Abs-19

Silvie Cinková, Martin Holub, and Lenka Smejkalová

Maintaining consistency of monolingual verb entries with inter-annotator agreement

There is no objectively correct way to create a monolingual entry of a polysemous verb. By structuring a verb into readings, we impose our conception onto lexicon users, no matter how big a corpus we use in support. How do we make sure that our structuring is intelligible for others?

We are performing an experiment with the validation of the fully corpus-based Pattern Dictionary of English Verbs (Hanks & Pustejovsky, 2005), created according to the lexical theory Corpus Pattern Analysis (CPA). The lexicon is interlinked with a large corpus, in which several hundred randomly selected concordances of each processed verb are manually annotated with numbers of their corresponding lexicon readings ("patterns"). It would be interesting to prove (or falsify) the leading assumption of CPA that, given the patterns are based on a large corpus, individual introspection has been minimized and most people can agree on this particular semantic structuring. We have encoded the guidelines for assigning concordances to patterns and hired annotators to annotate random samples of verbs contained in the lexicon. Apart from measuring the interannotator agreement, we analyze and adjudicate the disagreements. The outcome is offered to the lexicographer as feedback. The lexicographer revises his entries and the agreement can be measured againg on a different random sample to test whether or not the revision has brought an improvement of the interannotator agreement score. A high interannotator agreement suggests that lexicon users are likely to find a pattern corresponding to a random verb use of which they seek explanation. A low agreement score gives a warning that there are patterns missing or vague.

We focus on machine-learning applications, but we believe that this procedure is of interest even for quality management in human lexicography.

References:

Hanks Patrick Wyndham, Pustejovsky James: A Pattern Dictionary for Natural Language Processing in Revue française de linguistique appliquée 10 (2). 2005.