A Collaborative Model of Treebank Development

David Bamman[1], Marco Passarotti[2], Gregory Crane[1], Savina Raynaud[2]
[1]Latin Dependency Treebank, The Perseus Project, Tufts University (Boston - USA)
[2]*Index Thomisticus* Treebank, Catholic University of the Sacred Heart (Milan – Italy)

Latin Dependency Treebank[1]
Works from Classical era
http://nlp.perseus.tufts.edu/syntax/treebank

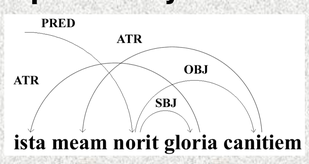
Date	Author	Words	Sentences
1st c. BCE	Cicero	2,119	127
1st c. BCE	Caesar	1,486	71
1st c. BCE	Sallust	12,891	717
1st c. BCE	Vergil	2,647	179
4th-5th c. CE	Jerome	8,382	405
	Total	**27,525**	**1,499**

Latin
- Rich morphology
- Free-word-order
- Non-projective dependencies
- Much homonymy
- Wide diachrony

***Index Thomisticus* Treebank**[3-9]
Thomas Aquinas Opera Omnia
http://gircse.marginalia.it/~passarotti/

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

Dependency Grammar



"that glory would know my old age"
(Prop. I.8.46)

A Single Annotation Manual[2]

For a true dataset compatibility extended to the level of the individual syntactic decisions

Difficulties

- No Latin treebanks available: no prior established guidelines to rely on
- Data from different eras
- No native speakers
- Different annotation feeling due to different schools/backgrounds
- Diverse syntactic constructions
- Lexical differences
- Many idiosyncratic constructions

Resolution of conflicts

- DG theory[6-11]
- Pinkster's Latin Grammar[10]
- PDT[4]
- Explicitness of annotation
- Information retrieval/extraction compatibility
- Consistency with similar constructions
- Real examples from both the treebanks

Advantages

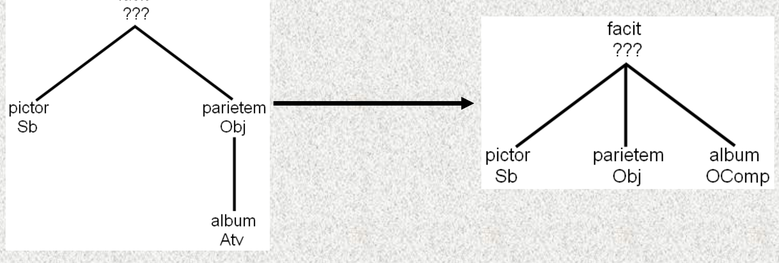
- More informed annotation decisions based on a variety of examples (e.g. constructions more common in one treebank than in the other)
- Linguistically reasoning each decision: treebanks not only as labelled data that can be used for training statistical tools and obtaining precision/recall scores in testing
- Combining datasets to train statistical dependency parsers[5-8]

Annotation samples

Some cases where our collaboration lead to revise our previously different annotation schemes

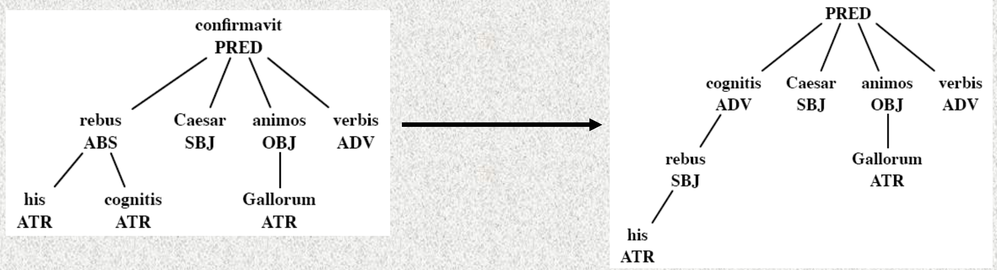
Object complements (OComp)

"pictor facit parietem album"
"the painter makes the wall white"
(Thomas, *Super Sententiis Petri Lombardi*, I, Dist. 17, Qu. 1, Art. 1, Resp. 5, 3-1, 3-4)



Ablative absolute

"his rebus cognitis Caesar Gallorum animos verbis confirmavit"
"these things known, Caesar cheered the Gaul's minds with his words"
(Caes. Gal. 1.33.1)

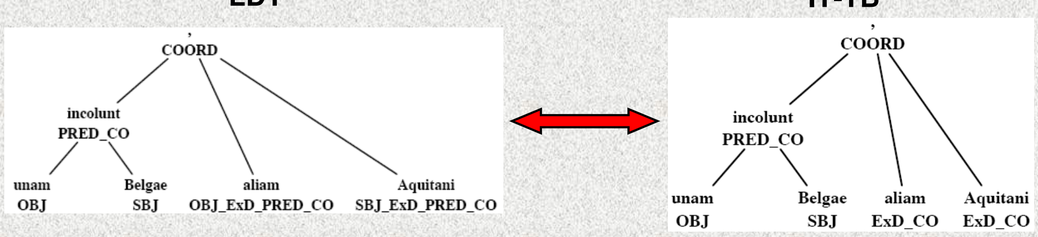


Differences

Ellipsis

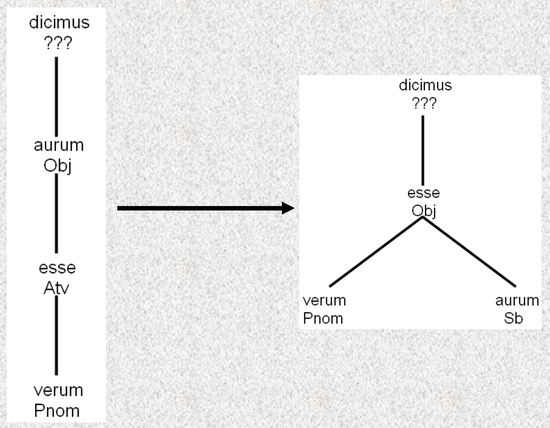
"unam incolunt Belgae, aliam [incolunt] Aquitani"
"one the Belgae inhabit, another the Aquitani"
(Caes. Gal. 1.1)

LDT **IT-TB**



Accusative + Infinitive

"dicimus verum esse aurum"
"we say that gold is genuine"
(Thomas, *Super Sententiis Petri Lombardi*, I, Dist. 8, Qu. 1, Art. 3, Sol., 29-6, 30-4)



"Bridge" structures (AuxP, AuxC)

"cum Germanis contendunt"
"they contend with the Germans"
(Caes. Gal. 1.1)

Deletion of some syntactic functions

G, COMP, IOBJ, REL, SPCH, EXCLAM, J, ABS

Future

- Staff exchanges
- Annotation back-off (comparison)
- Joint workshops
- Constituency checking in related annotation[7]
- Formal matters: official partnership

References

[1] D. Bamman. and G. Crane. The Latin Dependency Treebank in a cultural heritage digital library. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), Prague, Czech Republic, June 2007*. Association for Computational Linguistics, 2007, pp. 33-40.

[2] D. Bamman, M. Passarotti, G. Crane, and S. Raynaud. *Guidelines for the syntactic annotation of Latin treebanks*, version 1.3. Technical report, Tufts Digital Library, Medford, 2007, http://nlp.perseus.tufts.edu/syntax/treebank/1.3/docs/guidelines.pdf.

[3] R. Busa. *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, 1974-1980.

[4] J. Hajic, J. Panevova, E. Buranova, Z. Uresova, and A. Bemova. *Annotations at analytical level: Instructions for annotators* (English translation by Z. Kirschner). Technical report, UFAL MFF UK, Prague, Czech Republic, 1999.

[5] R. McDonald, F. Pereira, K. Ribarov, and J. Hajic. Nonprojective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 523-530.

[6] I. Mel'cuk. *Dependency Syntax: Theory and Practice*, State University of New York Press, New York, 1988.

[7] M. Dickinson and W. D. Meurers. Detecting Errors in Discontinuous Structural Annotation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05).* Ann Arbor, Michigan, 2005.

[8] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kübler, S. Marinov, and E. Marsi. *Maltparser: A language-independent system for data-driven dependency parsing.* «Natural Language Engineering», 13(2): pp. 95-135.

[9] M. Passarotti. Verso il Lessico Tomistico Biculturale. La *treebank* dell'*Index Thomisticus*, in R. Petrilli, D. 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*, Roma, Aracne Editrice, Pubblicazioni della Società di Filosofia del Linguaggio, 04, 2007, pp. 187-205.

[10] H. Pinkster. *Latin Syntax and Semantics.* Routledge, London, 1990.

[11] L. Tesnière. *Élements de syntaxe structurale*, Editions Klincksieck, Paris, 1959.