TMI? Visualisation as Research Instrument for Computational Philology

Elli Bleeker, Bram Buitendijk, Ronald Haentjens Dekker, Astrid Kulsdom
R&D Huygens ING - KNAW HuC

This paper investigates the role of digital visualisation in the field of computational philology. We use the the representation of the results of a new collation tool in order to address the more general issue of using data visualisations to advance textual research. After a brief discussion of the collation tool, we problematise the visualisation of its information-rich output. By drawing on and synthesizing methods from different fields of research (philology as well as information visualisation), we investigate how visualising the results of algorithmic analysis of text can be used to further enhance our understanding of textual variation.

Collation, literally meaning "placing side-by-side", is a long-standing scholarly method of comparing two or more versions (called **witnesses**) of a text and find out where the witnesses diverge and converge. Automated collation, in general terms, entails, the processing of two or more strings of characters (each representing a witness) and subsequently aligning the matching readings. Traditionally, most collation methods concentrate on textual differences across documents, called **interdocumentary variation**. The tools typically take TEI-XML files as input, but they collate the witnesses on a plain-text level only. Transforming TEI-XML files into character strings conveniently removes the need to deal with issues like overlap on a programmatic level, but it inevitably entails information loss: **intradocumentary variation** (i.e., variation within the text of an individual witness) and structural variation (paragraphs, chapters, etc.) are generally ignored.[1]

Indeed, the complex nature of texts in the humanities poses a set of interesting challenges for modelling, processing, and representation. Texts in the humanities are often not straightforward nor linear: they can be described as a "complicated web of interwoven and overlapping relationships of

[1] Although a number of tools retain certain markup elements in order to visualise revisions in the collation result, e.g. the BDMP's implementation of CollateX or Juxta Commons (see https://collatex.net/doc/ and http://www.juxtasoftware.org/juxta-commons/) , but these elements play no (analytical) role for the alignment.

elements and structures" (Vanhoutte 2007). Hence a model that takes into account these nonlinear and structural features would be able to provide a better and more detailed representation of the nature of a text. To this end, the R&D team of the Huygens Institute recently developed a collation tool, HyperCollate, that examines textual variation in an inclusive way using a hypergraph model for textual variation. HyperCollate treats texts as a network or graph so that it can natively process intradocumentary variation and store multiple hierarchies. As such, it uses the valuable intelligence expressed by markup to improve the analysis of textual variation.

A full discussion of the operations of HyperCollate, and how it handles intradocumentary and structural variation, has been presented elsewhere (Bleeker *et al.* 2018); this contribution focuses on the complexities of visualising the outcome of HyperCollate as it is illustrative for data visualisation in computational philology. We call the result of the HyperCollate's alignment a **collation hypergraph**. This collation hypergraph contains a wealth of information about the individual witnesses as well as the result of the comparison between these witnesses, on the level of the semantic markup as well as the structural markup. Undoubtedly, the collation hypergraph contains more intelligence than would be desirable to show from an information perspective: it can show intradocumentary and interdocumentary variation on a structural level and on the level of the textual content. Even when the input witnesses are short and simple fragments of a TEI-XML file, the resulting collation hypergraph quickly becomes too large and complex to visualise. This confronts us with a number of tricky challenges, leading up to the central question of this paper: in what way(s) can the information about textual variance be visualised?

If we consider the existing visualisations used in philology and textual scholarship, the answer appears to depend on the user's needs and interests. Following that line of reasoning, the most effective visualisation of HyperCollate's output requires a selection of interconnected textual aspects. In line with Patrick Sahle's *Textrad* (2013) and the editorial orientations described by Van Hulle and Shillingsburg (2015), we maintain that there are several perspectives on text and that these are determined by the scholar's research interests. A useful representation of the collation hypergraph, then, would vary according to the prevalent scholarly perspective. We address this issue accordingly by offering multiple visualisations, each one focusing on different aspects of the text. Users can define what aspects of text and markup they are interested in - and thus which aspects do not have to be visualised - and subsequently select a visualisation that supports their research into the text's variance. This approach necessitates a structural discussion of what different perspectives entail, clear

communication on the influence of a particular perspective on both the text and the selective visualisation, and providing insight into reusing the collation hypergraph for further processing.

In this respect, we recognise that digital visualisations can also serve the computational side of "computational philology", for instance by communicating the operations of HyperCollate through visualising the algorithmic decisions taken by the program, which would facilitate the interpretation of the result and enable the reusability of the output. A collation tool produces a result, regardless of a user's understanding of its operations. This black-box aspect may result in some suspicion among textual scholars regarding the use of tools (cf. Andrews 2017), but more importantly it hinders an understanding of the result. The parameters of HyperCollate are a direct reflection of the user's theory of text and research objectives, and a comprehensive understanding - albeit on a conceptual level - of the tool's decision-making process strengthens the research methodology. This may be effectuated by visual representations of certain intermediate steps, e.g. in the form of a decision tree, but here, too, the complexity and richness of the information hinders an informative visualisation.

In order to conduct a productive discussion on the visual representation of textual variation we need to understand the complexities of what exactly we want to visualise, i.e., the nature of text and the input of a collation tool. We contend that the visual representation of collation result is inherently different than archaeological objects due to the nature of humanities text, and because collation also includes a philological analysis of the textual variation. HyperCollate's inclusive approach to processing textual variation in combination with a visual representation of scholarly perspectives on text compels us to reflect on what we understand text to be.

Bibliography

- Bleeker, Elli, Bram Buitendijk, Ronald Haentjens Dekker, and Astrid Kulsdom. 2018. "Including XML Markup in the Automated Collation of Literary Text". Paper presented at XML Prague 2018, 9-11 February 2018. http://archive.xmlprague.cz/2018/files/xmlprague-2018-proceedings.pdf
- Jessop, Martyn. 2008. "Digital Visualization as a Scholarly Activity". In *Literary and Linguistic Computing* 23:3, pp. 281-293.
- Kosara, Robert. 2007. "Visualization Criticism - The Missing Link Between Information Visualization and Art". Paper presented at the 11th International Conference of Information Visualization, 4-6 July 2007.

- Sahle, Patrick. 2013. *Digitale Editionsformen. Teil III: Textbegriffe und Recodierung*. Schriften des Instituts für Dokumentologie und Editorik 7-9. Norderstedt.
- Vanden Moere, Andrew and Helen Purchase. 2011. "On the Role of Design in Information Visualization". In *Information Visualization*, 10:4, pp. 356-371.
- Vanhoutte, Edward. 2007. "Electronic Textual Editing: Prose Fiction and Modern Manuscripts: Limitations and Possibilities of Text-Encoding for Electronic Editions". *The Text Encoding Initiative* 2 http://www.tei-c.org/About/Archive_new/ETE/Preview/vanhoutte.xml
- Van Hulle, Dirk and Peter Shillingsburg. 2015. "Orientations to Text, Revisited". In *Studies in Bibliography* 59:1, pp. 27-44.