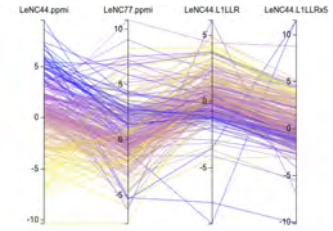
Friday, 26.09.14
10:20 – 11:00

#1



#2

Thomas Wielfaert, Kris Heylen, Dirk Speelman & Dirk Geeraerts
Quantitative Lexicology and Variational Linguistics (QLVL), University of Leuven

Parallel Coordinates as a complementary tool for exploring word similarity matrices

Visualising multivariate linguistic data can be challenging, especially when multiple solutions have to be captured in a single visual representation. These multivariate data can originate from popular statistical techniques for linguistics such as factor analysis or multidimensional scaling and different solutions can be generated for different parameter settings or different subcorpora. An example of the latter is Hilpert (2011), who uses interactive scatter plots, to visualise English verbs in diachronic subcorpora. In these motion charts, meaning changes over time can be visually tracked through moving data points. This approach is feasible for a limited number of solutions and data points, but for larger datasets, patterns in the changes become hard to see. We therefore, propose parallel coordinates as a complementary tool to visualise differences between scatter plots.

Our study explores how distributional semantic models, as developed in computational linguistics, can support theoretically interested linguists in largescale, corpusbased analyses of word meaning. Distributional models are based on the Firth's (1957) idea that 'You shall know a word by the company it keeps' and they use collocate frequencies as features to statistically model word meaning and semantic similarity. Already widely used in computational linguistics as black box models of semantics (see Turney & Pantel 2010), our approach aims to use visualisation techniques to make the models more transparent so that they can be used for an in depth, lexicological analysis of word meaning. In our case study, we selected 10 polysemous nouns from the Dutch electronic dictionary ANW (Algemeen Nederlands Woordenboek) and manually sense disambiguated a random sample of their occurrences (n >= 200) in two large scale Dutch newspaper corpora. The samples were then modelled with tokenlevel distributional models and visualised with our own tool resulting in interactive scatter plots of the nouns' occurrences (see Fig. 1 and Anonymised 2013 for a description of the visualisation tool).

Although the Firthian idea of modelling semantics through context is conceptually straightforward, distributional models are extremely parameterrich with e.g. many ways to select context features, weight collocates or measure similarity. Therefore, models have to be calibrated and multiple solutions have to be compared. Here, we compare models that use different collocational strength measures, Pointwise Mutual Information (PMI) and LogLikelihood Ratio (LLR), and different context window settings. Although the scatter plots produced by our visualisation tool can provide useful insights in the structure of a single model, comparing different parameter settings through different scatter plots quickly becomes impossible as more alternative solutions are added. Therefore, we use parallel coordinates plots as an alternative to compare different models in a single visualisation. Such a plot (see Fig. 2) allows us to get an insight in how the noun occurrences are plotted in a different position over the different solutions: Models with parallel coordinates capture the same semantic structure, whereas those with crossing coordinates organise the occurrences differently. The lexicologist can now also immediately see and further analyse which individual occurrences and which groups of occurrences are treated differently by the models.

References

Firth, J. R., 1957. Papers in linguistics, 1934-1951. Oxford University Press, London.
Hilpert, Martin. 2011. Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. In: International Journal of Corpus Linguistics Vol 16 (4), 435-461.
Turney, Peter D. and Patrick Pantel. 2010. Looking at word meaning. From Frequency to Meaning: Vector Space Models of Semantics. In: Journal of Artificial Intelligence, Vol 37: 141–188.