Università della Svizzera italiana

Text Analysis and Spatial Data 2022/2023

Final Project Report

Text and Spatial Analysis of Swiss Parliamentary Speeches

Abstract:
This report examines the Swiss parliamentary speeches across various dimensions. Through text and spatial analysis techniques, further insights on languages, cantons, congresspeople, factions and topics covered have been extracted. A variety of datasets have been employed throughout this project; the most important between all of these is the SwissParliamentarySpeeches dataset, this file stores all speeches given in the Swiss Parliament since 1999. This data is contained in the swissparl.csv. Please, note that all the required datasets are inside the folder LucaPernigoTASD-Datasets, which has been shared on OneDrive.

Author: Luca Pernigo

1 Introduction

Before delving into the analysis, some basic notions regarding the swiss political system are provided.
The Federal Assembly, also known as the Swiss parliament, holds the legislative power in Switzerland. The Swiss parliament is bicameral. Its lower house is the National Council, which is composed of 200 seats. The Council of States is its upper house and comprises 46 members. Elections for both the two houses are held once every four years. Seats in the National Council are distributed in proportion to population according to the largest remainder method; this does not hold for cantons that are entitled to just one seat, i.e. those cantons that are really small in terms of population. Since the considered dataset includes data from 1999 until 2019, the legislatures from the 46th to the 50th will be considered in this report. Additionally, information regarding the structure of the swissparl.csv dataset is shown in Table 1

Variable	Description
s_transcript_raw	Raw transcript of the speech as extracted from the website of the Swiss Parliament.
s_transcript	Speech without unspoken content (brackets, page references)
s_speaker_name	Name of the speaker
s_speaker_council	Council of the speaker
s_speaker_canton	Canton of the speaker
s_speaker_faction	Faction affiliation of the speaker
s_speaker_president	Dummy variable, whether speaker is also the president of the council
s_link	Link to the original transcript on the website of the Swiss Parliament
s_link_video	Link to video of the speech on the website of the Swiss Parliament
s_language_cld2	Main language detected by cld2
s_length_words	Length of the non-raw transcript in words
b_number	ID of the discussed item of business
b_title_german	Title of the discussed item of business in german
b_title_french	Title of the discussed item of business in french
b_link	Link to the item of business on the website of the Swiss Parliament
d_session	Session of the debate
d_date	Date of the debate (YYYY-MM-DD)
d_time	Start time of the debate (HHhMM)
d_council	Council of debate
d_preliminary	Dummy variable, Whether transcripts are preliminary or not . if equal to 1 the transcripts of the debate are preliminary.
scraping_timestamp	Time of scraping

Table 1. Swiss parliamentary speeches dataset structure

Other dataset that have been used in the project

- name_gender.csv: dataset containing a vast amount of first names and their corresponding gender.
- map: folder containing world's geometry.
- census: dataset containing Switzerland's population census.
- je-d-17.02.02.02.01.04: dataset containing info about gender distribution between Swiss parliamentary members.
- ra-fraktionen-statistik-1912-d: dataset containing distribution of parliamentary seats per parties for all legislatures after year 1911.

2 Parliamentary Speeches Language Distribution

Article 4 of the Swiss Federal Constitution establishes that German, French, Italian and Romansh are the national languages of Switzerland. Thus, before carrying out any textual or spatial analysis it is worth exploring a bit the dataset along the language dimension. This section deals with the usage distribution of those languages within the Swiss Federal Assembly. Because of Romansh is in practice never or almost never used, this analysis concentrates on just German, French and Italian.
Note that the info regarding the language for each transcript was already available in the dataset, inside the **s_language_cld2** column. In order to retrieve such information, the Compact Language Detector 2 model(CLD2) has been applied to the **s_transcript** column, i.e. the column of speeches.
Following, the percentages of languages used in all speeches between 1999 and 2019 are visualized in Figure 1.

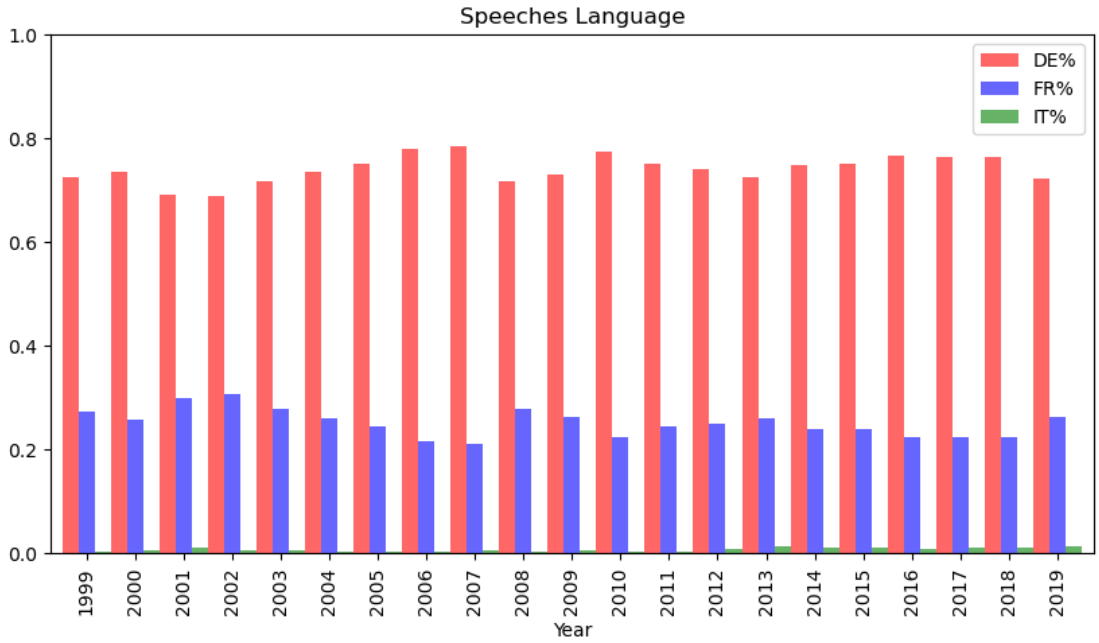

Figure 1. Language of speeches

By looking at this chart, it is clear that the most used language in the Federal Assembly is German, followed then by French. Usage of Italian is practically zero. Even though this may seem reasonable, take into account that circa 8.2% of Swiss citizens have Italian as their first tongue. Nevertheless, note that over this twenty years period, usage of Italian is relatively increased; consider that in 1999 its percentage of use was 0.001978 while in 2019 it was 0.014726, that is almost a tenfold increase. Finally, it can be concluded that practically speaking, the languages of the Swiss parliament are German and French.

3 Parliamentary Speeches by Canton

Next in order to inspect how speeches are distributed according to the speakers' cantons, an heatmap map for each of the legislatures between 1999 and 2019 has been created. Since the map dataset contained the geometry of the world, the first step was to filter everything out apart from Switzerland. After that, the geometry column was transformed to the relevant coordinate reference system(crs) for Switzerland. This was done by setting the epsg number equal to 21781, i.e. in accordance with the epsg.io number for Switzerland. Given that in Switzerland seats for the National Assembly are assigned to each canton based on its number of inhabitants, it would not make any sense to compare the number of speeches per canton in absolute terms. It is more interesting to inspect how each canton made use of the number of seats it was assigned relatively to its population. Therefore, the ratio between number of speeches and canton's population has been considered in order to analyze which were the most represented cantons in relative terms. Given that Switzerland carries out a population census every year, the census of year 2000 has been used for the 46th, 47th and 48th legislatures while the census of year 2010 has been used for the 49th and 50th legislatures. The resulting images are reported in Figure 2

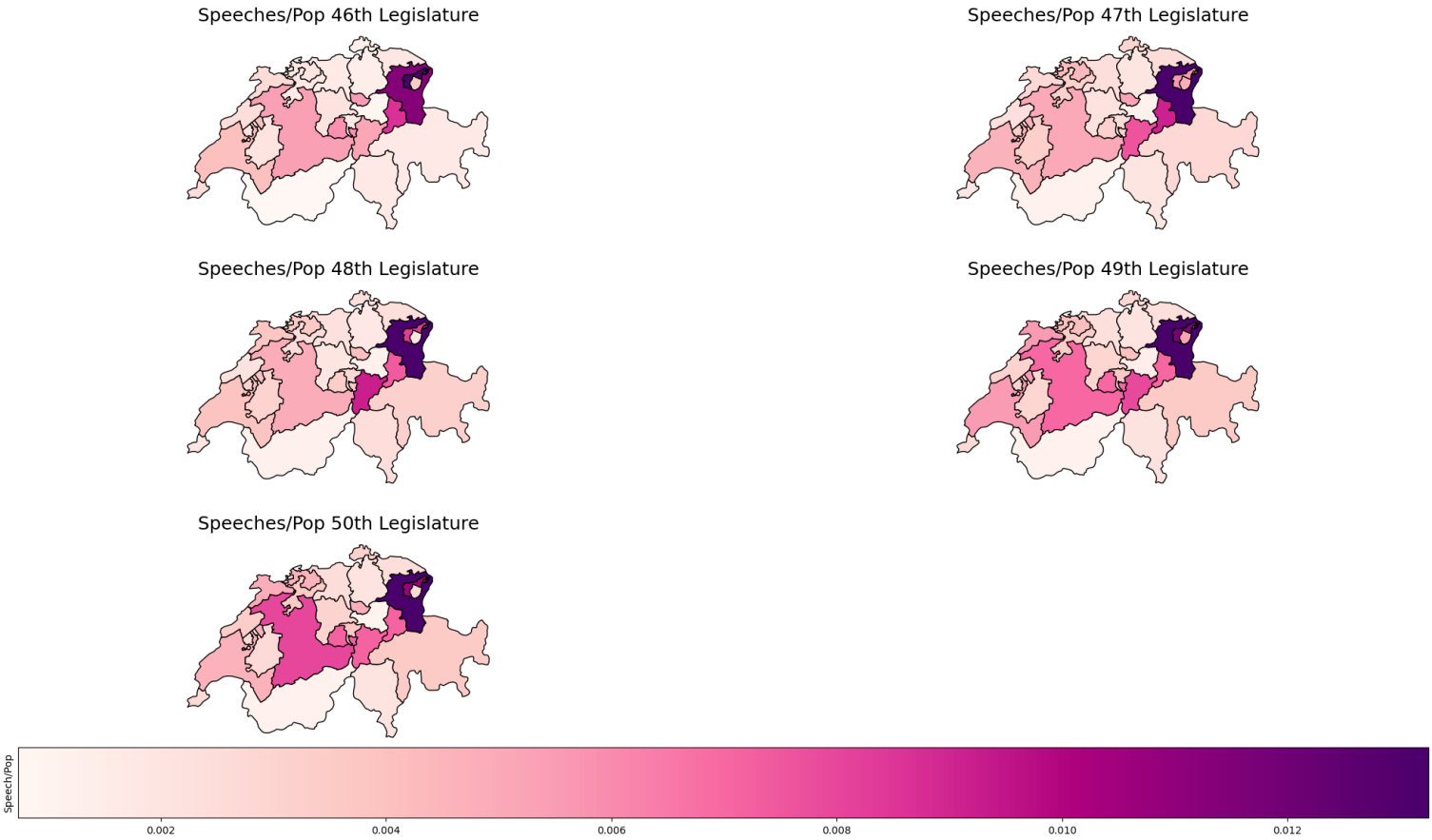

Figure 2. Heatmap map Speeches by Canton

Considering leglislatures from the 46th until the 50th, we can see that the cantons with the most number of parliament members who spoke up relative to the number of assigned seats by law were: San Gallen, Glarus, Uri and Bern. This could be interpreted as an indicator that inhabitants of these cantons have been represented more in relative terms over these twenty years; Moreover, note that canton San Gallen has been consistently the best in terms of this ratio indicator over these five legislatures.

4 Parliamentary Speeches by Gender

This subsection deals with analysing the distribution of gender within the Swiss Federal assembly. Since the sex of parliament members was not encoded in the original dataset, a gender classifier was trained and then applied to our dataset. The data that has been feed to the model is available on Kaggle at the following link name_gender dataset. Before proceeding, there are a couple of things to note. Firstly, we are only interested in the name of the congressperson; thus the **s_speaker_name** is broken on the space character and the last element of the resulting list is kept, in this way we have a column with just the name of each speaker. Secondly, before starting the training, it was necessary to change some labels of the name_gender dataset; this because names in this dataset were labelled according to the English speaking naming convention. To give you an example, in the original dataset "Patrice" is labelled as a female name, however in the french speaking world the name "Patrice" is considered masculine. Thus, before proceeding with the training of the model it was necessary to check the correctness of the labels and update the ones that did not match the Swiss naming convention.
Once done, the dataset was divided between training and testing, saved in spacy format and then trained in accordance with Spacy3 guidelines.
Hereafter, in Table 2, it is reported the final performance of the name-gender trained model.

E	#	LOSS	TEXTCAT	CATS_SCORE	CATS_MICRO_P	CATS_MICRO_R	SCORE
0	0		0.25	50.06	60.59	60.59	0.57
0	200		36.50	76.92	78.73	78.73	0.77
0	400		27.81	79.41	80.71	80.71	0.79
1	600		26.02	79.90	81.56	81.56	0.80
1	800		22.57	80.30	81.86	81.86	0.81
2	1000		20.02	80.19	81.69	81.69	0.80
3	1200		17.04	80.57	82.02	82.02	0.81
4	1400		12.41	80.31	81.92	81.92	0.81
5	1600		8.83	79.33	80.97	80.97	0.80
6	1800		6.36	79.73	81.37	81.37	0.80
8	2000		4.61	79.24	80.98	80.98	0.80
10	2200		3.53	79.34	81.09	81.09	0.80
13	2400		2.67	79.64	81.33	81.33	0.80
15	2600		2.21	79.51	81.27	81.27	0.80
18	2800		1.91	79.73	81.42	81.42	0.80

Table 2. name_gender Model Performance

After that, the latest trained model(model-last) has been called in the notebook and used to classify gender of congresspeople based on their first name.
With this additional information, it is now possible to drill down on the gender dimension of the parliamentary speeches. To get started, let us explore the gender distribution of speeches within the Swiss Parliament by considering Figure 3

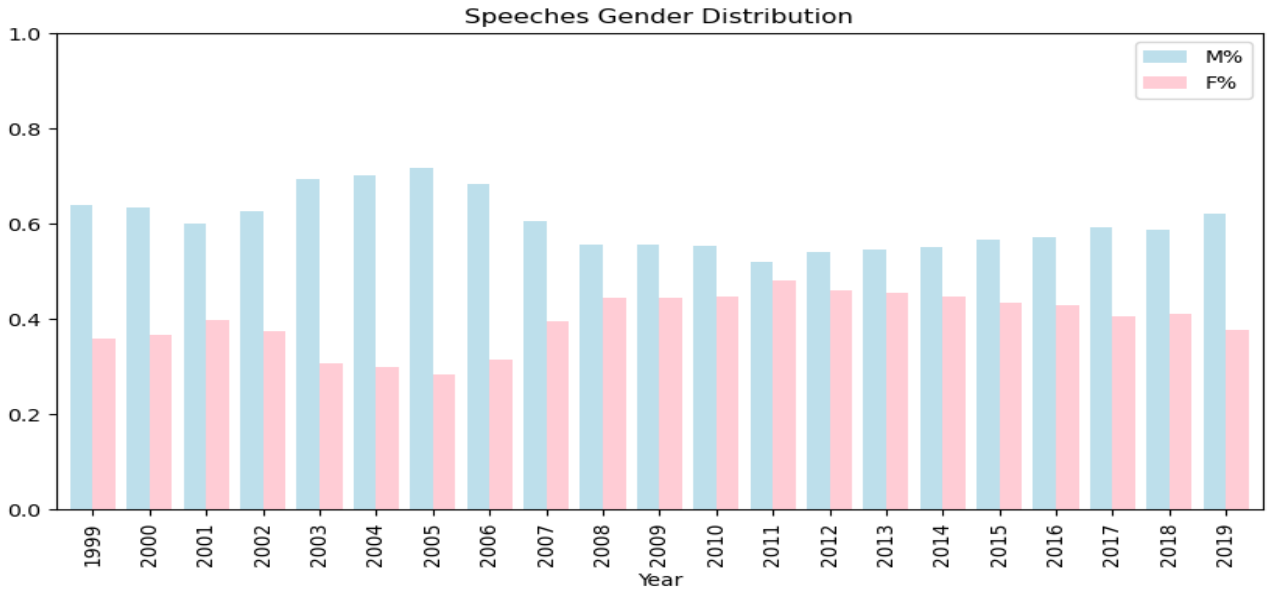

Figure 3. Speeches gender distribution

More or less the percentages of gender distribution by parliamentary speeches remained pretty stable over these years, with men accounting for circa 65% of congress speeches while women the other 35%.
Furthermore, by looking at Figure 3 we can get a feeling on how seats were distributed over the legislatures between 1999 and 2019. These numbers are not bad considering that the average of women in worlds parliaments is 26%; nonetheless, even though the proportion of seats held by women in national parliaments is a relevant theme, it is more interesting to consider whether men or women made out the most of the chance of representing Swiss citizens. In order to measure whether men or women participated most in parliamentary sessions, the ratio between the number of speeches and the number of seats per gender has been considered. Those results are visualized in Figure 4

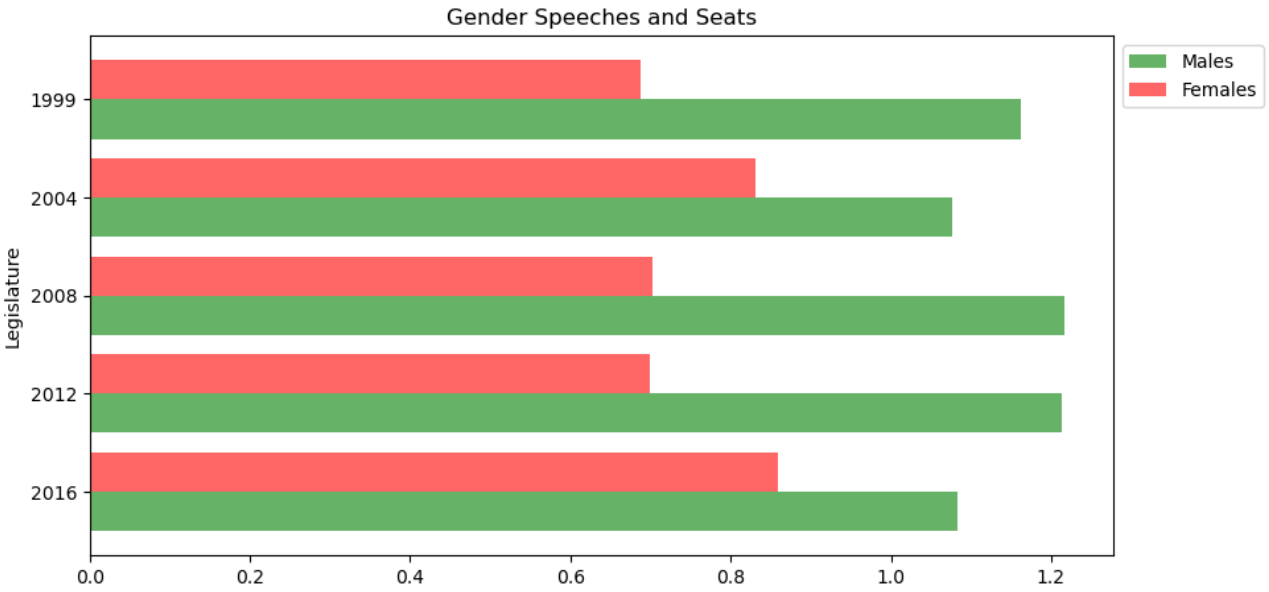

Figure 4. Gender Speeches/Seats

Thus, it can be concluded that male congressmembers participated relatively more to assemblies in terms of speaking with respect to their female counterparts. Note, that this result could be due to several different heterogeneous causes, one such cause could be the fact that parties in Switzerland are mostly led by men.

5 Parliamentary Speeches by Faction

We are now concerned with analysing the speeches' dataset with respect to the faction dimension.
Since in the starting dataset parties are identified by a code, it was necessary to map these identifiers to the corresponding factions' names. Therefore a mapping dictionary based on data available at the Swiss Parliament's web service was retrieved by using the request library. Once the data had been requested, it was then cleaned and formatted in order to match the format adopted by the Swiss parliamentary speeches dataset. Next, the string column containing parties' codes was mapped into a new column with the parties' names. After having carried out this text processing step, the data for the legislatures from the 46th to the 50th were grouped by faction and then analysed. In Figure 5 are reported the ratios between the number of speeches given by each faction and their corresponding number of seats within the parliament.

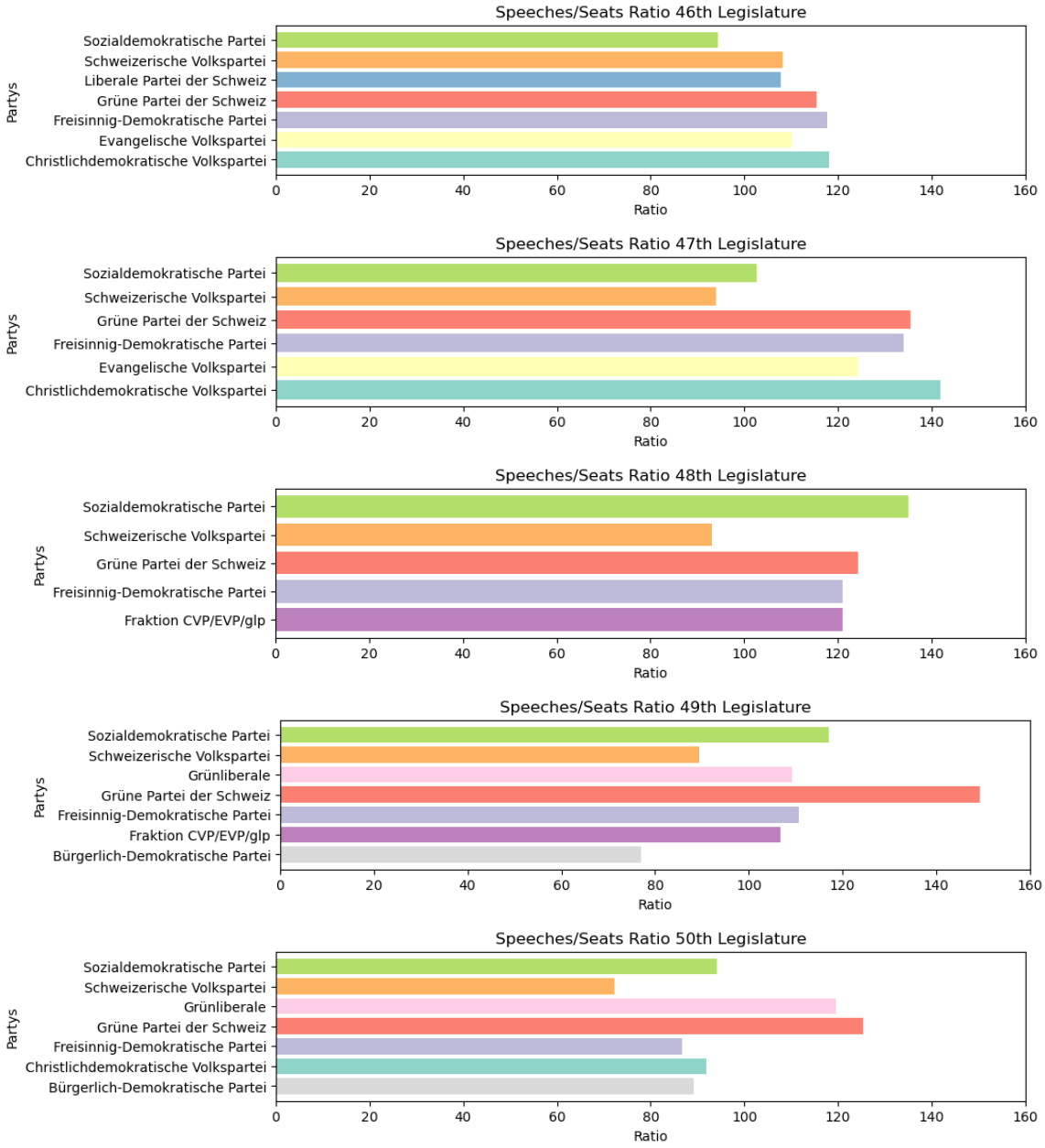

Figure 5. Faction Speeches/Seats

From the above chart, it can be concluded that the faction that participated relatively the most, over the considered twenty years, was the Grüne Partei der Schweiz . While, on the other hand the party that participated relatively less is the Schweizerisch Volkspartei. These two results are really consistent over the various legislatures considered in this report. Finally, note that even though its relatively low participation over these twenty years, the Schweizerisch Volkspartei is still the main party in Switzerland.

6 Topic Analysis

Here, the German speeches of the dataset were analysed by using the Latent Dirichlet Allocation algorithm(LDA). Before applying this algorithm, text had to be preprocessed. Note that in this section and the section after this, the analysis is carried out only for the German speeches, thus the spacy nlp for German has been used. To get started, the de_core_news_md pipeline, made available by spacy was loaded. Then, speeches were processed by tokenizing, removing stopwords tokens, removing undesired tags and lemmatizing.
Next, it was time to create the two inputs that the LDA topic models takes, i.e. the dictionary and the corpus. In the process of creating the vocabulary, low-frequency and high-frequency tokens have been filtered out. Both those tokens that appear in more than 40% of the total text and those that appear in less than 10 documents have been removed. Furthermore, within the remaining tokens, only the top 1000 most frequent are kept. That is the size of the dictionary was limited to max 1000 tokens.
Then the corpus was generated by combining the dictionary of above with the function doc2bow.
After that the LDA model was applied to the speeches data by choosing to iterate over the corpus 50 times, by specifying the number of topics to be 5, by dividing the work between 4 workers and by setting the pass parameter to 10. The model's output is depicted in Figure 6.

```
[(0,
  '0.060*"franken" + 0.037*"million" + 0.034*"prozent" + 0.017*"milliarde" + 0.013*"bund" + 0.011*"kosten" + 0.010*
"projekt" + 0.010*"armee" + 0.009*"geld" + 0.008*"bereich"'),
 (1,
  '0.052*"schweiz" + 0.020*"land" + 0.015*"schweizer" + 0.013*"international" + 0.010*"staat" + 0.008*"bereich" + 0
.008*"eu" + 0.008*"unternehmen" + 0.008*"schweizerisch" + 0.008*"ausland"'),
 (2,
  '0.028*"kanton" + 0.018*"sagen" + 0.017*"initiative" + 0.012*"herr" + 0.011*"frage" + 0.009*"frau" + 0.009*"probl
em" + 0.009*"geben" + 0.007*"kind" + 0.007*"sehen"'),
 (3,
  '0.041*"motion" + 0.027*"kommission" + 0.019*"frage" + 0.014*"bericht" + 0.013*"parlament" + 0.013*"herr" + 0.012
*"--" + 0.009*"postulat" + 0.009*"annehmen" + 0.008*"beantragen"'),
 (4,
  '0.041*"artikel" + 0.031*"kommission" + 0.030*"antrag" + 0.028*"minderheit" + 0.023*"mehrheit" + 0.021*"absatz" +
0.015*"ständerat" + 0.014*"vorlage" + 0.014*"gesetz" + 0.012*"nationalrat"')]
```

Figure 6. Top10 words for each topic

In order to better interpret these values, each of the resulting 5 topics have been assigned a label based on their most relevant terms. Table 3 shows the distribution of these topics over all speeches.

Topic	Label	Count
0	Investments and Finances	73809
1	National and International Affairs	26256
2	Initiatives for Men/Women/Children	14224
3	Legislative Matters	3318
4	Other Debates	1481

Table 3. Speeches grouped by topic

Finally, some visualizations of the analysis' results are provided in Figure 7 and Figure 8

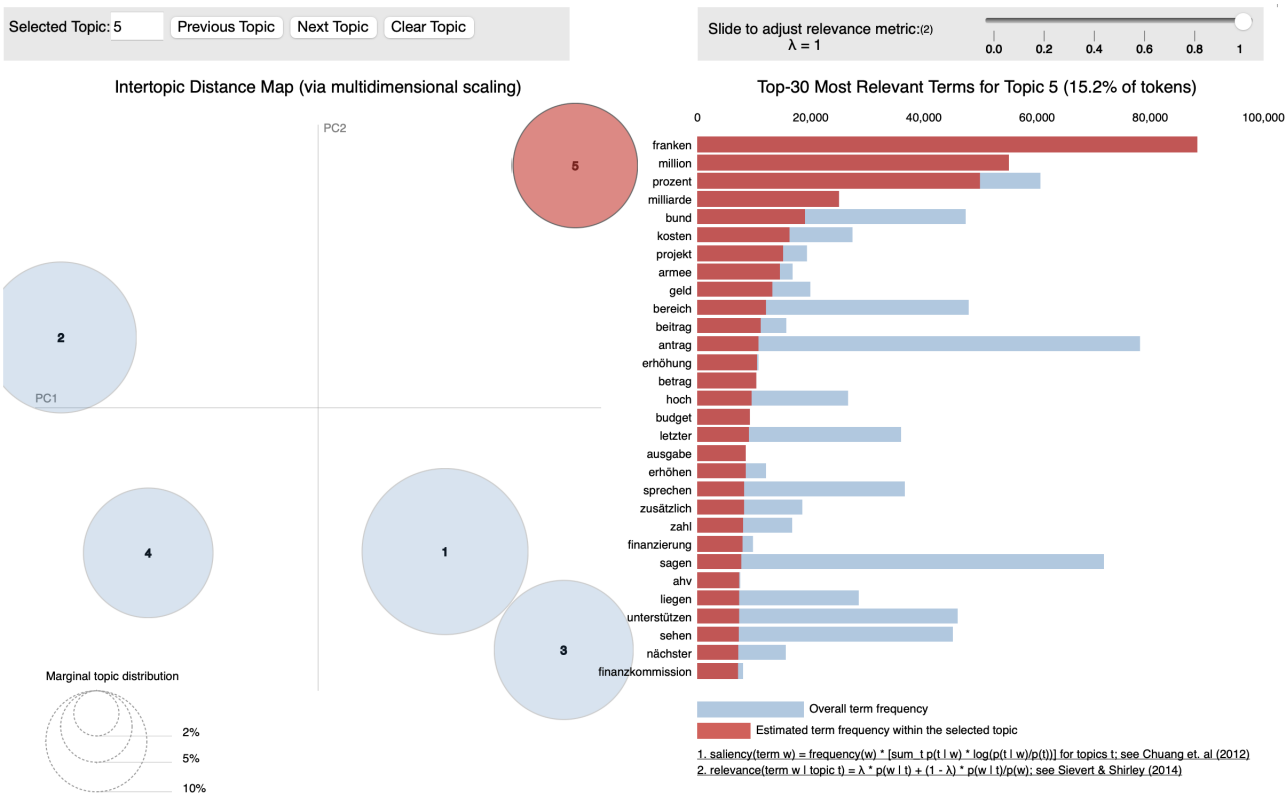

Figure 7. Topics Visualization

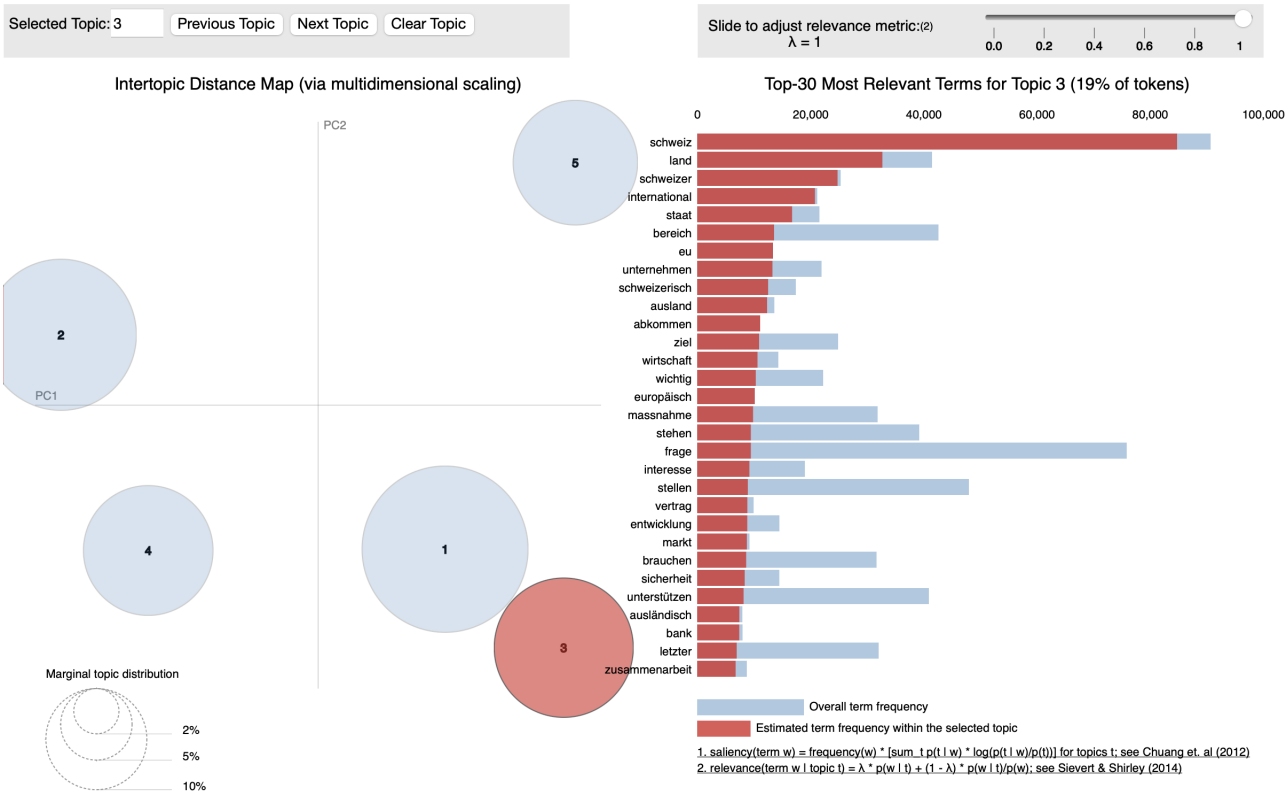

Figure 8. Topics Visualization

7 President vs Parlamentary Members

Finally, we are concerned with comparing presidents' speeches and parliament members' speeches through their wordclouds.
Note that from the previous section, speeches had been already tokenized, cleaned from stop words and lemmatized. Thus, data was already in the right format for creating the corresponding wordclouds.
Following, in Figure 9 are reported the corresponding word clouds for the Presidents' and Congresspeople's speeches.

Figure 9. WordClouds

By comparing these two pictures, note that parliamentary members like to use frequently different verbs e.g. 'sagen', 'sprechen', 'sehen', 'führen', etc. Moreover, congresspeople use specific words and frequently use terms like 'Millionen', 'Milliarden', 'Franken'. On the other hand presidents stay really general and use mostly political jargon 'Abstimmung', 'Beschluss', 'Mehrheit', 'Antrag', etc. Conversely parliamentary members use less frequently political terms compared to presidents; the only political terms they frequently use are 'Artikel' and 'Absatz'.
Finally, notice that the word "Schweiz" is more frequent/bigger in congresspeople speeches rather than in president speeches. All these differences, are likely to be because of the particular duties presidents have to fulfill.

8 Conclusion

The goal of this report, was to explore the Swiss parliament's dataset through various techniques covered during the course, both from the spatial and the textual realms. The first insight that has been extracted is that canton SanGallen was the canton that represented relatively more its citizens over the twenty years considered. Another result is that male congresspeople gave relatively more speeches with respect to female parliamentary members. The Grüne Partei der Schweiz was the party that consistently participated the most to congress speeches. Furthermore, majority of parliamentary speeches can be categorized in one of the following topics: Investments&Finances, National&International affairs, Initiatives oriented to the Swiss population or Legislative matters.
Finally, note that all the required code is inside the LucaPernigoTextAnalysisProject.ipynb file; this makes this type of analysis highly reproducible, i.e. it is easy to carry out the same analysis for other parliaments provided their data is available. Hence, a comparison between parliaments(national, cantonal, regional) of different countries could be then easily addressed; e.g. it might be interesting to carry out this type of analysis for each state member of the European Union.