Adding political orientation metadata to ParlaMint corpora

Katja Meden
Dept. of Knowledge Technologies,
Jožef Stefan International Postgraduate School,
Jožef Stefan Institute, Slovenia
katja.meden@ijs.si

Jure Skubic
Institute of Contemporary History,
Ljubljana, Slovenia
jure.skubic@inz.si

Tomaž Erjavec
Department of Knowledge Technologies,
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Abstract

Parliamentary debates are an important source for political discourse research as well as research in other disciplines. The ParlaMint project aims to create comparable corpora of parliamentary debates which, through unified encoding, provide a comprehensible resource to support such research. Within these corpora, speeches are attributed to speakers, and speaker metadata, including temporal affiliations with different organizations such as parliamentary groups and political parties. This paper discusses the addition of metadata on the political orientation of parties and parliamentary groups to the ParlaMint corpora. The paper explains our two sources for this information, namely the Chapel Hill Expert Survey Dataset and Wikipedia, the process of data collection and its subsequent encoding in the corpora. Furthermore, the paper presents an analysis of the extent of the added metadata, along with an example of exploratory data analysis. It also outlines the distribution of utterances across political orientation categories within ParlaMint, offering a comprehensive overview of the diverse perspectives and ideologies within the corpora. The inclusion of this supplementary metadata could prove valuable for parliamentary data research, while the methodology developed could be used to add further metadata to the ParlaMint corpora.

1 Introduction

Parliaments are of interest to the humanities and social sciences as they shape legislation that affect people's daily lives and are a source of power for MPs and other politicians (Bischof & Ilie, 2018). Parliamentary speeches and parliamentary data are of great importance for analysis at the (inter)national level and are an interesting topic for various research projects. In addition to transcripts of parliamentary debates, metadata (such as age, gender, party affiliation, political orientation, political role, etc.) are crucial for the study of parliamentary discourse, as they provide useful additional information that can be used in parliamentary discourse research and provide even more relevant and reliable research results.

The ParlaMint[1] projects, funded by CLARIN, aimed to create comparable and uniformly encoded corpora of speeches in European parliaments and make them openly accessible. In ParlaMint I (2020-2021), corpora for 17 European parliaments were created, made available, and used in research and education (Erjavec et al., 2022). The project continued as ParlaMint II (2022-2023), providing 12 new corpora, adding newer transcripts, improving the annotation schema and validation, machine-translating the corpora into English, and expanding the corpus metadata.

The additional metadata added to the corpora consists of information on whether and when a speaker is or was a minister and the political orientation of the parliamentary group or political party to which the speaker belongs. Both of these additions have been suggested by researchers (cf. Fišer and Pahor

[1] https://www.clarin.eu/parlamint

De Maiti, 2021) who had experience in using ParlaMint corpora so that analyses could take these further variables into account. But while the information on who is a minister is an objective and verifiable fact that can be easily found, political orientation is a much more controversial piece of information.

2 Political orientation

Political orientations (or political positions) are an interesting research concept in the social sciences, understood as a set of ethical ideals, principles, or doctrines of a social movement, institution, class, or large group that explain how society should function and provide a political and cultural blueprint for a particular social order (Blattberg, 2001). They are concerned with the allocation and use of (political) power and are usually pursued by political parties.

Political orientation can refer to any number of dimensions but is most often characterized and classified on a political left-to-right spectrum, usually represented with geometric axes corresponding to independent political dimensions (Heywood, 2021). The left-to-right (LR) dimension is one of the most common dimensions used as a measure of social, political, and economic stance. Originally, the terms ”left” and ”right” were used to describe the nature and ideological beliefs of political parties: ”left” as the ”parties of movement,” which are radical, progressive, and liberal, and ”right” as the ”parties of order,” which are conservative, traditional, and authoritarian (Knapp & Wright, 2006), and such classification has, although in various forms, been retained until today. Left-right conceptualization is often considered controversial not only in terms of being defined as too simplistic and insufficiently representative to describe variations in political beliefs but also in terms of dimensionality. Most commonly LR divide is understood as unidimensional (structured by socio-economic issues) whereas some authors opt for multidimensionality where despite the importance of the socio-economic content, the left-right divide also correlates with other, non-economic issues (such as religious or ”new politics” issues) (Freire, 2015). Despite said controversies, left-right conceptualization is still the most common way to describe the ideological position of political parties and their members. The division into ”left” and ”right” has formed a categorization of ideologies, a tool for classifying political orientation, a communication code, and an instrument for guiding voters in interpreting decisions and political phenomena (Freire, 2015).

The left-right characterization of political parties plays a crucial role in theorizing about many different aspects of democratic processes (Gabel & Huber, 2000), and sociology and political science have adopted and used it despite various scholarly reservations. Some disciplines, such as history, however, often refrain from using the left-right political spectrum to characterize the ideological beliefs of political parties.

Data on political positions are often collected by conducting expert surveys, analyzing the positions of party supporters in mass surveys, or analyzing party manifestos. In expert surveys, experts provide estimates of the left-right position of parties by ranking them on a predetermined political position scale. According to Gabel and Huber, 2000, such surveys are useful but have several limitations, the most common being irregular implementation. Analysis of partisan orientation in mass surveys (Eurobarometer, World Values Survey, etc.) is more common but often provides incomplete data because they are available only for a limited number of countries. In recent years, scholars have attempted to overcome these problems by extracting party positions from party manifestos. Several studies (Gabel & Huber, 2000; Heywood, 2021) conclude that such data are very useful because they provide comparable means of assessing party positions over a long period of time in different countries. Data from party manifestos are also consistent with parties' self-positioning on the left-right spectrum and provide useful insights into how parties view themselves in terms of their political ideology.

Most work in NLP attempts to determine political orientation directly from texts (whether from political tweets (Cohen & Ruths, 2021) or parliamentary debates (Yan et al., 2017)) and thus focuses on individual speeches. Unlike related work, we have instead focused on providing information about the political orientation of a political party rather than speeches and thus took the political orientation of a speech to follow from membership in a particular party to which the speaker belonged at the time of their speech. However, as mentioned earlier, the addition of metadata labels with information about the political orientation of individual political parties collected from a combination of sources can add value to the already extensive corpora and facilitate future research.

3 Data sources

The information on the political orientation of political parties contained in the ParlaMint corpora was gathered from three sources:

1. the Chapel Hill Expert Survey Europe (CHES Europe) (Jolly et al., 2022)[2];
2. Wikipedia entries on political parties; and
3. the corpus compilers' knowledge of political parties and their orientations.

We discuss each one in turn.

Chapel Hill Expert Survey: The CHES datasets contain expert data with built-in contextual and domain knowledge. They contain data on parliamentary political parties from countries, primarily from the EU, their attitudes toward European integration and specific EU policies, and on more specific topics such as corruption and anti-Islam rhetoric. We used two CSV files provided by CHES, namely the 1999-2019 trend file[3], which gives the values of the variables according to the covered years, and CHES 2019[4], which adds data for Norway, Iceland, and Turkey, as these were not covered in the CHES 1999-2019 trend file. This also means that these three corpora do not contain diachronic data.

The union of both CHES files provides 85 distinct variables on a given (political) position for each party and year covered, with most having a real value on the scale from 0 to 10, e.g. the variable `lrgen` measures the party's position in relation to its overall ideological stance on a scale from 0 (extreme left) to 10 (extreme right), with 5 representing the centre position. This wealth of data could be of great value to political scientists basing their research on the ParlaMint corpora. However, the CHES information also has drawbacks which can be seen especially in its coverage:

- CHES does not cover all ParlaMint corpora, in particular Bosnia, Serbia and Ukraine, as they are not part of the EU (candidate countries), nor Catalonia and Galicia, as they are not countries but autonomous regions;
- Many political parties included in ParlaMint could not be identified in the CHES dataset: of the 576 political parties belonging to the countries covered by CHES and that are included in ParlaMint, only 237 (41%) could be matched with a CHES party identifier;
- Even for the parties that are identified, CHES only covers the period to 2019, while ParlaMint extends to 2022; furthermore, not all variables are covered for all years, nor do the two input files share all the variables.

Wikipedia: The second source and type of data included is Wikipedia, in particular the data on the left-right spectrum of political orientation. This data was gathered by manually searching for the political parties' Wikipedia pages, which typically list their political orientation in the infobox of the Wikipedia article, although, for some, a more detailed examination of the Wikipedia article was required. We based our research on the English versions of the Wikipedia pages. When we could not find relevant information on the English page, we searched and translated the Wikipedia pages in the native language of the party's country. However, if there was no Wikipedia article for a particular political party or the political orientation information was not available there (in English or native language), we checked other sources (e.g. the websites of national parliaments) and extracted the information from there, also preserving the URL. It should be noted, however, that such cases were rare. Wikipedia uses values ranging from far-left to far-right, where in total, we identified 13 different values within the left-right scope, as well as 5 additional values which refer to specific political orientations outside the left-right scope, which are shown in Table 1.

[2] https://www.chesdata.eu/ches-europe

[3] https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv

[4] https://www.chesdata.eu/s/CHES2019V3.csv

Abbreviation	**Value**
FL	Far-left
LLF	Left to far-left
L	**Left**
CLL	Centre-left to left
CL	Centre-left
CCL	Centre to centre-left
C	**Centre**
CCR	Centre to centre-right
CR	Centre-right
CRR	Centre-right to right
R	**Right**
RRF	Right to far-right
FR	Far-right
BT	**Big tent**[5]
PP	**Pirate Party**[6]
SY	**Syncretic politics**[7]
SI	**Single-issue politics**[8]
NP	**Nonpartisanism**[9]

Table 1: Political orientation values, identified in the Wikipedia data.

The information from Wikipedia covers the ParlaMint political parties and parliamentary groups quite well: out of 932 such entities currently defined in ParlaMint, only 20 (2.2%) could not be assigned a left-right orientation.

Encoder classification: The third source of data were the encoders (i.e. compilers of the corpus), who, if they so decided, entered their classification on the left-right orientation, which was mainly so as to be able to mark the political parties that were not covered by Wikipedia. Currently, only three of the partners made use of this option.

The combination of sources proved useful in several aspects: The CHES datasets provided us with expert data on many dimensions associated with the political orientation of parties on a numerical scale and also offered the possibility of tracing changes in the political orientation of a particular party over the years, provided that the party had been a member of parliament for several years. However, as mentioned, its coverage is limited. Therefore, the second source, Wikipedia, has much greater coverage, even if the data is not as reliable as that of the CHES expert dataset, and also gives us only one dimension or political orientation, i.e. its category on the left-to-right scale.

[5] A big tent party, or catch-all party, is a term used in reference to a political party's policy of permitting or encouraging a broad spectrum of views among its members. https://en.wikipedia.org/wiki/Big_tent.

[6] Pirate Party refers to political parties that support civil rights, direct democracy, encourage innovation and creativity, free sharing of knowledge, information privacy, free speech, anti-corruption, net neutrality and oppose mass surveillance, censorship and Big Tech. https://en.wikipedia.org/wiki/Pirate_Party.

[7] Syncretic politics refers to politics that combine elements from across the conventional left–right political spectrum. https://en.wikipedia.org/wiki/Syncretic_politics.

[8] Single-issue politics refers to a political stance that is based on one essential policy area or idea.https://en.wikipedia.org/wiki/Single-issue_politics.

[9] Nonpartisanism refers to a political stance that does not agree with the current political party system. https://en.wikipedia.org/wiki/Nonpartisanism.

4 Data encoding

The task of encoding the data was divided into two parts: The first part consisted of the automatic extraction of the values from the CHES dataset for each political party included in the ParlaMint corpora. After the initial extraction of the CHES data, the identifiers of the parties in the dataset (CHES_ID) were automatically matched with the abbreviations from ParlaMint (PM_ID) using the following heuristics:

- Exact match: if the ParlaMint abbreviation was an exact match to the CHES identifier, the matching values were given in the corresponding fields (PM_ID and CHES_ID);
- Fuzzy match: an attempt was made to match the ParlaMint abbreviation without punctuation; if a fuzzy match was found, the matching values were given in the corresponding fields;
- Multiple matches: if multiple matches were found, all ParlaMint party abbreviations were output in separate rows with identical CHES-related columns;
- No match found: if no match was found for a CHES_ID, the PM_ID in the corresponding row was given a value of ”-” for ”unknown”.

For all ParlaMint parties for which no match was found, additional rows were added to the TSV; these contain the PM_ID, with all other CHES-related columns having the value ”0”. The second part consisted of manually editing the automatically generated TSV files to match the ParlaMint parties with the CHES parties in cases where no automatic match was found, but one was present. Special attention had to be paid to parliamentary groups that did not correspond to a single party but included several parties with possibly different political orientations - we handled such cases by inserting the value of the closest political party (if such a party existed) or we did not insert the value at all if no party corresponded well to the parliamentary group.

Since the ParlaMint corpora are encoded in XML according to the Text Encoding Initiative (TEI) Guidelines, the structures encoding the added metadata can be quite complex. Therefore, to simplify the process of adding metadata and make it less error-prone, we did not require the orientation data to be entered directly into XML but prepared tabular TSV files for each country that were pre-populated with the abbreviations of all political parties.

The Wikipedia URLs and the orientation data as well as the encoder orientation data were then added in Excel, possibly with comments, and the files were saved as TSV[10]. An XSLT script then takes the TSV files and the XML corpus file with the organisational data and inserts the new data into the XML file. A similar procedure was applied to the CHES data: Here, too, the CHES CSV files were converted to TSV, the party abbreviations from CHES were mapped semi-automatically in Excel to the ParlaMint party identifiers, the results were saved as TSV and again inserted into the XML files.

Figure 1 gives an example of the political orientation encoding. It should be noted that the CHES variables as well as the Wikipedia and encoder left-right orientations are pointers to taxonomy categories, which give the name and explanation of the reference, e.g. similarly to the categories and explanations presented in Table 1.

5 Metadata analysis

This section presents statistics of the added political orientation metadata, first examining the coverage of the CHES and Wikipedia TSV files separately to determine the coverage of both datasets, particularly with regard to missing values. With regard to the completeness of the CHES dataset, we first examined the percentage of available data for each CHES value (85 values in total) per ParlaMint country, the results are shown in Figure 2.

Austria (AT) stands out as the country with the most comprehensive variable coverage, with certain variables reaching up to 97% of the values (e.g., `lrgen`, `lrecon`, `eumember`, or `galtan`). Following closely behind are Estonia (EE) and Lithuania (LT), where the best-covered variables range between

[10] The TSV files are available on the ParlaMint GitHub page at the following link.

```
<org role="parliamentaryGroup" xml:id="MR">
  <orgName full="abb">MR</orgName>
  <orgName full="yes">Mouvement Réformateur</orgName>
  <idno type="URI"
  subtype="wikimedia">https://en.wikipedia.org/wiki/Reformist_Movement</idno>
  <state type="politicalOrientation">
    <state type="encoder" source="#GrietDepoorter" ana="#orientation.CRR">
       <note xml:lang="en">Orientation determined by encoder, using own
       knowledge of the parliamentary group.</note>
    </state>
    <state type="Wikipedia"
     source="https://en.wikipedia.org/wiki/Reformist_Movement"
     ana="#orientation.CR">
      <note xml:lang="en">From 1992 the Reformist Movement (MR) consisted of:
      FDF, MCC, PRL and PFF.
      In September 2001, FDF decides to leave the alliance and chooses a
      new name, becoming DeFI.</note>
    </state>
  </state>
  <state type="CHES" key="106" n="MR" from="2002" to="2018"
    source="https://www.chesdata.eu/s/1999-2019_CHES_dataset_meansv3.csv">
    <state type="variable" ana="#ches.lrgen">
      <state type="value" from="2002" to="2005" n="6.35"/>
      <state type="value" from="2006" to="2009" n="6.67"/>
      <state type="value" from="2010" to="2013" n="7.0"/>
      <state type="value" from="2014" to="2018" n="7.0"/>
    </state>
    ...
    <state type="variable" ana="#ches.vote">
      <state type="value" from="2002" to="2005" n="10.1"/>
      <state type="value" from="2006" to="2009" n="11.4"/>
      <state type="value" from="2010" to="2013" n="9.28"/>
      <state type="value" from="2014" to="2018" n="9.6"/>
    </state>
  </state>
</org>
```

Figure 1: Encoding of political orientation in ParlaMint.

92% and 94%. Conversely, Turkey (TR) and France (FR) exhibit the lowest variable coverage. For FR, the variables with the most data only encompass 27%, while for TR the coverage does not exceed 12%. This data scarcity for TR may not be surprising considering that the country was only included in the 2019 edition of the CHES surveys alongside Norway (NO) and Iceland (IS). In comparison, however, NO and IS have some variables that are still relatively well covered (between 75% and 80%).

In general, the variables with the most comprehensive coverage in the dataset are `year`, `lrgen`, `lrecon`, `galtan` (party's position in relation to its views on social and cultural values), `eu_position` (overall orientation of the party leadership towards European integration), `eu_dissent` (degree of dissent on European integration) and `eu_salience` (relative importance of European integration in the party's public stance), (which all account for 68.03% when calculating the percentage of available data per variable), while the variables with the least available data `galtan_sd` (standard deviation of expert placement of the party in 2019 concerning its views on democratic freedoms and rights), `lrecon_sd` (standard deviation of the party's expert ranking in 2019 in relation to its ideological stance on economic issues) with 1.92% of available data and `eu_econ_require` (party's position on fulfilling the economic requirements of EU membership), `eu_googov_require` (party's position on fulfilling the good governance requirements of EU membership) and `eu_political_require` (party's position on fulfilling the political requirements of EU membership) with only 0.44% of available data[11]. One of the reasons for the low coverage of the

[11] Expanded definitions for the variables can be found in the 1999-2019 Chapel Hill Expert Survey (CHES) trend file

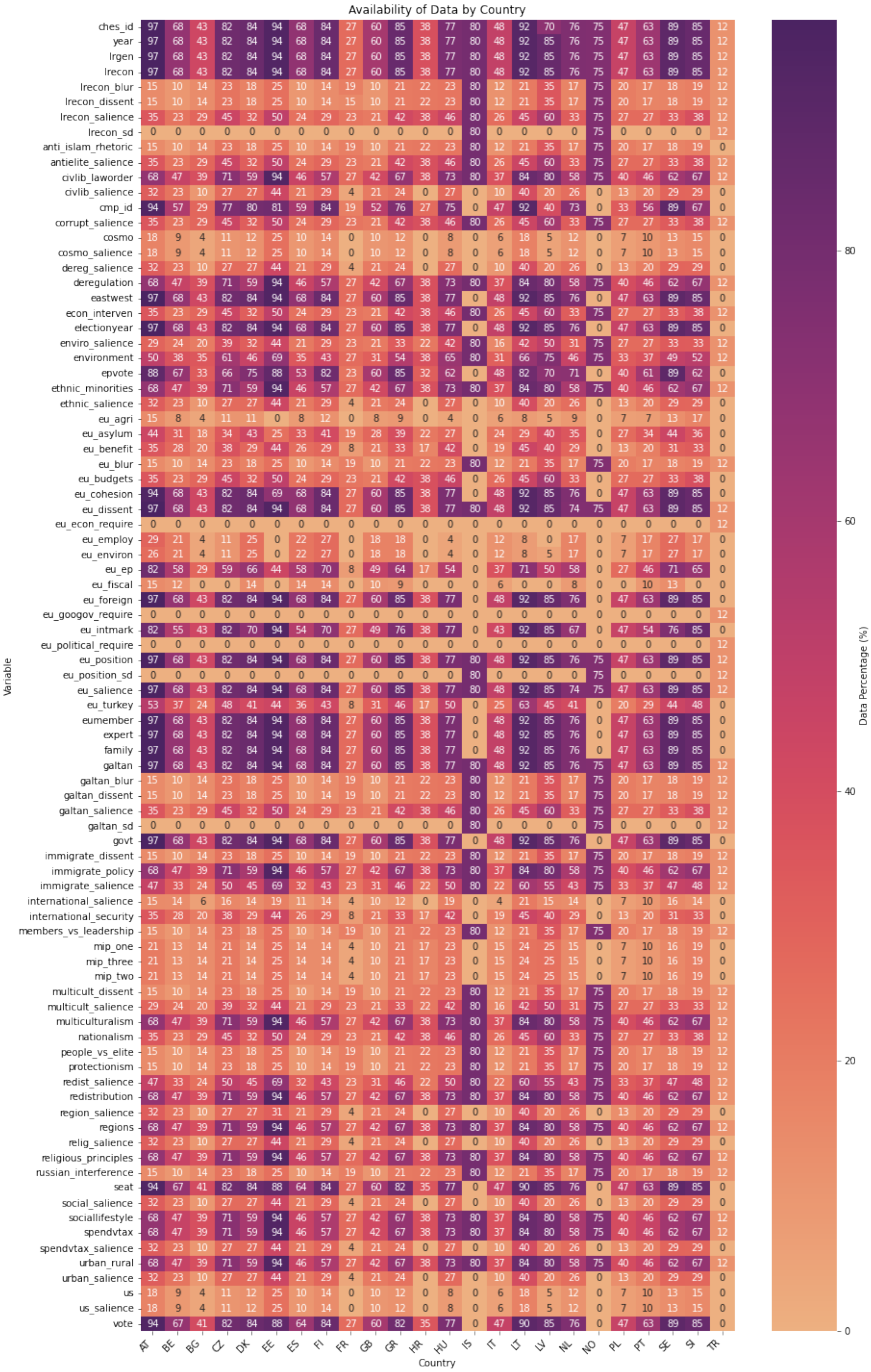


Figure 2: Percentage of data available for each variable in the CHES dataset per an individual ParlaMint country, variable definitions are available in the 1999-2019 CHES Codebook.

aforementioned variables with the lowest coverage is the fact that these variables are only included in the 2019 CHES dataset and were not measured in any other year/survey. In contrast, the percentage of missing data for the Wikipedia values on political orientation is only 15.25%, which provides good coverage for further analysis.

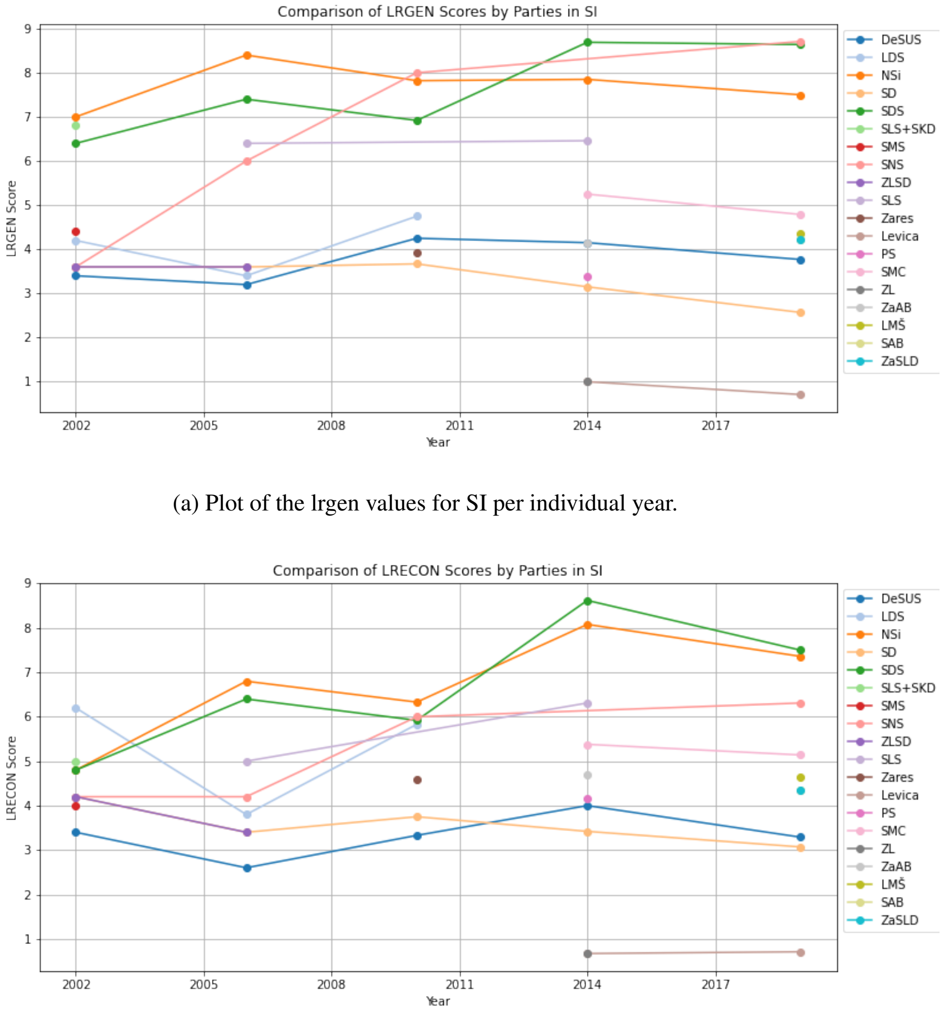
In analyzing the content of the CHES variables, we examined the aforementioned `lrgen` variable (indicating the party's position in a given year concerning its overall ideological stance) and the `lrecon` variable (indicating the party's position in a given year in terms of its ideological stance on economic issues) for the CHES dataset, visualizing some data trends for country comparisons as well as examining individual countries (and their political parties), in particular creating line charts for each country (provided that they were included in CHES dataset) to visualize changes in `lrgen` and `lrecon` values to examine trends in parties' political orientations over time. An example of such an analysis of the variables `lrgen` (Figure 3a) and `lrecon` (Figure 3b) per year for ParlaMint-SI is shown in Figure 3, which allows a comparison of the values in the case of Slovenian political parties for several years in the period from 2002 to 2019. In figure 3a, a distinction between (centre-) left and (centre-) right can be seen, with some of the notable examples, such as the political parties SDS (Slovenian Democratic Party) and NSi (New Slovenia – Christian Democrats) on the far right of the spectrum, DeSUS (Democratic Party of Pensioners of Slovenia) and SD (Social Democrats) on the left and ZL (United Left) and Levica (The Left, successor to United Left) on the far-left, with one exception - the political party SNS (Slovenian National Party) starts with a value of 3.6, a relatively (centre-)left value in 2002, which rises sharply to a value of 8.7 by 2019, surpassing the SDS (value 8.64) as one of the most right-wing political parties in Slovenia.

Similar distributions can also be seen in Figure 3b, where the distribution of `lrecon` values is relatively similar to that of `lrgen` values, which could indicate a possible correlation between the parties' general ideological position and their economic policies. This type of analysis could be extended further by comparing the scores with, for example, the variable `family`, which indicates the ideology of a single political party (where, for example, the SDS is noted as a conservative party, while the SNS is labelled as a radical right-wing (Rad right) political party, despite having a very similar, almost identical `lrgen` score). However, as shown, the data set is very limited for specific variables or countries, so any analysis should be carried out carefully and the coverage of the selected countries and/or variables should be checked. In the case of several countries (e.g. NO, IS, HR and TR), the data points are very limited and often cover only one or two years, making accurate analysis impossible.

In addition to the analysis of the CHES dataset, we also performed a more in-depth analysis of the Wikipedia dataset, for which we aggregated the per-speech metadata of all corpora, (utterance ID, speaker name, their political party at the time of speaking, L-R orientation from Wikipedia etc.) into one large TSV file comprising almost 8 million lines. The entire dataset consists of 7,995,766 utterances, 22,641 unique speakers, 774 unique political parties, and 54 identified political orientation categories. However, the dataset had to be filtered due to the problem of multiple affiliations of speakers, where in some cases a small number of utterances belong to speakers who are noted as members of multiple political parties, due to the fact that the information was either not processed correctly when the dataset was created or the data was not coded correctly, due to a 1-day overlap when a person changed their political party[12]. The problem manifested itself in problematic values for political orientation such as "`Big tent;Centre`" or even "`Centre-leftCentre-leftCentre-left-centre-left`". However, such values account for only 1.63% of the whole dataset (or, more specifically, 1,63% of all utterances 2.48% of total unique speakers and 7.75% of total unique parties) and these were not considered in the statistical analysis.

The filtered dataset, where the problematic values were omitted, consists of 22,078 unique speakers, 714 unique political parties, 18 orientations (preserving the Wikipedia-extracted values in Table 1) and 7,865,408 unique utterances in total. In particular, we examined the number of utterances and the number of speakers, the parties from which the utterances were spoken, and which were linked to specific political orientations for the entire ParlaMint dataset, as shown in Table 2.

[12] This problem will be corrected with the next maintenance version of the ParlaMint dataset.



(a) Plot of the lrgen values for SI per individual year.

(b) Plot of the lrecon values for SI per individual year.

Figure 3: Example of an analysis of the variables *lrgen* (a) and *lrecon* (b) for the political parties of ParlaMint-SI. The diagrams show the changes in the general and economic political orientation of the political parties for a period between 2002 and 2019.

The table shows the number of utterances, speakers and parties per individual political orientation as given in the Wikipedia values we extracted. One of the pieces of information included is also the number of utterances that do not contain any information on political alignment (in the table labelled as Missing data). This is either because this information was not available for a particular party, or in cases where the speaker does not belong to a political party (e.g. a guest speaker). Of the other categories with available data, `Centre-right` is the orientation with the largest number of spoken utterances, followed by `Centre-left` (which also contains the largest number of political parties and spoken utterances, making centre-left speakers the most vocal), `Right` and `Centre`. This is generally not too surprising, as these are relatively common categories when it comes to the political spectrum between left and right. Of the more nuanced political orientations (which tend to be less present in the left-right spectrum), `Centre-left to left`, closely followed by `Centre-right to right` and `Right to far-right` seem to predominate in terms of spoken utterances (and the large proportion of active speakers). Finally, looking at the distribution of spoken utterances for the orienta-

Table 2: Summary of political orientation statistics - an overview number of speakers, political parties, and utterances that belong to individual political orientation categories. Political orientations are based on Wikipedia-extracted values and range from far-left to far-right, with additional categories for other political alignments outside the left-right scope (Big tent, Pirate party, Single Issue Politics, Syncretic politics).

Political Orientation	**Utterances**	**Speakers**	**Parties**
Missing data	692,341	7121	147
Far-left	49,293	106	9
Left to far-left	176,621	245	16
Left	198,034	534	58
Centre-left to left	406,568	815	40
Centre to centre-left	269,422	730	62
Centre-left	1,517,916	3623	107
Centre	644,572	2076	196
Centre to centre-right	331,666	831	40
Centre-right	1,743,189	3621	256
Centre-right to right	401,967	1225	41
Right	759,368	1616	52
Right to far-right	385,432	1058	47
Far-right	76,019	322	21
Big tent	175,955	980	24
Pirate Party	10,950	29	1
Single Issue Politics	19	2	2
Syncretic Politics	26,076	53	5
Total	7,865,408	24987	1124

tions outside the left-right range, it can be seen that relatively many spoken utterances (and a large number of speakers) come from the political parties of the `Big Tent`, while for the parties of the `Single Issue Politics` only two speakers from two different parties can be found in the data set. Furthermore, the political orientation ”non-partisanship” (NP) does not appear in the ParlaMint corpus, or rather, no utterance was produced by a speaker belonging to a non-partisan political party.

6 Conclusions

We presented ongoing work to add political orientation metadata to the ParlaMint II parliamentary corpora. We have captured the political orientation of more than 350 European political parties by relying on two highly informative data sources, the Chapel Hill Expert Survey dataset and the Wikipedia pages of the respective parties, facilitating manual annotation of the political orientation on individual speeches from the corpora.

We faced several challenges and conceptual constraints, such as dealing with the political orientation of parties that were derived from others or were renamed. Regarding the CHES dataset, it could be argued that the dataset is somewhat sparse and “outdated” as it was last updated in 2019 and therefore does not provide information on the political orientation of parties formed after 2019[13]. As we initially only collected data for political orientation (i.e., only the `lrgen` variable, before deciding to integrate the entire CHES dataset) we identified this as a potential problem, which was addressed by using Wikipedia as a secondary source.

[13] However, a series of new CHES surveys have just been published, which will provide new data for the period up to 2022.

Contrary to our initial assumptions when comparing numerical values, we found that the Wikipedia data was highly consistent with the CHES variable `lrgen` and no major discrepancies were found between the comparison of the two sources. We attribute this to the fact that we had originally chosen to label the parties in more detail (e.g. left to centre-left) rather than simply left/centre/right. This allowed us to bridge minor differences between the two data sets. Example: When CHES indicated political orientation as centre-left and Wikipedia indicated left to centre-left, we understood this not as a contradiction, but as two alternative ways of labelling party orientation. If CHES labelled a particular party as left and Wikipedia as centre, this was understood as an inconsistency and we had to adjust our workflow accordingly. However, this was done during the initial compilation of the dataset and would require further testing to fully confirm.

We are aware that the political orientation of parties does not necessarily coincide with the personal orientation of the speaker belonging to the respective party and also recognize that people's ideological beliefs, as well as what they say, are often fluid and therefore difficult to capture. Nevertheless, the method that we have employed does give each speech its implied political orientation. The analysis carried out so far first gave us an insight into the composition of the metadata sets, both for the CHES and for the values extracted from Wikipedia, particularly with regard to data availability (especially for the CHES dataset). While the Wikipedia values have a much better coverage compared to the CHES data, there are still a large number of utterances (Table 2) that do not contain information on political orientation (the reason for this could be that the speakers do not belong to a political party or the information was not available for that particular party). On the other hand, even if its coverage is problematic, the CHES dataset still contains enough data on some country and/or variables so that the analysis can be performed without concerns about the balance of the dataset. One such example we presented in the analysis is tracking changes in general political orientation and economic policy orientation (`lrgen` and `lrecon` variables, respectively) for the period between 2002 and 2019 for Slovenian political parties. Finally, the analysis of the distribution of utterances between political orientation categories (from Wikipedia) for the entire metadata of the ParlaMint corpus provided a more comprehensive picture of the political landscape within the ParlaMint corpora, as it shows the distribution of political orientations among speakers and parties, indicating the diversity of perspectives and ideologies within political discourse.

In the future, we would like to gain further insights into the data by extending the current analysis to include the analysis of individual corpora using the CHES variables. At the time of writing, a new set of CHES datasets has just been released, alleviating some of the limitations in data availability and providing new variables for new types of analysis. In addition, we would like to expand our current analysis to focus more on the exploration of the content present in the corpora. Specifically with regard to political orientation, we would also like to enable a comparison of the speeches of left/right or centre-leaning speakers (or political parties) with each other to see whether they speak according to their political alignment or rather according to the political orientation of the political party to which the speaker belongs – instead of relying solely on the speaker's metadata, we could use various NLP-based techniques to analyse the speeches, statements or topics discussed and infer the speaker's current political stance, which may differ from the political orientation of the party the speaker belongs to. This type of analysis could then also be done for specific topics (e.g., attitudes toward European integration) that are included in the CHES metadata. In addition, the metadata will be used as part of the shared task on ideology and power identification in parliamentary debates[14], which will be part of the Touché lab[15] at the CLEF 2024[16] conference[17]. Lastly, we hope to include additional metadata useful to humanities and social scientists using ParlaMint corpora for their research, such as V-Dem[18] (Coppedge et al., 2021) and Party Facts[19] (Döring & Regel, 2019) datasets.

[14] https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html

[15] https://touche.webis.de/clef24/touche24-web/index.html

[16] http://clef2024.clef-initiative.eu/

[17] For simplicity, only the left-to-right labels will be used, flattening the fine-grained annotations but still making use of it.

[18] https://v-dem.net/

[19] https://partyfacts.herokuapp.com/documentation/about/

References

Bischof, K., & Ilie, C. (2018). Democracy and discriminatory strategies in parliamentary discourse. *Journal of Language and Politics*, *17*(5), 585–593. https://doi.org/https://doi.org/10.1075/jlp.00017.edi

Blattberg, C. (2001). Political philosophies and political ideologies. *Public Affairs Quarterly*, *15*(3), 193–217.

Cohen, R., & Ruths, D. (2021). Classifying Political Orientation on Twitter: It's Not Easy! *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 91–99. https://doi.org/10.1609/icwsm.v7i1.14434

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Alizada, N., Altman, D., Bernhard, M., Cornell, A., Fish, M. S., et al. (2021). V-dem dataset v11. 1.

Döring, H., & Regel, S. (2019). Party facts: A database of political parties worldwide. *Party Politics*, *25*(2), 97–109. https://doi.org/10.1177/1354068818820671

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., … Fišer, D. (2022). The ParlaMint corpora of parliamentary proceedings [https://doi.org/10.1007/s10579-021-09574-0]. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0

Fišer, D., & Pahor De Maiti, K. (2021). "First, I'm a Female Politician, Not a Male One, and Second …": A Corpus Approach to Parliamentary Discourse Research. *Contributions of contemporary history*, *61*(1), 144–179. https://doi.org/10.51663/pnz.61.1.07

Freire, A. (2015). Left–right ideology as a dimension of identification and of competition. *Journal of Political Ideologies*, *20*(1), 43–68.

Gabel, M. J., & Huber, J. D. (2000). Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science*, 94–103.

Heywood, A. (2021). *Political ideologies: An introduction*. Bloomsbury Publishing.

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, *75*, 102420. https://doi.org/https://doi.org/10.1016/j.electstud.2021.102420

Knapp, A., & Wright, V. (2006). *The government and politics of France*. Routledge.

Yan, H., Lavoie, A., & Das, S. (2017). The perils of classifying political orientation from text. *Linked Democracy: Artificial Intelligence for Democratic Innovation*, *858*, 8.