Course project
Petra Lehtoruusu
013304939
5.1.2019

Project: Visualising *Fältskärns berättelser* with Voyant tools

Research question

For my project I created visualizations of Z. Topelius' historical novel The Surgeon's Stories (Fältskärns berättelser) using Voyant tools.

Originally published as a feuilleton between 1851 and 1864 the novel consists of five parts (referred to as cycles by the author) and 15 stories plus a frame narrative of the old field surgeon narrating them. Althogether the novel (in the first printed editions) contained over 1400 pages and covers a period of 161 years of Finland's (and Sweden's) history narrated through the stories of the fictive Bertelsköld and Larsson families.

My aim was to use computational tools for a distant reading of the novel that would show the main themes of the entire novel as well as allow for a comparison of the main themes in the different sections and stories. I initially intended to use a library for topic modeling, but after closer examination of the options chose to use Voyant tools because it is easy to use and can be used to create multiple kinds of highly readable visualizations.

Creating and preprocessing the corpus

The starting point for my analysis was the recent critical edition published by SLS found here: https://www.sls.fi/sv/utgivning/faltskarns-berattelser. As browsing the Epub-version was quite slow I initially intended to use the PDF-publication which would have reqired quite a lot of preprocessing (removing the editors' introduction and commentaries, page numbers as well as text in headers and footers). Lucily I found that the EPB-version had a small icon for "Lästext" in the top left corner that opened a new window containing the text of one of the section I was reading. This function allowed me to copy the text of each of the 15 stories into plain text files that I could upload to Voyant as my corpus. The corpus can be viewed here: https://github.com/lehtoruu/Digital-humanities-project/tree/master/Corpus

I did not preprocess my corpus into lower case, remove punctuation, etc. because while running a few tests with the first story it seemed to me Voyant was actually preprogrammed to do this kind of processing. What my also showed me was an acute need to use stopwords, as my initial test cirrus cloud of the first story looked like this:

I found Voyant had a ready list of stopwords in Swedish and edited it by adding some terms that disturbed my analysis, most of them older orthographic forms for terms already found on the pre-existing list in their current form (af, der, hafva, hvad, öfver, hvilka, ers, äfven, voro, sjelf hela, j).

Visualisation

I chose to create several visualisations of the novel and its different parts to allow for comparison of relevant themes in the entire text, each section and each story. For this visualized the lists of most frequent terms as lists and cirrus clouds.

To detect connections between themes I created charts of collocate terms for the most frequent terms Links tool. Then I used the Trends tool to visualize the relative appearances of terms in different sections.

In addition to studying the most common terms of the novel I also explored the potential of the DreamScape visualization tool that creates maps based on what places are mentioned in a text. This seemed interesting to me as the novel has since its appearance been interpreted as a representation of Finnish and Swedish national history. I have previously noted that its cast of characters and plot are in fact interspersed with international events and therefore had a hunch that mapping the places discussed in the text and their occurrencies might make the international aspects of the plot more tangible.

I repeated the same visualizations for the entire novel text, for each section and for each story separately to see how the results compare to each other and how the themes vary along the novel. The results are visible here: https://github.com/lehtoruu/Digital-humanities-project/tree/master/Voyant%20analysis/Voyant%20results

Analysis

Compared to my reading of the novel text the lists of frequency of different terms created with the Terms tool (and visualizable as cirrus Clouds) seem to indicate themes fairly well on superficial

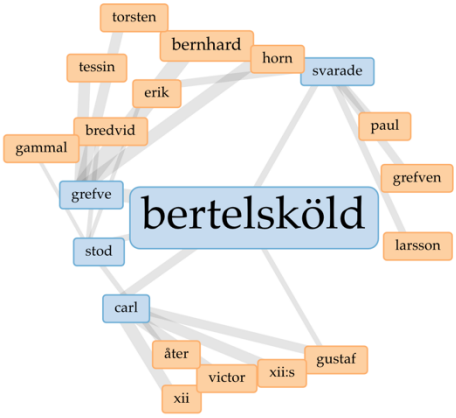
level. They for example show the relevance of Swedish kings and the Bertelsköld family in the entire novel. In the analyses of single stories the term counts show which characters are relevant. They also show significant changes in subject matter between the stories: for example the first story has among its 20 most frequent terms male first names and words related to kings and nationalities while the second story's five most frequent terms contain female names and titles as well as references to different ages that seem to point to entirely different subjects. The main link seems to be the character Regina whose name figures prominently in both stories.

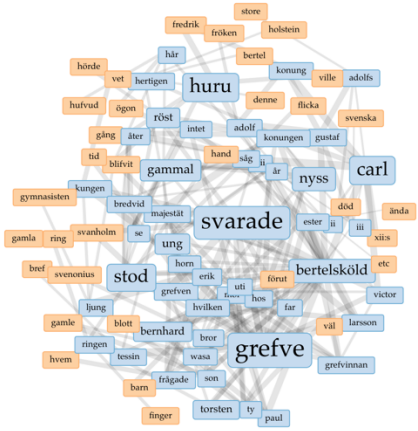
1st story		2nd story	
Term	Count	Term	Count
konungen	134	bertila	45
fröken	64	regina	37
gustaf	61	fru	36
regina	54	gamla	33
konungens	53	fröken	32
fältskärn	40	gamle	32
såg	40	märtha	29
svenska	40	larsson	28
bertel	39	meri	28
majestät	37	såg	28
adolf	36	se	26
heliga	36	väl	26
stod	34	barn	25
finnarne	33	åter	24
ryttare	32	korsholm	23
väl	32	ord	23
ögonblick	31	blott	22
hand	28	ögon	22
ord	28	uppå	22

However, the fact that Swedish words are often inflected created some problems in the analyses so that in the results for example "konungen" and "konungens" or "grefve" and "grefven" appeared as separate words. Eliminating this would require tokenizing the corpus so that the words would appear in their basic form. Also, my analyses of the different stories are not however entirely comparable, as I kept editing my list of stopwords and adding new terms to it while whenever I found it necessary. I however did not have the patience to repeat all previous analyses whenever I did this. Similarly, with the Trends tool I did not mechanically apply it to the five most prominent terms but skipped over some uninteresting ones or added others from further on in the list of frequency if I noticed an interesting term. So in many ways this project was more of a test on what must be taken into account in this kind of research than actually valid research.

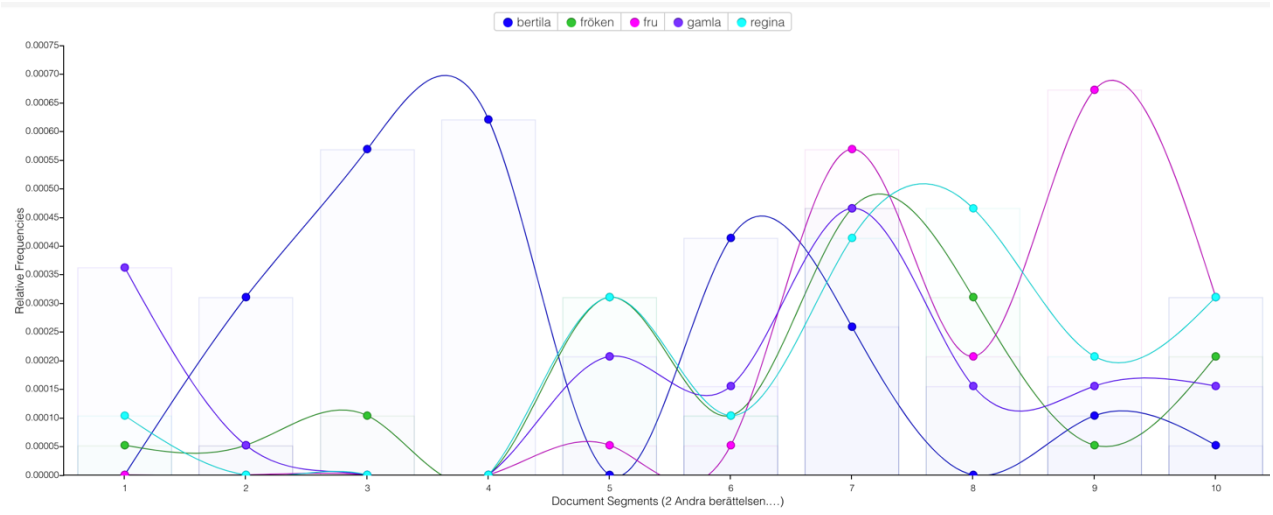
The collocate graphs created with Links tool did not seem very useful, since the tool was based on the terms appearing near each other in the text and therefore seemed to find recurring patterns like linking a character's first name or title to their last name. To be able to link different enties to

each other I attempted increasing the number of terms taken into account but this resulted in word clouds that were too large and messy to be very readable. Here are a few examples around the name Bertelsköld (and its occurrences throughout the novel). The first graph is based on observing a context of 3 words appearing in both directions from the name and the second has a context of 17 words to each direct ion.

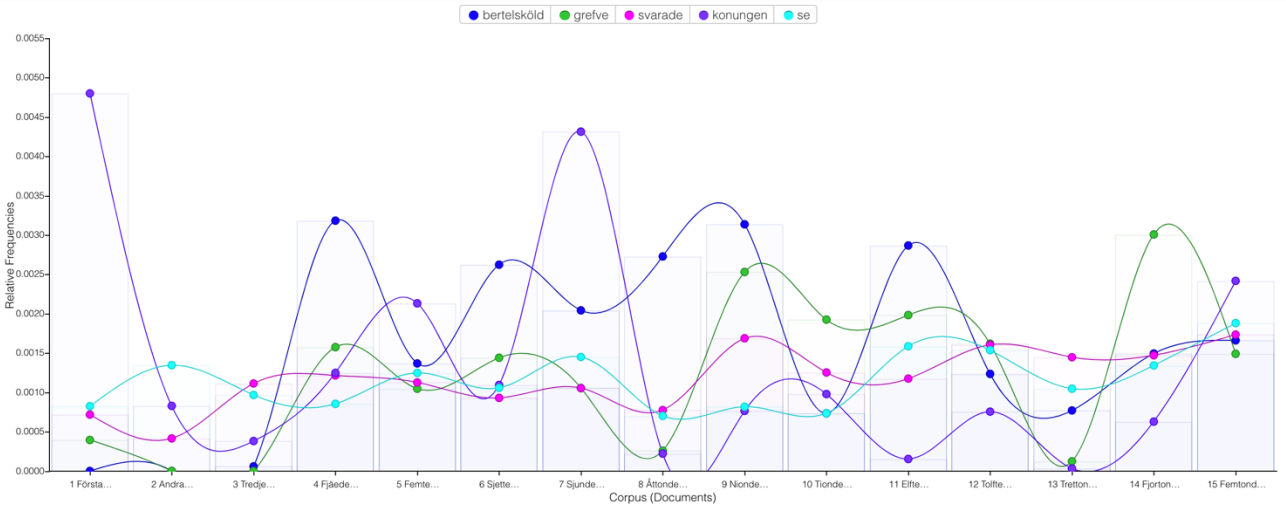




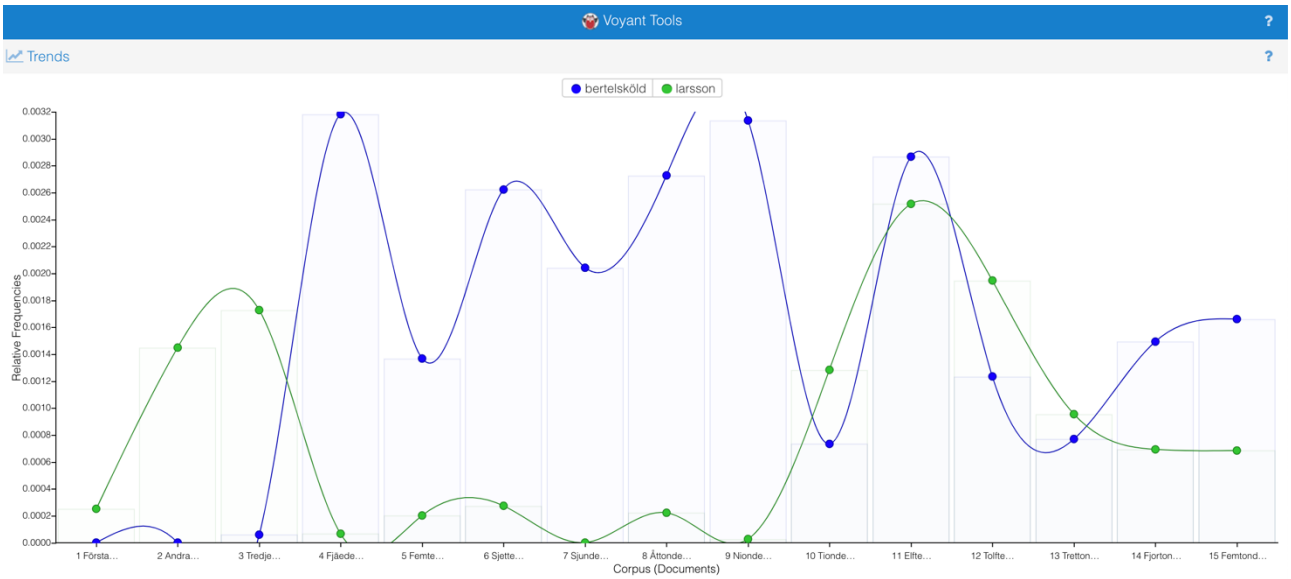
Instead I found the Trends tool to be useful in attempting to visualize the co-apppearances of different themes in the novel. It chops the text into segments and charts the appearances of chosen terms per text segment thereby and forms graphs on their relative frequencies in each segment. This seemed useful in visualizing which terms appear in in the same part of the narrative and therefore are likely to be connected. However, discovering connections requires the text to be cut into sufficiently small segments. In visualizing a single story it was for example possible to see at what parts of the story certain characters interact and at what parts they are followed separately from each other. This example of occurrencies of the most frequent terms in story 2 for example shows that the terms “fröken” and “regina” tend to appear in the same parts of the story, as do “gamla” and “fru” until Segment 8, whereas “bertila” appears in segments different from the other frequent terms. So we can suppose there is little interaction between the characters Regina and Bertila, while the terms “fröken” and “Regina” are related to same plot developments (the character is actually often referred to as “fröken Regina” or by using either term separately).



However, as the analysed segments grew larger the results became more trivial as is shown by this example of trends of occurrences of the most common terms in the different stories of the novel:



However, by carefully choosing the terms analysed even large corpuses can yield interesting results as shown by this graph of the occurrences of the names “Larsson” and “bertelsköld” in the different stories that tells us which family forms the protagonist of each story. The information however can not be taken at face value as the graph indicates a strong emphasis on the Larsson family in the beginning of the novel. This however is caused by a bias in the data: the families themselves are only founded at the end of the third story whereas the first three stories follow the adventures of their founding fathers as young men, one of them called Bertel and the other Larsson. If mentions of the character Bertel who is the actual protagonist of the first section of the novel would be included the beginning of the graph would look very different.



The Dreamscape tool seemed useful for the purpose of mapping the global extension of a narrative. However, it is based on identifying place names and calculating their frequencies of occurrence which means it must be borne in mind is not able to differentiate between a narrative

taking place in a certain location and a location simply mentioned in the narrative as a port of origin of a vessel or birth place of a character, discussed in the dialogue, etc. It also might not recognize all place names as such or to misinterpret some terms as place names: for example in the fifth story there is a chapter taking place in Åland, but the Dreamscape shows no references to either Mariehamn or Kastelholm (both mentioned in the text). On the other hand the place Falkenbergs in Western Sweden seemed to occur relatively too often, probably because the novel contained a character with the surname Falkenberg whose Swedish genitive form is "Falkenbergs". Voyant also mentions that DreamScape is a new and experimental tool and therefore may not be very reliable.

However, the most significant source of biases and errors in my project was its high reliance on manual work, especially copypasting texts and links, uploading data and processing and sorting screenshots. As a result, a single error in saving or copying the data can significantly mess up my results if I for example accidentally store data related to one story in the folder dedicated to another; forget to copy a link or even fail to copy some parts of the novel text to my corpus. While sorting my results in files for returning the project I noticed several errors and had to repeat quite a few stages. In a programming project for example errors in the code would have been instantly visible and ones the code functions, it functions with inhuman accuracy.