Published in: Visualisierungsprozesse in den Humanities. Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu 2017). Tagung vom 17. bis 19. Juli 2017. - Zürich: Universität Zürich, 2017

Visual Correlation for Detecting Patterns in Language Change

Peter Fankhauser, Marc Kupietz

Der Spiegel und Die Zeit are two of the most influential weekly news magazines in Germany. Their digitized corpus currently ranges from 1953 to 2017, and consists of 519 Mio tokens with at least 5 occurrences. In this paper we introduce a visualization of language change for this corpus. To this end we visually correlate two factors: Frequency change and distributional semantics of words.

Frequency change is visualized by means of color ranging from violet for decreasing frequency to red for increasing frequency. The corpus is divided into 13 periods of 5 years each[1]. For each word we calculate the slope of the generalized linear fit of the logistic transform of its relative frequencies in each time slice (Zuraw 2003) and map it to the color range. Thereby words with similar slope are colored similarly.

Semantics of words is visualized by positioning them in two dimensions such that words with similar usage 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). To calculate individual word embeddings for each period, we follow the approach of Dubossarsky et al. (2015) and Kim et al. (2014): The embeddings for the first year are randomly initialized and the embeddings for each subsequent period are initialized with the previous embeddings. With this approach, the embeddings are comparable across periods. 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).

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 period (here: 2014-2017).

[1] The visualization itself is currently restricted to the 30.000 most frequent types.

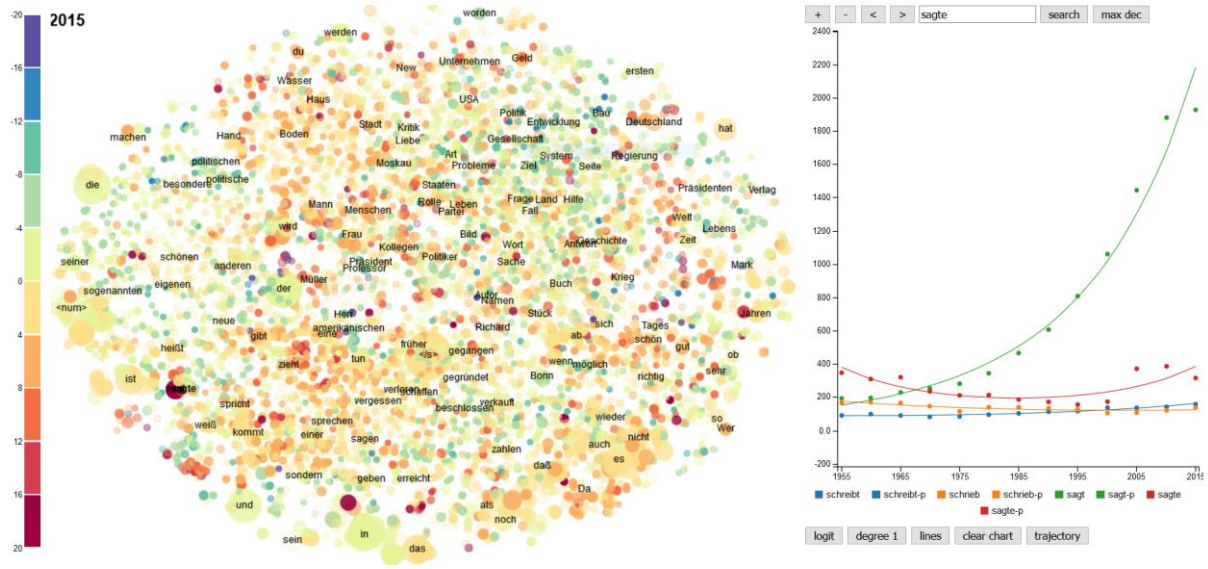

Figure 1. Overall visualization

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 depicted trends are examples of a general decrease of past-tense for communicative verbs like "sagte", "schrieb" (said, wrote) in favour of the temporally unmarked present tense.

Figure 2. Semantic neighbourhood of "Sicherung" 1955 vs. 2015

Figures 2 and 3 show two zoomed in regions for an early period to the left and the most recent period to the right. Both depict frequency change of words with similar usage context. Figure 2 shows a general decrease of nominalization with "-ung" (green/blue/violet) whereas Figure 3 indicates a general increase of superlatives of positively connotated adjectives (orange) – with some notable exceptions (e.g. "modernste"). Both examples also depict higher level patterns in language change: On the one hand, frequency change is typically

correlated with similarity in usage context, i.e., words close to each other have a similar frequency slope, on the other hand a semantic neighbourhood with decreasing frequency typically becomes less productive, i.e., consists of fewer types (Figure 2), and vice versa (Figure 3).

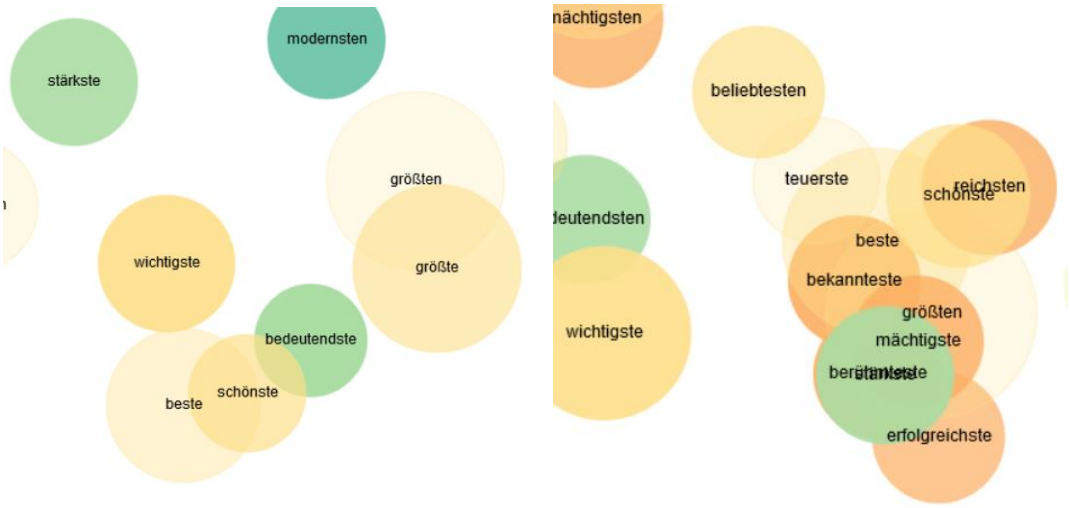

Figure 3. Semantic neighbourhood of "beste" 1960 vs. 2015

Thus, visually correlating two fundamental factors involved in language change - frequency and usage similarity - reveals a new macroanalytic perspective on language change that can be used as starting point for a more detailed analysis.

References

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

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. In *Journal of Machine Learning Research 1*, 1-48.

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