Visualization of data generated from the analysis of classical antiquity technical texts

Robert Boloc and Ramon Masià Fornos

to be published at http://www.digitalhumanities.org/dhq/

Abstract

The most common formats in which we can find corpus data are textual and oral. Since the advent of the age of computers, the corpus contents mankind has been generating and collecting have been growing exponentially. This makes the task of analysing the corpus more difficult and creates the need for more advanced tooling to aid with the linguistic studies based on the analysis of the corpus.

This project aims to create such tools to aid with the research of a specific corpus: the classical antiquity technical texts. More specifically it provides two interactive graphical visualisations that the user can filter and configure. These visualisations are built with the latest web technology standards so that only a browser is required in order to access and use them.

The first visualisation is a timeline of the contents on which the user can see the distribution of works or characters or the accumulated distribution of either works or characters. The view can be filtered by multiple content, genre, style, author or work. The user can also see statistics of what is being visualised and choose to see just exact or approximate data.

The second visualisation is a sunburst chart also known as a multi-level pie chart that allows viewing the dataset in a hierarchical form. The user can choose to see either the word, chars or chars without spaces count and the order in which the data is being organised and displayed from the content, genre or style perspective. The chart is zoomable and is accompanied by a legend, breadcrumbs and statistics.

Both visualisations use the same dataset of 131 works. The visualisations have been built to be extensible, so that is possible to add more contents or alter the existing ones without the need to update the code.

We believe that the tools will be useful to researchers trying to extract information from this corpus and will provider them with quick and visual information about the data and the way it has evolved during the time.

1. Introduction

Although Chomsky arguments against the use of corpus data for linguistic studies [1], corpus analysis is still one of the most used techniques in linguistics.

Corpus linguistic data can be found in many different formats, but the most common is textual and oral. This data was traditionally stored in books or recordings, but nowadays the most common format is digital. The increase of computing power in the last 50 years has contributed at increasing the volume of corpus data mankind is generating and storing. The volume at which the data increases is proportional to the increase of computing power [2].

One problem that arises with this increase in volume of data, for corpus linguistics researchers, is being able to analyse these datasets quickly and efficiently in order to extract valuable information [3]. Different methods of easing this task have been devised. Techniques such as filtering the data previous analysis, sorting by a predetermined criteria or sampling the data [4] help with the task but change the original dataset, and doing so can bias the results of the research. Other techniques such as graphical visualisations allow using complete datasets to extract information quickly using charts and statistical information.

The aim of this project is to implement such a graphical visualisation tool. Because there are many corpus datasets available and trying to create a generic one size fits all solution would be a complex task, the scope of this project is to implement a data visualisation of a specific corpus, the classical antiquity technical texts.

The chosen corpus is composed of about 2.5 million tokens generated from only 3000 different words and includes texts generated from -300 BC to 700 AD. The visualisation tool will allow analysing the data on a timeline and on a sunburst chart, filtering and mixing the data being visualised and will include statistical information.

We hope that this tool will provided researchers valuable information about the corpus and allow them to observe patterns and relationships either in the whole dataset or in specific aspects.

2. Method

Given the practical nature of the project the chosen methodology for research and implementation is the Systems Development Life Cycle [5]. We can divide this method into four stages:

Analysis

We started by researching the state of the art technologies being used for data visualization in the field of corpus analysis and outside of this field. Based on this research and the data we are using in this project we developed a research proposal scoped on the visual analysis of classical antiquity technical texts, and the generation of a tool to aid visualising this data.

Design

In this stage we designed wireframes for the proposed visualizations, decided on how much information to display to the user and in which format. We also created a repository to store the implementation.

Implementation

Using the outputs of the analysis and design stages we implemented the data visualisation tool using the D3 library, third party libraries and vanilla JavaScript, HTML and CSS.

Testing

This stage consists of testing that the implementation output satisfies the requirements we set in the analysis and design stages. When issues were encountered in this phase we would step back into the implementation phase, work on the issues and continue with the testing phase until the outcome was successful.

3. Results

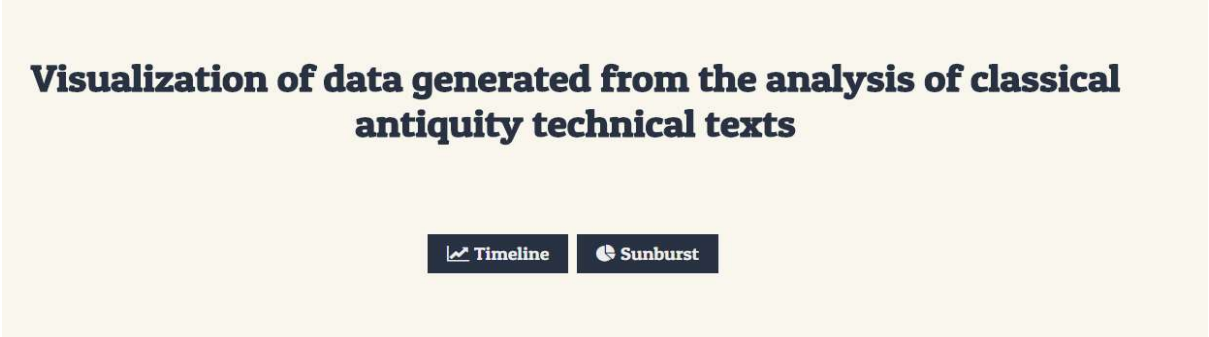

Fig. 1. Welcome screen

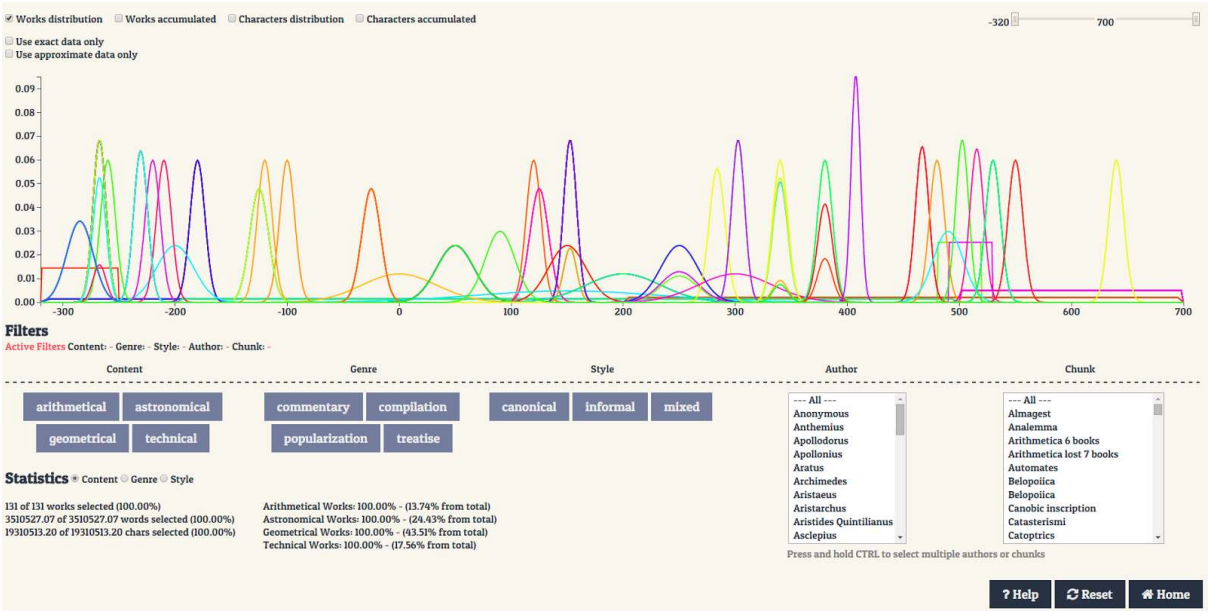

Fig. 2. Timeline visualisation

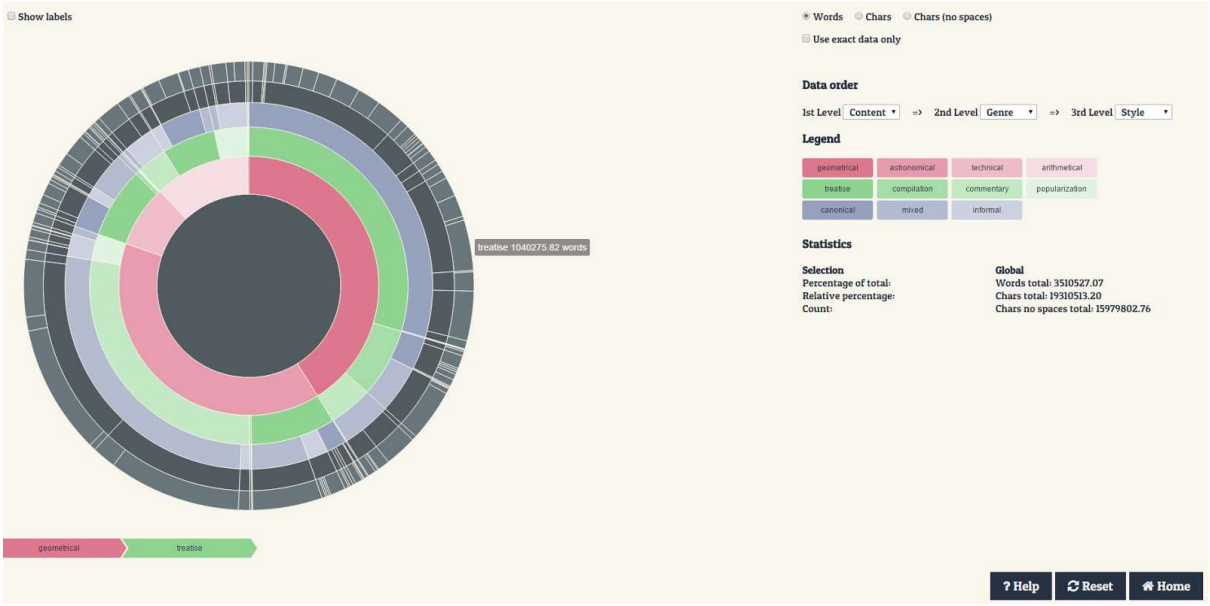

Fig. 3. Sunburst visualisation

The outcome of this project is in form of two data visualisations (Figure 2 and 3) that the user can manipulate and interact with. Both visualisations display the corpus data in a different way in order to cover most of the use cases for the end user.

Both visualisations are built using web technologies and follow the latest standards so that the user can interact with the charts just by using an up to date browser.

The reason to make the visualisations interactive as opposed to static is to allow the user to view the same data from different perspectives [6] and to allow them to explore the complete dataset for themselves.

To access the visualisations the user is presented with a welcome screen (Figure 1) from which he or she can choose which visualisation to access. The design and color scheme of the visualisations and welcome screen has been thought and implemented so that is not distracting, is pleasant and does not cause strain to the eyes when viewed for a large amount of time.

The visualisations can be publicly accessed here [8]

4. Discussion

The human body has excellent sensorial capacities and a big part of the cells related to perception belong to the vision. This is the reason we acquire more information through the vision then through all the other sense combined [7]. This is a great argument to use visual tools to analyse complex information and is one of the pillars of this project.

By mapping complex and large data on a visual timeline or chart we simplify the process of capturing information, allow easy pattern recognition and reduce the time it take to analyse the data.

There are a big number of possible visualisations that can be implemented: bar charts, pie charts, scatter plots, text clouds, etc. and we analysed the possibilities based on other works in the corpus visualisation field [9]. The outcome of this analysis has been that the best suited visualisations for our task were the timeline bar chart and the sunburst or multi-level pie chart. This decision was taken based on the structure of our dataset and the expected output of the data.

Implementing these two charts allow two different angles when analysing the corpus data and suit different research purposes. We hope to satisfy a broad spectrum of users with the visualisations and cater for their needs.

From a technical point of view implementing the two charts has created a few challenges as some of the features we desired to implement were not available in D3 or any other library so we had to

5. Conclusion

Extracting relevant information from large amounts of data is a difficult task, but tools such as the ones implemented by this project allow simplifying the process and reduce the amount of time needed for the extraction.

Implementing the solution using popular technologies, and web accessible such as D3 allow the tool to be extensible and easy to work on for future improvements. D3 is very powerful and the de-facto library when implementing web visualisations [10] and so we think it is the right choice for this project.

We hope the visualisations will be useful and help researchers or general users with their tasks.

6. Reference List

[1] McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press

[2] Rayson, P., Mariani, J. (2009). Visualising corpus linguistics. *CL2009 Proceedings of the Corpus Linguistics Conference.* Liverpool

[3] Garside, R., Leech, G. N., and McEnery, T. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman

[4] Rayson, P., Mariani, J. (2009). Visualising corpus linguistics. *CL2009 Proceedings of the Corpus Linguistics Conference.* Liverpool

[5] Oates, B. (2006). *Researching information systems and computing* (pp. 39-40). London: SAGE.

[6] Murray S. (2013). Interactive Data Visualization for the Web. O'Reilly Media, Inc

[7] Siirtola, H., Nevalainen T., Säily T., and Räihä K.J. (2011) *Visualisation of Text Corpora: A Case Study of the PCEEC.* Viewed on the 25th of October 2015, at http://www.helsinki.fi/varieng/series/volumes/07/siirtola_et_al/

[8] http://37.139.10.110/pfm/index.html

[9] Talianová, D. (2014). *Visualization of Corpus Data*. Brno: Masaryk University

[10] Zhu, N. Q. (2013). *Data visualization with D3.js cookbook*. Birmingham: Packt Publishing.