Visualizing Literary Data

Erik Tjong Kim Sang
Meertens Institute, Amsterdam
erik.tjong.kim.sang@meertens.knaw.nl

23 May 2016

Nederlab project

The Nederlab project aims at collecting and making available all texts relevant for the culture and language of The Netherlands

This involves texts of about the year 800 until now

These historical texts are being made available to the research community on a website, together with analysis tools

Extensive metadata are available for the texts: document types and dates, authors and even text analysis

Visualizations

Searches in the text collection produce large sets of numbers

We created visualizations for these number sets to enable researchers to examine large result sets

The visualizations can also be used as a user-friendly way for adding further restrictions to search results

For this purpose, the visualizations have been made interactive

Challenges

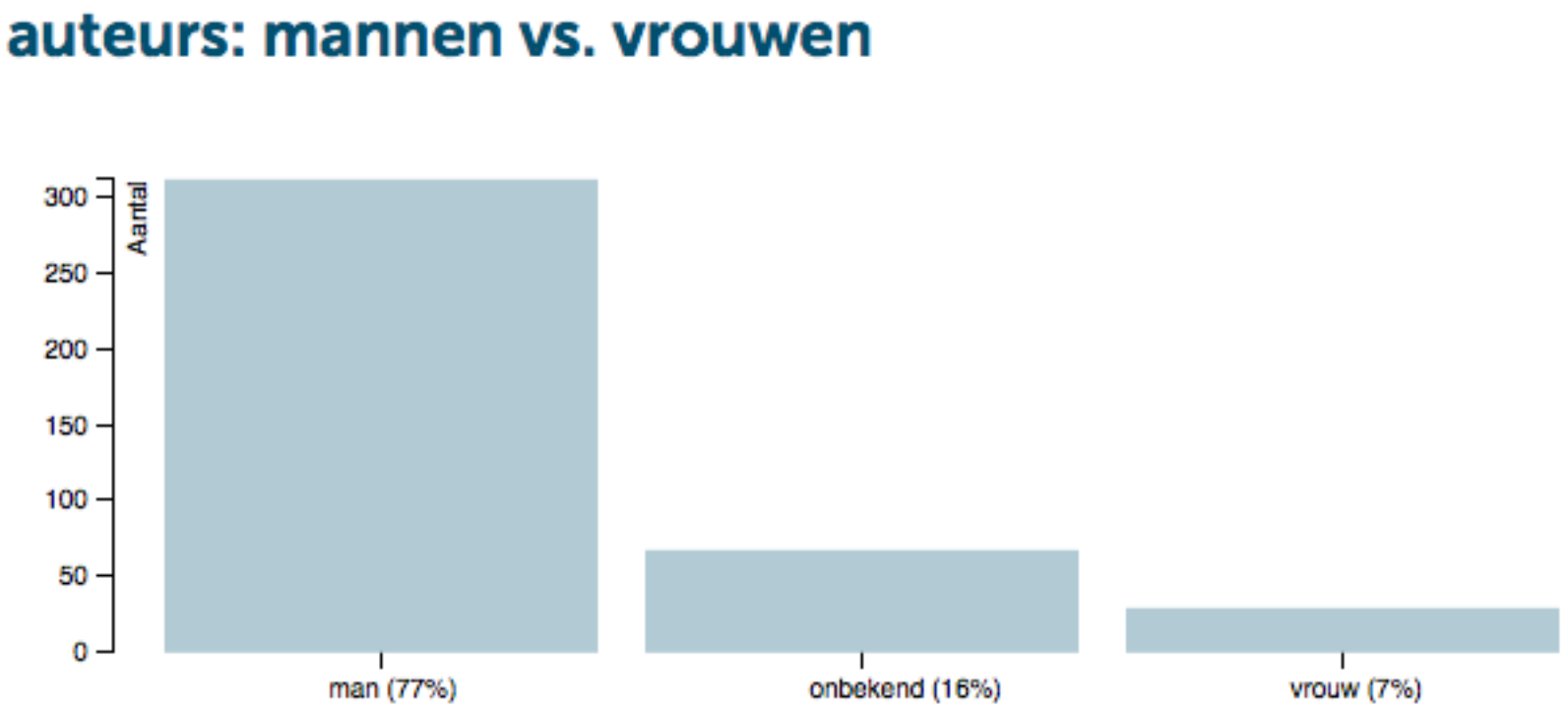
The design of the visualizations turned out to be more complex than expected

Basic questions like *How frequent was the word "war" in the various years of the nineteenth century?* can be answered in different ways

Each different method for answering a question requires a different visualization

Providing different visualizations for one questions make the user interface more complex: do we want this?

Complexity example: male vs. female authors



Design questions

How do we deal with:

- documents with unknown authors?
- authors with unknown gender?
- documents with more than one author?
- authors that contributed to more than one document?

Design choices

- documents with unknown authors:
  treat as document with one author with unknown gender
- authors with unknown gender:
  include in unknown authors count
- documents with more than one author:
  count each author seperately
- authors that contributed to more than one document:
  count each document of each author seperately

Effects of the design choices

Because documents are counted more than once if they have more than one author, absolute counts for male, female and unknown authors may exceed the number of documents

- 10 books, all written by male duos, have 20 male authors

Since a document may be written by combination of male, female and authors of unknown gender, restricting the search result to one of the gender may produce unexpected results

- after selecting books written by females from two books, one written by a male and the other by a male-female duo, the result only has 50% female authors

Effects on users

We chose simplicity over functionality

The consequence of this is that the data visualizations only offer limited options for changing data interpretations

Users that want visualization relying on alternative data interpretations will have the opportunity to download the (numeric) data and process them with their own software

Case study

In 1885 there was a schism in the Dutch literary magazine *De Gids* which led to the publication of the competitor *De Nieuwe Gids* in 1885-1894

As a case study, we would like to answer the following questions:

- Can we find demographic differences between the author groups?
- Did two magazines write about different topics?

Demo: author demographics comparison

Topics: birth decades, male-female ratio, birth provinces, interactivity

Content comparison

Summarizing the content of selections of our corpus is technically challenging

For example: we cannot get a frequency list of words in a search result if it contains more than fifty documents

Therefore we do not offer visualizations based on content

Users can download frequency lists based on samples and perform the analysis offline, see the paper for an example

Concluding remarks

We have created visualizations of literary data to improve the quality of the research analysis and to make it easier to inspect aspects of the data

Designing the visualizations proved to be challenging

We limited the functionality of the visualizations in order to avoid interface complexity

The visualizations deal with metadata; the system architecture offers only limited data for text visualizations

THE END