0/doc/manuals/en/t-layer/html/index.html.

Jarmila Panevová. 1974-1975. On verbal Frames in Functional Generative Description. *Prague Bulletin of Mathematical Linguistics*, 22: 3-40. Part II published in PBML, 23: 17-52.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP Framework. *Proceedings of IceTAL, 7th International Conference on Natural Language Processing, Reykjavík, Iceland, August 17, 2010*. 293-304.

Savina Raynaud. 2008. The basic syntagmatic act is predication. *Slovo a slovesnost*, 69(1-2): 49-67.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in its Semantic and Pragmatic Aspects*. D. Reidel, Dordrecht, NL.

John Sinclair. 2004a. Intuition and Annotation – the Discussion Continues. Karin Aijmer & Bengt Altenberg (eds.), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23)*. Rodopi, Amsterdam, 39-59.

John Sinclair. 2004b. Corpus and Text: Basic Principles. Martin Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, Oxford, 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/

Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. J. Benjamins, Amsterdam Philadelphia.