A Collaborative Model of Treebank Development

David Bamman[1], Marco Passarotti[2], Gregory Crane[1], and Savina Raynaud[2]

[1]Tufts University
The Perseus Project

[2]Catholic University of the Sacred Heart - Milan
Department of Philosophy

Abstract

We describe here a collaboration between two separate treebank projects annotating data for the same language (Latin). By working together to create a common standard for the annotation of Latin syntax and sharing our annotated data as it is created, we are each able to rely on the resources and expertise of the other while also ensuring that our data will be compatible in the future.

1 Introduction

Latin has been used as a productive language for over two thousand years. The duration of this lifetime has created enough distinguishable areas of scholarship that a single project is unlikely to build a treebank containing both Vergil's *Aeneid* (written in the first century BCE) and Johannes Kepler's *Astronomia nova* (published in 1609). One reason for this is the unique role that treebanks play within the humanities: while NLP-oriented researchers may build a treebank from newswire for such tasks as training automatic parsers and inducing grammars, traditional humanists are interested in the texts themselves, and will build a treebank consisting entirely of the Bible (for instance) in order to study the specific use of syntax within. We must expect and encourage different research groups to create individual treebanks containing texts from these different eras.

The development of more than one treebank for any given language, however, has the potential to lead to balkanization, with each individual project working independently and pursuing its own research agenda. This diversity is of course necessary for scientific progress, but it can also lead to a proliferation of annotation styles and datasets that are ultimately incompatible. The adoption of common structural standards such as XCES

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

(Ide, Bonhomme, and Romary, 2000) and infrastructure (CLARIN, 2007) mitigates this to a certain extent, but true dataset compatibility also extends to the level of the individual syntactic decisions themselves. While such compatibility is not always possible, the benefits of working together are significant. We here present a case study of such a collaboration.

2 The Treebanks

Our two groups are each independently creating a treebank for Latin – the Latin Dependency Treebank (LDT) (Bamman and Crane, 2006; Bamman and Crane, 2007) on works from the Classical era, and the *Index Thomisticus* (IT-TB) (Busa, 1974–1980; Passarotti, 2007) on the works of Thomas Aquinas. The composition of both treebanks is given in Tables 1 and 2.

Date	Author	Words	Sentences
1st c. BCE	Cicero	2,119	127
1st c. BCE	Caesar	1,486	71
1st c. BCE	Sallust	12,891	717
1st c. BCE	Vergil	2,613	179
4th-5th c. CE	Jerome	8,382	405
	Total	27,491	1,499

Table 1: LDT composition.

Date	Author	Words	Sentences
13th c. CE	Aquinas	17,966	818
	Total	17,966	818

Table 2: IT-TB composition.

These projects are the first of their kind for Latin, so we do not have prior established guidelines to rely on for syntactic annotation. Since we are both working within the theoretical framework of Dependency Grammar, we have each independently based our annotations on that used by the Prague Dependency Treebank (PDT) (Hajič et al., 1999) while tailoring it for Latin via the grammar of Pinkster (Pinkster, 1990). Adopting an annotation style wholesale, however, is easier said than done. Since nearly all Latin available to us is highly stylized, we are constantly confronted with idiosyncratic constructions that could be syntactically annotated in several different ways. These constructions (such as the ablative absolute or the passive periphrastic) are common to Latin of all eras. Rather than have each project decide upon and record each decision for annotating them, we decided to pool our resources and create a single annotation manual (Bamman et al., 2007) that would govern both treebanks.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

3 Annotation Standards

The creation of this common standard has been vital for the evolution of both of our projects. First and most importantly, it ensures that the treebanks we each create will be annotated in the same way. Both of our individual annotation styles have undergone significant revisions in order to converge on a common ground. Early in our collaboration this involved large-scale reassessments – dropping syntactic functions (the LDT, for instance, once had dedicated tags for indirect objects, ablative absolutes, and complements) or changing the representation of entire constructions (e.g., object complements or accusative + infinitives in the IT-TB). Its effects, however, extend well beyond compatibility. Since we are working with dialects of Latin separated by fourteen centuries, this collaboration has allowed us to base our syntactic decisions on a variety of examples from a wider range of texts. Our individual workflows are each independent of the other, but as both projects annotate more data, we each come across sentences that push the limits of our existing annotation standards: here our collaboration begins. After one group identifies a syntactic construction in its data for which the current annotation standards are insufficient, we both search our respective corpora for similar constructions and then come to a common solution by consulting with each other and with outside advisors (typically via email). Once we come to an agreement on annotation, we include it as part of the guidelines.

The diversity in our projects allows different annotation problems to surface with our individual texts. Two examples can illustrate this.

Ex. 1: Diverse syntactic constructions. Reflexive passives (in which an action is expressed without specifying the agent responsible for it) are much more common in later Latin (Medieval and beyond) than in Classical Latin, but are still present in all eras. In the course of annotating, the IT-TB uncovered eight examples of the reflexive passive in its data, while there were no examples in the LDT. By using the data from the IT-TB, we were able to revise our guidelines in order to codify the annotation and can now refer to that decision whenever we encounter it in our Classical texts.

Ex. 2: Diverse annotator errors. Since our individual annotators are working with different texts, they make different varieties of errors. By expanding our common guidelines to include more detailed descriptions of how to avoid such errors in the future, both groups benefit. For example: early in our development, the annotators for the LDT would frequently vary in their annotation of indirect questions. By focusing especially on this

problem and including it in the guidelines' appendix,[1] we are able to refer annotators from both projects to its solution.

Figure 1 presents two sentences annotated under these guidelines, one from each project.

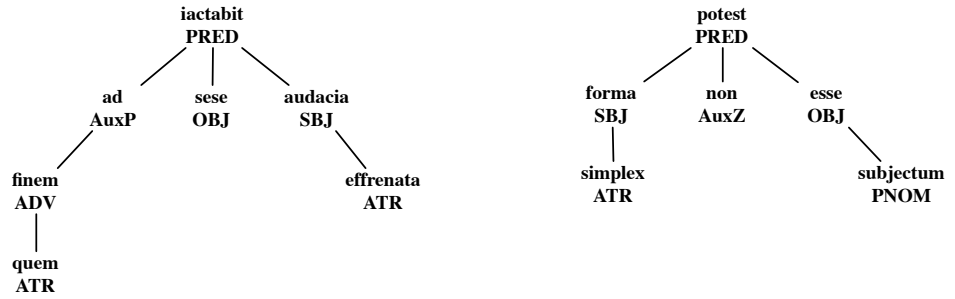

Figure 1: Left: Dependency tree of *quem ad finem sese effrenata iactabit audacia* ("to what end does your unbridled audacity throw itself?"), Cicero, *In Catilinam* 1.1, from the LDT. Right: Dependency tree of *simplex forma subjectum esse non potest* ("the simple form cannot be the subject"), Thomas Aquinas, *Scriptum super Sententiis*, Liber I, Quaestio 1, Articulus 4, Argumentum 1, from the IT-TB.

4 Differences

While we both adhere to these common standards in all other respects, we do differ in the annotation of a single construction: ellipsis. Since its inception, the LDT has annotated ellipsis in a manner that attempts to preserve the structure of the underlying sentence with a complex syntactic tag, while the IT-TB has followed the PDT convention of attaching an orphan to its head with the relation ExD. This difference can be seen in the differing annotations provided in figure 2.

While the edge labels we assign to these orphans are different, the structure of the tree is not, and our data is still compatible since the formalism used by the LDT can always be reduced to that used by the IT-TB.

5 Data

The data that each of our projects produces plays an important role in our future development, since it can supply the training data we need for

[1]The final section of the annotation guidelines ("How To Annotate Specific Constructions") specifically addresses syntactic problems as they are known in traditional Latin grammars – e.g., "relative clauses," "indirect questions," "the ablative absolute," "accusative + infinitive constructions," etc.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476

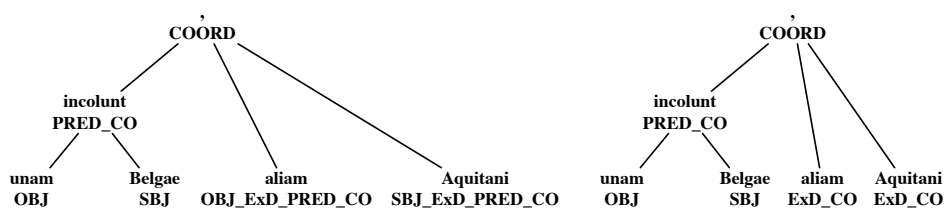

Figure 2: Dependency tree of *unam incolunt Belgae, aliam Aquitani* ("one the Belgae inhabit, another the Aquintani") (Caes. *B.G.* 1.1): on the left is the annotation by the LDT, on the right that by the IT-TB.

automatic syntactic parsing. By at least partially parsing our texts automatically, we can increase the efficiency of our annotators, but statistical dependency parsers such as MaltParser (Nivre et al., 2007) and MSTParser (McDonald et al., 2005) generally perform best with larger amounts of data. By combining our datasets – both annotated under the same general guidelines – we are able to double the size of our training data for such parsers.

6 Future

Collaborating has allowed both of our projects to accomplish more than if we each worked alone. This type of collaboration paves the way for a more distributed method of treebank building. By creating a communal standard for the annotation of Latin syntax and making our data freely available,[2] we hope to encourage other research groups working in different eras of Latin to collaborate with us, and hope to be a positive example for groups working in other languages as well.

7 Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484) and the National Science Foundation (BCS-0616521) provided support for this work.

References

Bamman, David and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78, Prague. ÚFAL MFF UK.

[2]The LDT data can be found online at http://nlp.perseus.tufts.edu/syntax/treebank, and the IT-TB data can be found at http://circse.marginalia.it/~passarotti.

Bamman, David and Gregory Crane. 2007. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 33–40, Prague. Association for Computational Linguistics.

Bamman, David, Marco Passarotti, Gregory Crane, and Savina Raynaud. 2007. Guidelines for the syntactic annotation of Latin treebanks, version 1.3. Technical report, Tufts Digital Library, Medford, http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf.

Busa, Roberto. 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 SI.* Frommann-Holzboog, Stuttgart-Bad Cannstatt.

CLARIN. 2007. http://www.mpi.nl/clarin/.

Hajič, Jan, Jarmila. Panevová, Eva Buráňová, Zdenka Urešová, and Alla Bémová. 1999. Annotations at analytical level: Instructions for annotators (English translation by Z. Kirschner). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary. 2000. XCES: An XML-based encoding standard for linguistic corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, pages 825–830, Athens.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 523–530.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

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

Pinkster, Harm. 1990. *Latin Syntax and Semantics.* Routledge, London.

ISSN 1736-6305 Vol. 1
http://dspace.utlib.ee/dspace
/handle/10062/4476