We would like to demonstrate how a corpus linguistic approach coupled with vector space models (among others Manning et al. (1999)) can be used in the study of taxonomies of concepts, which are usually taken to consist of 3 levels: a level of superconcepts (e.g. VEHICLE), a level of basic concepts (e.g. CAR) and one of subconcepts (e.g. VOLKSWAGEN). A recurring aspect in the study of taxonomies is *informativeness*, which deals with the mutual similarity between its members. It is hypothesized (Murphy (2002)) that this similarity decreases when descending down a taxonomy: 2 randomly chosen vehicles show less similarity than 2 randomly chosen cars, which in turn show less similarity than 2 randomly chosen Volkswagens.

Our goal is to test a corpus linguistic variant of (a part of) this hypothesis: we want to show that the surrounding contexts of use of basic level terms exhibit less similarity than those of their subconcepts. This experiment implies some terminological translation. A collection of members of a concept can be mirrored by the uses of its corresponding term in a corpus. Each use has its particular context which can serve as a representation of the member. Sagi et al. (2009) provide us with a means of constructing vectors from such a collection of context words. Measuring similarity between concept members thereupon becomes a matter of measuring distances between vectors.
To make this concrete we worked with the Leuvens Nieuws Corpus, which consists of 6 Flemish newspapers from the period 1999-2005 and totals roughly 750 million words. We picked out 3 basic concepts from the semantic field of transportation, 'auto' ('car'), 'fiets' ('bike') and 'vliegtuig' ('plane') and about 500 subconcepts. For this collection of terms we could indeed establish a tendency for basic concepts to be less internally coherent than their subconcepts.

Preference: oral

Key words: categorization corpus linguistics lexical semantics

References

Murphy, Gregory L. *The Big Book of Concepts*. Cambridge, Mass.: MIT Press, 2002.

Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass. ; London: MIT Press, 1999.

Sagi, Eyal, Stefan Kaufmann, and Brady Clark. *Semantic Density Analysis: Comparing Word Meaning Across Time and Phonetic Space*. Vol. Proceedings of the Workshop on Geometrical Models of Natural Language Semantics, 2009.