Workshop: *Data visualization in Corpus Linguistics: Critical reflections and future directions*

Workshop convenors

Ole Schützler & Lukas Sönning (University of Bamberg)
ole.schuetzler@uni-bamberg.de
lukas.soenning@uni-bamberg.de

Workshop summary

Data visualization is a vital element of quantitative research (Cleveland 1993, 1994). In the domain of statistics, we can distinguish graphs for data analysis and data presentation (Fienberg 1979). While the former communicate between researcher and data, the latter aim to convey findings to an audience. As graphical representations tap into the human visual system – an enormously powerful pattern-finding device – they can reveal structure in the data in a compelling and accessible way. With corpus-based research typically involving the quantitative analysis of a complex mixture of conditions, there is little room for doubting the relevance of data visualization for our field. This workshop addresses the role of graphical techniques for corpus data analysis and presentation by critically reflecting the state-of-the-art and examining avenues for future practice. To inspect current usage patterns, we conducted a small-scale survey of n = 131 corpus-based articles published in linguistic journals between 2015 and 2017 (see Table 1). Overall, two-thirds of papers made use of graphs for data presentation (89% featured tables). As for the graph types employed, Figure 1 indicates a predominance of classical variants (bar chart: 50% of articles; line plot: 34%). While most display types belong to the common core of statistical graphics, use is also made of less familiar forms. Among these are mosaic charts and various tools from the domain of statistical/machine learning (e.g. conditional inference trees, dendrograms, phylogenetic trees). This snapshot suggests that corpus linguistics strongly relies on well-established graph schemas (Pinker 1990), with novel techniques (as yet) playing a minor role.

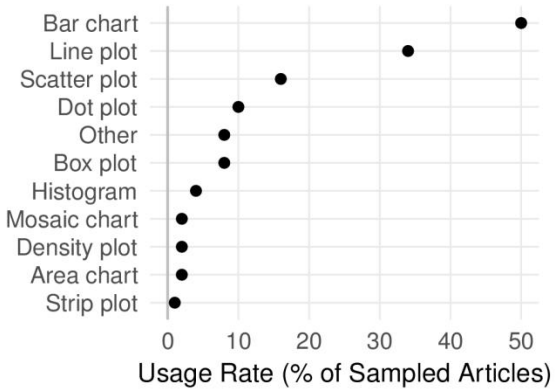

Figure 1. Current usage rate of graph types in leading linguistic journals. The plot shows the percentage of corpus-based articles (n = 131) featuring at least one instance of the respective forms. Note the category 'Other', which includes novel 'non-core' display types.

It is the aim of this workshop to stimulate discussion about best practices in corpus data visualization, including analysis and presentation graphics. We argue that it is genuinely worth considering whether and how issues raised in the literature on statistical graphics (e.g. Tukey 1977, 1993; Cleveland 1993, 1994; Tufte 2001; Kosslyn 2006; Ware 2013; Unwin 2015) equally apply to our field and where discipline-specific adaptations are needed. Concentrating on the goals of corpus linguistics, this workshop welcomes contributions that

- discuss principles of visual perception, reflecting on strengths and weaknesses of graphical forms

- offer comparative evaluations of display types
- explore means of increasing the informativity and processability of displays
- relate to the audience by considering issues of graphicacy/display familiarity, suggesting implications for the methodological training of (corpus) linguists
- discuss the application of graphical means of statistical inference to corpus data
- reflect on principles and tools specific to particular subject-matter/methodological applications
- address the sensitivity of graphs to aspects of data scaling
- offer avenues towards best practices in the visualization of corpus data

Call for papers

Abstracts of approximately 400 words (excluding references) should be sent to both ole.schuetzler@uni-bamberg.de and lukas.soenning@uni-bamberg.de. The deadline for abstract submission is 10 December 2017. Notifications of acceptance will be sent out before Christmas.

References

Cleveland, William S. 1993. *Visualizing data*. Summit: Hobart Press.

Cleveland, William S. 1994. *The elements of graphing data*. Summit: Hobart Press.

Fienberg, Stephen E. 1979. Graphical methods in statistics. *American Statistician* 33(4). 165–178.

Kosslyn, Stephen M. 2006. *Graph design for the eye and mind*. Oxford: Oxford University Press.

Pinker, Steven. 1990. A theory of graph comprehension. In Roy Freedle (ed.), *Artificial intelligence and the future of testing*, 73–126. Hillsdale: Erlbaum.

Tufte, Edward R. 2001. *The visual display of quantitative information*. Cheshire: Graphics Press.

Tukey, John W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tukey, John W. 1993. Graphic comparisons of several linked aspects: Alternatives and suggested principles. *Journal of Computational and Graphical Statistics* 2. 1–33.

Unwin, Antony. 2015. *Graphical data analysis with R*. Boca Raton: CRC Press.

Ware, Colin. 2013. *Information visualization: Perception for design*. Amsterdam: Elsevier.

Table 1. Journals used for the survey and composition of the sample; *n* denotes for each journal the number of corpus-based empirical studies that occurred in the sampled time period (i.e. among the 20 to 30 most recent contributions).

Journal	*n*
Corpus Linguistics and Linguistic Theory	26
International Journal of Corpus Linguistics	20
English Language and Linguistics	16
English World-Wide	15
Language Variation and Change	13
ICAME Journal	13
Cognitive Linguistics	9
Language	6
Applied Linguistics	5
Journal of Memory and Language	4
Natural Language and Linguistic Theory	2
Language Learning	2