What a parsed corpus is and how to use it

Anthony Kroch and Beatrice Santorini
University of Pennsylvania

In this presentation, we will show what information syntactic annotation adds to electronic text corpora and how such annotation can be used in the analysis of historical material. Using the phrase structure annotation scheme developed for the Penn Parsed Corpora of Historical English (Kroch and Taylor 2000, Kroch et al. 2004, 2010) and the York Corpora (Pintzuk and Plug 2002, Taylor et al. 2003, 2006), also the basis of the annotation in sizable corpora of French (Martineau et al. 2008), Icelandic (Wallenberg et al. 2010), and Portuguese (Galves and Britto 2002) as well as smaller corpora of German, Ancient Greek, and Spanish, we will show what information the annotation provides and what it leaves out. We will discuss how the problem of structural ambiguity can be handled and why the trees assigned to the sentences of a corpus cannot provide a complete structural description. We will also give a brief overview the methods used in the creation of the corpora and how they can be effectively searched for grammatical options and for quantitative analysis.

Once the methodological preliminaries have been addressed, we will give a concrete illustration of the use of parsed corpora by presenting a case study of the verb-second phenomenon. We will give a structural analysis of the verb-second phenomenon in English, discuss the grammatical issues that are at stake and show how quantitative measures support a certain approach to these issues. Extending the work of Speyer (Speyer 2010) and in opposition to standard approaches, we will give evidence that Old English was not a verb-second (V2) language, even though the corpus contains many clauses with the verb in second position. In contrast, the Northern dialect of Middle English did obey the V2 constraint (Kroch and Taylor 1997, Kroch et al. 2000), presumably due to influence from Old Norse. If time permits, we will go on to demonstrate how grammatical and quantitative analysis of the French corpus reveals both differences and similarities between verb-second word order in English and in French. Among the facts that we will address will be the variability between verb-second and verb-third orders that both languages exhibit and the time course of the loss of verb-second order in the two cases.

References

France Martineau et. al. 2009. *Corpus MCVF, "Modéliser le changement: les voies du français"*. University of Ottawa, first edition. F. Martineau CRSH grant project director.

Charlotte Galves and Helena Britto. 2002. *Tycho Brahe Corpus of Historical Portuguese*. Department of Linguistics, University of Campinas, http://www.ime.usp.br/ tycho/corpus/index.html, first edition.

Anthony Kroch, Beatrice Santorini, and Lauren Delfs. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*. CD-ROM, first edition.

Anthony Kroch, Beatrice Santorini, and Ariel Diertani. 2010. *Penn-Helsinki Parsed Corpus of Modern British English*. CD-ROM.

Anthony Kroch and Ann Taylor. 1997. Verb movement in Old and Middle English: Dialect variation and language contact. In Ans van Kemenade and Nigel Vincent, editors, *Parameters of morphosyntactic change*, pp. 297–325. Cambridge University Press, Cambridge.

Anthony Kroch and Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English*. http://www.ling.upenn.edu/hist-corpora/, second edition.

Anthony Kroch, Ann Taylor, and Donald Ringe. 2000. The Middle English verb-second constraint: A case study in language contact and language change. In Susan C. Herring, Pieter van Reenen, and Lene Schoesler, editors, *Textual parameters in older languages*, Current issues in linguistic theory 1950, pp. 353–391. John Benjamins, Amsterdam/Philadelphia.

Susan Pintzuk and Leendert Plug. 2002. *York-Helsinki Parsed Corpus of Old English Poetry*. Oxford Text Archive, first edition.

Augustin Speyer. 2010. *Topicalization and Stress Clash Avoidance in the History of English*, volume 69 of *Topics in English Linguistics*. De Gruyter Mouton. ISBN number 978-3-11-022023-0.

Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. 2006. *Parsed Corpus of Early English Correspondence*. Oxford Text Archive, first edition.

Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. 2003. *York-Toronto-Helsinki Parsed Corpus of Old English Prose*. Oxford Text Archive, first edition.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2010. *Icelandic Parsed Historical Corpus (IcePaHC)*. University of Iceland, version .2 edition.