Henna Männistö
Computational literacy
Final project

1. Research aims and the dataset

The aim of this project is to find out what kind of differences emerge from the Finnish government programs from the past ten years, released between 2014–2013. In total this includes five programs, and all of the programs are originally in PDF format and found directly from the Finnish Government (Valtioneuvosto) website, which lists all government programs released since 1917. I have chosen to include five programs in my analysis in order to keep the data size manageable. However, two of the programs, those by Marin and Rinne, are essentially the same in content, since the difference has to do with a change of prime minister rather than the whole government.

My points of interest regarding this data lay in semantic differences, such as what kind of content words the programs use regarding domains like the environment, European Union or the economy, as well as in syntactic questions about the sentence complexity. As I could find no comparable previous research regarding the language used in government programs, I intend to go about the project in a data driven manner, that is to say I have no prior hypothesis to test but rather I intend to see what kinds of phenomena emerge from the data, a method known as unmotivated looking (Hoey & Kendrick, 2017, p. 5). I have opted to use the Voyant Tools (Sinclair & Rockwell, 2016) as the main tool for analysis and additionally used Excel as a support tool in the progress.

Regarding the methodological pipeline and the topic of syntactic complexity a somewhat similar project was executed by Frangen (2020), where she looked at the language complexity between fake news and fact checked real news. However, as her dataset consists of texts in English, which typologically differs from Finnish, the same exact parameters can't be directly used to evaluate the complexity of Finnish text. In the case of Finnish, Karlsson and Widberg (2010, p. 93–94) have defined syntactic complexity to be dependent on the following factors: sentence length, number of verbs in a sentence, number of commas in a sentence, the number of relative pronouns and subordinate conjunctions in a sentence. Out of these factors the ones that are realistically possible to analyse with the tools chosen for this project are sentence length as well as the number of relative pronouns (mikä, mitkä, mitä, minkä, mihin, missä, mistä, mille millä, miltä, joka, jota, jonka, johon, jossa, josta, jolle, jolla, jolta, jotka, joita, joiden, joihin, joissa, joista, joille, joilla, joilta) and subordinate conjunctions (että, jotta, koska, jos, ellei, jollei, mikäli, kun, kunhan, kunnes, joskin, vaikka/vaikkakin).

2. Data processing

I started by downloading the PDF files of the programs from the government website and roughly manually looked through them to get a sense of their lengths and formats. As already mentioned, the programs by Marin and Rinne can be considered to be nearly identical, and I was not able to spot major differences by eye when skimming through. However, I decided to analyze them both as separate entities in case of some subtle

differences. Secondly, another important feature to note is that the programs by Sipilä and Stubb are significantly shorter than the rest. While the page counts of the Rinne, Marin and Orpo programs exceed 200, those of Stubb and Sipilä do not even reach 100, with the total length of the Stubb program being only 10 pages (with cover and end pages included) and that of Sipilä only 74.

2.1 Voyant Tools

I input all the programs separately into Voyant tools in their original format. I also input all the five programs as a corpus for comparative analysis. Summary of the analysis summary of each program is as follows

	Stubb	Sipilä	Rinne	Marin	Orpo
Avg. words per sentence	14.9	21.4	14.7	14.8	15.8
Vocabulary density	0.641	0.392	0.343	0.343	0.291
Total word count	2 077	19 024	40 409	40 501	56 633

Table 1: Summary of the statistics of the government programs

Because for all of the documents the most common words were non-content words such as “sekä”, “on” and “ja”, next I used a list of stopwords in Finnish found on Github and implemented them through the stopword settings on Voyant Tools for each separate program file as well as the whole corpus including all the files.

I divided the topics of interest into 3 categories within which I searched for potentially interesting keywords from the corpus that included all the programs. The search terms used are roots of words to account for related compound words and word declension:

- Topic 1: International cooperation
  euroop*, nato*, kansainväli*, yk
- Topic 2: Environment
  ekologi*, ympäristö*, ilmasto*, luonnon*
- Topic 3: Economy
  talous*, velka*, sääst*, työlli*

The numbers of the programs as indicated in the figures below are as follows (note that this is not their chronological order):

1) Stubb 2) Rinne 3) Marin 4) Sipilä 5) Orpo

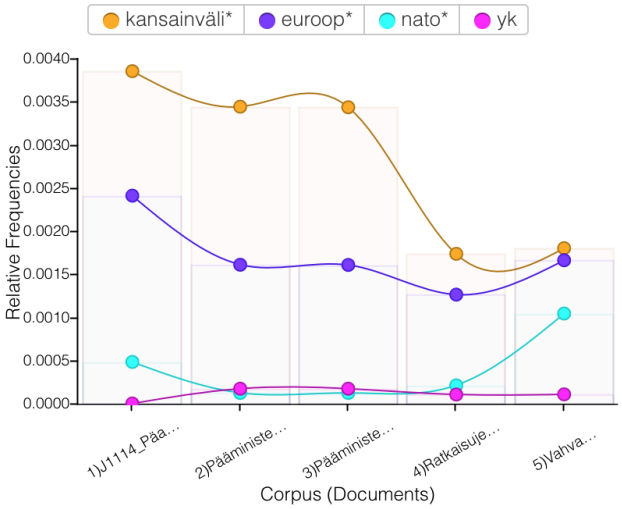

Figure 1: Relative frequencies of keywords in topic 1

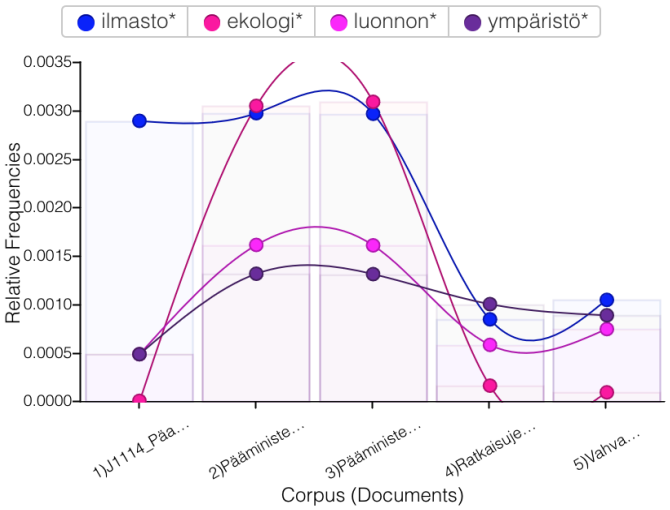

Figure 2: Relative frequencies of keywords in topic 2

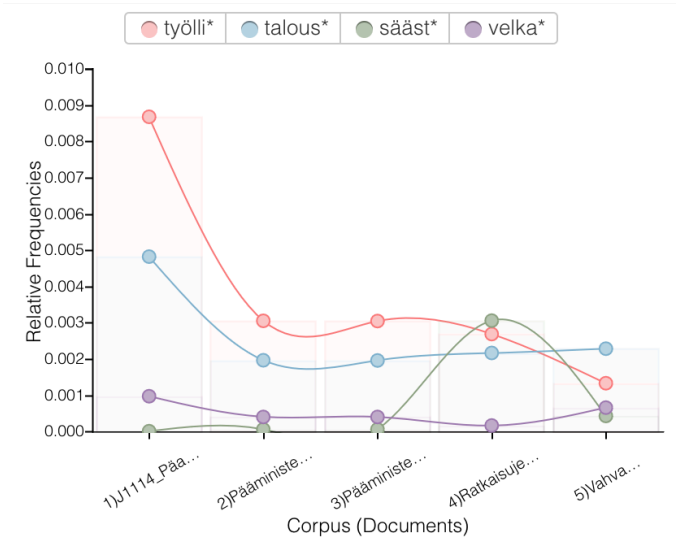

Figure 3: Relative frequencies of keywords in topic 3

2.2 Excel

After the keyword comparison by themes I looked at the most common words in the programs separately after implementing the stopword filter. From the lists given by Voyant tools I exported the lists to an Excel sheet. I also exported to a separate Excel sheet the counts of subordinate conjunctions and relative pronouns per program and then calculated how many were used per sentence on average. The formula used to calculate the average amount of subordinate conjunctions and relative pronouns per sentence was:
Average number of X per sentence = total number of X / (total words / average sentence length). The average numbers were rounded by 5 decimals.

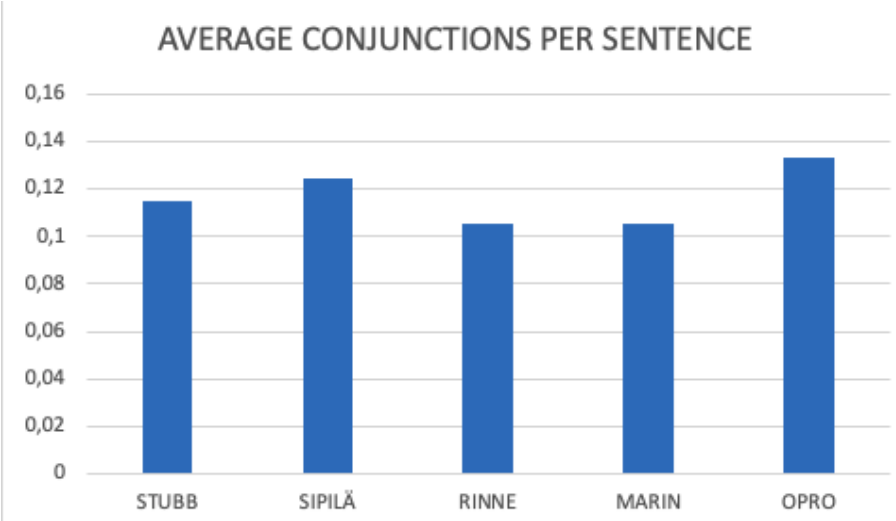

Figure 4: Comparison of the average of conjunctions per sentence in each program

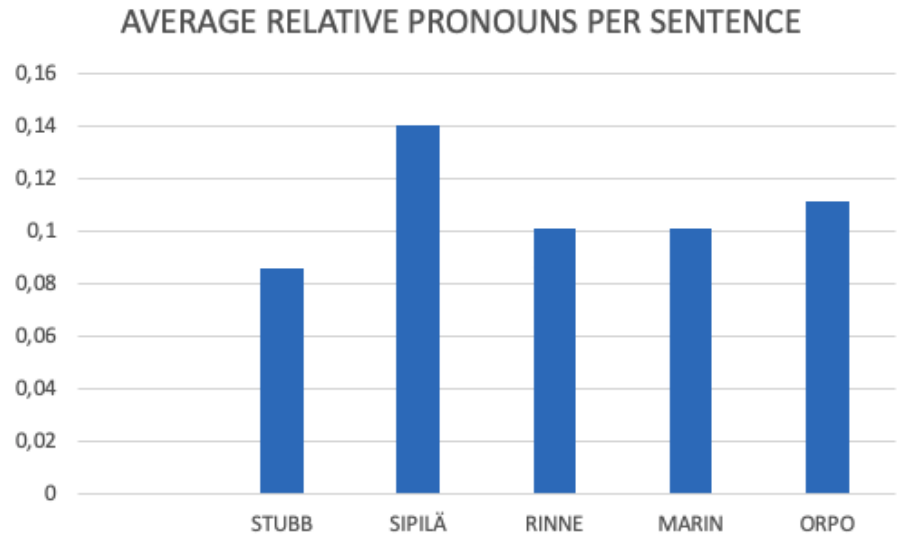

Figure 5: Comparison of the average of relative pronouns per sentence in each program

3. Results and discussion

The results show surprising differences between the programs regarding the vocabulary used and themes discussed, but the differences for the most part fall in line with the expected political commitments of the respective governments. However, on the other hand, most of the programs are quite similar in their language complexity and the calculations do not deviate much from those of political party programs (puolueohjelma) analyzed by Karlsson and Widberg (2010, p. 94). Apart from the clear differences in program lengths as seen in Table 1, differences can be spotted when looking at the list of most common words (with stop words removed) of each program. Compared to the average, the program by Sipilä seems to use many more numbers, as of the top 50 words around half are numbers. By contrast, the program by Stubb has the fewest numbers showing up in the top words list, but it also scores the highest in terms of vocabulary density. This might be due to the fact that it is by far the shortest among the programs which could thus drive up the information rate. Another interesting factor visible in Table 1 is the fact that the average sentence length of the Sipilä program is much higher than the others. Furthermore, when compared to the sentence length of the party programs in Karlsson and Widberg (2010), it seems that the sentences in the government programs are slightly longer than those of political party programs.

Regarding the topics outlined in 2.1, the results show considerable variance. As visible in figures 1–3, the topics, it seems that the Stubb program has a relatively high relative frequency of at least one of the key words in each topic. This might correlate with the high vocabulary density, since within that program more information is packaged into a shorter space. This is interesting when comparing with the Sipilä program, where nearly all keywords in the topics have a lower frequency than the rest of the programs. This could be explained by the fact that when looking at the list of most common words, among the high amount of numbers that appear at the top of the list, many of the words that do appear are connected to economic and monetary themes.

Perhaps the most difficult one of the programs to interpret in terms of the analysis of the top words and the keyword analysis by topic is the Orpo program. On the top words list many of the words that appear in the top 100 are verbs of which most are very vague verbs, such as “selvitetään”, “vahvistetaan”, “kehitetään”, “edistetään”. Furthermore, in the topics from figures 1–3 most keywords have a lower relative frequency than in the other programs, which could be explained by the fact that it is by far the longest of the programs. What is surprising about that is that the keywords in topic 3 have a lower frequency than in all the other programs, although the Orpo government has been labeled the most politically right-wing in Finland in nearly a century (Pelli, 2023). However, one clear indication of a political event visible in Figure 1 is the increase in use of the keyword “nato”.

It is clear that the scope of the project was a major limitation to the results I was able to get. It would be interesting to look at a wider set of programs from a longer timespan and thus get a clearer picture of diachronic change, since in this study the differences seem for the most part to reflect political differences and priorities among the governments. Furthermore, it turned out that using the original PDF-files as the material caused some bias in the data as well. For example, in the case of the program by Orpo, the top of every page in the program has the text “Vahva ja välittävä Suomi – Pääministeri Petteri Orpon hallituksen ohjelma

20.6.2023", which then caused all these words to rise to the top of the word list listing the top words. The lists could have also appeared cleaner and easier to read if more preprocessing had been done to remove all numbers from the text. The scope of the project also limited the number of topics chosen to look at. It is clear that there are more topics covered in a government program beyond the three chosen here, but the aim of choosing these particular topics was to illustrate the potential differences among the programs rather than uncover all of them.

As regards to the pipeline used in this project, using Voyant Tools as the main tool for analysis was made possible by the comparatively small amount of data. However, if there were more data (ie. more programs) to analyze, automating the process even more by means of a coded program would make the workflow smoother, as with the current approach there was an unexpected amount of manual steps to perform in moving the data between Excel and Voyant tools, as well as performing calculations on the retrieved data. Some of this might, however, be due to the limitations in my skill to use the application, despite the fact that my skills to use the applications developed significantly by doing the project itself.

4. References

Diaz, G. Stopwords ISO, GitHub repository, https://github.com/stopwords-iso/stopwords-fi/blob/master/stopwords-fi.txt . Accessed 16.12.2024

Frangen, L. (2020). Digital Humanities Project: Comparing Language Complexity in Fact-Checked Fake and Real News. Zenodo. https://doi.org/10.5281/zenodo.4327219

Hoey, E. M., & Kendrick, K. H. (2017). Conversation analysis. *Research methods in psycholinguistics and the neurobiology of language: A practical guide*, 151-173.

Karlsson, F., & Widberg, M. (2010). Puolueohjelmien kieliopillinen kompleksisuus. *Sananjalka*, *52*(1), 89-103.

Pelli, P. (27.4.2023). Tutkijat: Orpon nelikko olisi oikeistolaisin hallitus sitten 1930-luvun. https://www.hs.fi/politiikka/art-2000009546586.html . Accessed 20.12.2024

Sinclair, Stéfan and Geoffrey Rockwell, 2016. *Voyant Tools*. Web. http://voyant-tools.org/. Accessed 12th December 2024.