# Visualizing collation results

*Elisa Nury*[1]

*Paper presented at 'Digital Scholarly Editing: Theory, Practice, Methods' DiXiT Convention, Antwerp, October 5-7, 2016.*

In the recent years, the use of automatic collation tools such as CollateX has increased, and so has the need to display results in a meaningful way. Collation is admittedly more than just a record of variant readings (Macé *et al.* 2015, 331). It frequently incorporates additional notes and comments, 'paratextual' elements such as changes of pages or folia, gaps, lacunae, and so on. This combination of variants and paratextual material produces a large amount of complex collation data, which are difficult to read and interpret. Therefore, the editor needs to visualize and analyse the collation results as a whole, and not only variant by variant. A good visualization should offer a way to check collation against the actual witnesses, whether they are manuscripts or printed editions. In addition, the editor should be able to interact with the collation to analyse readings and variants. Collation could be filtered, so as to find patterns of agreements or disagreements between those witnesses, which can indicate how they are related to each other. Visualization and manipulation of collation results are thus essential in order to use collation for further research, such as studying the manuscript tradition and creating a *stemma codicum*.

To tackle these issues faced by an editor, I would like to present a method of visualization of collation results. The method of visualization consists of two aspects: first, a description of the collation table displayed in HTML; second, a Jupyter notebook[2] where the editor can interact with the collation through a Python script, for instance to select agreements between a group of witnesses against another group, or to make small corrections in the alignment. The case study to which the visualization was applied is the *Declamations* of Calpurnius Flaccus. It is a classical literary text in Latin from the second century. The witnesses, four manuscripts and two critical editions, have been encoded in TEI P5 and then pre-tokenized for collation with CollateX into a JSON format that allows to record a reading together with more detailed information.

---

1 King's College London; elisa.nury@kcl.ac.uk.

2 https://jupyter.org/ (Accessed November 8, 2016).

## The Collation Table

The collation table is a visualization that is user-friendly for scholars who do not work with CollateX or any computer-supported collation program. The table typically represents each witness on a separate line or column, with their text aligned when it matches, and blank spaces inserted where a part of the text is missing in a witness. In its most simple form, the table will show only plain text from the witnesses. Enhancements have been proposed to improve this basic table, for instance with colours to indicate the places where a variation occurs: in the Digital Mishnah demo, variant locations are highlighted in grey.[3] Another example is the Beckett Archive project, where deletions are represented with strikethrough and additions with superscript letters.[4] However, other elements are still missing from those helpful visualizations, such as, for instance, the changes of folia mentioned earlier. The reason for recording folia changes is mainly for checking purposes. If the editor or a reader wants to check the accuracy of the transcription for a particular reading, it will be much easier to find the reading back in the manuscript knowing the folio where it appears. How could this or similar paratextual elements be integrated into collation results? One solution is to take advantage of the JSON input format of CollateX.

By default, CollateX can take as input plain text transcriptions of the witnesses to collate. The texts will be split into 'tokens', smaller units of text, at whitespaces. This is the tokenization stage.[5] The collation is then performed on these tokens, which are usually the words of the text. However, it is also possible to 'pre-tokenize' the transcriptions. The JSON format, in particular, allows to record not only the plain text words (t), but also other properties, such as a normalized form of the word (n). There is no limitation to the token properties that can be added: they simply will be ignored during the collation stage, but still be available in the end results. In order to integrate folio location, links to digital images and editorial comments, I transformed the XML TEI transcriptions of Calpurnius Flaccus into pre-tokenized JSON. The tokens include, beside the (t) and (n) properties, a (location) property, eventually a (link) and/or a (note) property. Below is an example of a collation table for the *Declamations* of Calpurnius Flaccus, making use of those properties. I have created this table by comparing the agreements of normalized readings among the different witnesses, with the help of the Jupyter notebook described below.

There are four manuscripts in this collation: B, C, M and N. Each manuscript is divided into two witnesses according to the different hands that wrote the text. For example, B1 is the first hand of manuscript B and B2 is the second hand who made corrections to the text of the first hand. There are also two editions in the collation. The *editio princeps* was first published in 1580 by the French scholar Pierre Pithou and reprinted in 1594. The second is the critical edition of Calpurnius Flaccus published by Lennart Håkanson in 1978. The last column, ID, represents a way to identify rows in the table. There are a few items highlighted in the table of Figure 1:

3 See the demo here: http://www.digitalmishnah.umd.edu/demo (Accessed November 4, 2016).

4 See the news update of September 17, 2014: http://www.beckettarchive.org/news.jsp (Accessed November 4, 2016).

5 See CollateX documentation: http://collatex.net/doc/#input (Accessed November 4, 2016).


*Figure 1: collation table extract.*

- The (i) symbol next to a reading: on click, it can reveal/hide editorial comments that were made during the transcription, especially regarding problematic passages. Here the comment is related to an unknown abbreviation that was not resolved with certainty by the editors of Calpurnius.
- The (⁝) symbol in the ID column: on click, it will reveal/hide locations of the readings for each witness in the row.
- The location is in the form of 'folio number:line number'. The unknown abbreviation mentioned above appears in folio 148r, line 8, of manuscript B. Since there is a digital facsimile available for manuscript B, the location will also link to the image of the page.
- Green and red colours: the coloured lines next to a reading show agreement (green) or disagreement (red) with the same reading of a base text, chosen among the witnesses. Here the base text is the text printed in the edition of Håkanson. As a result, the readings Håkanson rejected because he considered them to be errors are shown in red, while the reading he accepted as true are shown in green. The pattern of colours would of course be different if another text, such as Pithou's edition, had been selected as the base text.

The purpose of the colours is to detect relationships between witnesses, according to the (neo) Lachmannian[6] method of text editing. Lachmann's method focuses on common errors shared by a group witnesses in order to postulate relationships between those witnesses: a group of witnesses are likely to be related if they (1) agree on readings that (2) they do not share with the other witnesses, and especially (3) when they agree in errors, i.e. when they share readings that have no manuscript authority[7] and do not represent the original text. Using the red/green colour scheme was inspired in part by Stemmaweb, a tool for creating

6 Here neo-Lachmannism refers to the improvements to Lachmann's method brought by Pasquali and other Italian scholars, who took Bédier's criticism into account and incorporated the study of the textual tradition and material documents (the manuscripts themselves) to the creation of stemmata. In this sense, neo-Lachmannism is also based on common errors shared by witnesses.

7 A reading with manuscript authority is 'a reading that may have reached us through a continuous sequence of accurate copies of what the author wrote back in antiquity and may therefore be authentic and (by definition) right' (Damon 2016, 202-203).

stemmata with computer algorithms (Andrews 2012).[8] Being able to find common errors in the collation results would be especially useful for a scholar preparing a critical edition. For this purpose, I have prepared a python script, in a Jupyter notebook, which lets users filter the collation in order to find witnesses that agree with each other, and not with others.

## The Jupyter Notebook

The Jupyter notebook offers an interactive way to explore collation results thanks to widgets, which are components of a user interface such as buttons, textboxes, and so on. Here I will show only one example of interaction, the selection of agreements between witnesses (Figure 2).

This small extract from the Jupyter notebook shows two selection widgets in the form of dropdown menus. In the first menu, the user selects a list of witnesses to see where they agree with each other. In the second menu, the user can select another list of witnesses. The resulting collation table will show readings where the first witnesses agree with each other and not with the witnesses in the second list.

The selection of witnesses in this example will result in the collation table that was presented in Figure 1. The table shows the agreements of B and C (both hands) against readings of M, N and P1594. This does not mean that M, N and P1594

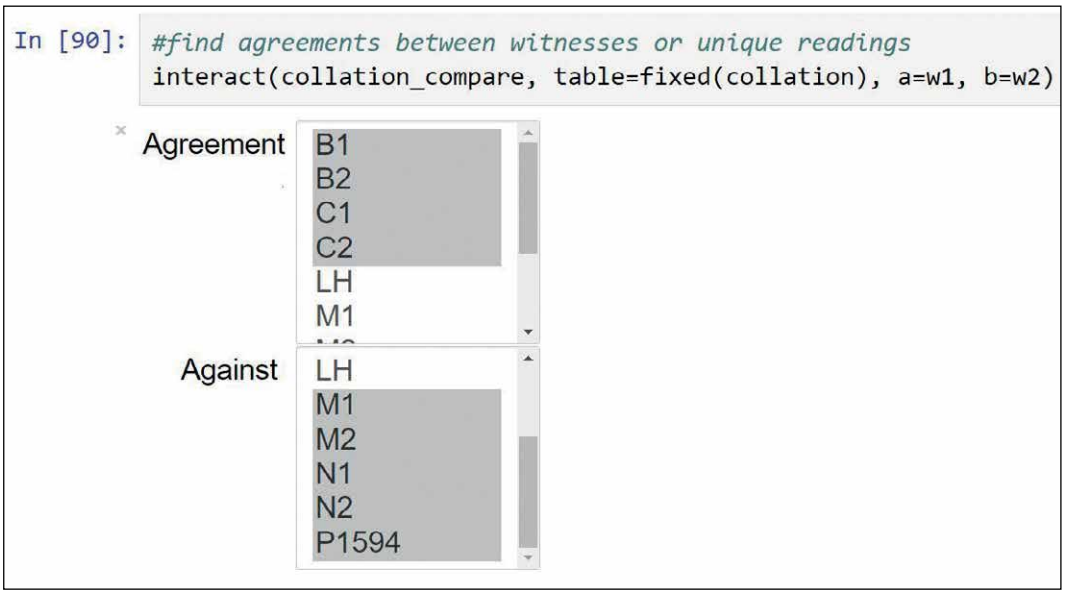


*Figure 2: widgets from the Jupyter notebook.*

8 Stemmaweb makes it possible to visualize collation tables where readings are highlighted in green or red. A green highlight means that the reading is consistent with a given stemma hypothesis. A red highlight means that the reading is not consistent with that same stemma hypothesis. In a similar way, the collation table highlights readings that are consistent or not with a 'text hypothesis,' the text that was selected as a base text.

agree together, only that they are different from B and C for the readings displayed in the table. The full table shows many examples of B and C agreeing together, and especially agreeing in errors. It shows that, at least according to Håkanson, they are related witnesses. In fact, editors of Calpurnius Flaccus recognize that B and C form a closer group, while M and N form another group of manuscripts (Sussman 1994, 19). The *editio princeps* of Pithou, on the other hand, is believed to be related to manuscript N.

The notebook offers more interactions with the collation results. Beside searching for agreements, it is possible to modify the witnesses' alignment, add notes to readings or search for a specific reading. Although the rest of the notebook is not discussed here, the code is made entirely available on Github.[9]

In conclusion, the two examples of visualization described in this paper demonstrate how to make use of collation in an electronic format for further research. CollateX results can be improved with the use of pre-tokenized JSON, as it was already done by other projects such as the Beckett Archive. However, it is possible to integrate more information into the collation with JSON tokens: elements such as location of a word in the manuscript, or editorial comments, are important aspects of collating texts and there is no reason to discard them in a computer-supported collation. As shown in the collation table, the use of a few symbols allows one to make those elements easily available without overcrowding the results. The use of colours is a straightforward way to reveal groups of witnesses that agree with one another and thus help draw conclusions about the manuscript tradition. The collation table and Jupyter notebook presented here hopefully will provide suggestions on how to make available the extra material that is not yet exploited fully in collation visualizations.

## References

Andrews, Tara. 2012. 'Stemmaweb – A Collection of Tools for Analysis of Collated Texts.' https://stemmaweb.net.

Damon, Cynthia. 2016. 'Beyond Variants: Some Digital Desiderata for the Critical Apparatus of Ancient Greek and Latin Texts.' In *Digital Scholarly Editing. Theories and Practices*, edited by Matthew James Driscoll and Elena Pierazzo, 201-218. Open Book Publisher.

Macé, Caroline, Alessandro Bausi, Johannes den Heijer, Jost Gippert, Paolo La Spisa, Alessandro Mengozzi, Sébastien Moureau and Sels Lara. 2015. 'Textual criticism and text editing.' In *Comparative Oriental Manuscript Studies: An Introduction*, 321-466. Hamburg: Tredition. http://www1.uni-hamburg.de/www/COMST/comsthandbook/321-466 Chapter 3.pdf.

Sussman, Lewis A. 1994. *The Declamations of Calpurnius Flaccus: Text, Translation, and Commentary*, edited by Lewis A. Sussman. Mnemosyne Bibliotheca Classica Batava. Leiden: New York: E. J. Brill.

9 https://github.com/enury/collation-viz (Accessed November 3, 2016).