Corpus Linguistics, Treebanks and the Reinvention of Philology

David Bamman, Gregory Crane

Perseus Project, Department of Classics
Tufts University, Medford MA, 02140USA
david.bamman@tufts.edu
gregory.crane@tufts.edu

Abstract: The fields of corpus and computational linguistics address fundamental goals – and challenge us to rethink the structure – of humanistic research. All work with historical languages is, in some sense, an exercise in corpus linguistics. The Greek and Latin Treebanks illustrate changes in intellectual practice. Linguistic annotation of historical corpora serves a different community and offers a different combination of challenges and opportunities. On the one hand, historical languages such as Greek and Latin have, by definition, no native speakers. At the same time, these corpora have been, and remain, objects of intensive study. The Greek and Latin Treebanks thus have spawned three areas of activity, each of which differs from what we find in corpus linguistics and which collectively constitute a new form of intellectual activity, one that draws upon both the most traditional goals of philology and upon emerging fields such as corpus and computational linguistics.

1 Introduction

Humanists in general and students of the Greco-Roman world in particular have been working with digital materials for a generation but the emerging digital world has, in this first generation, so far exerted relatively little effect upon the goals, practices and general intellectual culture of the humanities. Students of the past have used new tools to ask the same questions and to enhance well-established activities – they have used their large collections as giant concordances and email has accelerated, rather than changed, the flow of electronic publication. The Greek and Latin Treebanks being developed by the Perseus Project at Tufts University has begun to reflect more fundamental changes. Treebanks are collections of text with extensive morphological, syntactic and similar categories of annotation and are familiar instruments for corpus and computational linguistic research. In building Treebanks for historical languages such as Greek and Latin, we found a new intellectual space that combined elements from computational and corpus linguistics and from the ancient discipline of philology. The paper below outlines work on the Treebanks and then describes the implications of this work for Greek, Latin and other historical languages.

2 Syntactic Analysis

The resurgence of statistical methods in computational linguistics over the past twenty-five years has given rise to a great investment in the creation of treebanks – large, syntactically annotated corpora. Much of the work has focused on English (Marcus et al. 1993) and other modern languages, including Czech (Hajic 1998), German (Brants et al. 2002), Spanish (Moreno et al. 2000), French (Abeillé et al. 2000), Italian (Montemagni et al. 2000) and Japanese (Kurohashi and Nagao 1998), but several have arisen recently for historical languages as well, including Middle English (Kroch and Taylor 2000), Early Modern English (Kroch et al. 2004), Old English (Taylor et al. 2003b), Medieval Portuguese (Rocio et al. 2000), Ugaritic (Zemánek 2007), and several Indo-European translations of the New Testament (Haug and Jøhndal 2008).

Funding from the National Science Foundation allowed us to begin developing a treebank for Classical Latin in 2006. The results of this work directly led to private funding for the development of a 400,000-word treebank for Ancient Greek poetry. As of July 2010, we have publicly released over 280,000 syntactically annotated words from these two languages (230,953 words of Ancient Greek and 53,143 words of Latin).[1] Since Latin and Ancient Greek are both highly inflected languages with a high degree of variability in word order, we have based our annotation style on the dependency grammar used by the Prague Dependency Treebank (Hajic 1998) for Czech (another non-projective language), which has since been widely adopted by a number of annotation projects for other languages, including Arabic (Hajic et al. 2004), Slovene (Džeroski et al. 2006) and Modern Greek (Prokopidis et al. 2005). Figure 9 illustrates one such dependency tree in the Ancient Greek Dependency Treebank, taken from the first line of Homer's Iliad.

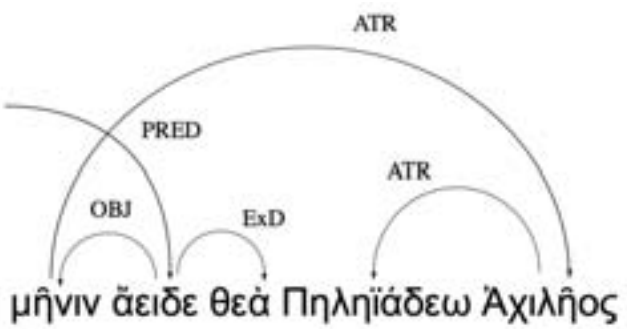

Figure 9: Dependency tree of μῆνιν ἄειδε θεὰ Πηληϊάδεω Ἀχιλῆος ("Sing, goddess, of the rage of Achilles, the son of Peleus"), Homer, *Iliad* 1.1. Arcs are drawn from heads to their dependents.

[1] All syntactically analyzed data is publicly available at:
http://nlp.perseus.tufts.edu/syntax/treebank/

3 Annotation Infrastructure

The efficient annotation of Latin and Ancient Greek is hindered by the fact that no native speakers exist and the texts that we have available are typically highly stylized in nature. To help this with problem, we have embedded our annotation environment within the Perseus Digital Library. Established in 1987 in order to construct a large, heterogeneous collection of textual and visual materials on the archaic and classical Greek world, Perseus today serves as a laboratory for digital library technologies and is also widely used by students, academics and others to access information on the Greco-Roman world (Crane 1987a, Crane 1987b, Crane 1998, Crane et al. 2006, Crane et al. forthcoming).

The scholarship that has attended historical texts since their writing has produced a wealth of contextual materials to help non-native speakers understand them, including commentaries, translations, and specialized lexica. The Perseus reading environment presents the Greek or Latin source text and contextualizes it with these secondary publications along with a morphological analysis of every word in the text and variant manuscript readings as well. Figure 10 presents a screenshot of the digital library with a syntactic annotation tool built into the interface. In the widget on the right, the source text in view (the first chunk of Tacitus' *Annales*) has been automatically segmented into sentences; an annotator can click on any sentence to assign it a syntactic annotation. Here the user has clicked on the first sentence (*Vrbem Romam a principio reges habuere*); this action brings up an annotation screen in which a partial automatic parse is provided, along with the most likely morphological analysis for each word. The annotator can then correct this automatic output and move on to the next segmented sentence, with all of the contextual resources still in view.

Our collaboration with the Alpheios Project has also allowed us to integrate a graphical treebank editor into our annotation process to make the construction of trees more intuitive and to provide annotators with greater flexibility as to their preferred input method. Figure 11 shows a tree in the process of being constructed, with a single word (Romam) being dragged onto its syntactic head.

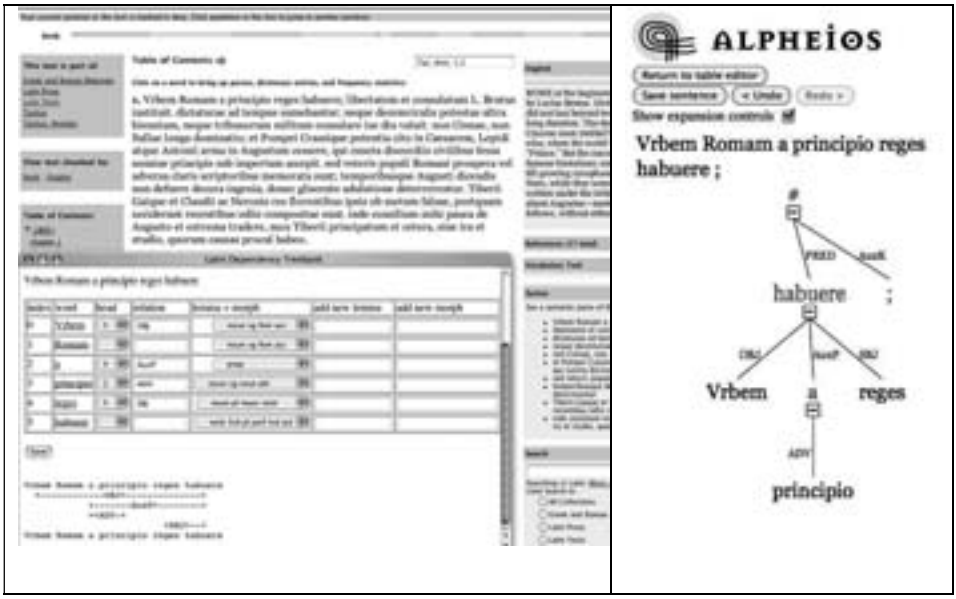

Figure 11: (left) A screenshot of Tacitus' *Annales* from the PDL; (right) Alpheios graphical treebank editor.

In addition to providing morphological analysis and digitized dictionaries such as Lewis and Short's Latin Dictionary or the LSJ, Perseus' translations and commentaries are also especially helpful in the understanding of a text. By situating our annotation environment in the middle of these contextualizing resources, we are providing support for non-native speakers of the language to maximize their contributions to the treebank, and are lowering the barriers of entry for contributing to our work. Such contextual information has greater impact on beginning students than on experts, but is of use to any annotator who wants to consult the published interpretations of authorities in the field.

By embedding our annotation environment within this online infrastructure, we have been able to build a network of annotators who are distributed not only across the United States but across the world as well (our annotators are based not only at Tufts University, the University of California-Berkeley, the University of Pennsylvania, and many other institutions in between, but are also based in Hungary, the United Kingdom and Australia). Ongoing collaboration with several Classics professors has allowed us to introduce treebanking into classrooms at Tufts University, the University of Missouri-Kansas City, Furman University, The College of the Holy Cross, and the University of Nebraska-Lincoln.

4 Methods of Annotation

In developing our work on the Latin and Ancient Greek Dependency Treebanks, we have leveraged three different method of annotation. The "classroom" production method involves soliciting annotations from students in class (e.g., a Greek course on Homer's *Iliad*), which are then reconciled by the professor; the "standard" production method involves soliciting annotations from two independent and heavily trained annotators, whose differences are then reconciled by a third; and the "scholarly" method follows the tradition of creating a critical edition, in which a single scholar with extensive training in the subject area creates a syntactic annotation for a work and is solely responsible for it as an act of interpretation.

4.1 "Classroom" Production Method

We have supported the use of treebanking in classrooms in six universities across the United States – Tufts, Brandeis, the College of the Holy Cross, Furman University, the University of Missouri at Kansas City, and the University of Nebraska at Lincoln. The primary motivation for this work has been pedagogical, since instructors and students both find the act of treebanking useful for learning complex grammatical phenomena. In addition to this fundamental utility, we have also leveraged the resulting annotations as raw material for our published treebanks. Under this method, the students provide multiple primary streams of annotation that the professor, as an expert, is then responsible for correcting and submitting.

In a study to evaluate the potential for this kind of contribution, we evaluated the annotations of a group of thirteen undergraduates at the College of the Holy Cross. Unlike the annotators in the standard production model, who undergo months of training with constant feedback on their performance, this group was provided with only limited training by their professor and access to an online handbook of annotation guidelines. While the overall inter-annotator accuracy averaged only 54.5% due to the different skill levels of students in the class, we importantly found that different students have (naturally) different skill sets – while they all can perform very well on some tasks (such as attributive modification, with an average 91.9% F-measure across the entire class), on other tasks the accuracy varies widely. Figure 12, for example, charts these users' ability to correctly identify participial attachment (i.e., distinguishing an adverbial use of a participle such as "*Reclining* on the bed, I read the book," from an attributive one, such as "the *wandering* king"). Here we see a much wider range of accuracy (reported again as F-measure), from 0% (user 12) all the way to 89.0% (user 3).

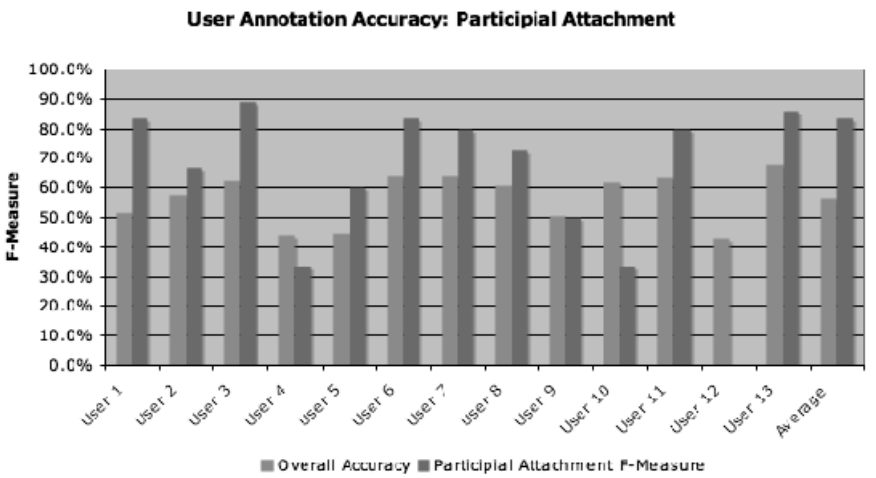

Figure 12: User annotation accuracy for participial attachment.

One pedagogical reward of incorporating treebanking into the classroom is the ability to automatically identify the strengths and weaknesses of individual students – Figure 12, for example, identifies that student 12 clearly needs more assistance in comprehending participial attachment in Greek. We can leverage this work simultaneously for the production of high-quality syntactically annotated data in two ways: first, a professor can either correct the streams of annotation produced by the students in the class, and submit that as the final, finished annotation (in which the entire class and the professor is acknowledged as the owner); second, we use the classroom annotation to help the professor identify the best-performing students, who can then go on to receive more training and provide the primary annotations in the "standard" model). "Standard" Production Method

Under the standard model of treebank production, the annotators who contribute to our existing Greek and Latin treebanks undergo extensive training with constant feedback on their performance. The backgrounds of these annotators range from advanced undergraduate students to recent PhDs and professors, with the majority being students in graduate programs in Classics. In addition to an initial training period, annotators are actively engaged in new learning by means of an online forum2 in which they can ask questions of each other and of project editors; this allows them to be kept current on the most up-to-date codifications to the annotation guidelines while also helping bring new annotators up to speed. Two independent annotators annotate every sentence and the differences are then reconciled by a third. This reconciliation (or "secondary" annotation as it is encoded in the XML release) is undertaken by a more experienced annotator/editor, typically a PhD with specialization in the particular subject area (such as Homer).

[2] The Latin and Ancient Greek forums can both be found here: http://treebank.alpheios.net/forum/

Expert analyses, however, are slow and expensive to create, especially given the difficulty and historical distance of Classical texts. The Penn Treebank can report a productivity rate of between 750 and 1000 words per hour for their annotators after four months of training (Taylor et al. 2003a) and the Penn Chinese treebank can report a rate of 240-480 words per hour (Chiou et al. 2001), but there are no native speakers of historical languages such as Greek. Our annotation speeds are therefore significantly slower, ranging from 92 words per hour to 224, with an average of 130.

4.2 "Scholarly" Production Method

With our treebank of the complete works of Aeschylus, we investigated a new mode of production: that of a single scholar completing a syntactic annotation for an entire work and treating it as an self-standing interpretation of the text.

The motivation for this work is the fundamentally different nature of historical treebanks compared to modern ones. While an article from the Wall Street Journal is certainly more representative of how native English speakers actually speak than Homer's epic Iliad is for ancient Greeks, the Iliad has been a focused object of study for almost 3,000 years, with schoolchildren and tenured professors alike scrutinizing its every word, annotating its syntax, semantics and other linguistic levels either privately in the margins of their books or as published commentaries. While ambiguity is of course present in all language, the individual ad hoc decisions that annotators make in resolving syntactic ambiguity when creating modern treebanks have, for heavily studied Classical and other historical texts, been debated for centuries; dissertations and entire careers have been made on the study of a single work of a single author.

Figures 12, 13 and 14, for example, illustrate the complexity that surrounds textual interpretation of a single of Aeschylus' Agamemnon (τὸν φρονεῖν βροτοὺς ὁδώσαντα, τὸν πάθει μάθος θέντα κυρίως ἔχειν, "[Zeus] ... who put men on the path of wisdom, who established that the law 'learning through suffering' shall be in force," lines 176-8).

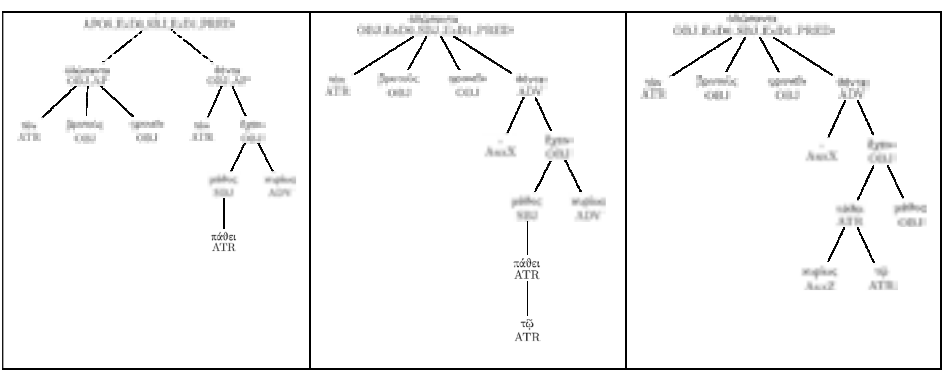

Figure 13: Three different interpretations of a sentence from Aeschylus' *Agamemnon* as machine-actionable syntactic analyses. Syntactic tree of Ag. 176-8 (Denniston-Page, Fraenkel, and Bollack).

Though the formula πάθει μάθος ("learning through suffering") is both quoted and commented upon in many general introductions to the theater of Aeschylus (it was even quoted by Robert F. Kennedy in his speech on the assassination of Martin Luther King Jr. (Kennedy 1968), both the text and syntactic interpretation of the sentence are highly controversial (Bamman et al. 2009). The three most recent commentaries on the play – Fraenkel (Fraenkel 1950), Denniston-Page (Page 1957) and Bollack (Bollack 1981) – have adopted three very different solutions based on their own weighing of the philological evidence, each resulting in a markedly different syntactic tree.

The variety of textual and syntactic interpretations for just these three lines of Aeschylus begins to point out the shortcomings of a standard treebank production model for texts of ongoing scholarly debate. In creating an annotated corpus of a language for which no native speakers exist (and for which we subsequently cannot rely on native intuitions), we are building on a mountain of prior scholarship that has shaped our fundamental understanding of the text.

5 Conclusion

The Greek and Latin Treebanks are not simply databases for research but the catalysts for new intellectual life. Two implications in particular stand out. First, they open for undergraduates in Greek and Latin opportunities similar to those familiar to their counterparts in many of the sciences for making tangible contributions. Second, the Treebanks are not just databases of industrially produced data but repositories of machine-actionable interpretations. A syntactically analyzed sentence is a new form for the publication of scholarly conclusions about language – a form that is itself largely language independent. Whether the researcher's preferred language of publication is English or German, Arabic or Chinese, the parse tree looks the same. The Greek and Latin Treebanks thus have opened new possibilities for students and for advanced researchers to participate more fully in the sudy of Greco-Roman culture than was feasible in print.

Literaturverzeichnis

Abeillé, A., L. Clement, A. Kinyon, and F. Toussenel (2000), "Building a Treebank for French," in: Proceedings of the Second Conference on Language Resources and Evaluation (Athens), pp. 87-94.

Bamman, D., F. Mambrini and G. Crane (2009), "An Ownership Model of Annotation: The Ancient Greek Dependency Treebank," in: Proceedings of the Eighth International Workshop on Treebanks and Linguistics Theories.

Bollack, J. and P. Judet de La Combe (1981), L'Agamemnon d'Eschyle: le texte et ses interprétations. Presses universitaires de Lille, Lille, 1981.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith (2002), "The TIGER Treebank," in: Proceedings of the Workshop on Treebanks and Linguistic Theories (Sozopol, Bulgaria).

Chiou, F., D. Chiang, and M. Palmer (2001), "Facilitating Treebank Annotation Using a Statistical Parser," in: Proceedings of the First International Conference on Human Language Technology Research HLT '01, pp. 1-4.

Crane, G. (1987), "Clay Balls and Compact Disks: Some Political and Economic Problems of New Storage Media," Favonius Supplement, 1, pp. 1-6.

Crane, G. (1987), "From the Old to the New: Integrating Hypertext into Traditional Scholarship," in: Hypertext '87: Proceedings of the 1st ACM conference on Hypertext, pages 51-56.

Crane, G. (1998), "New Technologies for Reading: The Lexicon and the Digital Library," Classical World, pages 471-501.

Crane, G., D. Bamman, L. Cerrato, A. Jones, D. M. Mimno, A. Packel, D. Sculley, and G. Weaver (2006), "Beyond Digital Incunabula: Modeling the Next Generation of Digital Libraries," in: Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), pp. 353-366.

Crane, G., D. Bamman, and A. Jones (forthcoming), "Philology in an Electronic Age," in: Greek Lexicography After Liddell and Scott. Pre-publication version: http://geryon.perseus.tufts.edu/data/Papers and Props/Philology.pdf

Page, D. (1957) Aeschylus Agamemnon. Edited by the late John Dewar Denniston and Denys Page. Clarendon Press, Oxford.

Džeroski, S., T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky and A. Žele (2006), "Towards a Slovene Dependency Treebank," in: Proceedings of the Fifth International Conference on Language Resources and Evaluation, ELRA, Genoa.

Fraenkel, E. (1950), Aeschylus. Agamemnon. Clarendon Press, Oxford, 1950

Hajič, J. (1998), "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank," In : E. Hajičová, editor, Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová. Prague, Charles University Press.

Hajič, J., O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška (2004), "Prague Arabic Dependency Treebank: Development in Data and Tools," In Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools.

Haug, D.T.T., and M.L. Jøhndal (2008), "Creating a Parallel Treebank of the Old Indo-European Bible Translations," In Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008).

Kennedy, Robert F. Statement on the assassination of Martin Luther King, Indianapolis, Indiana, April 4, 1968

Kroch, A., and A. Taylor (2000), Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/

Kroch, A., B. Santorini, and L. Delfs (2004), Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1

Kurohashi, S., and M. Nagao (1998), "Building a Japanese Parsed Corpus while Improving the Parsing System," Proceedings of the First International Conference on Language Resources and Evaluation (Granada).

Marcus, M. P., M. A. Marcinkiewicz and B. Santorini (1993), "Building a Large Annotated Corpus of English: The Penn Treebank," Computational Linguistics 19, pp. 313-330.

Montemagni, S., F. Barsotti, M. Battista, N. Calzolari, O. Corazzari, A. Lenci, A. Zampolli, F. Fanciulli, M. Massetani, R. Raffaelli, R. Basili, M. T. Pazienza, D. Saracino, F. Zanzotto, N. Mana, F. Pianesi, and R. Delmonte (2000), "The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation," in: Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000).

Moreno, A., R. Grishman, S. López, F. Sánchez and S. Sekine (2000), "A Treebank of Spanish and its Application to Parsing," in: Proceedings of the Second Conference on Language Resources and Evaluation.

Prokopidis, P., E. Desipri, M. Koutsombogera, H. Papageorgiou, and S. Piperidis (2005), "Theoretical and Practical Issues in the Construction of a Greek Dependency Treebank," In Proceedings of the 4th Workshop on Treebanks and Linguistic Theories (TLT), pages 149–160.

Rocio, V, M. A. Alves, J. Gabriel Lopes, M. F. Xavier and G. Vicente (2000), "Automated Creation of a Medieval Portuguese Partial Treebank," in Anne Abeillé (ed.), Treebanks: Building and Using Parsed Corpora (Dordrecht: Kluwer Academic Publishers), pages 211-227.

Taylor, A., M. Marcus, and B. Santorini (2003), "The Penn Treebank: An Overview." In: Anne Abeille, editor, Treebanks: Building and Using Parsed Corpora, pages 5-22. Kluwer Academic Publishers.

Taylor, A., A. Warner, S. Pintzuk and F. Beths (2003). York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York.

Zemánek, Petr (2007), "A Treebank of Ugaritic: Annotating Fragmentary Attested Languages," In Proceedings of the Sixth Workshop on Treebanks and Linguistic Theories (TLT2007), pages 213–218, Bergen.