The Development of the *Index Thomisticus* Treebank Valency Lexicon

Barbara McGillivray
University of Pisa
Italy
b.mcgillivray@ling.unipi.it

Marco Passarotti
Catholic University of the Sacred Heart
Milan, Italy
marco.passarotti@unicatt.it

Abstract

We present a valency lexicon for Latin verbs extracted from the *Index Thomisticus* Treebank, a syntactically annotated corpus of Medieval Latin texts by Thomas Aquinas.
In our corpus-based approach, the lexicon reflects the empirical evidence of the source data. Verbal arguments are induced directly from annotated data.
The lexicon contains 432 Latin verbs with 270 valency frames. The lexicon is useful for NLP applications and is able to support annotation.

1 Introduction

Over the last decades, annotated corpora and computational lexicons have gained an increasing role among language resources in computational linguistics: on the one hand, they are used to train Natural Language Processing (NLP) tools such as parsers and PoS taggers; on the other hand, they are developed through automatic procedures of linguistic annotation and lexical acquisition.

The relation between annotated corpora and computational lexicons is circular: as a matter of fact, if linguistic annotation of textual data is supported and improved by the use of lexicons, these latter can be induced from annotated data in a corpus-based fashion.

In the field of cultural heritage and in particular that of classical languages studies, much effort has been devoted throughout the years to the digitization of texts, but only recently have some projects begun to annotate them above the morphological level.

Concerning lexicology and lexicography of classical languages, a long tradition has produced and established many dictionaries, thesauri and lexicons, providing examples from real texts. Nevertheless, nowadays it is possible and indeed necessary to match lexicons with data from (annotated) corpora, and viceversa. This requires the scholars to exploit the vast amount of textual data from classical languages already available in digital format,[1] and particularly those annotated at the highest levels. The evidence provided by the texts themselves can be fully represented in lexicons induced from these data. Subsequently, these lexicons can be used to support the textual annotation itself in a virtuous circle.

This paper reports on the creation of a valency lexicon induced from the *Index Thomisticus* Treebank, a syntactically annotated corpus of Medieval Latin texts by Thomas Aquinas. The paper is organised as follows: section 2 describes the available Latin treebanks, their annotation guidelines and gives some specific information on the Index Thomisticus treebank; section 3 deals with the notion of valency, while section 4 describes the state of the art on valency lexicons; section 5 illustrates the procedures of acquisition and representation of our valency lexicon; finally, section 6 draws some conclusions and describes future work.

2 Latin Treebanks

Latin is a richly inflected language, showing:

- discontinuous constituents ('non-projectivity'): this means that phrasal constituents may not be continuous, but broken up by words of other constituents. An example is the following sentence by Ovid (Metamorphoses, I.1-2): "In nova fert animus mutatas dicere formas corpora" ("My mind leads me to tell of forms changed into new bodies"). In this sentence, both the nominal phrases "nova corpora" and "mutatas formas" are discontinuous;
- moderately free word-order: for instance, the order of the words in a sentence like "au-

[1] See, for instance, the Perseus Digital Library (Crane et al., 2001), or data repositories such as LASLA (Denooz, 1996).

Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education –LaTeCH – SHELT&R 2009, pages 43–50,
Athens, Greece, 30 March 2009.

daces fortuna iuvat" ("fortune favours the bold") could be changed into "fortuna audaces iuvat", or "fortuna iuvat audaces", without affecting the meaning of the sentence.

These features of Latin influenced the choice of Dependency Grammars (DG)[2] as the most suitable grammar framework for building Latin annotated corpora like treebanks.

While since the 1970s the first treebanks were annotated via Phrase Structure Grammar (PSG)-based schemata (as in IBM, Lancaster and, later on, Penn treebanks), in the past decade many projects of dependency treebanks development have started, such as the ALPINO treebank for Dutch (Van der Beek et al., 2002), the Turin University Treebank for Italian (Lesmo et al., 2002), or the Danish Dependency Treebank (Kromann, 2003). On the one hand, this is due to the fact that the first treebanks were mainly English language corpora. PSG were a suitable framework for a poorly inflected language like English, showing a fixed word-order and few discontinuous constituents. Later on, the syntactic annotation of moderately free word-order languages required the adoption of the DG framework, which is more appropriate than PSG for such a task. On the other hand, Carroll et al. (1998) showed that inter-annotator agreement was significantly better for dependency treebanks, indicating that phrase structure annotation was requiring too many irrelevant decisions (see also Lin, 1995).

Although much Latin data is nowadays available in digital format, the first two projects for the development of Latin treebanks have only recently started: namely the Latin Dependency Treebank (LDT) at the Tufts University in Boston (within the Perseus Digital Library) based on texts of the Classical era (Bamman, 2006), and the Index Thomisticus Treebank (IT-TB) at the Catholic University of the Sacred Heart in Milan, based on the *Opera omnia* of Thomas Aquinas (Passarotti, 2007).

Taking into account the above mentioned features of Latin, both the treebanks independently chose the DG framework as the most suitable one for data annotation. The same approach was later on followed by a third Latin treebank now available, which is ongoing at the University of Oslo in the context of the PROIEL project (Pragmatic Resources in Old Indo-European Languages): the aim of PROIEL is the syntactic annotation of the oldest extant versions of the New Testament in Indo-European languages, including Greek, Latin, Gothic, Armenian and Church Slavonic (Haug and Jøhndal, 2008).

[2] With Tesnière (1959) as a common background, there are many different current DG flavours. See for instance the following: Dependency Unification Grammar (Hellwig, 1986), Functional Generative Description (Sgall, Hajičová and Panevová, 1986), Meaning Text Theory (Mel'čuk, 1988), Word Grammar (Hudson, 1990).

2.1 Annotation Guidelines

Since LDT and IT-TB were the first projects of their kind for Latin, no prior established guidelines were available to rely on for syntactic annotation.

Therefore, the so-called 'analytical layer' of annotation of the Prague Dependency Treebank (PDT) for Czech (Hajič et al., 1999) was chosen and adapted to specific or idiosyncratic constructions of Latin. These constructions (such as the ablative absolute or the passive periphrastic) could be syntactically annotated in several different ways and are common to Latin of all eras. Rather than have each treebank project decide upon and record each decision for annotating them, LDT and IT-TB decided to pool their resources and create a single annotation manual that would govern both treebanks (Bamman et al., 2007a; Bamman et al., 2007b; Bamman et al., 2008).

As we are dealing with Latin dialects separated by 13 centuries, sharing a single annotation manual is very useful for comparison purposes, such as checking annotation consistency or diachronically studying specific syntactic constructions. In addition, the task of data annotation through these common guidelines allows annotators to base their decisions on a variety of examples from a wider range of texts and combine the two datasets in order to train probabilistic dependency parsers.

Although the PROIEL annotation guidelines are grounded on the same grammar framework as the LDT and IT-TB, they differ in a number of details, some of which are described in Passarotti (forthcoming).

2.2 The *Index Thomisticus* Treebank

The *Index Thomisticus* (IT) by Roberto Busa SJ (1974-1980) was begun in 1949 and is considered a groundbreaking project in computational linguistics. It is a database 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 corpus is morphologically tagged and lemmatised.

Early in the 1970's Busa started to plan a project aimed at both the morphosyntactic disambiguation of the IT lemmatisation and the syntactic annotation of its sentences. Today, these tasks are performed by the IT-TB project, which is part of the wider 'Lessico Tomistico Biculturale', a project whose target is the development of a lexicon from the IT texts.[3]

Presently, the size of the IT-TB is 46,456 tokens, for a total of 2,103 parsed sentences excerpted from the *Scriptum super Sententiis Magistri Petri Lombardi*.

3 Valency

As outlined above, the notion of valency is generally defined as the number of complements required by a word: these obligatory complements are usually named 'arguments', while the non-obligatory ones are referred to as 'adjuncts'. Although valency can refer to different parts of speech (usually verbs, nouns and adjectives), scholars have mainly focused their attention on verbs, so that the notion of valency often coincides with verbal valency.

Valency is widely used in DG formalisms, but it also figures in PSG-based formalisms like HPSG and LFG.

While Karl Bühler can be considered as the pioneer of the modern theory of valency,[4] Lucien Tesnière is widely recognised as its real founder. Tesnière views valency as a quantitative quality of verbs, since only verbs constrain both the quantity and the quality (i.e. nouns and adverbs) of their obligatory arguments; through a metaphor borrowed from drama, Tesnière classifies dependents into *actants* (arguments) and *circonstants* (adjuncts): "Le noeud verbal […] exprime tout un petit drame. Comme un drame en effet, il comporte obligatoirement un procès, et le plus souvent des acteurs et des circonstances. Transposés du plan de la réalité dramatique sur celui de la syntaxe structurale, le procès, les acteurs et les circonstances deviennent respectivement le verbe, les actants et les circonstants" (Tesnière, 1959: 102).[5]

Arguments can be either obligatory or optional, depending on which sense of the verb is involved. For example, the *seem* sense of the verb *appear* requires two obligatory arguments in active clauses, as in the following sentence: "That lawyer appears to love his work". Here the second argument ("to love his work") cannot be left out without changing the meaning of the verb. On the other hand, optional arguments are recorded into the verbal argument structure itself, althought they may not appear at the clausal level. For instance, in the following sentence the object required by the verb *eat* is missing, but the sentence is still acceptable: "He eats (something)".

Optionality can also act at the communicative level as well as at the structural one. For instance, adjuncts can be necessary for communicative intelligibility in particular contexts, as in the following sentence: "I met James at the Marquee club", where the locative adverbial ("at the Marquee club") is required to answer a question like "Where did you meet James?". On the other hand, structural optionality depends on the features of the language and applies at the clausal level. For instance, as a poorly inflected language, English requires the subject of a predicate to be expressed in declarative and interrogative main clauses, so that a sentence like the following is ungrammatical if the subject is missing: "[I] slept all morning".

Given the so-called "syntax-semantics interface" (Levin, 1993), arguments are generally associated with a predicate sense rather than a predicate form, and are structured in sequences called 'subcategorization frames' (SCFs) or 'complementation patterns'. For example, there is a semantic difference between the *bill* sense and the *attack* sense of the verb *charge* in English, as in the following sentences:

- (a) "The hotel charges 80 euros for a night".
- (b) "The army charged the enemy".

In these sentences, the two predicate senses show two different SCFs:

- (a) [Subj_NP, Pred, Obj_NP, Obj_PP-for]
- (b) [Pred, Obj_NP]

Arguments are also selected by verbs according to lexical-semantic properties, called 'selectional preferences' (SPs) or 'selectional restrictions'. For example, a sentence like "*The train flew to Rome" is ungrammatical, since it violates

[3] http://itreebank.marginalia.it.

[4] In the *Sprachtheorie*, he writes that "die Wörter einer bestimmten Wortklasse eine oder mehrere Leerstellen um sich eröffnen, die durch Wörter bestimmter anderer Wortklassen ausgefüllt werden müssen" (Bühler, 1934: 173) ("words of a certain word-class open up around themselves one or several empty spaces that have to be filled by words of certain other word-classes"; our translation).

[5] "The verbal node expresses a whole little drama. As a drama, it implies a process and, most of the times, actors and circumstances. Transposed from the dramatic reality to structural syntax, the process, the actors and the circumstances respectively become the verb, the actants and the circumstants" (our translation).

the SP of the verb *fly* on its subject and can only be accepted in a metaphorical context.

4 Valency Lexicons

Over the past years, several valency lexicons have been built within different theoretical frameworks: these lexicons have an important role in the NLP community thanks to their wide applications in NLP components, such as parsing, word sense disambiguation, automatic verb classification and selectional preference acquisition.

As shown in Urešová (2004), a valency lexicon can also help the task of linguistic annotation (as in treebank development), providing annotators with essential information about the number and types of arguments realized at the syntactic level for a specific verb, along with semantic information on the verb's lexical preferences.

In the phase of lexicon creation, both intuition-based and corpus-based approaches can be pursued, according to the role played by human intuition and empirical evidence extracted from annotated corpora such as treebanks.

For instance, lexicons like PropBank (Kingsbury and 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 from corpora.

On the other side, research in lexical acquisition has recently made available a number of valency lexicons automatically acquired from annotated corpora, such as VALEX (Korhonen, et al., 2006) and LexShem (Messiant et al., 2008). Unlike the fully intuition-based ones, these lexicons aim at systematically reflecting the evidence provided by data, with very little human intervention. The role of intuition is therefore left to the annotation phase (where the annotator interprets the corpus data), and not extended to the development of the lexicon itself.

Corpus-based lexicons show several advantages if compared with traditional human-developed dictionaries. Firstly, they systematically reflect the evidence of the corpus they were extracted from, while acquiring information specific to the domain of the corpus. Secondly, unlike manually built lexicons, they are not prone to human errors that are difficult to detect, such as omissions and inconsistencies. In addition, such lexicons usually display statistical information in their entries, such as the actual frequency of subcategorization frames as attested in the original corpus. Finally, they are less costly than hand-crafted lexical resources in terms of time, money and human resources.

While several subcategorization lexicons have been compiled for modern languages, much work in this field still remains to be done on classical languages such as Greek and Latin. Regarding Latin, Happ reports a list of Latin verbs along with their valencies (Happ, 1976: 480-565). Bamman and Crane (2008) describe a "dynamic lexicon" automatically extracted from the Perseus Digital Library, using the LDT as a training set. This lexicon displays qualitative and quantitative information on 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 such a large corpus, their approach finds the most common arguments and the most common lexical fillers of these arguments, thus reducing the noise caused by the automatic pre-processing of the data.

5 The *Index Thomisticus* Treebank Valency Lexicon

We propose a corpus-based valency lexicon for Latin verbs automatically induced from IT-TB data. The automatic procedure allows both the extension of this work to the LDT (thanks to the common annotation guidelines) and the updating of the lexicon as the treebank size increases.

First, we automatically extract the arguments of all the occurrences of verbal lemmata in the treebank, along with their morphological features and lexical fillers.

In the IT-TB, verbal arguments are annotated using the following tags: Sb (Subject), Obj (Object), OComp (Object Complement) and Pnom (Predicate Nominal); adjuncts are annotated with the tag Adv (Adverbial). The difference between Obj and Adv corresponds to the that between direct or indirect arguments (except subjects) and adjuncts. A special kind of Obj is the determining complement of the object, which is tagged with OComp, such as *senatorem* in the phrase "aliquem senatorem facere" ("to nominate someone senator"). Conversely, the determining complement of the subject is tagged as Pnom, as in "aliquis senator fit" ("someone becomes senator").[6]

[6] As in the PDT, all of the syntactic tags can be appended with a suffix in the event that the given node is member of a coordinated construction (_Co), an apposition (_Ap) or a parenthetical statement (_Pa).

In order to retrieve the arguments realised for each verbal occurrence in the treebank, specific database queries have been created to search for the nodes depending on a verbal head through the functional tags listed above.

The head-dependent relation can be either direct or indirect, since intermediate nodes may intervene. These nodes are prepositions (tag AuxP), conjunctions (tag AuxC) and coordinating or apposing elements (respectively, tags Coord and Apos).

For example, see the following sentences:

- [1] "primo determinat formam baptismi;"[7] ("at first it determines the form of the baptism;")
- [2] "ly aliquid autem, et ly unum non determinant aliquam formam vel naturam;"[8] ("the 'something' and the 'one' do not determine any form or nature")

Figure 1 reports the tree of sentence [1], where the Obj relation between the verbal head *determinat* and the dependent *formam* is direct.

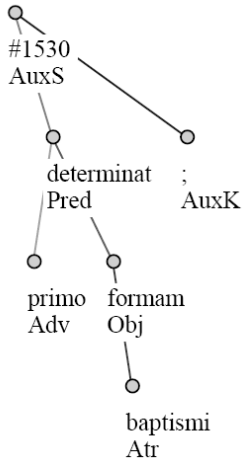

Figure 1.
Tree of sentence [1]

Figure 2 shows the tree of sentence [2]. In this tree, two coordinated subjects (*aliquid* and *unum*) and two coordinated objects (*formam* and *naturam*) depend on the common verbal head *determinant* through two different Coord nodes (*et* and *vel*)[9].

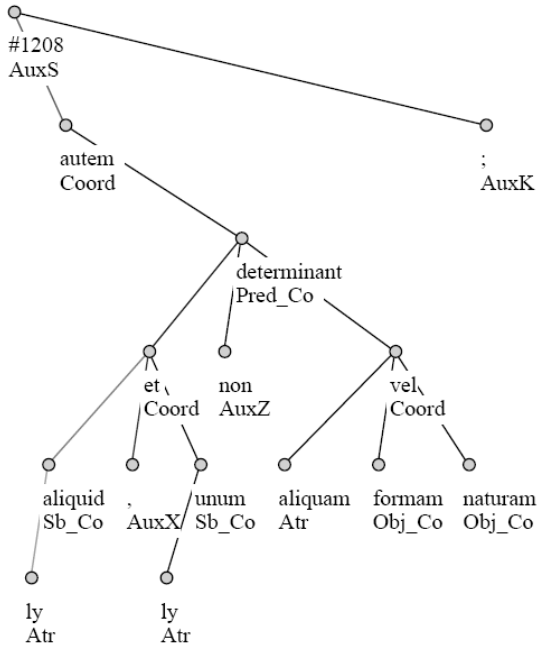

Figure 2
Tree of sentence [2]

In the case of indirect relation, the intermediate nodes need to be detected and extracted, in order to be inserted into the lexicon as subcategorization structures containing the syntactic roles of the verbal arguments. To represent these structures, we distinguished two major types of them: subcategorization frames (SCFs) and subcategorization classes (SCCs).

An SCF contains the sequence of functional labels of verbal arguments as they appear in the sentence order, whereas an SCC reports the subcategorization elements disregarding their linear order in the sentence. SCFs and SCCs play a different role in our lexicon. On the one hand, SCFs are very detailed patterns useful for diachronic and/or comparative studies on linear order. On the other hand, SCCs are more general and make the data in the lexicon comparable with the subcategorization structures as usually defined in the literature and in other valency lexicons. For each of these structures we then created the following sub-types, ranging from the most specific to the least specific one.

SCF_1: subcategorization frame marking the full path between the verbal head (referred to as 'V') and each of its argument nodes in the tree. SCF_1 also assigns the same index to those argument nodes linked by coordinating or apposing elements. For instance, the SCF_1 of the verbal

[7] Thomas, *Super Sententiis Petri Lombardi*, IV, Distinctio 3, Quaestio 1, Prologus, 41-6, 42-2. The edition of the text recorded in the IT is Thomas (1856-1858).

[8] Thomas, *Super Sententiis Petri Lombardi*, III, Distinctio 6, Quaestio 2, Articulus 1, Responsio ad Argumentum 7, 4-5, 6-1.

[9] Following PDT-style, the distributed determination *aliquam*, which modifies both the coordinated objects *formam* and *naturam*, depends on the coordinating node *vel*. For more details, see Hajic et al. (1999), 236-238.

head *determino*[10] in sentence [1] is 'V + Obj', while in sentence [2] is '(Coord)Sb_Co$^{(1)}$ + (Coord)Sb_Co$^{(1)}$ + V + (Coord)Obj_Co$^{(2)}$ + (Coord)Obj_Co$^{(2)}$'. In the latter, the intermediate nodes Coord are in square brackets and indices *1* and *2* link the coordinated nodes. These indices have been adopted in order to disambiguate subcategorization structures where more Obj_Co tags can refer to different verbal arguments. For instance, in a sentence like "I give X and Y to W and Z", both the tranferred objects (X and Y) and the receivers (W and Z) are annotated with Obj_Co. Using indices, the subcategorization structure of the verb *give* in this sentence appears as follows: 'Sb + V + (Coord)Obj_Co$^{(1)}$ + (Coord)Obj_Co$^{(1)}$ + (Coord)Obj_Co$^{(2)}$ + (Coord)Obj_Co$^{(2)}$'. The indices cannot be applied *a priori* to subsequent arguments, since Latin, allowing discontinuous constituents, can show cases where coindexed nodes are separated by other lexical items in the linear order.

SCC_1: the subcategorization class associated with SCF_1. The SCC_1 of the verb *determino* in [1] is '{Obj}', while in [2] is '{(Coord)Sb_Co$^{(1)}$, (Coord)Sb_Co$^{(1)}$, (Coord)Obj_Co$^{(2)}$, (Coord)Obj_Co$^{(2)}$}'.

SCF_2: a subcategorization frame containing only the labels and the indices of the arguments, but not the full path. So, the SCF_2 of *determino* in [1] is 'V + Obj', while in [2] is 'Sb_Co$^{(1)}$ + Sb_Co$^{(1)}$ + V + Obj_Co$^{(2)}$ + Obj_Co$^{(2)}$'.

SCC_2: the subcategorization class associated with SCF_2. For *determino*, this is '{Obj}' in [1] and '{Sb_Co$^{(1)}$, Sb_Co$^{(1)}$, Obj_Co$^{(2)}$, Obj_Co$^{(2)}$}' in [2].

SCC_3: a subcategorization frame containing only the argument labels. The SCC_3 of *determino* is '{Obj}' in [1] and '{Sb, Obj}' in [2], showing that in this sentence *determino* is used as a biargumental verb, regardless of the number of lexical fillers realised for each of its arguments at the surface level.

6 Conclusion and future work

Presently, the size of the IT-TB valency lexicon is 432 entries (i.e. verbal lemmata, corresponding to 5966 wordforms), with 270 different SCF_1s. In the near future, the lexicon will be enriched with valency information for nouns and adjectives.

The corpus-based approach we followed induces verbal arguments directly from annotated data, where the arguments may be present or not, depending on the features of the texts. Therefore, the lexicon reflects the empirical evidence given by the data it was extracted from, encouraging linguistic studies on the particular language domain of our corpus.

In addition to the syntactic information reported in the different types of SCFs and SCCs, it is possible at each stage to include both the morphological features and the lexical fillers of verbal arguments, helping define verbal selectional preferences.

The lexicon may also be useful for improving the performance of statistical parsers, enriching the information acquired by parsers on verbal entries. On the other hand, moving from parser performance to lexicon development, the lexicon can be induced from automatically parsed texts when an accurate parsing system is available.

The syntactic and lexical data recorded in the lexicon are also important in further semantic NLP applications, such as word sense disambiguation, anaphora and ellipsis resolution, and selectional preference acquisition. Following a widespread approach in valency lexicons, a close connection between valency frames and word senses will be followed in the description of lexicon entries: this means that each headword entry of our lexicon will consist of one or more SCFs and SCCs, one for each sense of the word.

We plan to make the lexicon available online through a graphical interface usable also during the annotation procedures, as has been already done for the PDT via the tree editor TrEd.[11] In this way, the consistency of the annotation process can be tested and enforced thanks to the information stored in the lexicon.

In order to test the accuracy of our system, it will be also necessary to evaluate the quality of our valency lexicon against the Perseus "dynamic lexicon", Happ's list and other existing resources for Latin, such as traditional dictionaries and thesauri. A comparison with the lexicon by Perseus is also very interesting in a contrastive diachronic perspective, as it may show important linguistic differences between Classical and Medieval Latin.

Acknowledgments

We would like to thank Paolo Ruffolo for his help in designing the database architecture.

References

[10] *Determino* is the lemma of both the wordforms *determinat* (sentence [1]) and *determinant* (sentence [2]).

[11] TrEd is freely available at http://ufal.mff.cuni.cz/~pajas/tred/.

David Bamman. 2006. The Design and Use of Latin Dependency Treebank. In Jan Hajič and Joakim Nivre (eds.), *TLT 2006. Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories. December 1-2, 2006, Prague, Czech Republic*, Institute of Formal and Applied Linguistics, Prague, Czech Republic, 67-78.

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

David Bamman, Marco Passarotti, Gregory Crane and Savina Raynaud. 2007a. *Guidelines for the Syntactic Annotation of Latin Treebanks*, «Tufts University Digital Library». Available at: http://dl.tufts.edu/view_pdf.jsp?urn=tufts:facpubs:dbamma01-2007.00002.

David Bamman, Marco Passarotti, Gregory Crane and Savina Raynaud. 2007b. A Collaborative Model of Treebank Development. In Koenraad De Smedt, Jan Hajič and Sandra Kübler (eds.), *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories. December 7-8, 2007, Bergen, Norway*, Northern European Association for Language Technology (NEALT) Proceedings Series, Vol. 1, 1-6.

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. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). May 28-30, 2008, Marrakech, Morocco*, European Language Resources Association (ELRA), 2008.

Karl Bühler. 1934. *Sprachtheorie: die Darstellungsfunktion der Sprache*, Jena: Gustav Fischer, Stuttgart.

Roberto Busa. 1974–1980. *Index Thomisticus: sancti Thomae Aquinatis operum omnium indices et concordantiae, in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur quaeque / consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa SJ*, Frommann-Holzboog, Stuttgart-Bad Cannstatt.

Gregory R. Crane, Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeff A. Rydberg-Cox, David A. Smith and Clifford E. Wulfman. 2001. Drudgery and deep thought: Designing a digital library for the humanities. In *Communications of the ACM*, 44(5), 34-40.

John Carroll, Ted Briscoe and Antonio Sanfilippo. 1998. Parser Evaluation: a Survey and a New Proposal. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998). May 28-30, 1998, Granada, Spain*, 447-454.

Joseph Denooz. 1996. *La banque de données du laboratoire d'analyse statistique des langues anciennes (LASLA)*. « Le Médiéviste et l'ordinateur », 33, 14-20.

Jan Hajič, Jarmila Panevová, Eva Buráňová, Zdeňka Urešová and Alla Bémová. 1999. *Annotations at Analytical Level. Instructions for annotators*, Institute of Formal and Applied Linguistics, Prague, Czech Republic. Available at: http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/a-layer/pdf/a-man-en.pdf.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alla Bémová, Veronika Kolárová-Reznícková and Petr Pajas. 2003. PDT-VALLEX: Creating a Large Coverage Valency Lexicon for Treebank Annotation. In Joakim Nivre and Erhard 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 University Press, Växjö, Sweden, 57-68.

Heinz Happ. 1976. *Grundfragen einer Dependenz-Grammatik des Lateinischen*, Vandenhoeck & Ruprecht, Goettingen.

Dag Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Language Technology for Cultural Heritage Data Workshop (LaTeCH 2008), Marrakech, Morocco, 1st June 2008*, 27-34.

Peter Hellwig. 1986. Dependency Unification Grammar, In *Proceedings of the 11th International Conference on Computational Linguistics*, Universität Bonn, Bonn, 195-198.

Richard Hudson. 1990. *English Word Grammar*, Blackwell Publishers Ltd, Oxford, UK.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to Propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas – Gran Canaria, Spain.

Anna Korhonen, Yuval Krymolowski and Ted Briscoe. 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, Italy.

Matthias T. Kromann. 2003. The Danish Dependency Treebank and the underlying linguistic theory. In Joakim Nivre and Erhard 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 University Press, Växjö, Sweden.

Leonardo Lesmo, Vincenzo Lombardo and Cristina Bosco. 2002. Treebank Development: the TUT Approach. In Rajeev Sangal and Sushma M. Bendre (eds.), *Recent Advances in Natural Language Processing. Proceedings of International Conference on Natural Language*

Processing (ICON 2002), Vikas Publ. House, New Delhi, 61-70.

Beth Levin. 1993. *English verb classes and alternations: a preliminary investigation*, University of Chicago Press, Chicago.

Dekang Lin. 1995. A dependency-based method for evaluating broadcoverage parsers. In *Proceedings of the IJCAI-95*, Montreal, Canada, 1420-1425.

Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*, State University Press of New York, Albany/NY.

Cedric Messiant, Anna Korhonen and Thierry Poibeau. 2008. LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008). May 28-30, 2008, Marrakech, Morocco*, European Language Resources Association (ELRA), 2008.

Jarmila Panevová. 1974-1975. *On Verbal Frames in Functional Generative Description*. Part I, «Prague Bulletin of Mathematical Linguistics», 22, 3-40; Part II, «Prague Bulletin of Mathematical Linguistics», 23, 17-52.

Marco Passarotti. 2007. Verso il Lessico Tomistico Biculturale. La treebank dell'*Index Thomisticus*. In Raffaella Petrilli and Diego Femia (eds.), *Il filo del discorso. Intrecci testuali, articolazioni linguistiche, composizioni logiche. Atti del XIII Congresso Nazionale della Società di Filosofia del Linguaggio, Viterbo, 14-16 Settembre 2006*, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 04, Roma, 187-205.

Marco Passarotti. Forthcoming. Theory and Practice of Corpus Annotation in the *Index Thomisticus* Treebank. In *Proceedings of the Conference 'Trends in Computational and Formal Philology - Venice Padua, May 22-24, 2008'*.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson and Jan Scheffczyk. 2006. *FrameNet II. Extendend Theory and Practice*. E-book available at http://framenet.icsi.berkeley.edu/index.php?option=com_wrapper&Itemid=126.

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

Lucien Tesnière. 1959. *Éléments de syntaxe structurale*, Editions Klincksieck, Paris, France.

Thomas Aquinas. 1856-1858. *Sancti Thomae Aquinatis, doctoris angelici, Ordinis praedicatorum Commentum in quatuor libros Sententiarum magistri Petri Lombardi, adjectis brevibus adnotationibus*, Fiaccadori, Parma.

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

Leonoor Van der Beek, Gosse Bouma, Rob Malouf and Gertjan van Noord. 2002. The Alpino Dependency Treebank. In Mariet Theune, Anton Nijholt and Hendri Hondorp (eds.), *Proceedings of the Twelfth Meeting of Computational Linguistics in the Netherlands (CLIN 2001)*, Rodopi, Amsterdam, 8-22.