TreeNet - a computational system for discovering constructions in historical parsed corpora

Gard B. Jenset (independent)

Keywords: parsed corpora, constructions, computational resource, N-grams

Constructions have long been argued to hold a central role in accounts of language (Fillmore, Kay, and O'connor 1988; Goldberg 1995), and construction-based approaches to historical linguistics are well established (Bardal et al. 2015; Hilpert 2013; Bardal et al. 2012). However, identifying constructions automatically in corpora can be methodologically challenging, especially in historical linguistics (Zehentner 2020). Quantitative and corpus-based approaches are undoubtedly more challenging in historical linguistics, which might be a contributing factor to the lower degree of reliance on these approaches in the field (Jenset and McGillivray 2017).

The presentation describes and motivates TreeNet, a computational system for discovering and identifying constructions in historical parsed corpora. Extracting constructions, or construction-like patterns, from such corpora for further analysis can be a non-trivial computational task. TreeNet offers a simple Python package for carrying out this task. The system is inspired by, but different from, the StringNet lexical knowledgebase for Present-day English, which uses parts-of-speech annotated BNC data (Wible and Tsao 2011).

The presentation uses the the Penn-Helsinki Parsed Corpus of Middle English (ME) (Kroch and Taylor 2000) as a case study to explore constructions identified by the system. However, the system is readily extendable to other Penn-Helsinki phrase-structure style parsed corpora.

TreeNet constructs lexically anchored hybrid n-grams, in the form of syntactic annotation tags mixed with lexical items. For example, the ME existential construction below,

(1) ther ben dyuysiones of Grace (from *English Wycliffite Sermons*)
"there are types of grace"

can be generalised as *EX-there [BE] [NP]*.

These hybrid n-grams are pruned to exclude very infrequent patterns. Since they are based on a parsed corpus, TreeNet hybrid n-grams explicitly capture and represent syntactic regularities, as well as lexical items. TreeNet is released as open source code instead of as a static resource. This has a number of advantages, including support for openness and transparency in data processing, as well as being compliant with corpus license terms. The overall aim of the package is to further facilitate historical corpus-based research.

References

Bardal, Jóhanna, Elena Smirnova, Lotte Sommerer, and Spike Gildea. 2015. *Diachronic Construction Grammar*. Vol. 18. Amsterdam: John Benjamins Publishing Company.

Bardal, Jóhanna, Thomas Smitherman, Valgerur Bjarnadóttir, Serena Danesi, Gard B. Jenset, and Barbara McGillivray. 2012. "Reconstructing Constructional Semantics: The Dative Subject Construction in Old Norse-Icelandic, Latin, Ancient Greek, Old Russian and Old Lithuanian." *Studies in Language* 36 (3). John Benjamins Publishing Company: 511–47.

Fillmore, Charles J, Paul Kay, and Mary Catherine O'connor. 1988. "Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone." *Language*, 501–38.

Goldberg, A. E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. The University of Chicago Press.

Hilpert, Martin. 2013. *Constructional Change in English: Developments in Allomorphy, Word Formation, and Syntax*. Cambridge: Cambridge University Press.

Jenset, Gard B., and Barbara McGillivray. 2017. *Quantitative Historical Linguistics: A Corpus Framework*. Oxford: Oxford University Press.

Kroch, Anthony, and Ann Taylor. 2000. "The Penn-Helsinki Parsed Corpus of Middle English (PPCME2)." Department of Linguistics, University of Pennsylvania.

Wible, David, and Nai-Lung Tsao. 2011. "The Stringnet Lexico-Grammatical Knowledgebase and Its Applications." In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, 128–30.

Zehentner, Eva. 2020. "Cognitive Reality of Constructions as a Theoretical and Methodological Challenge in Historical Linguistics." *Belgian Journal of Linguistics* 34 (1). John Benjamins: 371–82.