Visualizing Language Change in a Corpus of Contemporary German

Peter Fankhauser and Marc Kupietz (Institut für Deutsche Sprache, Germany)

We introduce an approach to visualize language change in a large corpus of contemporary German newspapers, spanning the years 2000 to 2015, drawn from DeReKo (Kupietz et al. 2010). The visualization combines two factors involved in language change: Semantics and frequency change.

Semantics of words is visualized by positioning them in two dimensions such that words with similar co-occurrence contexts are positioned closely together. This is accomplished in two steps: First, word embeddings are computed with the structured skip-gram approach described in (Ling et al. 2015), which takes into account word order. To calculate individual word embeddings for each year, we follow the approach of Dubossarsky et al. (2015) and Kim et al. (2014): The embeddings for the first year are initialized based on newspapers from 1950 to 1999 the embeddings for each subsequent year with the embeddings of the previous year. With this approach, the embeddings are comparable across years. Second, the 200 dimensions resulting from the first step are further reduced to two dimensions using t-Distributed Stochastic Neighbor Embedding (Van der Maaten & Hinton 2008).

Frequency change is represented by color, ranging from violet for words with decreasing frequency to red for words with increasing frequency. To this end we calculate the slope of the generalized linear fit of a logistic transform on the relative frequencies in each year (Zuraw 2003) and map it to the color range. The goodness of fit for linear fits is 44.4% and for 2nd order fits 59.2%, thus, the colors characterize frequency change fairly well.

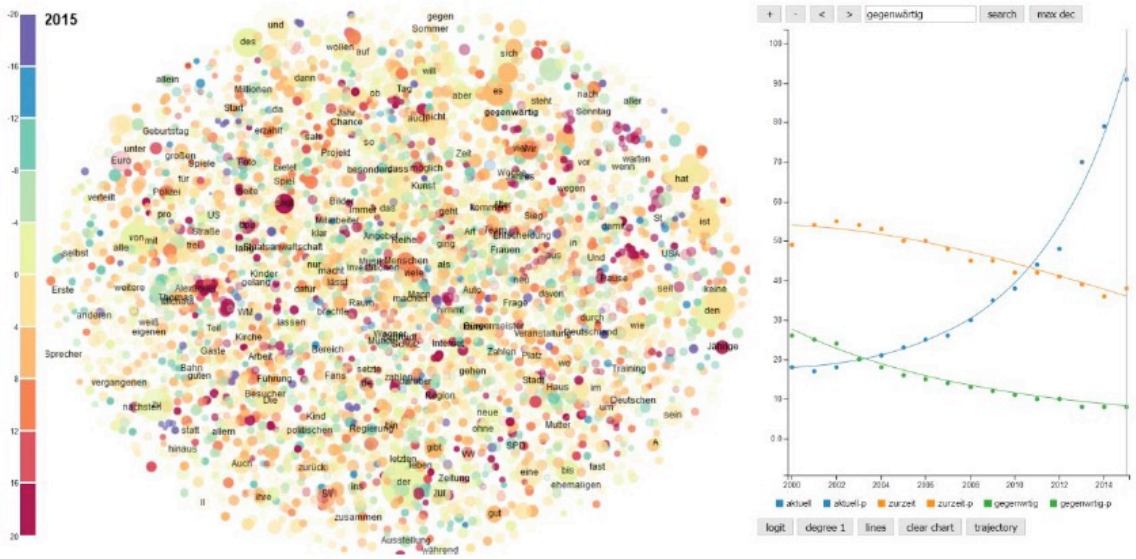

Figure 1. Overall visualization

Figure 1 gives an overview on the visualization. To the left, a bubble chart represents the color encoded semantic space of words, with the size of bubbles

proportional to the square root of the relative frequency in the chosen year (here: 2015). To the right frequency change of individual words is represented by simple line charts showing the fitted 2nd order polynomials of the logit transformed relative frequencies. The line chart also doubles as a selector for individual years.

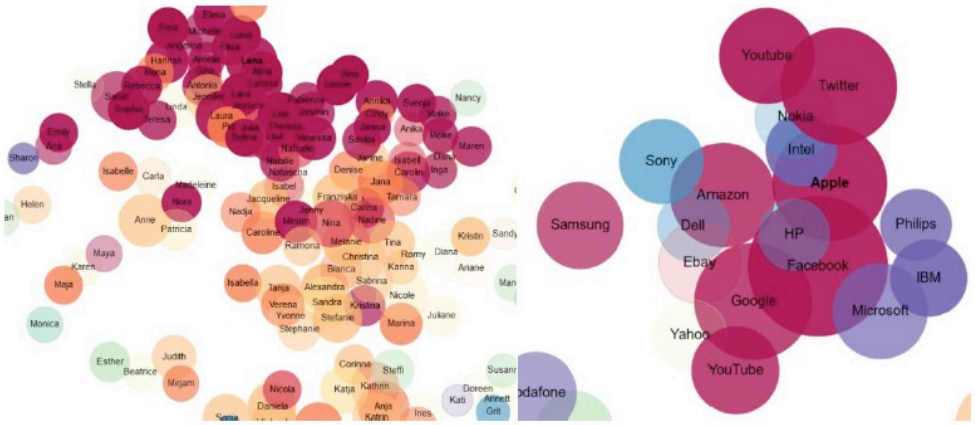

Figure 2. Two regions zoomed in

Figure 2 shows two zoomed in regions, female first names to the left, and brand names of tech companies to the right. In both cases we can see that frequency change is typically correlated with semantic similarity, i.e., words close to each other typically have a similar frequency slope.

Other uses of the visualization include detection of semantic change and replacement of obsolete words. The visualization has been implemented using the D3 framework (Bostock et al. 2011), postprocessing of word embeddings and fitting of the generalized linear models in R. The visualization is publically available at: http://corpora.ids-mannheim.de/diaviz/dereko.html

References

Bostock, M., Ogievetsky, V., Heer, J. (2011). *D3: Data-Driven Documents. IEEE Trans. Visualization & Comp*. Graphics (Proc. InfoVis).

Dubossarsky, H., Tsvetkov, Y., Dyer, C., Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In Pirrelli, Marzi & Ferro (eds.), *Word Structure and Word Usage*. Proceedings of the NetWordS Final Conference.

Kim, Y., Chiu, Y.I., Hanaki, K., Hegde, D., Petrov, S. (2014). *Temporal analysis of language through neural language models*. arXiv preprint arXiv:1405.3515

Kupietz, M., Belica, C., Keibel, H., Witt, A. (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, N. et al. (eds.): *Proceedings of the 7th conference on International Language Resources and Evaluation* (LREC 2010) (pp. 1848–1854). Valletta, Malta: European Language Resources Association (ELRA). (PID: http://hdl.handle.net/10932/00-0373-23CD-C58F-FF01-3)

Ling, W., Dyer, C., Black, A., & Trancoso, I. (2015). Two/too simple adaptations of word2vec for syntax problems. In *Proc. of NAACL*.

Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 1, 1-48.

Zuraw, K. (2003). Probability in Language Change. In Bod, Hay, Jannedy (Eds.) *Probabilistic Linguistics* (pp. 139-176). MIT Press.