The role of annotation scheme and parser accuracy in learning word representations[§]

Models of lexical semantics are estimated by observing contexts in which words appear. There are roughly two possibilities of constructing the contexts: linear (window of words) and syntactic. While there exists research comparing both, not much is known about how the *syntactic* representations alone are affected by the following two factors:

- choice of dependency annotation scheme
  - e.g. Penn Treebank convention versus Stanford dependencies: differ significantly in the set of labels as well as in the attachment rules
- parser accuracy
  - automatic syntactic analysis involves wrong annotations
  - how much we lose by not having a "perfect" annotation?
  - is the effect more severe when using parsed text for *training* a word model, or when parsing the test data on which to *apply* the word model?

Evaluation
Some manual qualitative analysis. Word models as features in a concrete prediction task: e.g. semantic role labeling.

Models
Any of:

- Distributional-semantic, vector space models
- (Neural-like) word embeddings
- Clustering
- Hidden Markov models

Language English, Dutch, ...

Parser Alpino, Malt, MST, Mate, Turbo, Stanford, ...

References Upon request.

[§]Short MA thesis proposal. Author: Simon Šuster, s.suster@rug.nl. 28. 11. 2014.