Visualization of Corpus Frequencies at Text Level

Stefan Fischer (Universität des Saarlandes, Germany), Peter Fankhauser (Institut für Deutsche Sprache, Germany) and Elke Teich (Universität des Saarlandes, Germany)

We present a tool for close reading of texts that improves the intuitive understanding of corpus-based frequency measures (e.g., probability in context) using colour and font size.

Visualization techniques facilitate the understanding of data and their application in the digital humanities is actively researched. In their survey, Jänicke, Franzini, Cheema, and Scheuermann (2016) give an overview of visualization tools and the features used. Font size is rarely used as a feature. Word clouds, for example, scale words based on frequency, but they do not maintain sentence structure. Prism (Walsh, Maiers, Nally, & Boggs, 2014) uses size for the visualization of annotation agreement, with frequently annotated words shown in a larger font.

Corpus analysis tools like the Corpus Workbench (Evert & Hardie, 2011) and CQPWeb (Hardie, 2012) are useful for querying corpora and finding interesting text instances. Subsequent interpretation, however, is often difficult. In particular, corpus positions that are numerically annotated (see Table 1, third column) can be challenging.

Word	POS	Bits
It	PP	5.55
Grows	VVZ	8.26
In	IN	1.67
dry	JJ	4.95
ground	NN	3.28
.	SENT	3.81

Table 1. Sentence with part-of-speech and surprisal annotation.

In our work, we study the effects of surprisal of words in context on the sentence and text levels. Surprisal of words is defined as the negative logarithm of their probability and is measured in bits (cf. Genzel & Charniak, 2002). To this end we developed an interface for close reading, which uses font size and colour for the visualization of surprisal. The tool is implemented based on D3 (Data-Driven Documents; Bostock, Ogievetsky, & Heer, 2011) and can be used either stand-alone or for the visualization of surprisal in text instances discovered using CQP.

Figure 1 shows three sentences from a historical corpus visualized with our tool. The individual words are scaled based on surprisal. Pointing the mouse at the last occurrence of *Arsenick*, all of its occurrences in the text are highlighted by underlining. In addition, a pop-up window shows further information about the word. From the pop-up, one can see that the current word has a value of 19.24 bits and that *Arsenick* occurs six times in total (N=6). The values of all occurrences of *Arsenick* are in the range between 13.9 and 20.84 bits, the average being 19.15 bits. As a further help, words with low surprisal (e.g., function words) can be shown in a less intensive colour based on a threshold.

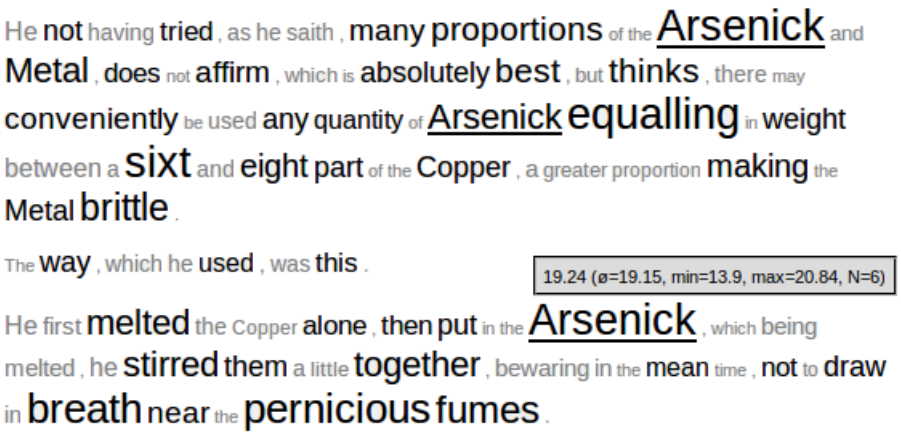

Figure 1. Visualization of three sentences using underlining, font size and colour.

In summary, our tool can be used to spot outliers (instances with very high/low frequencies), which might be worth investigating further (e.g., in a concordance). The pop-up can provide further information about the variability of a word. Finally, one can recognize trends and distributional patterns that are not obvious from a tabular display as in Table 1.

References

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics, 17*(12), 2301–2309.

Evert, S., & Hardie, A. (2011). Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 conference*. Birmingham: University of Birmingham.

Genzel, D., & Charniak, E. (2002). Entropy Rate Constancy in Text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 199–206). Stroudsburg, PA: Association for Computational Linguistics.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380–409.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2016). Visual Text Analysis in Digital Humanities. *Computer Graphics Forum*.

Walsh, B., Maiers, C., Nally, G., & Boggs, J. (2014). Crowdsourcing individual interpretations: Between microtasking and macrotasking. *Literary and Linguistic Computing*, *29*(3), 379–386.