Space vector models and empirical Saussurean semantics

In a well-known passage, Saussure states that: “In the language itself, there are only differences” (Saussure 1986:118 [1916:166]). Semantic analysis on this basis seems to be inevitably limited to restricted domains (such as pronoun systems) in which finite sets of oppositions can be identified. However, recent developments in computational linguistics may assist in increasing the possibilities for semantics which is empirically based and respects Saussure’s insight.

Space vector models (Clark 2015; Turney & Pantel 2010) of word meaning use models of text in which each word is located a in multi-dimensional space; words which are used in similar contexts are close to each other in the spatial model and words which rarely co-occur in the text are far apart. Given a sufficiently large text sample, a model can be constructed which approximates the Saussurean ideal of showing the differences between every lexical element of a language. Implementations of algorithms to produce such models are now easily available (Mikolov et al. 2013).

In this paper, we present initial results of semantic analysis using space vector models. This case study took 22 verbs used to describe events of cutting and breaking (as identified in the study by Majid, Boster & Bowerman 2008; Majid et al. 2007). A 20 dimensional model was built using the entire contents (more than 500 million words) of the Corpus of Contemporary American English (Davies 2008). A vector matrix for the 22 cut/break words to was extracted from the model. The matrix was then the basis for hierarchical clustering analysis, resulting in the dendrogram in Figure 1 which shows seven clusters as the most parsimonious grouping of the data.

We suggest that the dendrogram shows two aspects of the value of these methods in semantic analysis. Firstly, the clustering reflects semantic intuitions in most cases. For example, the first split in the clustering contrasts words which can be viewed as more basic, such as *cut* and *break* themselves, against more specific words, such as *slash* and *hack*, which are hyponyms of the first group. As another example at a lower level of the clustering, the somewhat archaic words *hew* and *cleave* group together. However, there are anomalies in the clustering: for example, *saw* is in the first main group discussed, and *scythe* does not group with *hew* and *cleave*. Secondly, the lowest level of clustering show us the words which are closest to each other in the model, allowing us to ask what conceptual differences are relevant in distinguishing these words. An interesting example of this is the group *slice, peel* and *chop*. Intuition might suggest that *slice* and *chop* would be close to each other with *peel* denoting a rather different type of cutting. But in the model, *peel* and *chop* are closest with *slice* grouping together with them at the next level in the hierarchy.

The anomalies in these results suggest that the next step in applying these methods is to use them in association with collocational analysis. The space vector model is built from co-occurrence of words, therefore a phenomenon such as the relation seen here between *peel* and *chop* may be based on a commonality in what entities the activity is applied to rather than intrinsic properties of the activity.

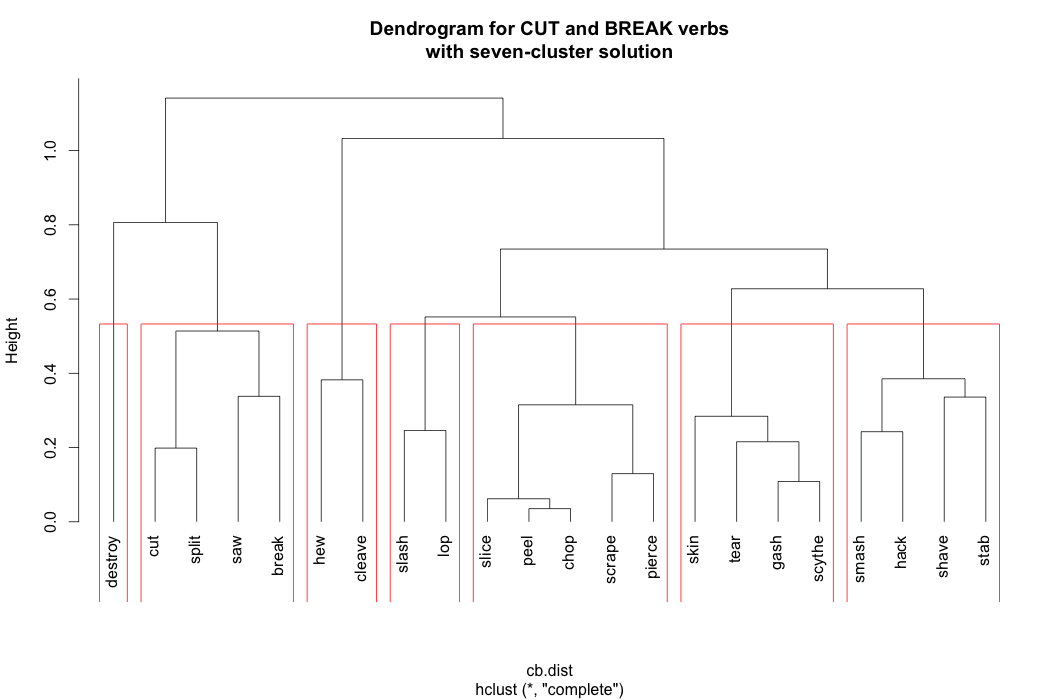


Figure 1 – Hierarchical cluster analysis of 22 verbs of cutting and breaking

References:

Clark, Stephen. 2015. Vector Space Models of Lexical Meaning. In Shalom Lappin & Chris Fox (eds.), *The Handbook of Contemporary semantic theory*, 493–522. Second Edition. Hoboken: John Wiley & Sons.

Davies, Mark. 2008. The Corpus of Contemporary American English: 520 million words, 1990-present. http://corpus.byu.edu/coca/.

Majid, Asifa, James S Boster & Melissa Bowerman. 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition* 109(2). 235–250.

Majid, Asifa, Melissa Bowerman, Miriam van Staden & James S Boster. 2007. The semantic categories of cutting and breaking events: A crosslinguistic perspective. *Cognitive Linguistics* 18(2). 133–152.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saussure, Ferdinand de. 1986. *Course in general linguistics*. (Ed.) Charles Bally, Albert Sechehaye & Albert Riedlinger. (Trans.) Harris. LaSalle, Ill: Open Court.

Turney, Peter D & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research* 37(1). 141–188.