Leaving Behind the Less-Resourced Status.
The Case of Latin through the Experience of the *Index Thomisticus* Treebank

Marco Passarotti
Università Cattolica del Sacro Cuore
Largo Gemelli 1, 20123 Milan, Italy
E-mail: marco.passarotti@unicatt.it

Abstract

Despite its key role in the history of computational linguistics, thanks to the pioneering work by Roberto Busa SJ on the *Index Thomisticus*, Latin can still be considered as a less-resourced language. Although during the last decades several Latin texts have been digitized, only a few of them have been linguistically tagged, while most still lack linguistic tagging at all. However, while the less-resourced status affects historical languages in general, over the past few years a number of language resources for Latin and other historical languages have been started, among which are several treebanks. Presenting the experience of the *Index Thomisticus* Treebank project and, particularly, its valency lexicon, this paper reports some general insights about the creation and use of language resources for less-resourced languages, showing that, although creating from scratch a language resource for a less-resourced language still remains a labor-intensive and time-consuming task, today this is simplified by exploiting the results provided by previous similar experiences in language resources development.

1. Introduction

Despite its key role in the history of computational linguistics, thanks to the pioneering work by Roberto Busa SJ on the *Index Thomisticus* (IT; 1974-1980), Latin can still be considered as a less-resourced language, lacking powerful NLP tools and a broad suite of state-of-the-art language resources (LRs) such as annotated corpora and lexica.

However, while the less-resourced status affects historical languages in general (because of reasons such as being not commercially interesting or lacking native speakers), over the past few years a number of LRs for Latin and other historical languages have been started. Among these LRs are treebanks for Middle, Early Modern and Old English, Early New High German, Medieval Portuguese, Ugaritic, Ancient Greek and several translations of the New Testament into Indo-European languages[1].

As far as Latin is concerned, while trying to meet the needs of the research community working on Latin to have better access to and understanding of textual data, we realized that basic Latin LRs and NLP tools were missing. This is the reason why in 2005 we designed a set of basic LRs and technologies for Latin and started to create a Latin treebank based on the IT data.

Moreover, the collaboration with other similar projects and the exploitation of tools developed over the years for the creation and use of LRs established a kind of virtuous circle for the development of further NLP tools and LRs for Latin, such as lexica. Indeed, the relation between annotated corpora and lexical resources should be circular: while linguistic annotation of textual data is supported and improved by the use of basic lexical resources, these latter can be induced from annotated data in a corpus-driven fashion. This is what we experienced in the *Index Thomisticus* Treebank (IT-TB) while creating a Latin valency lexicon from the annotated data.

Presenting the experience of the IT-TB project, this paper reports some general insights about the creation and use of LRs for less-resourced languages. The paper is organized as follows: section 2 describes the state of the art of the available LRs and NLP tools for Latin; section 3 presents a basic language resource kit for Latin; section 4 deals with some of the main features and achievements of the IT-TB project, presenting the data, the annotation style, the parsing procedures and, particularly, the valency lexicon; finally, section 5 draws some general conclusions on the creation and use of LRs for less-resourced languages and provides an outlook on the next steps of the project.

2. Survey of LRs and NLP tools for Latin

Although during the last decades several Latin texts have been digitized [2], only a few of them have been linguistically tagged, while most still lack linguistic tagging at all.

Only recently (namely, in 2005) two projects started to develop Latin treebanks. These are the IT-TB by the Catholic University in Milan on texts from the IT (McGillivray et al., 2009)[3] and the Latin Dependency Treebank (LDT) by the Perseus Digital Library in Boston on texts of the Classical era (Bamman & Crane, 2007).

[1] For references, see Bamman et al. (2009).

[2] See for instance the Perseus Digital Library at Tufts University in Boston, or the textual databases by CTLO in Turnhout (Centre "Traditio Litterarum Occidentalium") and by LASLA at the University of Liège (Laboratoire d'Analyse Statistique des Langues Anciennes).

[3] Busa started early in the '70s to plan a project aimed at the syntactic annotation of the IT data. Today, the IT-TB project has undertaken this task as part of the wider "Lessico Tomistico Biculturale" project (LTB), whose goal is the development of a Thomistic lexicon grounded on the IT data.

Later on, a third Latin treebank was started at the University of Oslo as part of the project PROIEL (Pragmatic Resources in Old Indo-European Languages), which is aimed at the syntactic annotation of the oldest extant versions of the New Testament in Indo-European languages: Latin, Greek, Gothic, Armenian and Old Church Slavonic (Haug & Jøndal, 2008).

The size of these treebanks is presently around 80,000 annotated words for IT-TB, 55,000 for LDT and 100,000 for the Latin section of the PROIEL corpus.

In regard to Latin lexical resources, many Latin dictionaries and lexica are today available on-line or on CD-ROM. Some of the most relevant are the Lewis-Short dictionary provided by Perseus, the *Thesaurus Linguae Latinae* from the Bayerische Akademie der Wissenschaften in Munich, the *Thesaurus Formarum* (TF-CILF) from the CTLO and the *Neulateinische Wortliste* by Johann Ramminger (http://www.lrz-muenchen.de/~ramminger/). Presently, the main project aimed at developing a Latin lexical resource is Latin WordNet (Minozzi, 2008), which is integrated within the wider MultiWordNet project (http://multiwordnet.fbk.eu).

However, although WordNet is a lexical resource that can be used for NLP tasks such as information extraction, data mining, word sense disambiguation and topic classification, the available NLP tools for Latin are still far from providing automatic processing of such tasks. In this domain, three morphological analysers of Latin are presently available, namely LEMLAT (Passarotti, 2004), Whitaker's *Words* (http://archives.nd.edu/words.html) and *Morpheus* (Crane, 1991), this latter being first created for Ancient Greek in 1985 and extended to support Latin in 1996. Specific tools for morpho-syntactic disambiguation and Part-of-Speech (PoS) tagging have been developed by LASLA for the annotation of their textual database, while a first attempt at Latin dependency parsing is described by Koch (1993), who reports on the enhancement for Latin of an existing dependency parser. Finally, Koster (2005) describes a rule-based top-down chart parser, automatically generated from a grammar and a lexicon built according to a two-level formalism (AGFL: Affix Grammar over a Finite Lattice).

3. A Basic Language Resource Kit for Latin

In order to identify the best strategy to follow over the upcoming years to move Latin from being a less-resourced language to being a language with basic LRs, we first defined a basic minimal set of underlying LRs and tools that are considered necessary for language technology applications working on Latin.

To achieve this aim, we grounded our decisions on the BLARK concept (Basic Language Resource Kit) consisting in defining "for every language a specification of the minimum general text or spoken corpus, basic tools to manipulate it and skills required to be able to do any pre-competitive research for the language" (Mapelli & Choukri, 2003, p. 4).

Since spoken data is missing for Latin, we sketched a set of basic components comprised of technologies for written languages only. Among human language technologies (HLT), BLARK distinguishes between modules (software components used for the development of HLT applications), applications (which make use of HLT) and data (used to create, refine and evaluate the modules). Following these requirements, a BLARK-like set was sketched for Latin, consisting of the following components.

Modules:

- Text pre-processing (tokenization and named-entity recognition)
- Lemmatization: morphological analysis and morpho-syntactic disambiguation (PoS taggers)
- Syntactic analysis: parsers and shallow parsing
- Anaphora resolution
- Semantic and pragmatic analysis

Applications:

- Entering and acquiring information: typing, digitization, annotation, OCR systems[4]
- Document management: automatic and computer-assisted indexing
- Information retrieval and presentation

Data:

- Unannotated corpus of text
- Syntactically annotated corpus of text (treebank)
- Monolingual lexicon (valency lexicon)
- Semantically and pragmatically annotated corpus of text

Considering the state of the art of Latin LRs and NLP tools, the following were recognized as the components most urgently needed in order to meet the requirements of the basic set.

- Modules: NLP tools for the automatic processing of the morpho-syntactic and syntactic layers of annotation (PoS taggers and parsers)
- Applications: tools for data annotation and for information retrieval
- Data: a treebank and a valency lexicon

Although components like semantic and pragmatic analysis, as well as anaphora resolution, are part of the set, we deferred their development, since we believe syntax to be an essential level of analysis in view of such "higher" tasks.

Moreover, the present availability of data-driven and language-independent NLP tools, such as probabilistic PoS taggers and parsers, strengthened our idea of starting to build the basic kit for Latin beginning first of all with the development of a Latin treebank. Indeed, our short-term perspective was to exploit the data from the treebank in two ways: (a) to train NLP tools for morpho-syntactic disambiguation and syntactic analysis, and (b) to use the data as the basis for several subsequent layers of annotation, including anaphora resolution, and semantic and pragmatic analysis. Similarly, we wanted to enhance the syntactic annotation with valency

[4] Specific OCR systems for printed and handwritten characters are particularly required for digital and computational philology purposes.

information, in order to induce a valency lexicon from the treebank data.

4. The *Index Thomisticus* Treebank

4.1 The *Index Thomisticus*

Started by Roberto Busa SJ in 1949, the IT is a corpus containing 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 words, each morphologically tagged and lemmatized by hand. The corpus can be browsed on CD-ROM or on-line at the following address: http://www.corpusthomisticum.org.

4.2 Annotation Style

Since the *Index Thomisticus* Treebank and the Latin Dependency Treebank were the first projects of their kind for Latin, no prior established guidelines were available to rely on for syntactic annotation. Rather than have each treebank project decide upon and record each decision for annotating the data, the two projects decided to pool their resources and create a single annotation manual that would govern both treebanks (Bamman et al., 2007). Rather than design the manual from scratch, we chose to follow the annotation style developed for the 'analytical layer' by the Prague Dependency Treebank of Czech (PDT; Hajič et al., 1999)[5]. Only minor changes were applied, for the treatment of specific or idiosyncratic constructions of Latin (Bamman et al., 2008).

PDT is a dependency-based treebank with a three-layer structure, ordered as follows: (1) a morphological layer: lemmatization and full morphological annotation; (2) an 'analytical layer': dependency-based superficial (surface) syntactic annotation; (3) a 'tectogrammatical layer': annotation of the underlying meaning of the sentence, based on the Functional Generative Description framework (FGD; Sgall et al., 1986).

In the IT-TB and LDT projects, we have chosen the PDT annotation style for both linguistic and "structural" reasons.

As far as the former are concerned, Latin and Czech share some relevant properties such as being richly inflected, having a moderately free word-order and an high degree of synonymity and ambiguity of the endings, and showing discontinuous phrases (i.e. phrases broken up by words of other phrases: 'non-projectivity')[6]. Both languages have 3 genders (masculine, feminine, neuter), cases with roughly the same meaning and no articles.

As for the latter, the PDT three-layer structure is ideal both for our present needs and for the perspectives of development of new Latin LRs and tools. Indeed, the analytical annotation in PDT is not meant to be a layer standing on its own, but is intended as a technical step towards the tectogrammatical annotation. The strict relation between the overall structure of the annotation workflow in PDT and a sound background theory like FGD allows us to consider each single layer of annotation as one part of a general framework that is driven by a functional perspective aimed at understanding the underlying meaning of the sentence. This task is performed through topic-focus articulation tagging, ellipsis resolution and semantic role labelling, this latter making use of labels (called 'functors') such as Actor, Patient, Addressee, Origin, Effect and several kinds of free adverbials (temporal, local, causal, manner, etc.). Pragmatic tagging (topic-focus articulation), anaphora resolution and, ultimately, semantic analysis are just components of the basic kit of Latin LRs and tools that are still missing.

Moreover, the adoption of PDT as the main reference framework not only provided our annotation efforts with a sound theoretical background, but also gave us the opportunity to re-use tools for annotation and retrieval which had been developed by PDT for its own purposes. Particularly, for IT-TB manual and semi-automatic annotation we adopted the tree editor TrEd by Petr Pajas, while on-line browsing of the IT-TB data can be performed through the searching and viewing interface Netgraph (Mírovský, 2006) at the IT-TB website: http://itreebank.marginalia.it.

4.3 Parsing and PoS Tagging

After an early phase of manual annotation, we started to exploit the available annotated data to train and test a number of probabilistic dependency parsers. This was done in order to increase the quality and speed of the annotation process. Indeed, in this way annotators no longer have to draw trees from scratch, but need only to check the automatically produced trees and to manually correct mistakes.

In our recent work (Passarotti & Dell'Orletta, 2010), we describe a number of modifications that we applied to DeSR parser (Dependency Shift-Reduce; Attardi, 2006), including the design of a feature model specific to Medieval Latin as well as revision and combination techniques. Using a training set of 61,024 tokens (2,820 sentences), this improved the previously available accuracy rates, reaching 80.02% for LAS, 85.23% for UAS and 87.79% for LA[7].

Since the IT data are morphologically tagged, our first priority has been automatic syntactic parsing. However, we also started to train PoS taggers, in order to automatically perform morpho-syntactic disambiguation of the IT morphological lemmatization. Bamman and

[5] Although they differ in some details, the PROIEL treebank annotation guidelines are quite similar to those governing IT-TB and LDT. An automatic conversion procedure from PROIEL to the IT-TB and LDT annotation style is ongoing.

[6] The condition of projectivity in a dependency tree says that if a node *a* depends on *b* and there is a node *c* between *a* and *b* in the linear ordering, *c* depends (directly or indirectly) on *b*. The non-projective nodes are those where such condition is not met.

[7] LAS (Labeled Attachment Score) is the percentage of tokens with correct head and relation label; UAS (Unlabeled Attachment Score) is the percentage of tokens with correct head; LA (Label Accuracy) is the percentage of tokens with correct relation label (Buchholz & Marsi, 2006).

Crane (2008) report accuracy rates of around 95% in resolving PoS, reached with a PoS tagger (TreeTagger; Schmid, 1994) trained on a set of approximately 47,000 tokens from LDT. Our preliminary results for PoS tagging, using the HMM-based HunPos tagger (Halácsy et al., 2007) and the IT-TB training set (61,024 tokens), were the following: 96.75% in correctly disambiguating coarse-grained PoS + fine-grained PoS, and 89.90% if morphological features are also considered[8].

Given the small training set, these are quite high rates, resulting from the use of language-independent NLP tools that were not specifically designed for IT-TB purposes.

4.4 Valency Lexicon

The present availability of Latin treebanks fosters the creation of new lexical resources for Latin that match with the annotated data. Indeed, the evidence provided by such corpora can be fully represented in lexical resources induced from the data. Subsequently, such resources can in turn be used to support the annotation of new textual data.

In particular, the creation of a lexicon can be pursued by both intuition-based and data-driven approaches, according to the role played by human intuition and by the empirical evidence provided by annotated corpora such as treebanks.

For instance, lexica like PropBank (Kingsbury & Palmer, 2002), FrameNet (Ruppenhofer et al., 2006) and PDT-VALLEX (Hajič et al., 2003) have been created in an intuition-based fashion and then checked and improved with examples excerpted from corpora.

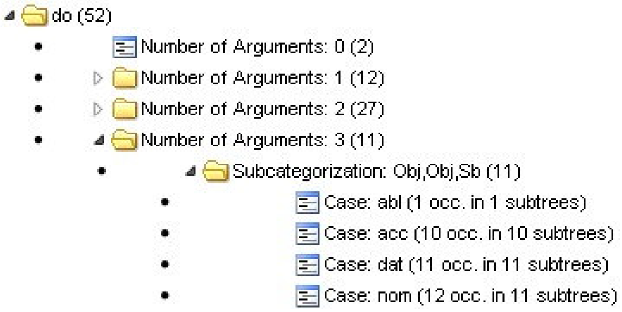
On the other hand, research in lexical acquisition has recently made available a number of data-driven valency lexica automatically acquired from annotated corpora, such as VALEX (Korhonen et al., 2006) and LexShem (Messiant et al., 2008).

In the IT-TB project we followed a data-driven approach, inducing a valency lexicon for Latin verbs from IT-TB data (McGillivray & Passarotti, 2009). The notion of valency is generally defined as the number of obligatory complements required by a word: these complements are usually named 'arguments', while the non-obligatory ones are referred to as 'adjuncts'. Although valency can be assigned to different PoS (usually verbs, nouns and adjectives), scholars have mainly focused their attention on verbs, so that the notion of valency often coincides with verbal valency. Presently, the size of the IT-TB valency lexicon is 432 entries (corresponding to 5,966 verbal occurrences in the treebank)[9]. The lexicon is automatically updated as the amount of the annotated data increases.

A similar approach has been pursued by LDT. Bamman and Crane (2008) describe a Latin 'dynamic lexicon' automatically extracted from the Perseus Digital Library, using LDT data as training set. The lexicon reports qualitative and quantitative information on the subcategorization patterns and selectional preferences of each word as it is used in every Latin author of the corpus. Relying on morphological tagging and statistical syntactic parsing of a large corpus (around 3.5 million words), only the most common arguments and the most common lexical fillers of these arguments are shown, thus reducing the noise caused by the automatic pre-processing of data.

While PDT, as a project, represents the main reference model for IT-TB, in the development of the valency lexicon we did not follow the same approach. Indeed, while PDT-VALLEX was created before the annotation of PDT started and the annotated data were linked subsequently to the single items in the lexicon, the IT-TB valency lexicon results from the opposite procedure. The lexicon is created in an annotation-driven fashion and the valency of a lexical item is defined as annotators get through its first occurrence in the data. Furthermore, since the IT-TB valency lexicon relies on data annotated on the analytical layer (and not on the tectogrammatical one), it just reports for each entry the number of the arguments occurring on the surface syntactic structure, while no information on semantic roles is provided[10].

This approach has pros and cons. On the one hand, not grounding the annotation decisions on a previously available valency lexicon developed in an intuition-based fashion can lead to inconsistencies in annotation, since annotators do not make their decisions about valency on the basis of one common lexicon. On the other, the exploitation of the data in our approach allows an in-depth evaluation of the quality of the annotation, making it possible to discover inconsistencies and to make decisions on unclear cases. At any rate, our choice to develop a valency lexicon from an available treebank was strictly motivated by the less-resourced status of Latin, which requires that the creation of a new LR results from exploiting as much as possible the available resources.



- do (52)
  - Number of Arguments: 0 (2)
  - Number of Arguments: 1 (12)
  - Number of Arguments: 2 (27)
  - Number of Arguments: 3 (11)
    - Subcategorization: Obj,Obj,Sb (11)
      - Case: abl (1 occ. in 1 subtrees)
      - Case: acc (10 occ. in 10 subtrees)
      - Case: dat (11 occ. in 11 subtrees)
      - Case: nom (12 occ. in 11 subtrees)

Figure 1: Mock-up of part of the entry for *do, -are*.

The lexicon will be soon made accessible at the IT-TB website through a user-friendly interface. Figure 1 shows

[8] In the IT tagset there are 5 different coarse-grained PoS tags and 14 fine-grained PoS tags.

[9] The only available list of Latin verbs which also provides their valencies is reported by Happ (1976, pp. 480--565), who created a lexicon on a basis of around 800 verbal occurrences excerpted from Cicero's *Orationes*.

[10] This means that, in the IT-TB valency lexicon, the arguments are distinguished according to their syntactic role, i.e. 'analytical' function: subject, (direct or indirect) object, nominal predicate and object complement.

an example of what a lexical entry looks like in the interface. It reports a part of the entry for the verb *do, -are* (*to give*). It is shown that in the IT-TB there are 52 instances of *do* , 2 of which occur with no arguments, 12 with 1 argument, 27 with 2 arguments and 11 with 3 arguments. In particular, the subcategorization pattern of all the 11 trivalent cases is formed by 2 objects and 1 subject; furthermore, information on the case of each argument is provided[11].

Clicking on each level of the lexical entry, all the sentences included in such level are shown and, for each of them, the place of its occurrence in the IT-TB and a subtree showing the verbal subcategorization pattern are provided. In each sentence, the verbal head and its arguments are highlighted (in different colours).

The lexicon can be browsed in many other ways, as for instance by number of arguments, subcategorization pattern, surface order of the arguments and lexical fillers. For example, the lexicon can be queried searching for all the verbal occurrences having 2 arguments, one of which is a subject and the other an indirect object (in dative), the latter occurring in preverbal position and being a form of a specific lemma[12].

5. Discussion and Conclusion

Although creating from scratch an LR for a less-resourced language still remains a labor-intensive and time-consuming task, today this is simplified by exploiting the results provided by previous similar experiences in LRs development. Such results, in terms of methods, data and tools, can be re-used with limited effort and applied to other languages.

Together with the use of those LRs and tools that are available for the less-resourced language in question, the re-use of previous research experience is even more helpful in those cases where they involve LRs for languages that share certain primary properties with the less-resourced language for which new LRs have to be created. These language relationships can be used for porting, in a a rapid and low-cost fashion, LRs and tools from one language to another, taking an approach to LRs creation and use that stands in the middle between knowledge-free approaches and knowledge-intensive ones.

The IT-TB is a case which shows how good results can be achieved in quite a short time, through adapting already existing language technologies developed over the years for well-resourced languages and particularly for Czech, which shares with Latin a number of linguistic properties. For instance, the re-use of tagging and browsing software applications that were created for PDT purposes allowed saving time and funds that otherwise should be spent to create such tools specifically for IT-TB.

Furthermore, the language-independent nature of available probabilistic NLP tools makes them extremely useful for the purposes of projects aimed at the creation of LRs for less-resourced languages, since there is no need to develop specific (usually, rule-based) NLP tools for the processing of just one language or, sometimes, for the aims of just one project.

Once a small amount of annotated data has been made available by the IT-TB project, this has been in turn used (a) to train probabilistic NLP tools, such as PoS taggers and parsers, achieving promising results despite a quite small training set, and (b) to induce another new LR for Latin, namely the IT-TB valency lexicon.

One of the advantages of working on a less-resourced language is the small number of people who are involved. Although these people usually work on different projects, they can easily collaborate to find common solutions to common problems. Such collaboration can start with the very beginning of the projects, as in the IT-TB and LDT cases, where common annotation guidelines were developed before the annotation of data was started. This allows the setting of standards that are really shared by the projects and not imposed on the projects in a top-down manner. In our case, collaboration with LDT is even further essential since Latin is a language with a long diachronic usage extending over more than two thousand years. While the two projects are dealing with Latin dialects separated by 13 centuries, sharing a single annotation manual proved to be very useful for comparison purposes, such as checking annotation consistency, making annotation decisions on a wider number and kind of examples, or diachronically studying specific syntactic constructions.

The overall design is important as well. Grounding a project on a sound theoretical framework (like FGD) and aiming at the creation of a pre-defined set of basic LRs motivates each step of the work, which is thus considered in a wider perspective.

Our goal in the near future is to apply named-entity recognition systems to the IT data and to enlarge the amount of analytically annotated data in IT-TB, relying on the good results provided by DeSR. Annotation of data at the tectogrammatical layer will be started as well, still grounding on PDT guidelines and using TrEd as annotation editor. This will also enrich the IT-TB valency lexicon, enhancing the current argument information with semantic roles (functors). Finally, since PROIEL is a multilingual resource providing syntactic annotation of the same texts from the New Testament in several different languages, the PROIEL annotated corpus is a good starting point for the development of a multilingual valency lexicon based on treebank data. This multilingual aspect will be further improved by linking the lexical entries of the IT-TB valency lexicon with the corresponding entries in the Latin WordNet and, from there, they will be linked to the WordNets of all the other languages included in the MultiWordNet project.

[11] In figure 1, 'abl' stands for 'ablative', 'acc' for 'accusative', 'dat' for 'dative' and 'nom' for 'nominative'. The ablative argument occurs in the passive use of *do*, whose agentive argument is a prepositional phrase headed by the preposition *a/ab* (*by*), which takes the ablative case.

[12] In those cases where an argument is not a single word or a prepositional phrase but a subordinate clause, the lexical filler reported in the lexicon is the verb heading this clause.

6. References

Attardi, G. (2006). Experiments with a Multilanguage Non-Projective Dependency Parser. In *Proceedings of the CoNLL-X*, pp. 166--170.

Bamman, D. & Crane, G. (2007). The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of LaTeCH 2007. Prague, Czech Republic*, pp. 33--40.

Bamman, D. & Crane, G. (2008). Building a Dynamic Lexicon from a Digital Library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2008).*

Bamman, D., Crane, G., Passarotti, M. & Raynaud, S. (2007). *Guidelines for the Syntactic Annotation of Latin Treebanks.* Technical report. Boston: Tufts Digital Library.

Bamman, D., Mambrini, F. & Crane, G. (2009). An Ownership Model of Annotation: The Ancient Greek Dependency Treebank. In *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT8). Milan, Italy*, pp. 5--15.

Bamman, D., Passarotti, M., Busa, R. & Crane, G. (2008). The annotation guidelines of the Latin Dependency Treebank and *Index Thomisticus* Treebank. The treatment of some specific syntactic constructions in Latin. In *Proceedings of LREC 2008. Marrakech, Morocco*.

Buchholz, S. & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *CoNLLX, SIGNLL, 2006*.

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

Crane, G. (1991). Generating and Parsing Classical Greek. *Literary and Linguistic Computing*, vol. 6, n. 4, pp. 243--245.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z. & Bémová, A. (1999). *Annotations at analytical level: Instructions for annotators*. Technical report. Prague, Czech Republic: ÚFAL MFF UK.

Hajič, J., Panevová, J., Buráňová, E., Urešová, Z., Bémová, A., Kolárová-Rezníčková, V. & Pajas, P. (2003). PDT-VALLEX: Creating a Large Coverage Valency Lexicon for Treebank Annotation. In *TLT 2003 – Proceedings of the Second Workshop on Treebanks and Linguistic Theories*. Vol. 9 of *Mathematical Modelling in Physics, Engineering and Cognitive Sciences*, pp. 57--68.

Halácsy, P., Kornai, A. & Oravecz, C. (2007). HunPos – an open source trigram tagger. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 209--212.

Happ, H. (1976). *Grundfragen einer Dependenz-Grammatik des Lateinischen*. Goettingen: Vandenhoeck & Ruprecht.

Haug, D.T.T. & Jøndal, M.L. (2008). Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of LaTeCH Workshop - LREC 2008. Marrakech, Morocco*, pp. 27--34.

Kingsbury, P. & Palmer, M. (2002). From Treebank to Propbank. In *Proceedings of LREC 2002. Las Palmas – Gran Canaria, Spain*.

Koch, U. (1993). *The Enhancement of a Dependency Parser for Latin*. Technical Report n° AI-1993-03, Artificial Intelligence Programs, University of Georgia.

Korhonen, A., Krymolowski, Y. & Briscoe, T. (2006). A Large Subcategorization Lexicon for Natural Language Processing Applications. In *Proceedings of LREC 2006. Genoa, Italy.*

Koster, C.H.A. (2005). Constructing a Parser for Latin. In *Computational Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science*. Berlin-Heidelberg: Springer, pp. 48--59.

Mapelli, V. & Choukri, K. (2003). *Report on a (minimal) set of LRs to be made available for as many languages as possible, and map of the actual gaps*. ENABLER project internal report, Deliverable 5.1.

McGillivray, B. & Passarotti, M. (2009). The Development of the *Index Thomisticus* Treebank Valency Lexicon. In *Proceedings of LaTeCH-SHELT&R Workshop 2009. March 30th 2009, Athens, Greece.*

McGillivray, B., Passarotti, M. & Ruffolo, P. (2009). The *Index Thomisticus* Treebank Project: Annotation, Parsing and Valency Lexicon. *Traitement Automatique des Langues*, 50 (2), pp. 103--127.

Messiant, C., Korhonen, A. & Poibeau, T. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of LREC 2008. Marrakech, Morocco*.

Minozzi, S. (2008). La costruzione di una base di conoscenza lessicale per la lingua latina: Latinwordnet. In *Studi in onore di Gilberto Lonardi*. Verona: Fiorini, pp. 243--258.

Mírovský, J. (2006). Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0. In *TLT 2006. Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories. December 1-2, 2006, Prague, Czech Republic*, pp. 211--222.

Passarotti, M. (2004). Development and perspectives of the Latin morphological analyser LEMLAT. In A. Bozzi, L. Cignoni & J.L. Lebrave (Eds.), *Digital Technology and Philological Disciplines. Linguistica Computazionale*, XX-XXI, pp. 397--414.

Passarotti, M. & Dell'Orletta, F. (2010). Improvements in Parsing the *Index Thomisticus* Treebank. Revision, Combination and a Feature Model for Medieval Latin. In *Proceedings of LREC 2010. La Valletta, Malta*.

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R. & Scheffczyk, J. (2006). *FrameNet II. Extendend Theory and Practice*. E-book available at http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126.

Schmid, G. (1994). *TreeTagger - a language independent part-of-speech tagger*. Available at http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

Sgall, P., Hajičová, E. & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: Reidel.