# Refining character relationships using embeddings of textual units

Guillaume Guex[1]

[1]Departement of Language and Information Sciences, University of Lausanne, Switzerland

**Abstract**

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

**Keywords**

LaTeX class, paper template, paper formatting, CEUR-WS

## 1. Introduction

Distant reading tools allow researchers, from various fields, to quickly gain knowledge on textual corpora without actually reading them. Purposes of these methods are various, but can be mainly categorized into two groups: in the first case, these methods are used in order to tag, classify, or summary large quantities of texts, in order to quickly structure information or to deliver a speech over the whole studied corpus. Methods in this case rely heavily on Big Data and make an extensive use of Machine Learning algorithms. In the second case, researchers use these methods to underline hidden structures in a particular text, helping them to refine their understanding of it and reinforce stated hypotheses. Methods in this setting can also rely on Machine Learning, but must typically be build with more caution and attention to details: corpus are smaller, analyses are closer to the work, and methods must be more transparent in order to appropriately interpret results.

Automatic extraction and analysis of *character networks* from literacy works typically belong in the latter group. These methods aim at representing various interactions occurring between fictional characters found in a textual narrative with a graph, thus showing explicitly the hidden structure of character relationships constructed by the author. This structure might allow to find hidden patterns within a book, which can highlight a particular genre or style.

*COMHUM 2022: Workshop on Computational Methods in the Humanities, June 9–10, 2022, Lausanne, Switzerland*

guillaume.guex@unil.ch (G. Guex)

0000-0003-1001-9525 (G. Guex)



## 2. Methodology

When building character networks from a textual narrative, the most widespread method consists in dividing the studied work into $n$ textual units $u_1, \ldots, u_n$, which can be, e.g., sentences, paragraphs, or chapters, and counting characters co-occurrences in these units. Usually, the text constituting these units is discarded and the resulting network displays edges which roughly represent an aggregated number of interactions between characters. However, by doing so, the aggregation occurs on various type of interactions and will give little information about the type of relationship which exist between characters. In this paper, we propose a data organization leading to various methods which takes into consideration the text contained in the unit, helping to refine understanding of characters and their relationships.

### 2.1. Data organization

A textual narrative divided in $n$ textual units $u_1, \ldots, u_n$ can be represented in a $n \times (1 + p + p^2)$ table, as shown in Table 2.1, where $p$ is the number of characters found in the narrative. Each line represent a textual unit, the first column is the text composing this unit, and the next $p$ columns contains variables $O_1, \ldots, O_p$ counting occurrences of each character in the unit. The $p \times p$ remaining columns contains variables $C_{ij}$, which count characters co-occurrences and can be defined with $C_{ij} = \min(O_i, O_j)$. Note that higher order co-occurrences could also be considered (e.g. variables denoting the co-occurrences of three characters), but it remains outside the scope of this article.

This data organization can seem trivial, but it already gives an orientation to the subsequent analyses and should be kept in mind of the practitioner. As a matter of fact, textual unit are now considered as *individuals* (in the statistical terminology), defined by their *variables* contained in the different columns. Of course, the whole

| text | Azelma | Babet | Brujon | ... | Babet-Brujon | ... |
|---|---|---|---|---|---|---|
| Il arriva [...] | 1 | 0 | 0 | ... | 0 | ... |
| Quand Éponine [...] | 0 | 1 | 1 | ... | 1 | ... |
| Éponine alla [...] | 0 | 1 | 0 | ... | 0 | ... |
| Si bien qu'en [...] | 0 | 2 | 2 | ... | 2 | ... |

**Table 1**
A snippet of the data table extracted from *les Misérables*. Rows are paragraphs, the first column contains the text of these paragraphs, the next $p$ columns counts character occurrences, and the last $p \times p$ columns character co-occurrences.

information is contained in the text column, but, for the moment, this column is raw and improper to any kind of statistical analysis. Its treatment will come later in sections (REF) and (REF). The current digression aims at showing that this dataset is oriented in way to see how the last $p + p^2$ variables, i.e. character occurrences and co-occurrences, have an influence over the text contained in the first variable. While the author uses its characters and character interactions in order to build its narrative, we, to a certain extent, work backward: we are searching how character appearances in the textual unit influence her or his choice of words. The choice to include character co-occurrences along with character occurrences is motivated by the fact that the interaction between two (or more) characters should be considered as an object of its own: this interaction is not necessarily the sum of its parts and gives a particular flavor to the unit. As stated before, for the moment, the column containing the text of the unit is improper for statistical analysis. Different approaches could be made in order to extract workable information from this column, such as sentiment analysis (REF), textometry (REF), or even Deep Learning methods. Here, we will restrain ourselves to methods using a *Bag-of-Path* approach.

## 2.2. The Bag-of-Path approach

detecting how character interactions gives

# 3. Results

# 4. Conclusion

# References