Visualizing Literary Data

Erik Tjong Kim Sang
Meertens Institute Amsterdam
Joan Muyskenweg 25, 1096 CJ Amsterdam, The Netherlands
erik.tjong.kim.sang@meertens.knaw.nl

Abstract

We look at different aspects of Dutch magazines, both from the fields of literary studies and linguistic studies. We explore the background of authors with respect to birth locations, ages and gender, and also in how language use in the magazines evolved over a period of several decades. We have created several interactive visualizations which enable researchers to browse and analyze text data and their metadata. The design of these visualizations was nontrivial: invoking questions about how to deal with missing data and documents with multiple authors. The data required for some of the visualizations useful for researchers, were infeasible for the software architecture to generate within a reasonable time-span. In a case study, we look at some of the research questions that can be answered by the data visualizations and suggest another data view that could be interesting for literary research. Interesting topics for future research rely heavily on improvements of the search architecture used and including extra annotation layers to our text corpora.

1. Introduction

Magazines are an interesting subject of study for both literary and linguistic researchers because they represent the spirit of time, both with respect to the topics covered and the language variants used. Magazine texts are even more valuable for research when they are available for long spans of time, like decades or even centuries, and when they are accompanied by metadata, covering aside from publication times information about the authors of the magazine articles, like their birth location, age and gender.

For the language Dutch, which is spoken in The Netherlands and the northern half of Belgium, a large collection of magazines is available in the Digital Library for the Dutch Literature (DBNL[1]), which contains digital versions of Dutch texts from the thirteenth century until today. The magazine collection includes the literary magazine De Gids, of which about 170 digital yearly editions are available, dating back to 1837. DBNL also offers biographical information about authors like dates and locations of birth and death.

The present (2016) websites of text collections like DBNL make it possible to search and find text segments that contain specific words or word variants. For researchers in the humanities this is not sufficient to answer their research questions. They frequently want to quantify and compare search results. For example, they would like to know how the frequency of a certain word in the seventeenth century compares with its frequency in the eighteenth century. Or they want to know if a gender and age bias which is present among a certain magazine, has disappeared in a new magazine that is a spin-off of this magazine. It is difficult, if not impossible, to answer these questions with a search interface which only returns text snippets as search results.

For such research goals, we developed interactive visualizations of text search results. The visualizations have two important goals. First, by quantifying the search results and presenting the numbers visually, they aim at improving the quality of the analyses that can be performed of this data. Second, by offering researchers the possibility to select sections of visualizations, they offer the opportunity to further zoom in into the search results and thereby increase the possibilities to browse the data.

We aim at visualizations which are as simple as possible and which give quick insights in certain aspects of the data. However some users will be interested in a more elaborate analysis of the data. Rather than making the visualizations more complex for all users, we aim at satisfying the information need of these users by enabling them to download the counts on which the visualizations were based.

After this initial introductory section, this paper contains five more sections. In the next section two, we will describe some related work. In section three we will outline the technical infrastructure required for being able to retrieve quantitative search results from our text corpus. In section four, we present the interactive visualizations we offer for exploring and analyzing the data. Section five contains a case study with these visualization. In section six, we conclude.

2. Related Work

Modern data visualizations can be traced back to the eighteenth century (Tufte, 2001). Back then, the prime data format used in visualizations were numbers, which will also be the main data format that we will work with. A useful library for numeric visualization, which we will use, can be found on the website `d3js.org`. However, our data is also suited for more text-friendly variants of visualizations, like word clouds, created by Jim Flanagan around 2002, and for geographic visualizations, like the visualizations offered by the website `openlayers.org`.

The source of the text data that we work with for this paper, is the website `dbnl.nl`. Earlier, a one-time analysis of two Dutch magazines present in DBNL was done and static comparisons and visualization of biographical author data were created (Zhang et al., 2013). We will use a vocabulary comparison method applied earlier to differences between standard Dutch and Surinamese Dutch, the variant of Dutch spoken in the South American country Suriname (Tjong Kim Sang, 2014): t-score (Church et al., 1991). We were also inspired by the Google Books Ngram Viewer (Google,

[1] dbnl.nl

```
{
 "condition": {
  "type": "equals",
  "field": "NLTitle_title",
  "value": "Amsterdam"
 },
 "response": {
  "facets": {
   "facetranges": [
    {
     "field": "NLTitle_yearOfPublicationMin",
     "start": "1600",
     "end": "2000",
     "gap": "100",
     "ex": "persons"
    }
   ]
  }
 }
}
```


Figure 1: Example of faceted search. This query looks for all documents with the word *Amsterdam* in the title and presents document counts per century in the results.


```
{
 "status": "ok",
 "facets": {
  "facetranges": {
   "NLTitle_yearOfPublicationMin": {
    "counts": [
     "1600", 9,
     "1700", 229,
     "1800", 61004,
     "1900", 127
    ],
    "gap": 100,
    "start": 1600,
    "end": 2000
   }
  }
 }
}
```


Figure 2: Output of the query presented in Figure 1: a list of document counts per century.

2010), which compares word frequencies of different years. Since we have centuries of text available, we should be able to generate similar graphs. Parallel to our work, people are working on creating visualizations of historical Dutch text data with the statistical package R (Komen, 2015).

3. Technical Infrastructure

Currently many people have experience in working with information extraction systems and therefore users have high expectations of search systems. Our user group, researchers in the humanities, expects to be able to search in large text collections and get adequate results in a fraction of a second. The technology for quickly retrieving relevant text snippets and document counts is available, for example in Apache Lucene and ElasticSearch (Gormley and Tong,

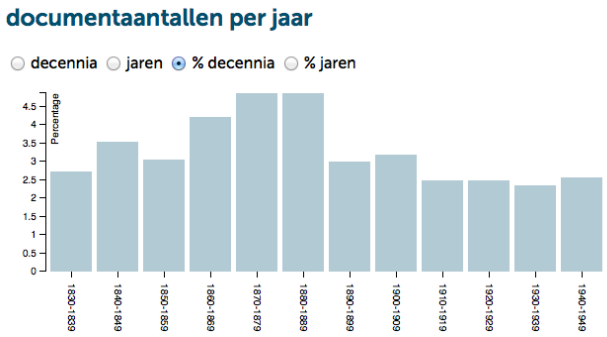

Figure 3: Visualization displaying the percentage of documents per decade containing a certain word. Users may select different output modes: absolute counts per year or decade, or relative counts (percentages) per year or decade.

2015), but we need more. The visualizations which we aim at require faceted search, that is search in which multiple filters are combined, and segmented output, output that is divided in several sections, for example corresponding with different time segments.

An engine for faceted search for our data was created in 2013 (Brouwer et al., 2013) in the Nederlab project (Brugman et al., 2016). The data is stored in several Apache SOLR indexes. An interface layer called the broker performs the interaction between the user and the indexes. This broker translates user queries written in JSON to one or more commands for the indexes. The index responses are analyzed by the broker and sent back to the user. Examples of a user command and the broker output for this command can be found in Figures 1 and 2, respectively. The query looks for documents with the word *Amsterdam* in a document title and requires the results to be presented in matching document counts per century. A lot of processing time is saved because the combination the broker and the indexes perform aggregating different relevant counts internally.

4. Visualizations

We compiled a list of data visualizations which could be useful for humanities researchers while browsing and analyzing text collections. We used JavaScript in combination with the visualization library D3 (Bostock, 2011) for implementing these online visualizations. The communication between the JavaScript code and the data run via the broker interface described in section 3..

4.1. Documents Counts per Time Unit

Our first goal was to create a visualization of the number of times that a word or phrase is used over time. However, the broker interface does not return word counts, apparently because of limitations of the underlying information system architecture: Solr and Lucene (Brouwer et al., 2015). Therefore, the only way to obtain the relevant numbers would be to retrieve all relevant documents and perform the word counting when the user requests this. This could lead to unacceptable time delays in the processing of queries and for this reason we have refrained from creating this visualization mode. However, it is on the top of

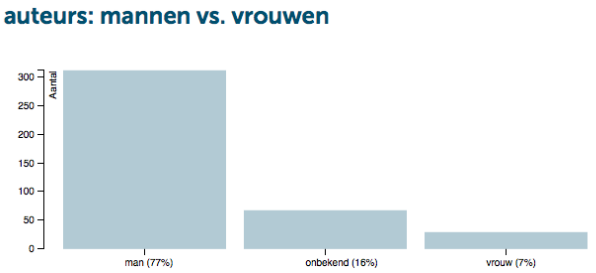

Figure 4: Bar plot showing the gender of the authors of documents that matched a query. There are three values: male (*man*), unknown (*onbekend*) and female (*vrouw*). The counts are document-based so authors that wrote multiple matching documents will be counted multiple times. Also, documents with several authors will contribute to the counts several times. Gender value unknown means either than the gender of an author is unknown or that the author of a document is unknown.

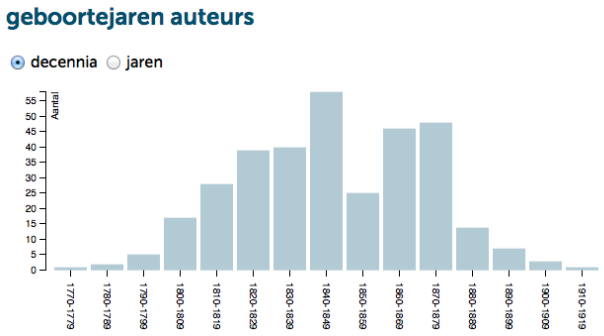

Figure 5: Bar plot displaying the number of times authors involved in documents matching a query were born in certain decades (*decennia*). Users may choose a plot with year (*jaren*) bars instead. The counts are document-based, so authors that are involved in multiple documents matching the query will be counted multiple times.

our wish list regarding extensions of our system, as many users have expressed their interest in obtaining overviews of word counts.

As an alternative for the word count visualization, we created a document count visualization, showing how many documents contain a certain word or phrase at least once, within a certain time frame. As basic time unit we chose one year but we also offered the possibility to examine the data in chunks of ten years, thus providing a smoother view for data that are available for longer time periods.

Figure 3 is an example of of a bar plot showing the percentage of documents that contain a certain word per decade in the period 1830-1949. In order to find the exact number associated with a bar, users can hover over the bar after which a small pop-up window will present the time period and the associated value. Furthermore users can choose between displaying absolute document counts per time frame and percentage counts. Since our document collection is all but balanced with respect to time, with most documents originating from recent times, the latter option often produces a fairer view.

Percentage counts in comparison with all available corpus material are not sufficient to answer all research questions. For example, researchers my want to know what percentage of articles of a magazine contained a certain word in various time periods. Or they could be interested the percentage of articles of a certain magazine written by female authors. Currently we compute percentage scores by comparing query results with overall collection counts, which works fine for studying magazine texts. However, a more general solution in which two arbitrary queries are compared, is preferable.

Clicking on a bar will impose a time-restricted filter on the query related to the time frame corresponding to the bar. For example, selecting the decade bar 1920-1929 for documents written by female authors of a certain magazine will restrict the query results to that time period. This way of interacting with the visualization enables users to quickly inspect several aspects of the corpus material.

The present visualization interaction allows users only to create time filters which correspond with a single bar in the graph. In order to select arbitrary time frames for inspection, we have experimented with two sliders under the graph: one restricting the number of years displayed in the graph and one for changing the start and the end time of the graph. However, since our users found the interaction with the sliders complicated and confusing, we have removed them from the visualization.

4.2. Gender of Authors

Next, we created a bar plot with counts of the gender of authors. Since our metadata is incomplete, we have three gender values: male, female and unknown. We needed to make a choice about how to count the genders: author-based, where each author is counted only once, or document-based, where authors of of documents that appear several times in the query result, are counted several times. We chose the second alternative because this provided the fasted query responses of the broker[2].

We also needed to choose how to deal with documents of which the authors were unknown. There were two ways to handle this: we could leave the unknown authors out of the visualization or we could assume documents with unknown authors were written by an unknown author with an unknown gender. After consulting with our users we chose the second option. This means that the bar for the value unknown in the graph represents counts for two different cases: a known author with an unknown gender or an unknown author.

Figure 4 is an example of a bar plot containing gender counts. There are three values: male (*man*), unknown (*onbekend*) and female (*vrouw*). The bars represent absolute counts but bar-count-based percentage counts have been added below the graph. Since documents can have multiple authors, the author counts may exceed the number of documents. This may confuse the user. An alternative would be to use fractions for authors of shared publications, like 0.75 and 0.25 for a document with three male and one female

[2] In more recent implementations, the broker responce speed is not influenced by the format of the gender visualization.

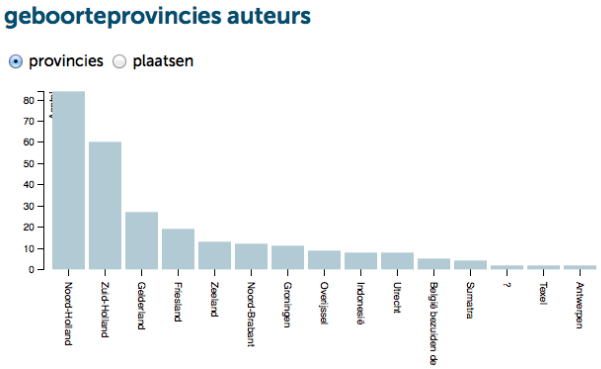

Figure 6: Birth provinces (*provincies*) of authors matching a certain query. Users can also select birth places (*plaatsen*). Provinces of birth and death include local and foreign ones. The question mark is nonstandard content of the birth location field. Provinces are sorted from the most frequent (left) to the least frequent (right). Provinces with a count of zero are not shown.

author. However, needing to deal with fractions of author could be difficult to explain to the users as well.

Like in the document count visualization, moving the mouse over a bar will display the gender and the exact count. Clicking on a bar will restrict the query results to that particular gender. However, here the fact that documents can have multiple authors causes unexpected query results. For example, a query for female authors usually generates a result including some male and unknown authors. The reason for this is that queries are document-based and documents written by females may include coauthors which are male or of which the gender is unknown.

The gender bar plot was an interesting case in which we found out that a seemingly basic view of the data, turned out to be far more complicated than was expected. One reason for this is that we had to work with a search architecture which did not always provided the numbers that we needed. The other was that our interpretation of what was the optimal way to visualize the numbers was sometimes different of the expectations of the users. The current visualization is the most simple one that our users could agree with[3].

4.3. Birth and Death Year of Authors

Bar plots for birth years or death years, can be drawn in the same way as in the visualization for document counts (see section 4.1.). Like in the visualization for gender, birth and death years of authors that are associated with more than one document in the query result, will be counted multiple times.

Figure 5 presents an example of a birth year visualization. Only absolute counts are visible in the plot. Percentage counts could be added in the same way as in the gender plot (Figure 4), by dividing the counts to the total of the

[3]The current gender visualization does not offer options for alternative data interpretations but a valid case could be made to offer choices between counting authors of multiple documents more than once or not and counting every author of a shared document once or only by a fraction.

Figure 7: Bubble map representing the birth locations of authors writing in the Dutch magazine De Gids. The number is each bubble represents how often an author of that city or region was involved as an author of an article in the magazine. Bubbles with numbers are combined when the user zooms out from the map, creating a regional overview. For example, the bubble with 226 represents the two cities The Hague (118) and Rotterdam (108). Together with Amsterdam (311), these cities are the most common birth places of authors of De Gids.

counts in the graph. Exact counts related to a bar can be inspected by moving the mouse pointer over the bar. Clicking on a bar repeats the query with an extra filter restricting the birth year/decade or the death year/decade of the author to the value corresponding to the bar. Because the search is document-based and documents may have more than one author, a birth/death time restriction may produce out-of-range results which correspond to coauthors of documents with matching authors.

4.4. Birth and Death Location of Authors

Birth and death locations can be visualized in the same way as birth years: with a bar plot. Figure 6 presents an example of a visualization of birth provinces. The provinces are sorted from the one with the highest count (left) to the one with the lowest count (right) while ignoring provinces which did not match the query. Provinces can be both local (i.e. located in The Netherlands or Belgium) or foreign. The question mark in the graph is a nonstandard way of indicating that the birth location of an author is unknown, usually the field is left blank. Users can choose to view counts for birth places rather then birth provinces. Like in the other visualizations, moving the mouse pointer over a bar will evoke a pop-up window showing the exact count corresponding to the bar. Clicking on a bar will start a new query with an extra filter restricting the birth province of the authors to the value corresponding to the bar. Again this

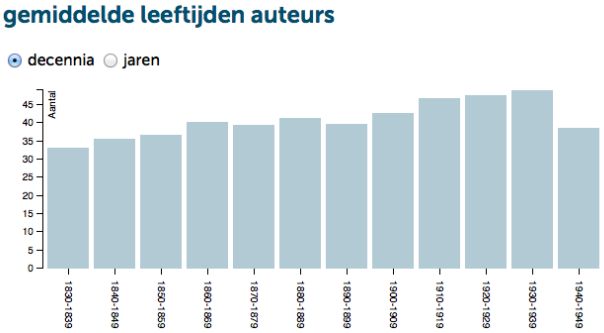

Figure 8: Visualization showing the average age of authors of a certain magazine per decade. The average age seems to move up slowly from 1830-1839 (33) to 1930-1940 (50) after which there is a big drop for the war decade 1940-1945 (38). Users can choose to view a bar plot for years (*jaren*) rather than for decades (*decennia*).

Figure 9: Word cloud for the query *Havelaar*, a novel title and a character name, in the articles of the Dutch literary magazine De Gids. The size of the font used for a word indicates its frequency in the query results. Striking words in the cloud are the names of the author of the novel: *Multatuli* and *Dekker*, as well as words common to literary works: *literatuur* (*literature*) and *roman* (*novel*). Word clouds for query results proved to be expensive to compute.

may lead to out-of-range birth province values, because the search is document-based and documents may have more than one author.

We have also experimented with a visualization based on a map. We used a standard vector data representation for maps with clustered numbers on the map which become more detailed when the user zooms in on the map (ol.Cluster of OpenLayers version 3.0). The example map in Figure 7 with birth locations of authors of the Dutch magazine De Gids makes clear that authors from Amsterdam (311) and The Hague/Rotterdam (226) contribute most to the magazine. Unfortunately, our users found this mapping visualization method confusing. Since we were unable to find a mapping visualization that suited their needs, we have stopped using them. The map was the only of the visualizations that was not (yet) interactive.

4.5. Average Age of Authors

For the publications for which we know both the publication year and the birth year of the authors, we can compute the age of the authors at publication time. The average of these numbers per time unit gives an insight in the average age of authors in different time periods . The visualization is similar to the document count visualization described in section 4.1.: a bar plot with time units years or decades. An example graph can be found in Figure 8.

The interaction with the bars is the same as in the document count graph described in section 4.1.: counts related to a bar can be inspected by moving the mouse pointer over the bar and clicking on a bar provides starts a new query with the birth year of the authors restricted to the value(s) corresponding to the bar.

4.6. Other Visualizations

Apart from the unrealized word count visualization (see section 4.1., there is one more visualization which we would like to create: one that shows which words occur more frequent in a query result than could have been expected. For example, an author may use certain words more frequently than others. By comparing the relative frequencies of the words in his work with the the frequencies of the words in all documents, we could find out his favorite words. A similar comparison could be made for documents of a certain time period.

For such a visualization, we need relative word frequencies: frequencies of words in a query result divided by their corresponding overall frequency. We only have word frequencies of query results available but not the overall frequencies. We work with a dynamic corpus, so the overall word frequencies are subject to change. Ideally they would have been computed at index time and made available, but this is not the case. This means that we need to compute these frequencies ourselves for each query. But since queries may result in tens of thousands of words, such computations will require a lot of time. Therefore we have refrained from creating this visualization.

As an alternative, we have generated a word cloud visualization of the content words in a query, that is the words minus the stop words (common words like the Dutch equivalents of *the*, *be*, *to*, ...). This offers a good summary of the words that be found in documents matching a query (see Figure 9). Users that are interested in a particular word in the cloud can restrict their query to documents containing that word by clicking on it in the word cloud.

In order to create the word clouds, all the words of all the documents in the query result need to be processed. This requires many computational resources. Queries with many results have even caused our broker search engine to crash. For this reason we have restricted the maximum number of documents used for creating the word clouds to ten. The sample of selected documents is chosen randomly. Our users did not like this fact, so they rejected this particular visualization. For that reason we have abandoned using the word cloud visualization.

5. Case Study

De Gids (*The Guide*) is a Dutch literary magazine which was first issued in the year 1837. The magazine still exists in 2016. Most of the issues are available in digital format on the website `dbnl.nl`. In 1885 a group of young Dutch authors started a new magazine De Nieuwe Gids (*The New Guide*) with a specific name reference to the older maga-

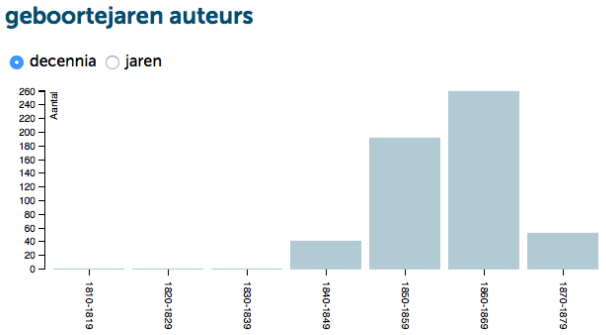

Figure 10: Overview of the decades of birth of the authors writing in the magazine De Nieuwe Gids in 1885-1894. The sequence of bars is shaped like a Poisson distribution with the maximum value in the decade 1860-1869. The author birth year distribution for the older magazine De Gids in the same time period looks similarly, but it has its peak in the decade 1840-1849.

eene haar zijne hem hare Aurelie had Guido Cecile Eduard ende was Jules den zich oom aan Ottilie welke jaren Grimm Staten Moli Bismarck Quaerts Japan hij hart Amerika Oranje Unie Molire eeuw negers Zuiden archieven brieven Jonker geschiedenis letterkunde dollars parlement oude vader Goncourts Coster godsdienst vrouw burcht Rensken dochter Dekker Tiryns re Lievendaal Flaubert Balzac in Noord Japansche Jacob Douwes gebied Holland president zoon zijnen haren Latijn grondwet Mainz der werd kon Zij vreemdeling Noorden blik zwaard mevrouw Zuid graan Amerikanen Armande Athene uitvinding Kees Europa U weerloosheid Kaakebeen middeleeuwsche Germaansche mijnheer Doopsgezinden Wilhelm sprak rijtuig Amerikaansche

Figure 11: The 100 most surprising words in the magazine De Gids in 1885-1894 in comparison with the words found in de editions of De Nieuwe Gids of the same period. The words are sorted by decreasing t-score.

zine. The editors of the new magazine were part of a new Dutch literary movement called Tachtigers (*From the Eighties*) which were organized in a group called Flanor. The new magazine lasted until 1943 but most of the original editorial staff left in 1894. Digital versions of the issues of De Nieuwe Gids can also be found on the website `dbnl.nl`.

It would be interesting to study the differences between the two magazines, especially for the rebel years period: 1885-1894. Did the authors really write about different topics in the new magazine? How did that influence their choice of words and topics? Were the authors of the new magazine younger than those of the older publication? What about the male-female ratio in the two magazines? Was there a geographical difference, for example by birth location, between the two magazines?

The visualizations presented in the previous sections make it easy to compare the two magazine section. We searched for all articles published in De Gids and De Nieuwe Gids in the years 1885-1894. We found 1034 articles for De Gids and 536 articles for De Nieuwe Gids. In the visualizations the most striking difference between the two magazines was the difference in age. The authors in De Gids were on average between 43 and 49 years old in the ten publication years while the average age of the contributors to De Nieuwe Gids was between 28 and 33. This difference is also visible in the graphs for birth years: most authors writing in De Gids in 1885-1894 were born in the decade 1840-1849 while most authors of De Nieuwe Gids in that time period were born in 1860-1869 (see Figure 10).

Another striking difference could be found in the gender visualization. 7% of the known authors in the older magazine De Gids proved to be female while only 2% of the known authors in the new De Nieuwe Gids were women. We had expected a higher female-male ratio in the new magazine but this expectation turned out to be wrong. The numbers in the gender visualizations were hard to use. The magazine De Gids had many authors with unknown gender (21%) and these made it difficult to see the actual male-female ratio: 93%-7% rather than 74%-5% as reported in the graph[4]. Having the counts for unknown gender in the graph makes it difficult to see the actual male-female ratio. However, the counts for unknown gender are in the graph because that was an explicit wish of our users. But it would have been better if displaying the numbers for unknown gender was optional.

Since most of the editorial staff of the new De Nieuwe Gids was born in and around Amsterdam, we expected that many of the authors of the magazine would be born there and in the associated province North Holland. This proved to be the case: 43% of the authors who wrote in De Nieuwe Gids in 1885-1894 were born in Amsterdam and more than half (58%) came from the province of North Holland. However, these numbers were not exceptionally higher than the corresponding numbers for the magazine De Gids: Amsterdam 33% and North Holland 45%. So one can argue that Amsterdam authors dominated De Nieuwe Gids but this was also true to a lesser extend for the magazine De Gids.

We compared the context of the two magazines by selecting ten articles of each volume and counting the words in those articles. Next we compared the word frequencies of the two magazines with the t-score (Church et al., 1991) which can be seen as a surprise factor: a high t-score indicates a surprisingly high frequency of a word in comparison with the frequency of the word in some reference corpus. In this case we calculated the surprise factor of a word in one magazine by comparing its frequency with its frequency in the other magazine.

Figures 11 and 12 contain an overview of the most surprising words in the two magazines. These include character names of stories (Aurelie in De Gids and Johannes in De Nieuwe Gids) which as expected because the two magazines publish different stories. De Gids seems to contain more articles about foreign politics: Bismarck, Japan,

[4]We assume that the distribution of unknown values is equal to that of the known values. Hence, from a 74-5-21 distribution we conclude that 74/(74+5)=93% are male and 5/(74+5)=7% are female.

Sokrates Johannes is Alkibiades ge menschen ik Heer een dingen Windekind uw dat Want mijn En politie Ktesippos wil mensch geneeskundige sentiment Thales zal inkomen Plato mooi wij kan oorzaak socialisten gij redeneering niet arbeiders dus Dit klasse begeerte deze zon zeggen doch Verstege sonnetten Winkel sonnet wezen Wistik spreken Amsterdam dit Shakespeare Marken hulp volgens ziel erg Houten Nu Hippias gemeente pct inkomstenbelasting Berthollet bestaan oorzaken socialistische Handelsblad arbeider heer en wet waarneming sociaal nous Dat brochure samenleving Robinetta lichaam stoffen kapitalistische Rochemont Agathon hebt God behandeling anders als begrip temperatuur socialisme boekje ding doelleer B elkaar woorden

Figure 12: The 100 most surprising words in the magazine De Nieuwe Gids in 1885-1894 in comparison with the words found in de editions of De Gids of the same period. The words are sorted by decreasing t-score.

Amerika, Zuiden (*South*) and grondwet (*constitution*). De Nieuwe Gids has more articles on ancient Greece (Socrates, Ktesippos, Thales, Plato and Hippias) and the Amsterdam area (Amsterdam, Marken and Houten). Both write about religion: godsdienst (*religion*) and Doopsgezinden (*Mennonites*), and God in De Nieuwe Gids. The latter magazine seems to write about left-wing politics: socialisten (*socialists*), arbeiders (*workers*) and klasse (*class*). Its word list also reveals an interest in poetry: sonnet and sonetten (*sonnets*).

The data visualization enabled us to get insights in the metadata differences between the two magazines without much effort. We believe a word cloud would be a good method to examine the vocabulary differences between the magazines, as shown with the t-score comparison in this section. Unfortunately our users did not agree. An alternative to a word cloud presentation would be to offer the top of the list of surprising words in tables with additional numeric information like t-score, absolute frequency and relative frequency. It would be interesting to be able to find topic differences in the same way but this requires a topic annotation of the corpus, which is currently unavailable.

6. Concluding Remarks

We have presented interactive visualizations which can be used to explore and analyze texts and literary metadata. These involve tracking frequencies of documents containing certain words over time and examining biographical data of authors. We used the visualizations to compare two competing Dutch magazines and found some interesting results.

There are several extensions we have in mind for future work. First, as explained in section 4.1., we would like to create graphs of word counts over time rather than document counts. However, in order for this to be possible in acceptable computing time, the software architecture which we rely on, will need to be changed. Currently people are working on this extension so we are optimistic about being able to create this visualization in the coming year.

The dependency on the architecture was one of the lessons we learned from this project. Even if interesting data is available for visualization, generating a visualization may take too much time or require too many computational resources. Ideally much of the visualized data is generated by the underlying search architecture so that little computation is required at run time. A good cooperation between the visualization team and the search team is required for achieving this.

As a topic for future work, we are also interested in building a visualization which presents summaries of the content of text collections. Since our users did not like word clouds we will need to consult them on this and find a representation of textual data that both suits them and is feasible computationally. We would also like to visualize author data on a map. A candidate for this is the bubble map in combination with a parallel coordinates plot, as shown in (Theron and Wandl-Vogt, 2014), which could interactively display birth or death location together with other biographical information in one view.

Syntactic annotations are a standard feature of our corpus which we have not yet used in our visualizations. It would be interesting to visualize aggregated information about the part-of-speech of words in the neighborhood of query words, like done in OpenSonar (Reynaert et al., 2014). With a visualization like the one in the interHist interface (Lyding et al., 2014), part-of-speech classes of several neighboring words could be inspected in one view.

One of the recurring questions of our users is about finding new and abandoned words in a certain time period. A new word can be identified from its frequency graph: it was never used before a certain year and becomes popular after that year. An abandoned word behaves in the opposite way. The question: *Which new words became popular in the nineteenth century?* can be answered but it requires a lot of computation which may need to be repeated when the text corpus changes. Still this would be an interesting search option to be offered to our users. A related but even more challenging question is *What new sense of a word became popular at a certain time?*. Methods for answering this question and visualizing the results, exist (Rohrdantz et al., 2011) and exploring this approach would be an interesting topic for future study.

7. Acknowledgments

The study described in this paper was enabled by support from the Dutch organization CLaRIAH (Common Lab Research Infrastructure for the Arts and Humanities[5]). We thank Nicoline van der Sijs for applying for funding for the project and René van Stipriaan for valuable user feedback on the visualizations. The paper was reviewed by two anonymous reviewers, who we thank for useful comments.

[5] clariah.nl

8. Bibliographical References

Bostock, M. (2011). D3 Data-Driven Documents. http://d3js.org/ Retrieved 22 March 2016.

Brouwer, M., Brugman, H., Kemps-Snijders, M., Kunst, J. P., van Peet, M., Zeeman, R., and Zhang, J. (2013). Providing searchability using broker architecture on an evolving infrastructure. unpublished manuscript.

Brouwer, M., Brugman, H., Kemps-Snijders, M., Kunst, J. P., van Peet, M., and Zeeman, R. (2015). MTAS: Extending Solr and Lucene with Scalable Searchability on Annotated Text. unpublished manuscript.

Brugman, H., Reynaert, M., van der Sijs, N., van Stipriaan, R., Tjong Kim Sang, E., and van den Bosch, A. (2016). Nederlab: Towards a single portal and research environment for diachronic dutch text corpora. In *Proceedings of LREC 2016*. ELRA, Portoroz, Slovenia.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In Uri Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

Google. (2010). Google books Ngram Viewer. https://books.google.com/ngrams Retrieved 22 March 2016.

Gormley, C. and Tong, Z. (2015). *Elasticsearch: The Definite Guide*. O'Reilly Media Inc.

Komen, E. R. (2015). *An insider's guide to the Nederlab R visualization webserver*. Technical report, Radboud University Nijmegen. Version 2.5, March 16. Retrieved March 22, 2016 from http://erwinkomen.ruhosting.nl/nedlab/2015_NederlabR webservice.pdf.

Lyding, V., Nicolas, L., and Stemle, E. (2014). 'interHist' - an interactive visual interface for corpus exploration. In *Proceedings of LREC 2014*, pages 635–641. Reykjavik, Iceland.

Reynaert, M., van de Camp, M., and van Zaanen, M. (2014). OpenSoNaR: user-driven development of the SoNaR corpus interfaces. In *Proceedings of COLING 2014*, pages 124–128. Dublin, Ireland.

Rohrdantz, C., Hautli, A., Mayer, T., Butt, M., Keim, D. A., and Plank, F. (2011). Towards Tracking Semantic Change by Visual Analytics. In *Proceedings of ACL 2011*, pages 305–310. Association for Computational Linguistics, Portland, Oregon.

Theron, R. and Wandl-Vogt, E. (2014). The Run of Exploration: How to Access a Non-Standard Language Corpus Visually. In *VisLR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources*. LREC, Reykjavik, Iceland.

Tjong Kim Sang, E. (2014). Finding Syntactic Characteristics of Surinamese Dutch. Technical report, Meertens Institute, Amsterdam, The Netherlands, March.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*. Graphics Press, Chesthire, Connecticut.

Zhang, J., Brouwer, M., Brugman, H., Droes, M., Kemps-Snijders, M., Kunst, J. P., van der Sijs, N., van Stipriaan, R., Sang, E. T. K., and Zeeman, R. (2013). *Visual Analytics in a Virtual Research Environment for Humanities*. Poster presented at the International UDC Seminar: Classification & Visualization, Interfaces to Knowledge, The Hague, The Netherlands.