Annotated Timelines and Stacked Area Plots for Visualization in Lexicography

Armin Hoenen, Academy of Sciences, Göttingen, hoenen@em.uni-frankfurt.de. Elexis Workshop, Galway 2018.

1 Introduction

The idiom 'a picture is worth a thousand words' points to a possible advantage of the use of visualizations in many fields, naturally including lexicography. Visualizations may convey complex patterns of information through man's primary sense – vision – very effectively, but at the same time bad visualizations may cause misunderstandings or may not be understandable at all. With the advent of the digital age, visualizations can be produced and proliferated with much more ease than before regarding costs (color, display), rendering (computation) and presentation (dynamicity). Their production has taken on an additional function: that of supporting 'Distant Reading' (Moretti, 2013) for data masses that are so huge, that a manual assessment would be impossible. For all these reasons, the author considers it worthwhile to survey visualisations for lexicography and lexicology, albeit with a look to the precedents and current usages in the field. As the outcome of this endeavour, here, two carefully chosen visualizations for lexicographic data from the project of ZHistLex (Center for historical lexicography) are being presented.[1]

2 Some Brief Comment on the History of Visualization in Lexicography

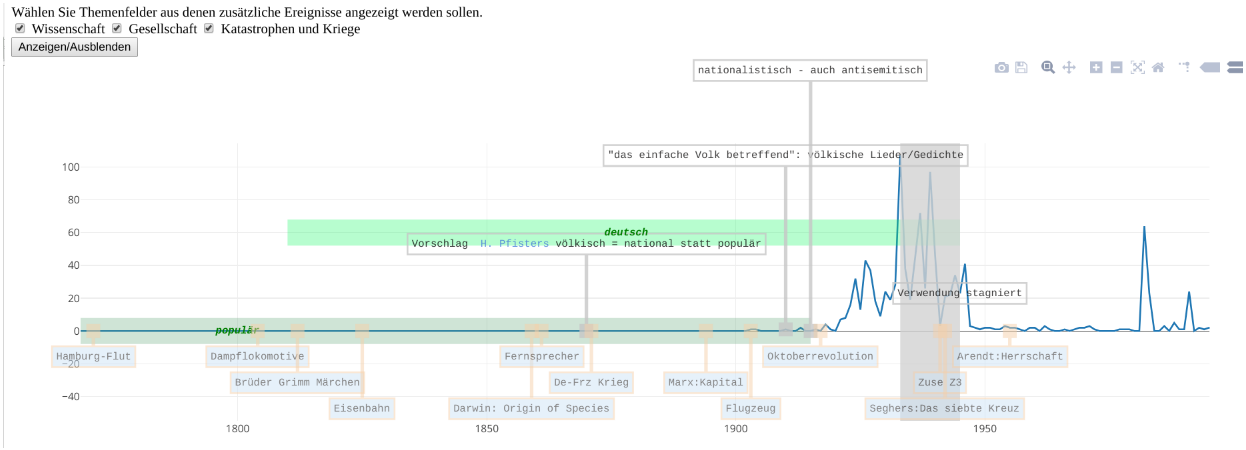
Modern systematical analytic visualization in general started relatively recently with William Playfair (1758-1823) (Rehbein, 2017, S.328). In lexicography and lexicology, in the print age, infographics have been hardly present in articles on lexemes, where information was rather codified into complex ordering schemas – onomasiological or semasiological, abbreviations and other script-centered devices. In the scientific lexicographical literature on the other hand, graphics have been used more, e.g. in Geeraerts (1997). In Hundsnurscher et al. (2002) for instance, some tables have been used which are in the view of the author genuine visualizations, albeit such where 'reading' is the primary entry-point.[2] In the digital era, consequently some publications have focussed on visualizations in lexicology and lexicography (Dixit and Karrfelt, 2016; Hilpert, 2011; Hamilton et al., 2014; Hoenen, 2018) where the generation of the concurrent graphics partly rely on large resources or annotation. Finally, modern online

[1] This paper draws on unpublished work on "völkisch" (Volker Harm) and on "gefälligst" (Thomas Gloning, Ralf Plate) in the ZHistLex project.

[2] The reason why wordclouds are so popular may have to do with this same entry-point or mode of access. Script itself is some kind of visualization.

dictionaries such as the OED[3] or DWDS[4] often offer at least some kind of visualization such as frequency curves or word clouds. Visualizations remain however, something explorable in more detail in the field and are relatively hard to evaluate (Morse et al., 2000; Elmqvist and Yi, 2015).

Visualisations for single lexemes may help novices and experts alike grasp the information on a lexeme's history.[5] Questions such as by which macro-phenomena its history has been influenced (amelioration, pejoration) or if it has undergone a semasiological narrowing or expansion could be supported by visualizations. Such visualizations may also help to preview histories of lexemes with yet unwritten articles. Finally in the writing process it could help lexicographers explore and spot new relations. Consequently, one may ask how to visualize aspects of a lexeme's history, which aspects could be informative and appropriately visualizable and which data would be necessary as input and so forth. Here, time is one of the most important dimensions for lexicographic visualizations when dealing with some form of diachronicity. The following two visualizations are the most promising results of a much larger survey on visualizations for the field.



Wählen Sie Themenfelder aus denen zusätzliche Ereignisse angezeigt werden sollen.
☑ Wissenschaft ☑ Gesellschaft ☑ Katastrophen und Kriege

Figure 1: Timeline for the German term 'völkisch'. Input is a list of pivotal events for a lexeme previously produced by the lexicographer while writing an article. Users can hide/unhide general historical events from politics (wars, revolutions), inventions, literature and philosophy as well as natural desasters extracted and filtered manually from the Wikipedia (below the x-axis). The frequency curve was generated on the basis of data from the DWDS.

2.1 Timeline+

Frequency plots are simplistic and display time on their x-axis, but annotated frequency plots can be much more informative. The first visualization in the current context thus links together frequency curves and timelines with annotated events. The javascript framework Plotly [6] is

[3] http://www.oed.com/

[4] https://www.dwds.de

[5] A visualization should not distract from the text of the article and should be understood as an addition by the users – not a substitute or placeholder.

[6] https://plot.ly/

used to display a frequency plot annotated with certain points in time with text content in boxes in order to link a lexeme's history with historical events (it is planned to include corpus comparisons, synonyms, collocations, word family members), see 1. The result is editable via the Plotly interface (requires login). With such a timeline, events which are decisive for the history of a lexeme may be clarified and the development of the lexeme can be contextualized additionally by general events which drove change in the real world and as a reflection also in the lexicon. For not yet written articles on certain lexemes the contextualization with the general events may be used for exploration. A Timeline+ can be automatically generated for any lexeme through a Java executable requiring as minimal input the data underlying a frequency curve. An optional csv-input file may specify the user annotations.[7]

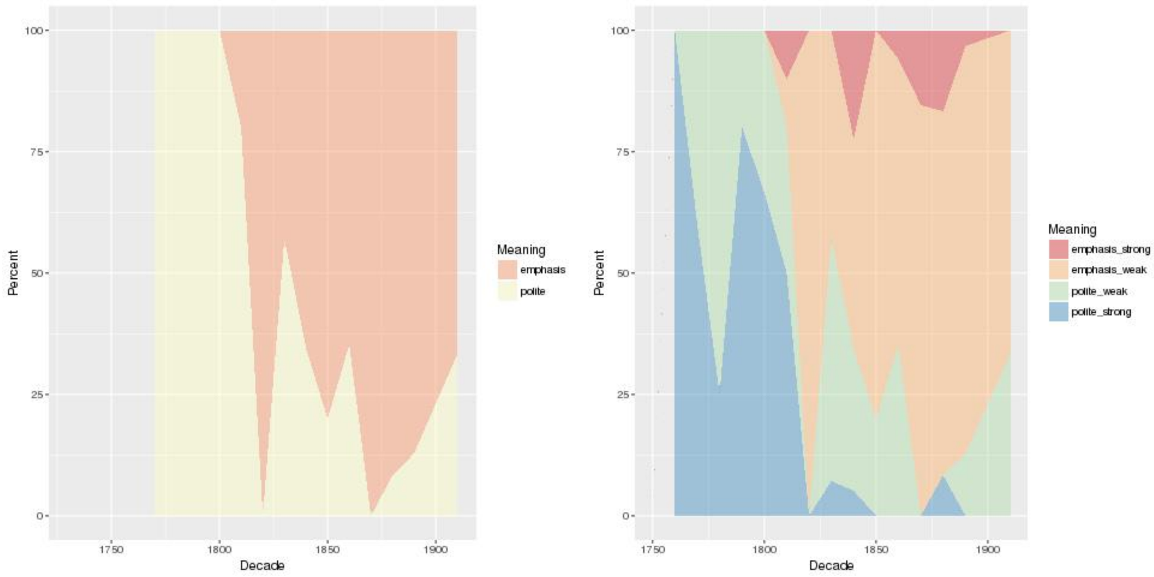


Figure 2: Stacked area plot for the German term 'gefälligst', with an older meaning implying politeness and a more modern meaning expressing emphasis produced with R, package fsmb. Sense of a word occurrence (here from DTA) should be disambiguated (manually or by WSDA). One sees how the new meaning rapidly takes hold.

2.2 Stacked Area Plot

In antiquity, Eusebius invented a tabulated comparative history with one column for Christians, Greeks and Jews (Rosenberg and Grafton, 2013). This became the most used elaborated format for historiography. Building on it, T. Jeffrey started using colors to distinguish empires and McNally (Rosenberg and Grafton, 2013, S.217) normalized and used breadth as relative governed surface area per empire. This was the successful *HistoMap* visualization. For lexicography, a stacked area plot which is similar to that format albeit with swapped axes and instead of empires displaying meanings (senses) of a single lexeme is proposed, see 2. This format requires the previous annotation of each occurrence of a lexeme with a meaning, a task which may be conducted manually or by means of technologies such as word sense disambiguation. The visualized example shows how the sense profile of a lexeme develops

[7] The executable and the code for the plot in the next section can be requested from the author directly.

where a new sense comes in and establishes a firm position within the spectrum of meanings. Larger numbers of such annotations may allow some investigation of types of processes in which new meanings enter spectra.

3 Conclusion

Two possible time-bound visualizations for lexicography and lexicology which both have a history making a good application plausible have been presented, the first as extending a visualization already used in digital lexica (frequency curve), the second by tracing its history to successful ancestors in visualization (HistoMap).

References

Dixit, C. and Karrfelt, F. (2016). Visualizing etymology: A radial graph displaying derivations and origins.

Elmqvist, N. and Yi, J. S. (2015). Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269.

Geeraerts, D. (1997). *Diachronic prototype semantics*. Clarendon Press.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2014). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics.*, pages 1489–1501. ACL.

Hilpert, M. (2011). Dynamic visualizations of language change. *International Journal of Corpus Linguistics*, 16(4):435–461.

Hoenen, A. (2018). Attempts at visualization of etymological information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Hundsnurscher, F., Augst, G., Splett, J., Gruaz, C., and Haßler, G. (2002). *Lexikologie/Lexicology. Ein internationales Handbuch zur Natur und Struktur von Wörtern und Wortschätzen.* Handbücher zur Sprach- und Kommunikationswissenschaft 21.1. de Gruyter.

Moretti, F. (2013). *Distant reading*. Verso Books.

Morse, E., Lewis, M., and Olsen, K. A. (2000). Evaluating visualizations: using a taxonomic guide. *International Journal of Human-Computer Studies*, 53(5):637–662.

Rehbein, M. (2017). Informationsvisualisierung. In *Digital Humanities*, pages 328–342. Springer.

Rosenberg, D. and Grafton, A. (2013). *Cartographies of time: A history of the timeline*. Princeton Architectural Press.